Effect of Robot Tutor Embodiment on Human Cognitive Gains

Natasha Bustnes
nbustnes@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Joon Jang
jiwoongj@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Prithu Pareek
ppareek@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Omkar Savkur
osavkur@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Ishraaq Shams
ishams@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

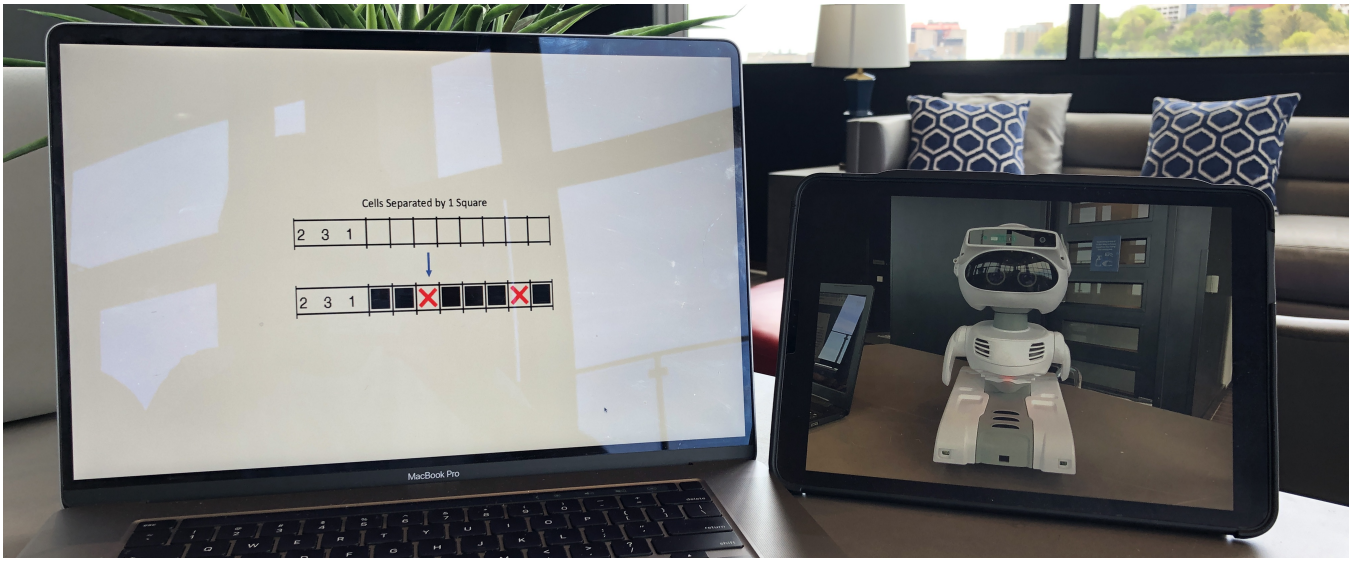


Figure 1: Visual Representation of a Robot Giving Advice on Nonograms

ABSTRACT

This paper replicated the study *The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains* by Leyzberg et al. [13] to explore the robustness of the results that a physically embodied robot tutor increases the cognitive gains of the participants. To measure this effect, participants completed 3 nonogram puzzles under three conditions: *(1) personalized visual advice with no robot, (2) personalized advice from a video representation of a robot, and (3) personalized advice from a physically present robot.* ANOVA tests were conducted to determine if there were significant differences in the improvement times from puzzle to puzzle between the three conditions. There was a significant improvement from the first puzzle to the last puzzle, but we did not have enough evidence to conclude that any difference exists in the improvement times between the conditions. This study was conducted in person during the COVID-19 pandemic, limiting the total number of participants to 15. Our participants were also limited to a relatively homogeneous population, mainly male undergraduate students at Carnegie Mellon University. Although this study was not able to show conclusive evidence of differences between cognitive learning gains as an effect of robot embodiment, robot embodiment is still an area worthy of continued research.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CMU 16467 HRI '21, May 2021, Pittsburgh, PA
© 2021 Copyright held by the owner/author(s).

CCS CONCEPTS

• **Human-centered computing** → *User studies*; *Auditory feedback*; **HCI theory, concepts and models**.

KEYWORDS

robot embodiment, feedback, cognitive gains

ACM Reference Format:
Natasha Bustnes, Joon Jang, Prithu Pareek, Omkar Savkur, and Ishraaq Shams. 2021. Effect of Robot Tutor Embodiment on Human Cognitive Gains. In *CMU 16467 HRI '21: Human Robot Interaction Replication Study, May 2021, Pittsburgh, PA*. 9 pages.

1 INTRODUCTION

Exploring the effects of embodiment in robot tutoring tasks is a key area of research in the field of Human-Robot Interaction. Conclusive results in this research can help determine how physically embodied robots can help in the education and elderly-care domains.

This research paper is a replication of the study The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains by Leyzberg et al. [13]. This study explored the effect of physical robot embodiment on participants in a cognitive skill learning task. The results showed that participants who received advice from a robot physically present outperformed participants in all other conditions. This yielded the conclusion that physical robot tutor embodiment positively correlates to learning gains. This finding can be implemented into our current education system for more effective tutoring schemes to improve learning. Leyzberg et al.'s [13] study introduced new ways of learning, and by replicating this study, we can confirm the evaluation. Our replication mainly differs from the original work in that there are significantly fewer test conditions and we use a different robot.

This study aims to replicate the result of the original study, that the mere physical presence of the robot shows a marked improvement in performance in cognitive learning, especially subliminally. Similar to the original study, cognitive improvement between puzzle are measured by the difference in solving times between the puzzles as well as self-reported measures of performance.

Our main research question was to see the effect, if any, of robot embodiment when advising participants. We predicted that our results would be similar to the original study's results, as we are replicating a subset of their conditions with a different robot. We hypothesized that participants would have increased learning when given advice from a physically embodied robot compared to a virtual representation or no robot at all.

2 BACKGROUND

Physically embodied robots provide obvious benefits to stakeholders in that they can perform tasks that involve physically manipulating their environment. Interestingly, however, there is significant evidence to believe that a human-robot interaction where the robot is physically embodied leads to improved reactions in non-manipulation tasks as well. In effect, the physical embodiment gives rise to the emergent phenomenon of improved interaction.

Research on robots in the context of assistive or rehabilitative therapy have shown this most explicitly, such as in studies conducted by Tapus et al. [19] where elderly individuals found conversing with a physically embodied Social Assistive Robot (SAR) agent to be more enjoyable than a virtual agent, or when Kozima et al.[11] and Pop et al. [16] found that a physically embodied robot proved successful in behavioral therapy and counseling for children diagnosed on the autism spectrum.

Even outside the contexts of disability and rehabilitation, multiple studies, both before and after Leyzberg et al.'s [13] work, have attempted to address whether the physical embodiment of a robot affects the interaction between its underlying system and human stakeholders. This effort encompasses explorations of more fundamental aspects of a users' subjective experience of a robot.

One such study, by Bainbridge et al. [1], found that people's perception of a robot agent with respect to trust and respect would be affected by a robot's embodiment, in that people engendered higher levels of trust and respect for the embodied robot, as opposed to a video display. Similarly, studies have shown other generally positive associations between embodied robots and anthropomorphism [10], and perception or helpfulness [22] versus an agent on a display, or one visualized in virtual space. These studies generally found that the physical embodiment and presence of a robot led to users perceiving the interaction with the underlying system to be more salient, lifelike, trustworthy, and helpful.

However, more recent work involving robot embodiment in more complex settings, with studies that aim to measure the effect of more complex interactions, has shown that the positive effects of robot embodiment may be context-dependent. Ceha et al. [3] demonstrated that an embodied robot agent which attempts to encourage curiosity to learn a new subject may cause information overload to its users when it attempts to explain the source of its curiosity. Short et al. [18] developed a SAR agent to encourage healthier food choices amongst children found that the adaptability of the robot's system to more complex levels of communication, such as humor, was most crucial for it to remain engaging to children throughout the study period. Meanwhile, Kidd et al. [9] found that people perceived an embodied robot agent to be not as engaging or social in a scripted exchange as opposed to another human, while more engaging than a virtual agent.

The fact that Leyzberg et al.'s [13] work demonstrated stronger learning effects from the physical embodiment of the robot agent connects it perhaps most closely with research work in HRI pertaining to education. The study followed those which investigated the role of robotic agents in the context of augmenting the learning effects from Intelligent Tutoring Systems (ITS) [15], and from Moundridou et al.'s [14] work which showed improved learning experience with a virtual agent vis-a-vis without, and brought physical embodiment as an important variable in the context of education.

Leyzberg et al.'s [13] study became widely cited as others sought to replicate effects of physical embodiment enhancing learning effects in domains other than puzzle games. Trinh et al.'s [20] RoboCOP demonstrated that a physically embodied robot speech coach provided both an enhanced coaching experience and improved performance results as compared to a virtual agent and real-time visualization of important speech-delivery metrics. Wijnen [23] similarly found that extending a learning system with a physically embodied SAR led to improvement in learning outcomes, while a later study by Leyzberg et al. [12] found the use of physically embodied robots led to enhanced results in a personalized learning context.

Subsequent work in the field of human-robot interaction regarding education has helped researchers better understand opportunities and constraints concerning the deployment of physically embodied robots in an educational environment. Serholt et al.'s [17] work revealed teachers' envisioning of robot teachers as augmenting the existing classroom structure, especially during group work, as well as concerns about equitable distribution of robot hardware through a survey study.

Meanwhile, Davison et al.'s [5] longitudinal study and Belpaeme et al.'s [2] synthesis of results from existing studies help provide nuance as to the attributes that physically embodied robots must have when used in settings with young children. In particular, the emotional and social aptitude of the agent underlying a physically embodied robot are frequently cited as important attributes in designing the human-robot interaction, with lack or overabundance of emotional and social support cited as distracting or otherwise detracting to the learning experience [2], [4], [7], [8], [21].

3 METHODS

3.1 Participants

There were a total of 15 participants in this study, all members of the Carnegie Mellon University community, and all either undergraduate or graduate students in the age range of 18-30. A majority of the participants - approximately two-thirds - identified as male, while the rest identified as female. Only a third of the participants indicated that they had prior experience solving nonograms and a majority indicated they had prior experience interacting with robots. Each participant was assigned to one of three groups: (1) *personalized visual advice with no robot*, (2) *personalized advice from a video representation of a robot*, and (3) *personalized advice from a physically present robot*. There were 5 participants in each group. There were no exclusion criteria for participants.

Figure 2: The Misty II robot used to provide advice to participants during the study

3.2 Nonograms

Like the original study by Leyzburg et al. [13], we chose to use the nonograms as the cognitive task of choice. This was for several reasons, including a desire to reach parity with the original study and because we believed nonograms to be fairly novel for most study participants, normalizing the difference in prior experience. Nonograms are commonly found in 10x10 or 8x8 grid formats, and we chose the latter as it provided puzzles with less inherent complexity.

Nonograms are a simple numbers-based game with some basic rules. The game board itself is a square grid of empty squares with the objective being to fill in the squares in the grid to respect the rules written on the side of the rows and on top of the columns. For a set of numbers next to a row, each number indicates the existence of a group (or stretch) of that many filled-in squares on that row. Note that there is at least one white square between each group of black squares.

For example, looking at the puzzle in Figure 3 the third row has the numbers "2 1 1". This means that the row must first contain a stretch of length two, followed by a stretch of length one, followed by another stretch of length one. Each of these stretches must be separated from each other by at least one white square. Notice how the groups do not necessarily have to start on the first column, or end on the last one.

These rules apply in a similar fashion to columns as well. The trick to figuring out a nonogram is to find a way to fill in the grid such that each rule on the rows and columns are all satisfied.

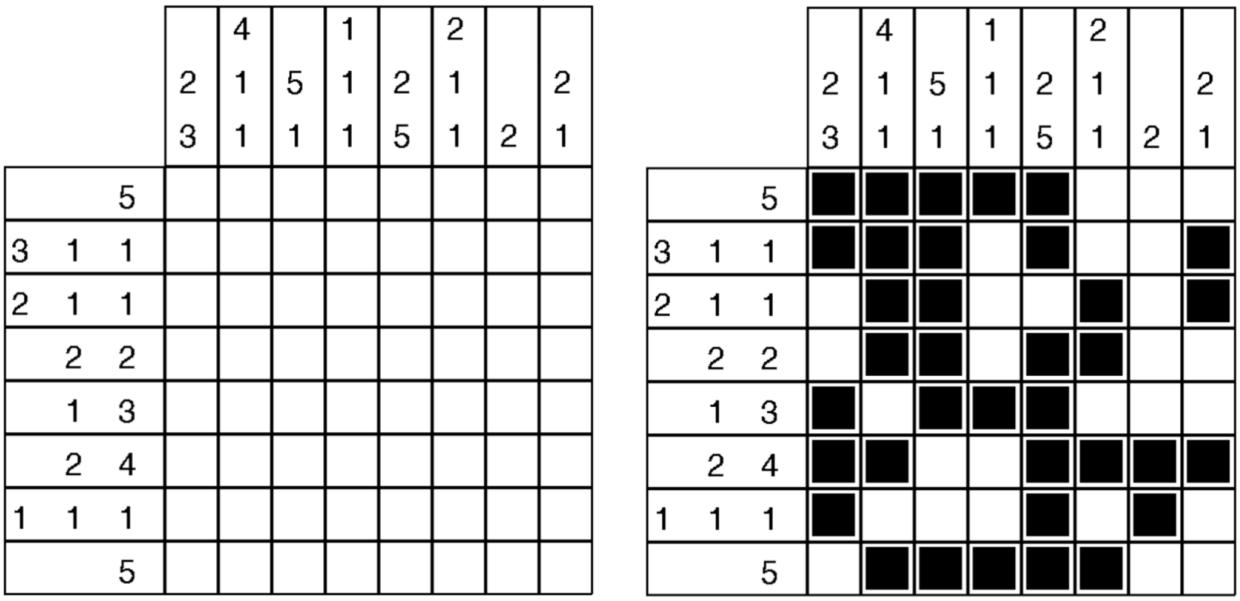
3.3 Study Design

During the user studies, participants were first asked to complete a pre-survey which asked them Likert-scale questions about their familiarity with robots in general and nonograms, as well as their comfort level with solving puzzles. Then, after a brief introduction to familiarize participants with the rules of nonograms, participants were instructed to solve 3 nonograms.

Participants had a time limit of 15 minutes for each with a 3-minute break in between puzzles. Similar to the study by Leyzburg et al. [13], the last puzzle was identical to the first, except it was rotated from the first puzzle by 90 degrees, functionally creating a nonogram of identical difficulty, but one which subjects would not necessarily recognize as such. Subjects were measured to see how much time they took for each puzzle. Puzzles timed out if more than 15 minutes had elapsed since starting the puzzle. In this case, the puzzle was taken to be completed at 15 minutes. After the puzzles, subjects were asked to offer their opinions of the advice and rate their performance in a post-survey.

3.4 Advice

While solving the three puzzles, participants were interrupted with advice multiple times either by a visual popup on the screen - in the *no robot* condition - or by a robot tutor - in the *video representation* and *physically present robot* conditions. The advice ranged from 7 to 26 seconds in length and consisted of a combination of visual and auditory advice. There was no auditory advice in the *no robot* condition.



(a) nonogram, unsolved

(b) nonogram, solved

Figure 3: An example of an unsolved and solved nonogram. The objective of the puzzle is to fill in the black squares in the grid to respect the rules laid out for each row and column. More details are in the Nonogram section.

Figure 4: Setup for the *physically present robot* condition

Figure 5: Setup for the *video representation of a robot* condition

In each puzzle, advice was given either when the participant had not made a move for 45 seconds or when they had completed 25%, 50%, and 75% of the puzzle, whichever came first. After the game provided three pieces of advice for a given puzzle, none would be given for the remainder of the puzzle. No advice was repeated more than once.

To give some personalization to the advice, we implemented heuristic functions for each advice that would check if our advice was applicable to the current puzzle state, and one advice from the list of applicable advice was chosen. If no valid advice existed, for the current board state, we randomly picked from a list of advice that had not been given to that participant yet.

The pieces of advice are ones that we felt were generally applicable to most nonogram puzzles and were found through online research. Most of them came from a European website on nonograms called *Techtonic Puzzel* [6].

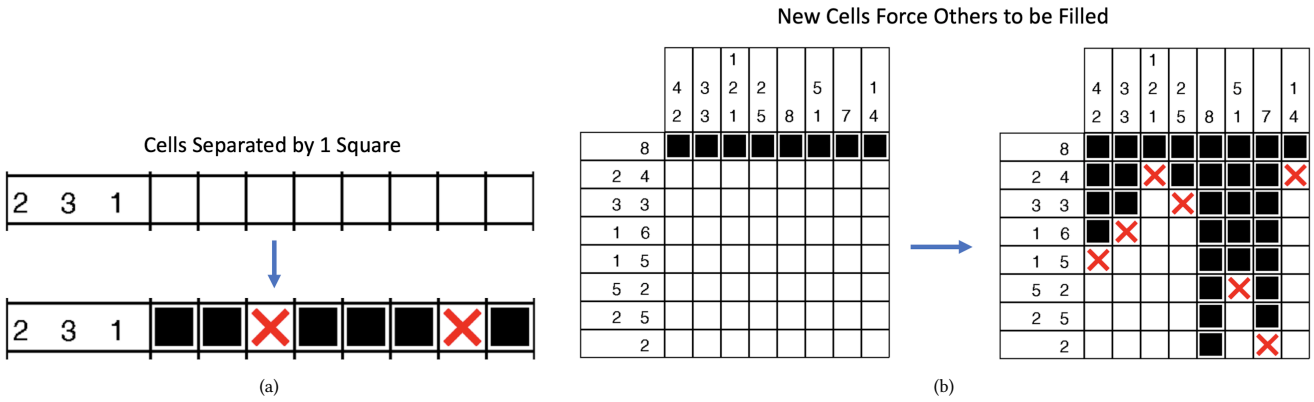


Figure 6: An example of some of pieces of advice given to participants while they are solving nonograms. These diagrams were also accompanied by verbal instructions by the robot in the *visual representation* and *physically present robot* conditions. An example of this verbal advice that the robot would give for 6(a) is as follows: "*Here's a piece of advice. For some rows and columns, if you add their numbers, plus any spaces that you have to have between the stretches and it equals 8, this means they're only one way to lay out the black squares. You can see in the example that the numbers add to 6, and we're required to have at least 2 crossed out squares, which add to 8.*"

For the *no robot* condition, the advice was simply a visual diagram displayed on the screen. For the *video representation* and *physically present robot* conditions, the visual advice was also accompanied by verbal instructions given by the robot.

3.5 Robot Tutor

For both of the *robot* conditions the Misty II robot (Figure 2) was used to provide advice to the participants as they were solving the nonograms. Before the first puzzle was shown, Misty would greet the subject then turn her head to face the computer screen while the participant solved puzzles.

Whenever it was time for Misty to give the participant advice, she would turn her head to face the participant, her light would turn from green to red, and she would start verbally giving the player advice. Once the advice was over, her light would turn back to green and she would turn her head to face the puzzle on the computer screen once again.

For the *physically present robot* condition, Misty would be sitting directly next to the computer screen at an angle facing the participant as shown in Figure 4. The audio for the advice would play directly through her speakers.

For the *video representation* condition, an image of Misty was displayed on an external monitor, with her voice played through a Bluetooth®speaker behind the monitor as seen in Figure 5.

4 RESULTS

Overall, as shown in Table 1, participants took an average of 532 seconds to complete the first puzzle, improved to 375 seconds for the second, and showed a slight decline in performance to 398 seconds for the third puzzle. Participants in the *no robot* condition seemed to have the greatest overall improvement in performance, followed by the *physically present robot* and then the *video representation* condition. However, participants in the *video representation* condition had the fastest completion for the first puzzle, while the *no robot* condition had the slowest completion time for this puzzle. Additionally, it is interesting to see that all conditions except video had a decline in performance between the second and third puzzles although there was still an overall improvement across all the puzzles.

To verify whether or not there was an improvement in the amount of time between the puzzles, we conducted a one-sided t-test with the null hypothesis being that there is no meaningful improvement between the two puzzles, and the alternative hypothesis is that the mean improvement is greater than zero. The resulting p-values of the t-test can be seen in Table 3.

At level of $\alpha = 0.05$, according to the table of p-values, it can be seen that there is enough evidence that participants with *no robot* condition and *physically present robot* condition improved from the first puzzle to the second and third puzzle. However, there is not enough evidence that participants in the video condition improved from puzzle to puzzle since the t-test yielded a p-value greater than $\alpha = 0.05$. Interestingly, across every condition, it can't be determined whether there was an improvement in time between the second and third puzzles. Looking at Table 1, the participants generally did worse from puzzle 2 to puzzle 3, explaining why the p-values are much higher.

Now that it has been shown that in some of the cases, there was an improvement from puzzle to puzzle, we would like to see whether the times of improvement differ among the three conditions. To accomplish this, several ANOVA hypothesis tests were conducted to analyze the improvement time between the puzzles. Tables 1, 3, and 4 show the results of the ANOVA tests for the improvements in time between puzzles 1 and 2, puzzles 2 and 3, and puzzles 1 and 3.

The null hypothesis for the ANOVA tests for the tables above is that the mean improvement times for each condition are the same, while the alternative hypothesis is that at least one of the conditions

Table 1: Mean Puzzle Solve Times, Improvements Between Puzzles

	Puzzle 1 (sec)	Puzzle 2 (sec)	Puzzle 3 (sec)
Overall Average	532.8±273.0	375.3±212.8	398.0±302.6
No Robot	645.6±302.6	383.4±211.6	472.6±394.9
Video Representation	475.8±285.2	412.0±296.0	388.6±293.8
Physically Present Robot	477.0±253.4	330.4±147.9	332.8±256.4

	1 - 2 Improvement (sec)	2 - 3 Improvement (sec)	1 - 3 Improvement (sec)
Overall Average	+157.5±346.2	-22.±369.9	+134.8±407.6
No Robot	+262.2±369.1	-89.2±448.0	+173.0±497.5
Video Representation	+63.8±411.0	+23.4±417.1	+87.2±409.5
Physically Present Robot	+146.6±293.4	-2.4±296.0	+144.2±360.4

Table 2: Survey Results

Pre-Survey	Prior Experience With Robots	Prior Experience with Nonograms	Comfort with Puzzles
Overall Average	4.5±1.8	2.5±1.9	5.3±1.1
No Robot	5±1	2.4±1.9	5.6±1.1
Video Representation	5±2	2.6±1.5	5.2±1.1
Physically Present Robot	3.6±2.1	2.6±2.6	5±1.2

Post-Survey	Self-Rating of Performance	Advice Helpfulness
Overall Average	4.5±1.5	4.07±1.8
No Robot	4.4±1.7	4.6±1.1
Video Representation	4.6±1.1	4.6±2.3
Physically Present Robot	4.6±1.8	3±1.6

Table 3: P-Values for One-Sided T-Test

	1 - 2 Improvement	2 - 3 Improvement	1 - 3 Improvement
No Robot	3.36e-05	0.815	0.012
Video Representation	0.171	0.421	0.224
Physically Present Robot	7.26e-05	0.514	0.032

Table 4: ANOVA Tests for Advice Helpfulness in Post-Survey Responses

Post-Survey Responses				
Source	dof	SS	MS	F
Treatments	2	8.5	4.27	1.406
Error	12	36.4	3.03	
Total	14	44.9		
P-value	0.28262			

has a different mean improvement time than the other conditions. Since we are testing at a level of $\alpha = 0.05$, we will reject the null hypothesis if the p-value retrieved from the ANOVA test is lower than the threshold of 0.05. From the results of the ANOVA tests in Table 5, the p-values for each of the ANOVA tests are nowhere close to the threshold. For example, the p-value for the test analyzing the improvement from puzzle 1 to puzzle 3 was about 0.93. Hence, the null hypothesis cannot be rejected, meaning that there is not enough evidence to claim that at least one of the conditions has a different mean improvement time than the others.

Looking at the pre-survey responses that the participants filled before and after solving the puzzles in Table 2, we see that people had similar prior experiences with robots with an average response of 4.5 out of 7. The participants also came in to the study with little experience with nonograms. In response to the pre-survey question rating their prior experience with the puzzle, the average response was around a 2.5 out of 7 on the Likert scale of 1-7. Although most participants didn't have too much experience with nonograms, when asked to rate themselves on how comfortable they are with puzzle games in general, the participants rated themselves highly, with an average response of 5.3 out of 7.

Table 5: ANOVA Tests for Improvement between Puzzles

Improvement from Puzzle 1 to Puzzle 2				
Source	dof	SS	MS	F
Treatments	2	99302.9	49651.5	1.101
Error	12	541386.8	45115.6	
Total	14	640689.7		
P-value	0.36404			

Improvement from Puzzle 2 to Puzzle 3				
Source	dof	SS	MS	F
Treatments	2	34797.7	17398.9	0.104
Error	12	2011551.2	167629.2	
Total	14	2046348.9		
P-value	0.90221			

Improvement from Puzzle 1 to Puzzle 3				
Source	dof	SS	MS	F
Treatments	2	19066.8	9533.4	0.0685
Error	12	1670977.6	139248.1	
Total	14	1690044.4		
P-value	0.93419			

For the post-survey results, on a Likert scale of 1-7, the participants in the *no robot* condition self-rated their performance at an average of 4.4, and the participants in the *video representation* and *physically present robot* conditions both self-rated their performance at an average of 4.6. This result is not significant, as there was a high standard deviation for all three distributions (2.8 for *no robot*, 1.3 for *video representation*, and 3.3 for *physically present robot*).

In addition to the mean improvement for each condition, an ANOVA test was used to determine whether the mean response to helpfulness of the advice from post-survey are the same or not. The results of this hypothesis test is in Table 4. Since the p-value is not less than the level of significance of 0.05, we fail to reject the null hypothesis that the mean survey responses for advice helpfulness are different from one another.

To make some sense of the results of the ANOVA tests, the box plot for the improvement time between puzzles 1 and 2 is shown in the Figure 7. It can be seen that the median value of the three conditions are different from each other, with the video condition having an improvement value of around 60 seconds, while the participants in the robot conditions having an average improvement of around 220 seconds. However, the variation of the data in the three conditions was relatively large and overlapped with each other. Since the variation of the values within the three conditions was large, it is less likely for the ANOVA test to result in rejecting the null hypothesis.

5 DISCUSSION

There was no significant difference in the improvement time of completing a nonogram between the three conditions. We do not have enough evidence to prove that any difference in improvement

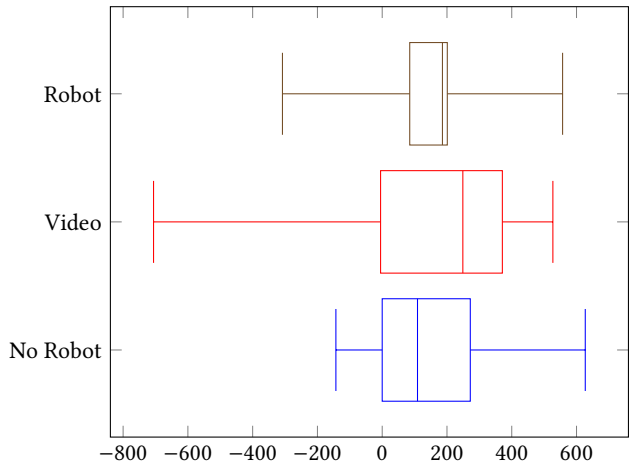


Figure 7: Boxplot for Improvement between Puzzles 1 and 3

times was due to the conditions. We can not conclude our hypothesis, as we predicted that the improvement time of the participants in the *physically present robot* condition would be more than the improvement time of the participants in the *video representation* and *no robot* conditions. In the paper by Leyzberg et al. [13] that we were trying to replicate, they found a significant difference between the improvement times of the participants in the *physically present robot* condition compared to the other conditions. Leyzberg et al. [13] had 100 participants, split into 5 conditions, with each completing 5 nonogram puzzles, resulting in a lower variation in improvement times within a particular condition. The participants in the original study completed more nonograms than in our study, allowing the researchers to collect more data. It also provide the participants with the opportunity to increase their familiarity with nonograms by the last puzzle, which would reduce variation in the data.

Past studies that have also looked at the effect of embodiment on learning have found that the physical embodiment of the robot can have mixed results on participants. Short et al. [18] found that a SAR agent had to be complex to deal with emotions, while Ceha et al. [3] and Kidd et al. [9] found that a robot agent may not be the most engaging with participants. On the other hand, Trinh et al. [20], Wijnen et al. [23], and another study by Leyzberg et al. [12] determined that physically embodied robots enhanced learning in the participants. Our results suggest that the robot might not have engaged with the participants enough to keep them interested and follow the advice it was giving. Perhaps the visual appearance of the robots used in the other studies made the participants more perceptible to the robots' advice.

The participants in the *video representation* and *physically present robot* conditions rated their own performance to be slightly higher than the participants in the *no robot* condition, albeit this result was not significant. Nevertheless, this difference was expected, as the robot provided hints that may have boosted the confidence of the participants.

However, there was a significant change in the completion time between the third and first puzzle for the *no robot* and *physically*

present robot condition, which means that those participants learned how to solve nonograms by the end of their three puzzles. However, the improvement from the first puzzle to the second puzzle was more pronounced in these cases. This is most likely because the participants had more experience with nonograms on the second puzzle and were beginning to understand the ideas behind the pieces of advice. There is also the possibility that fatigue started to affect the participants during the third puzzle, as the game code allocated 15 minutes for each puzzle before timing out. Having spent possibly the past 30 minutes working through nonograms, participants on their last puzzle might have been feeling burned out from all the strenuous thought required to determine how to fill in the grid, despite the fact that they had a 3 minute break between each puzzle.

While some participants did not feel tired after the first two puzzles, we suspect that the learning curve of nonograms plateaus very quickly, meaning that somebody with limited experience in the puzzle can understand enough to do well but not enough to improve significantly with only one more puzzle to work on. Perhaps with more experience, the participants could improve their nonogram-solving abilities.

From the post-survey feedback, while most participants liked the advice, several participants found that the personalized advice seemed to appear at random times, which ended up distracting the participants and caused them to lose their train of thought. This could have increased the completion times of the participants, as they had to pause and re-figure out what their next move was going to be once the advice section completed. There was a varied response for how helpful the robot and advice given was. Some thought that the order in which the advice was given could have been optimized, as they felt the more useful hints came only in the second puzzle. One participant said, *"Misty distracted me when I was making moves at times. A warning might have been nice before the robot spoke."* A general consensus was that the advice in the first two puzzles was useful, but the advice in the last round was not helpful. Some participants wanted Misty to interrupt less. One even wrote, *"there should be an option to tap out of the advice if you already know it so it doesn't distract you."*

Another explanation for the lack of improvement between the second and third puzzle could be because the participants did not learn any more hints on the third puzzle. The participants most likely already knew the strategies that were presented on the third puzzle, rendering the advice useless towards the end. The participants figured out the strategies before the robot told them the same thing. This also explains why some people had a negative view of the robot. Their last memory of interacting with Misty before filling out the post-survey was Misty giving them unhelpful advice that they already knew and disrupting their train of thought.

5.1 Limitations

Our study was limited by the ongoing COVID-19 pandemic, so we could only recruit our friends and people that we were already in contact with to be participants in our study. We allocated a maximum of 15 minutes per puzzle before it timed out to minimize the time for the study. We decided to have each participant solve 3 nonograms because we wanted to try to be respectful of everyone's time, and we thought that each participant solving 5 would be too arduous and time-consuming, especially if each puzzle could take a maximum of 15 minutes. Therefore, each trial in the study was allocated to a maximum of 1 hour, including set-up time and breaks. Both of these factors resulted in us being able to gather a total of 5 participants per condition. We were not able to reduce our variation in the completion time with so few participants. The distributions of the improvement times by condition (see Figure 7) seem to be somewhat identical in the ANOVA test because the variation of completion time was so large.

Our participants also came from a relatively homogeneous population, as most of our social circles overlap in demographics. The participants were mainly undergraduates, between the ages of 18 and 30, with a few Ph.D. students. Our participants consisted of 11 males and 4 females. Considering both of these factors, as well as the number of participants, we are not able to generalize our results to a larger population, especially since we did not sample enough people to satisfy the large-enough condition in the ANOVA test.

It is also worth noting that there was quite a bit of variation in the difficulty of the nonograms. Some participants found the puzzles too easy, while others found them too difficult. Since the first and third puzzles were reflections of each other, some participants were unable to complete either puzzle. In the post-survey, one participant noted that the *"puzzles were decently easy"*, while another wrote that the *"first and third puzzles were very hard (impossible? tempted to say impossible but I couldn't figure out how to prove it so I lean towards thinking I was missing something), middle one was easy."* The nonograms were selected to be difficult enough to take several minutes to complete but easy enough to be completed within 15 minutes. The actual difficulty of each puzzle differed based on the participant.

We also had a variability in location, as we could not get permanent a setup to run our experiment over the course of two weeks. This meant that we had to use whichever rooms were available, which might have introduced the variation in the participants' improvement times. Two participants completed their puzzles remotely, as they were not able to make it to the testing location. In these cases, we gave them remote access to the puzzles through Zoom. In one of these cases, for the *video representation* condition, we fed a live feed of Misty through the Zoom camera. These changes might have increased the difficulty of completing a nonogram, since there was a delay between clicking the mouse and the game responding.

6 CONCLUSION

The goal of this study was to determine if robot embodiment affected peoples' cognitive gains. Specifically, this study looked at their improvement in performance while solving the puzzle nonogram. Although we found some differences between the increase in participants' learning between the three conditions, they were not statistically significant. Rather than this study throwing doubt on Leyzberg et al.'s [13] study methodology and results, it is more likely that some of several confounding factors are to blame. These could range from the aforementioned study limitations, like limited sample size - resulting in us not being able to obtain results with any statistical power, to the design and timing of the hints.

In future studies, in an effort to reduce these confounds, participants should be given a warning when a hint is about to be given. Additionally, the order of the advice should be changed, and the number of advice possibilities should be increased. This would reduce the likelihood of a scenario in which the advice given is no longer useful during the third puzzle. Finally, the difficulty and type of puzzles should be varied in order to ensure that the results of Leyzberg et al. [13] are robust enough to be transferred to another learning task.

ACKNOWLEDGMENTS

To Professor Henny Admoni and the 16467 TAs, for helping shape and guide this project.

REFERENCES

[1] Wilma A. Bainbridge, Justin Hart, Elizabeth S. Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Munich, Germany, 701–706. https://doi.org/10.1109/ROMAN.2008.4600749

[2] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science Robotics* 3, 21 (Aug. 2018), eaat5954. https://doi.org/10.1126/scirobotics.aat5954

[3] Jessy Ceha, Nalin Chhibber, Joslin Goh, Corina McDonald, Pierre-Yves Oudeyer, Dana Kulić, and Edith Law. 2019. Expression of Curiosity in Social Robots: Design, Perception, and Effects on Behaviour. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300636

[4] D. Leyzberg, E. Avrunin, J. Liu, and B. Scassellati. 2011. Robots that express emotion elicit better human teaching. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 347–354. https://doi.org/10.1145/1957656.1957789 Journal Abbreviation: 2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI).

[5] Daniel P. Davison, Frances M. Wijnen, Vicky Charisi, Jan van der Meij, Vanessa Evers, and Dennis Reidsma. 2020. Working with a Social Robot in School: A Long-Term Real-World Unsupervised Deployment. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Cambridge United Kingdom, 63–72. https://doi.org/10.1145/3319502.3374803

[6] Danny Demeersseman. 2020. Nonogram Solving Techniques. http://www.tectonicpuzzel.eu/nonogram-solving-techniques-griddler-tips.html

[7] J. Kennedy, P. Baxter, and T. Belpaeme. 2015. The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 67–74. Journal Abbreviation: 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI).

[8] Kristin S. Jordan, Roxanna Pakkar, and Maja J Mataric. 2019. Improving Robot Tutoring Interactions Through Help-Seeking Behaviors. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, New Delhi, India, 1–6. https://doi.org/10.1109/RO-MAN46459.2019.8956370

[9] C.D. Kidd and C. Breazeal. 2004. Effect of a robot on user perceptions. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, Vol. 4. IEEE, Sendai, Japan, 3559–3564. https://doi.org/10.1109/IROS.2004.1389967

[10] Sara Kiesler, Aaron Powers, Susan Fussell, and Cristen Torrey. 2008. Anthropomorphic Interactions with a Robot and Robot–like Agent. *Social Cognition - SOC COGNITION* 26 (April 2008), 169–181. https://doi.org/10.1521/soco.2008.26.2.169

[11] H. Kozima, C. Nakagawa, and Y. Yasuda. 2005. Interactive robots for communication-care: a case-study in autism therapy. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE, Nashville, TN, USA, 341–346. https://doi.org/10.1109/ROMAN.2005.1513802

[12] 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, Bielefeld Germany, 423–430. https://doi.org/10.1145/2559636.2559671

[13] Daniel Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. 2012. The physical presence of a robot tutor increases cognitive learning gains, Vol. 34. Issue: 34.

[14] M. Moundridou and M. Virvou. 2002. Evaluating the persona effect of an interface agent in a tutoring system. *Journal of Computer Assisted Learning* 18, 3 (Sept. 2002), 253–261. https://doi.org/10.1046/j.0266-4909.2001.00237.x Publisher: John Wiley & Sons, Ltd.

[15] Roger Nkambou, Jacqueline Bourdeau, and Valéry Psyché. 2010. Building Intelligent Tutoring Systems: An Overview. In *Advances in Intelligent Tutoring Systems*, Roger Nkambou, Jacqueline Bourdeau, and Riichiro Mizoguchi (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 361–375. https://doi.org/10.1007/978-3-642-14363-2_18

[16] Cristina A. Pop, Ramona E. Simut, Sebastian Pintea, Jelle Saldien, Alina S. Rusu, Johan Vanderfaeillie, Daniel O. David, Dirk Lefeber, and Bram Vanderborght. 2013. Social Robots vs. Computer Display: Does the Way Social Stories are Delivered Make a Difference for Their Effectiveness on ASD Children? *Journal of Educational Computing Research* 49, 3 (Oct. 2013), 381–401. https://doi.org/10.2190/EC.49.3.f

[17] Sofia Serholt, Wolmet Barendregt, Iolanda Leite, Helen Hastie, Aidan Jones, Ana Paiva, Asimina Vasalou, and Ginevra Castellano. 2014. Teachers' views on the use of empathic robotic tutors in the classroom. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Edinburgh, UK, 955–960. https://doi.org/10.1109/ROMAN.2014.6926376

[18] Elaine Short, Katelyn Swift-Spong, Jillian Greczek, Aditi Ramachandran, Alexandru Litoiu, Elena Corina Grigore, David Feil-Seifer, Samuel Shuster, Jin Joo Lee, Shaobo Huang, Svetlana Levonisova, Sarah Litz, Jamy Li, Gisele Ragusa, Donna Spruijt-Metz, Maja Mataric, and Brian Scassellati. 2014. How to train your DragonBot: Socially assistive robots for teaching children about nutrition through play. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Edinburgh, UK, 924–929. https://doi.org/10.1109/ROMAN.2014.6926371

[19] Adriana Tapus, Cristian Tapus, and Maja Mataric. 2009. The role of physical embodiment of a therapist robot for individuals with cognitive impairments. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Toyama, Japan, 103–107. https://doi.org/10.1109/ROMAN.2009.5326211

[20] H. Trinh, R. Asadi, D. Edge, and T. Bickmore. 2017. RoboCOP: A Robotic Coach for Oral Presentations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (June 2017), 1–24. https://doi.org/10.1145/3090092

[21] Paul Vogt, Mirjam de Haas, Chiara de Jong, Peta Baxter, and Emiel Krahmer. 2017. Child-Robot Interactions for Second Language Tutoring to Preschool Children. *Frontiers in Human Neuroscience* 11 (2017), 73. https://doi.org/10.3389/fnhum.2017.00073

[22] Joshua Wainer, David J. Feil-Seifer, Dylan A. Shell, and Maja J. Mataric. 2007. Embodiment and Human-Robot Interaction: A Task-Based Perspective. In *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Jeju, South Korea, 872–877. https://doi.org/10.1109/ROMAN.2007.4415207

[23] Frances M. Wijnen, Daniel P. Davison, Dennis Reidsma, Jan Van Der Meij, Vicky Charisi, and Vanessa Evers. 2020. Now We're Talking: Learning by Explaining Your Reasoning to a Social Robot. *ACM Transactions on Human-Robot Interaction* 9, 1 (Jan. 2020), 1–29. https://doi.org/10.1145/3345508