Personalizing Robot Tutors to Individuals' Learning Differences

Daniel Leyzberg
Yale University
Dept. of Computer Science
daniel.leyzberg@yale.edu

Samuel Spaulding
Mass. Institute of Technology
MIT Media Lab
samuelsp@media.mit.edu

Brian Scassellati
Yale University
Dept. of Computer Science
scaz@cs.yale.edu

ABSTRACT

In education research, there is a widely-cited result called "Bloom's two sigma" that characterizes the differences in learning outcomes between students who receive one-on-one tutoring and those who receive traditional classroom instruction [1]. Tutored students scored in the 95^{th} percentile, or two sigmas above the mean, on average, compared to students who received traditional classroom instruction. In human-robot interaction research, however, there is relatively little work exploring the potential benefits of personalizing a robot's actions to an individual's strengths and weaknesses. In this study, participants solved grid-based logic puzzles with the help of a personalized or non-personalized robot tutor. Participants' puzzle solving times were compared between two non-personalized control conditions and two personalized conditions (n=80). Although the robot's personalizations were less sophisticated than what a human tutor can do, we still witnessed a "one-sigma" improvement (68^{th} percentile) in post-tests between treatment and control groups. We present these results as evidence that even relatively simple personalizations can yield significant benefits in educational or assistive human-robot interactions.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics; J.4 [**Computer Applications**]: Social And Behavioral Sciences—*Psychology*

General Terms

Experimentation

Keywords

Education, Robotics, HRI, Tutoring, Personalization, Assessment, ITS

HRI'14, March 3–6, 2014, Bielefeld, Germany.

http://dx.doi.org/10.1145/2559636.2559671.

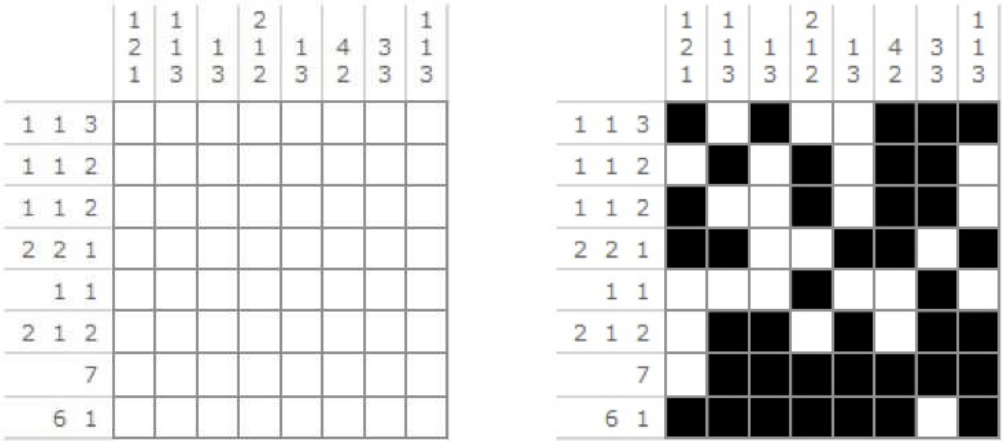
1. INTRODUCTION

There is a long-held belief in the HRI community that personalized interactions are likely to be more compelling or more useful than non-personalized interactions. Research in this area has shown that people behave more socially towards a robot that personalizes its interactions based on interaction history [2]. However, relatively little is known about personalization as a means to elicit behavioral change in educational or assistive HRI. To what extent does personalization improve the utility of such robots? Is it difficult or expensive to build systems that tailor their output to individuals' strengths and weaknesses in this context? Does the addition of personalization elicit desirable behavioral changes among users?

To explore these questions we chose a domain in which individual customizations are critical to the success of an interaction. Effective tutoring requires teachers to assess students' individual differences and align their curricula and methods to best suit an individual student's needs. Bloom's "two sigma" result and many years of studies indicate that individually tailored lessons are much more effective than traditional classroom lessons [3]. These results confirm that personalization is important in education, but comparing a classroom learning environment to a one-on-one learning environment leaves open the question of to what extent the difference in learning outcomes can be attributed to personalization alone.

In this work, we compare four conditions, all of which were one-on-one human-robot tutoring interactions. In two conditions, we personalized the lessons participants received based on an online assessment of the individual's skills. The content of each lesson was pre-recorded and identical in both personalized conditions; the only difference was the order in which the lessons were delivered. There are also two control conditions in this study: one with no lessons whatsoever, to establish a baseline of puzzle solving performance in our population, and one with the same pre-recorded lessons as in the two personalized conditions, but where the order is chosen entirely independently of the user's skills.

We find that, despite only changing the order of lessons, participants that received personalized lessons performed "one sigma" better than either control group in post-test puzzle solving time. This result indicates that even relatively simple personalizations can significantly improve the utility of an educational or assistive human-robot interaction.



(a) Sample nonogram puzzle, blank.

(b) Sample nonogram puzzle, solved.

Figure 1: A sample nonogram puzzle. The objective of nonograms is, starting with a blank board (see left figure), 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 (see right figure). For a more detailed explanation see Section 3.2.2.

2. RELATED WORK

Bloom's "two sigma" result inspired nearly three decades of research investigating the differences in learning outcomes between one-on-one tutoring and traditional classroom instruction. While there is no debate whether one-on-one tutoring benefits students, a recent review and meta-analysis [3] noted that the average benefit of personalization was closer to one sigma than to two sigma. Whether the benefits are one sigma or two, there is little doubt that individualized instruction has a substantial impact on students' learning gains.

In addition to human tutoring, another active area of research is creating and evaluating computerized tutoring systems [4]. The average effect size across studies evaluating learning gains made via computerized tutoring is that tutored students performed 0.76 sigma better than students who received only traditional classroom education, a close second to the effect of human tutors [3]. These learning outcome results, whether produced by human tutor or computer tutor, underscore the importance of personalization in the education domain. No previous work, however, has isolated the effect of personalization on human-robot interactions in an educational or assistive setting.

Previous work in personalization in HRI has been limited to measures of engagement and user satisfaction. Snackbot, a robot that personalized dialogue in reference to an individual user's history of snack choices, was found to be more engaging on several measures than a non-personalized version of the same robot, including an increased desire to use the robot and an increase in social behavior directed toward the robot [2]. Kidd's robot weight loss coach generates customized dialogue based on the progress of the user [5] but his research does not isolate the effect of personalization. In other work, users that were allowed to decorate, and thus personalize, their Roombas self-reported higher engagement with the robot and more willingness to use the robot in the future [6]. No earlier work, however, isolated the effects of personalization in HRI in an educational or assistive setting.

There has been other work on robot tutoring, however. One such platform is RUBI, a robot tutor designed to interact with 18 to 24 month old children [7]. RUBI is a humanoid with articulated arms, an expressive face, and a tablet embedded in its midsection on which it displays educational content like vocabulary lessons. A study of the iRobiQ, a humanoid similar in design to RUBI, provides some experimentally-derived guidelines for using robot tutors in classrooms [8]. Research has also been done on tutoring robots that operate as museum guides [9] or teleoperated instruments of a human teacher, such as the Huggable robot [10]. A long-term study of elementary students playing chess with an iCat explored how supportive students perceive different versions of the robot tutor to be [11]. No previous robot tutor research, however, isolates the role of personalization.

3. METHOD

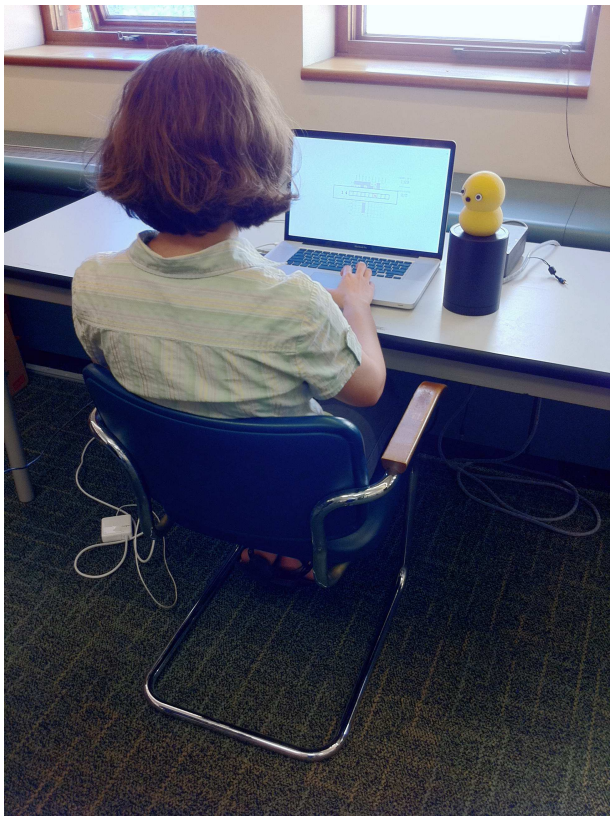
To assess the effect of personalization in an educational human-robot interaction, an experiment was conducted in which a robot tutor assisted participants in solving grid-based logic puzzles.

In previous work, the authors demonstrated that physically-embodied robot tutors produce greater learning gains than on-screen virtual agents delivering the same lessons[1] [12]. The present work isolates the effects of personalization in educational/assistive HRI by comparing robot tutors that provide individualized lessons to those that do not.

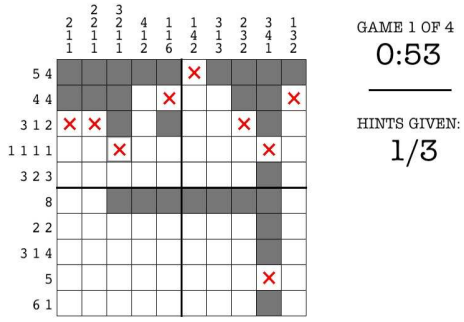
3.1 Participants

There were 80 participants in this experiment, between 18 and 40 years of age. Most participants were undergraduate and graduate students of Yale University, none of whom

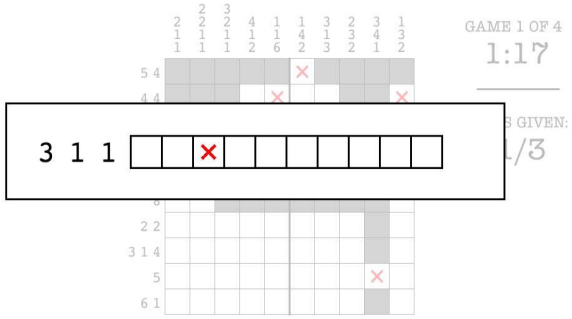
[1]Our earlier work on embodiment [12] used a similar apparatus and some data here were previously presented as part of that work. The current work introduces new data, new algorithms for personalization, and an experimental design that isolates the effect of personalization in educational/assistive HRI.



(a) Participants solve a nonogram puzzle on the computer as the robot (Keepon) analyzes the moves they make and, three times per puzzle, delivers brief lessons (21 – 47 sec.) about gameplay strategies.



(b) A screenshot of the nonogram puzzle user interface during gameplay. All boards start blank. Participants played four puzzles for a maximum of fifteen minutes per puzzle.



(c) A screenshot of the tutoring user interface. The robot 'speaks' in pre-recorded spoken messages and and moves in pre-recorded motions while coordinated visuals appear on screen.

Figure 2: Experiment apparatus and user interface screenshots.

were pursuing a degree in computer science. This study is a between-subjects design with four groups of 20 participants each, receiving either: (1) no lessons, (2) randomized-but-relevant lessons, (3) personalized lessons chosen by an additive skill assessment algorithm, or (4) personalized lessons chosen by a Bayesian network skill assessment algorithm. Exclusion criteria were a lack of English fluency or prior academic experience with robotics or artificial intelligence.

3.2 Apparatus

Participants in the experiment were asked to solve a set of logic puzzles on a computer. A robot tutor interrupted several times throughout each participants' session to deliver puzzle solving strategy lessons. In the two *personalized lessons conditions*, the robot used one of two skill assessment models (described in Section 4) to choose which among a set of pre-recorded lessons to give the participant, whereas in the *randomized-but-relevant lessons condition* the robot picked a random pre-recorded lesson among those applicable to the current state of the board but independent of the skills of the user. In the *no lessons condition* participants solved the puzzles with no help, the robot merely served as an announcer for the game. We compare the puzzle solving performance between these four groups to evaluate the effects of personalization in educational/assistive HRI.

3.2.1 Robot

The robot we used, Keepon, is a small yellow snowman-shaped tabletop robot, 11 inches tall, see Figure 2(a). We chose Keepon because it is particularly well suited to expressive non-threatening social communication [13, 14].

During the experiment, the robot played three roles. First, it refereed and acted as a host: 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, it "observed" the board during gameplay: the robot's body faced the screen and its head followed the location of the mouse cursor as the participants worked on the puzzles. Third, it delivered short puzzle-solving strategy lessons, three times per puzzle: it turned to face the participant and "bounced" its body up and down while playing one of several pre-recorded spoken messages with accompanying visuals appearing on screen, overlaid onto the puzzle interface, see Figure 2(a). If a lesson was going to be repeated, the robot would first apologize for repeating itself (i.e., "I'm sorry to repeat this hint but I think this will help.").

To focus our efforts on the personalization of the interaction, we did not use a vision system to detect state changes in the puzzle board. Instead, the robot had perfect knowledge of the game state via a software link to the puzzle interface.

3.2.2 *Puzzle*

To ensure the greatest likelihood of participants starting the study at the same skill level, we chose a puzzle game that is relatively obscure to American audiences: a grid-based fill-in-the-blanks puzzle called "nonograms" (also called "nonogram puzzles") that resemble crossword puzzles or Sudoku. Nonogram puzzles are a relatively difficult cognitive task, one that requires several layers of logical inferences to complete. Solving a nonogram puzzle of arbitrary size is an NP-complete problem [15].

The objective of nonograms is to, starting with a blank board, shade 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, and with at least one empty box between the sets of adjacent shaded boxes. For a sample puzzle and its solution see Figures 1(a) and 1(b).

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 has solved a nonograms puzzle when one finds a pattern of blank and shaded boxes such that all of the requirements for each row and column are satisfied. Some nonogram puzzles have more than one solution, but most only have one.

In a typical puzzle, making progress on any row or column depends on the boxes in the rows and columns that it intersects. One must infer some parts of rows or columns based on what one has already established by inference earlier in the game. When a player has reasoned that some box should be shaded, the player shades it; if he or she reasons that a box will definitely *not* be shaded, they mark such boxes with an 'X' for reference.

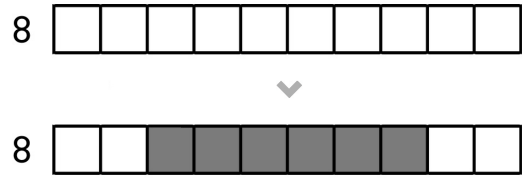
We created a full-screen nonograms interface that participants used via mouse and keyboard. The user interface included a game timer and a count of how many lessons (called "hints") the participant had received thus far and how many more they would receive for that puzzle, see Figure 2(b).

Participants were asked to solve four puzzles on a ten-by-ten grid with a time limit of fifteen minutes per puzzle. The four puzzles chosen were identical for all participants across all groups. The fourth puzzle was a copy of the first puzzle, though disguised by rotating the puzzle 90°, such that the column requirements were swapped with row requirements. This manipulation enabled us to make within-subjects comparisons about the extent to which each participant improved over the course of the study. There was no indication that any participant became aware of this manipulation.

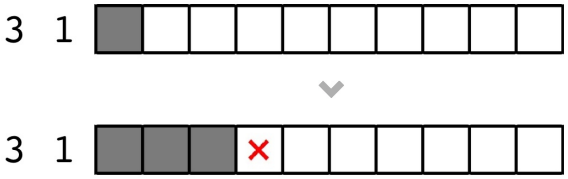
3.2.3 *Skills & Lessons*

Three times per puzzle, the robot interrupted the participant, paused the puzzle, and delivered a short lesson about nonograms solving strategies. 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 during the lesson as well as a set of coordinated robot motions specific to each lesson.

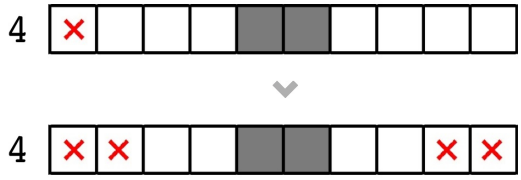
When beginning a lesson the robot would turn to face the participant and say "I have an idea that might help you," or "Here's another hint for you." During the lesson, the robot faced the participant and bounced up and down except when, in the course of the lesson, it referenced a visual presented on screen, at which point the robot briefly turned to face the screen. For instance, when the robot said "Like in this example..." or "As you see here...," the robot turned briefly to the screen and then back to the participant.



(a) In this row, there must be one long stretch. By the process of elimination one can infer that this stretch must occupy at least the middle six boxes, no matter where in the row it is placed.



(b) In this row, the first box is already shaded. Given that, and that the first stretch must be 3 boxes long, one can infer that the first three boxes must be shaded and the fourth must be crossed out.



(c) In this row, there is only one short stretch and some boxes are already shaded. One can infer that regardless of where that one stretch is placed, it cannot occupy the first two or the last two boxes in that row.

Figure 3: Examples of nonograms skills. Displayed are the contents of a row before and after each skill is applied. Although only rows are shown here, all nonograms skills apply to columns as well.

Ten puzzle-solving skills were identified based on the subjective experience of the authors. Each skill is defined as a set of row or column states in which one can logically fill in some remaining empty boxes based on what is already filled in. For example, a stretch of length 9 can fit in a blank row or column of 10 boxes in only one of two ways. Either it fills the first box and the next 8, but not the last box, or it fills the same middle 8 boxes and the last box but not the first one. In either case, the middle 8 boxes must be shaded. Following this pattern, one skills defined for this study is that, for an empty row or column with only one stretch requirement, n where $n > 5$, the center $(2n - 10)$ boxes are shaded. For example usages of this skill and two others, see Figure 3.

For each of these ten identified skills, there was one pre-recorded lesson. Three lessons were delivered per puzzle, for each of four puzzles; at least two lessons were repeated per

participant. Lessons were triggered either when a participant made no moves for 45 seconds or as he or she filled in the 25^{th}, 50^{th} or 75^{th} box on the board of 100. Participants were informed in the user interface of how many lessons were remaining for each puzzle.

The lessons were chosen based on the participant's experimental condition: either the lesson corresponding with the skill that had the lowest internal skill assessment score (in both *personalized lessons conditions*) or randomly chosen from among the applicable lessons to the current game board (in the *randomized-but-relevant lessons condition*). In all conditions, the only lessons eligible were ones that had an available application for the current board. This ensured that each lesson provided information that was actionable at the time the lesson was given.

3.2.4 Isolating Personalization

It was our intention to isolate the effect of personalization in an educational human-robot interaction. To achieve this, the *randomized-but-relevant control condition* chose lessons that were relevant to the participant's current board state but not necessarily the ones that were best suited to their skill level. This is intended to emulate a non-personalized classroom setting in which a teacher chooses an appropriate lesson for the class but not one that is necessarily tailored to any individual student's needs.

4. PERSONALIZED SKILL ASSESSMENT

The skill assessment algorithms presented here are not proposed as optimal solutions to the skill assessment problem more broadly. Instead, we are interested specifically in isolating what effect, if any, relatively simple personalizations have on education/assistive HRI.

To personalize the order of the lessons to suit the skills of individual participants, we created two relatively simple algorithms that assessed students' skills as they solved puzzles. Both algorithms take as input the moves participants make in each puzzle and produce as output a live updated vector of ten elements, each representing the likelihood that the participant has mastered one of ten predefined skills. In both *personalized lessons conditions*, the robot gave users the lesson that corresponded to their lowest scoring skill, of the subset of skills relevant to their current board state. The difference between the two algorithms is how the skill assessment scores were computed.

For our purposes, a skill i is defined as a function s_i that maps a potential state of the world ($w \in W$) from before an application of that skill to all potential resulting states after that skill is applied. A skill is *not* applicable to a state w, if and only if $s_i(w) = \{w\}$.

The skill functions are designed to be used in two ways:

* Skill functions are used to identify successful demonstrations of a skill. Skill i is said to be demonstrated at state w_t if $w_t \in s_i(w_{t-1})$.
* Skill functions are used to identify missed opportunities to demonstrate a skill. Skill i is said to have gone undemonstrated at previous state w_t if no action was taken and $s_i(w_t) \neq \{w_t\}$.

4.1 Additive Skill Assessment

In our first algorithm, we use a simple additive model to update a vector of skill assessments.

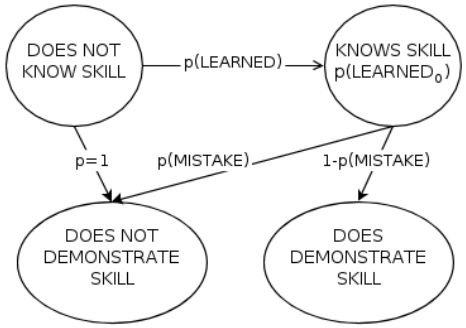


Figure 4: The Bayesian network skill assessment algorithm; an assessment of each skill is determined by an independent network in the form above. The upper nodes are internal states of the algorithm and the lower nodes correspond to the robot's observations of the user. For details see Section 4.2.

Given a set of skill definitions $s_i \in S$, this algorithm produces two internal Boolean functions for each skill i: a positive indicator p_i and a negative indicator n_i. p_i takes as input the previous and current world states and determines whether proficiency in skill i could have been responsible for this state transition ($w_t \in s_i(w_{t-1})$). n_i takes as input a world state and determines whether the i^{th} skill is applicable to that state ($s_i(w_t) \neq \{w_t\}$). The positive indicator functions are evaluated every time the state of the world changes. The negative indicator functions are evaluated every time the state of the world does not change for a given delay, and at regular time intervals thereafter until the user changes the state. In our implementation below, the initial delay was set to 3 seconds and the regular interval was 1 second. These time delays are dependent on the task; the ones used in this paper were chosen based on the authors' subjective experience with the task domain.

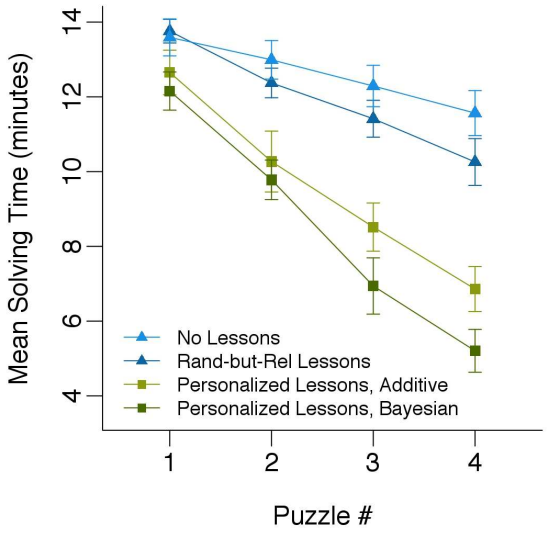
The skill assessment $a_{i,t}$ for each skill i is produced at time t as follows:

$$a_{i,t} = d + \sum_{j=0}^{t} (\omega_p p_i(w_{j-1}, w_j) - \omega_n n_i(w_j))$$

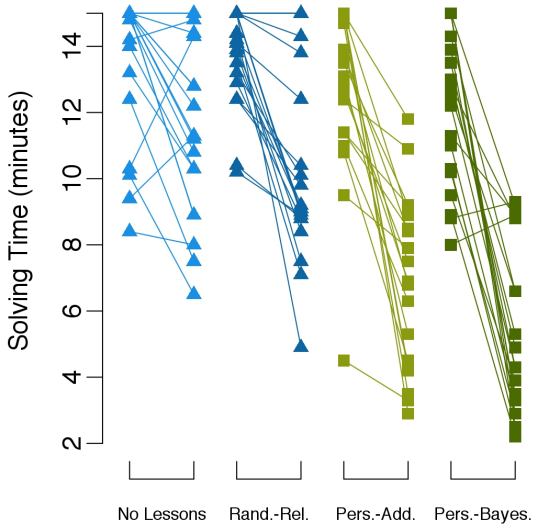
Each skill assessment $a_{i,t}$ starts at an initial seed value of $d = 50\%$, and is incremented or decremented by a linear combination of the positive and negative indicator signals. The relative weights of positive indications (ω_p) to negative indications (ω_n) will vary with expected relative frequencies of positive to negative indicators. In this application, we expected positive indications to be rare relative to negative indications; the weights we used were $\omega_p = 50\%, \omega_n = 1\%$. A floor of 0% and a ceiling of 100% is 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.

4.2 Bayesian Network Skill Assessment

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.



(a) Mean solving time per group, per puzzle. Participants who received personalized lessons solved each puzzle after the first significantly faster than participants in either of the control groups, see Table 1. Participants in both personalized groups performed approximately "one sigma" better on the final puzzle than participants in either control condition.



(b) Pre-test/post-test puzzle solving times for individual participants, separated by group. The fourth puzzle was the same as the first, but disguised by a 90° rotation. Participants receiving personalized lessons improved their same-puzzle solving time significantly more than participants in the control groups, $p < 0.01$.

Figure 5: Personalization produces greater learning gains: (a) Participants whose lessons were personalized solved the last three puzzles 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.

We addressed this weakness by offering a Bayesian network approach. Bayesian networks provide a way of modeling the relationship between observations and skill assessment in probabilistic terms. For each skill, we used the same graph structure with assumed independent variables (see Figure 4). Skills in this representation are categorized as either learned or not learned.

The following equation is used to estimate the student's proficiency with each skill individually:

$$p(L_t) = p(L_{t-1}|w_t) + (1 - p(L_{t-1}|w_t)p(L_0) \quad (1)$$

where $p(L_t)$ is the sum of the posterior probability that the rule was already learned, regardless of the current world state and the probability that the rule will make the transition to the learned state if it is not learned. In this experiment, we chose not to model the possibility of forgetting a skill. Our application domain was an interaction lasting less than one hour, thus we deemed forgetting a skill to be a relatively unlikely occurrence.

The two parameters are learned based on observations via an Expectation Maximization (EM) algorithm. Expectation Maximization requires starting points for each value it estimates; in the experiment described below, EM was implemented with seed values of 0 for p(LEARNED) and .5 for p(MISTAKE). EM alternates between performing an expectation step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization step, which computes parameters maximizing the expected log-likelihood. This method is commonly used to estimate the unknown parameters of a Bayesian network.

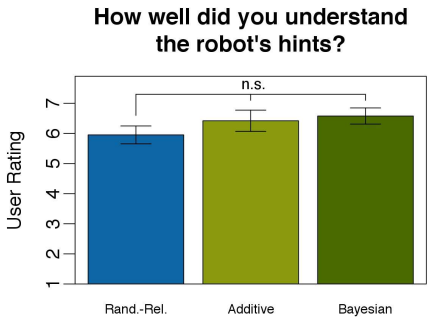
For both algorithms, the result is a vector of likelihoods that the tutor has about the users' proficiency with each individual skill. That information is used to select the most appropriate lesson to suit an individual's strengths and weaknesses.

4.3 Procedure

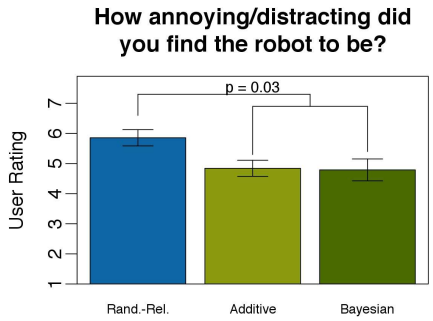
Before participating in this study, participants watched a five minute instructional video describing the rules of nonograms and how to use our computer interface. In this video, participants were encouraged to use logical reasoning to make moves in the game, rather than guessing. Afterwards, any questions were answered by an experimenter.

During the experiment, participants were alone in a room with the robot, the computer, and a video camera positioned behind them, see Figure 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 three open-ended questions and five Likert-scale questions. The questions were designed to assess whether participants perceived the lessons to be helpful, clear, and influential, as well as participants' perceptions of the robot. Participants rated how relevant the lessons were to them, how much the lessons influenced their gameplay, how well they understood the lessons, and how "smart/intelligent" and, separately, how "distracting/annoying" they perceived the robot to be.



(a) There is no significant difference between groups in how participants rated their understanding of the lessons.



(b) Participants in both *personalized lessons groups* rated the robot as significantly less "distracting/annoying" than those in the *randomized-but-relevant lessons group*, $p = 0.03$.

Figure 6: Results of two survey questions answered by participants upon completing the experiment.

5. RESULTS

The central hypothesis of this study is that personalized educational human-robot interactions will produce measurable behavioral differences when compared to non-personalized interactions. The behavioral measure is the length of time participants took to solve each of the four puzzles. For the purposes of calculating a mean, puzzles that were not completed within 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, varying from 29% to 38% in the first game to 9% to 17% in the fourth game.

	Game 1	Game 2	Game 3	Game 4
None	13.6 ± 2.2	13.0 ± 2.3	12.3 ± 2.5	11.6 ± 2.7
Rand-Rel	13.8 ± 1.4	12.5 ± 2.0	11.4 ± 2.3	10.3 ± 2.9
Additive	12.7 ± 2.6	10.0 ± 3.5	9.4 ± 3.0	7.6 ± 3.1
Bayesian	12.2 ± 2.3	9.8 ± 2.4	6.9 ± 3.4	5.2 ± 2.6

Table 1: Solving time means and standard deviations, in minutes. In each game except the first game, participants in both *personalized lessons groups* solved the puzzle significantly faster than participants in both the *randomized-but-relevant lessons* and *no lessons* groups ($p < 0.03$ for all).

Participants in both *personalized lessons groups* solved three of four puzzles significantly faster, on average, than those in either the *randomized-but-relevant lessons* or *no lessons* groups, $p < 0.03$ for all, see Table 1 and Figure 5(a). These results confirm the main hypothesis: personalized human robot interactions produced significantly improved user performance than non-personalized interactions.

Between *personalized lesson groups*, the *Bayesian group* did significantly better on the last puzzle than the *Additive group*, $t(37) = 0.05$.

In this study, the fourth puzzle was the same as the first, though disguised by a 90° rotation. There was no indication that any participant became aware of this manipulation. The difference in completion times between the first and fourth puzzles is a within-subjects measure of an individual participant's improvement over the course of the experiment. According to this metric, participants in either *personalized lessons group* improved ($M = 5.8$ minutes, $SD = 3.3$) their same-puzzle solving time significantly more than those in either *control group* ($M = 3.1$ minutes, $SD = 2.4$), $t(31) < 0.01$. See Figure 5(b).

Survey results indicate participants in the *personalized lessons groups* rated the lessons significantly more relevant to them ($M = 4.9, SD = 1.4$) than participants in the *randomized-but-relevant lessons group* ($M = 2.9, SD = 1.1$), $t(33) < 0.001$. There was no significant difference, however, in how participants rated their understanding of the lessons between groups, ($M = 5.4, SD = 1.5$) in the *personalized lessons groups* and ($M = 5.0, SD = 1.4$) in the *randomized-but-relevant condition*, $t(36) = 0.32$. Nor was there a significant difference in how participants self-assessed the degree to which their gameplay was affected, ($M = 4.3, SD = 1.3$) in the *personalized lessons groups* and ($M = 4.1, SD = 1.3$) in the *randomized-but-relevant condition*, $t(36) = 0.31$. Participants in the *personalized lessons groups* rated the robot as smarter or more intelligent ($M = 4.7, SD = 1.8$) than participants in the *randomized-but-relevant condition* ($M = 3.5, SD = 1.6$), $t(36) = 0.03$.

6. DISCUSSION

This study assesses whether personalization in educational or assistive HRI produces beneficial behavioral changes in users. The data indicate that personalization can lead to behavioral differences in users that result in more successful human-robot interactions.

In this puzzle task, we saw a "one sigma," or one standard deviation, improvement in participants final puzzle solving time from those received *personalized lessons* over those receiving *randomized-but-relevant lessons*, see Table 1. One sigma is more than the mean standard deviation effect size that most software Intelligent Tutoring Systems produce compared to classroom education alone, 0.76 sigma [3]. This may be due to the nature of the puzzle game that was used for this experiment. Success in nonograms requires several layers of logical inference, to which participants who received personalized lessons caught on more quickly than those in the control groups. An early lead allowed these participants to progress faster and perhaps feel more motivated.

The self-report survey data indicate that participants did not report having more difficulty understanding the lessons

presented to them in the *randomized-but-relevant lessons group* than in either *personalized lessons group*. All three groups rated their level of understanding fairly highly: a mean of 5.4 across *personalized lessons groups* and 5.0 in the *randomized-but-relevant lessons group* out of 7, $t(36) = 0.32$, see Figure 6(a). It is notable that the *randomized-but-relevant lessons group* reported a relatively high understanding of the lessons despite performing poorly. This may indicate that this population was reluctant to admit when they did not understand something.

Judging by survey free-response data, participants in this study ranged greatly in their own evaluation of the usefulness of the lessons. One participant in the *randomized-but-relevant lessons group* reported that: "Lessons were repetitive and a little distracting, even frustrating." In the *personalized lessons conditions*, some reported positive feedback while others reported only frustration. A participant who received personalized lessons, in response to a question about whether the lessons affected his/her gameplay replied, "Not really. I just learned by seeing what worked & what did not." However, judging by performance data, many participants who reported disregarding the content of the lessons, seemed to benefit from them just as much as the others. Some participants in the personalized groups claimed to be unaffected by the lessons but applied a lesson's content more frequently immediately after receiving lessons. From these survey results we conclude that a student's perceived value of a lesson may contradict with the lesson's measurable impact.

The free-response survey data also offers insight into the perception and impact of "easy" lessons. Many participants reported that they found some lessons "easy," "obvious," or "very obvious;" however, providing those lessons may not have been a waste. Eight participants reported a variant of the following sentiment: "Most hints I had previously explicitly figured out, though I found myself more actively seeking the pattern(s) suggested by a hint." This result highlights the benefit of personalization based on behavior rather than self-reported preferences.

7. CONCLUSION

In this paper we investigate the role of personalization in educational/assistive HRI. We compare participants' puzzle solving times across four one-on-one robot tutoring conditions, two of which were personalized to the learning progress of individual participants and two which were not personalized. We find that participants who received personalized lessons outperformed participants who received non-personalized lessons in a pre-test/post-test performance metric. We present these results as evidence that relatively simple personalizations can yield significant benefits in educational/assistive human-robot interactions.

8. REFERENCES

[1] B. S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring." *Educational Researcher*, vol. 13, no. 6, pp. 4–16, 1984.

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

[3] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," *Educational Psychologist*, vol. 46, no. 4, pp. 197–221, 2011.

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

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

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

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

[8] E. Hyun, H. Yoon, and S. Son, "Relationships between user experiences and children's perceptions of the education robot," *5th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 199–200, 2010.

[9] M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke, "Towards a humanoid museum guide robot that interacts with multiple persons," *5th IEEE-RAS International Conference on Humanoid Robots*, pp. 418 –423, dec. 2005.

[10] J. K. Lee, R. Toscano, W. Stiehl, and C. Breazeal, "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, aug. 2008.

[11] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Long-term interactions with empathic robots: Evaluating perceived support in children," *Lecture Notes in Computer Science*, vol. 7621, pp. 298–307, 2012.

[12] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati, "The physical presence of a robot tutor increases cognitive learning gains," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Austin,TX: Cognitive Science Society, 2012.

[13] H. Kozima, C. Nakagawa, and Y. Yasuda, "Interactive robots for communication-care: A case-study in autism therapy," *IEEE International Symposium on Robot and Human Interactive Communication*, 2005.

[14] D. Leyzberg, E. Avrunin, J. Liu, and B. Scassellati, "Robots that express emotion elicit better human teaching," in *6th International Conference on Human-Robot Interaction*. New York, NY, USA: ACM, 2011, pp. 347–354.

[15] C.-H. Yu, H.-L. Lee, and L.-H. Chen, "An efficient algorithm for solving nonograms," *Applied Intelligence*, vol. 35, no. 1, pp. 18–31, 2011.