Theory of Mind Assessment with Human-Human and Human-Robot Interactions

Trent Rabe[1(✉)], Anisa Callis[2], Zhi Zheng[3], Jamison Heard[3], Reynold Bailey[3], and Cecilia Alm[3]

[1] University of Colorado, Boulder, USA
trent.rabe@colorado.edu
[2] Pennsylvania State University, State College, USA
[3] Rochester Institute of Technology, Rochester, USA

Abstract. Human-robot interaction has played an increasingly significant role in more recent research involving the Theory of Mind (ToM). As the use of robot facilitators increases, questions arise regarding the implications of their involvement in a research setting. This work addresses the effects of a humanoid robot facilitator in a ToM assessment. This paper analyzes subjects' performances on tasks meant to test ToM as those tasks are delivered by human or robot facilitators. Various modalities of data were collected: performance on ToM tasks, subjects' perceptions of the robot, results from a ToM survey, and response duration. This paper highlights the effects of human-robot interactions in ToM assessments, which ultimately leads to a discussion on the effectiveness of using robot facilitators in future human-subject research.

Keywords: Social robotics · Human-robotic interaction · Theory of Mind

1 Introduction

Humanoid robots are increasingly applied as Socially Assistive Robots (SARs), such as tutors, assessment/intervention facilitators, and companions [1]. Using SARs as facilitators has become more relevant, due to the potential for lower cost, reducing social stigma, and better accessibility. However, previous studies have indicated that information delivery is different between robots and humans, which raises concerns about the underlying bias when comparing human- with robot-facilitated experiments [2]. This prompts important questions surrounding the proliferation of SARs, including how effectively SARs can be used in human subjects research. This paper researches the effectiveness of humanoid robot facilitators using Theory of Mind studies. The *Theory of Mind (ToM)* is the ability to understand and reason about the mental state of another person, including one's emotions, beliefs, and intentions [3]. ToM has been extensively explored in human-robotic interaction research, but has typically focused on

M. Kurosu (Ed.): HCII 2022, LNCS 13303, pp. 564–579, 2022.
https://doi.org/10.1007/978-3-031-05409-9_41

AI methods for providing a robot with ToM [4,5], instead of determining if a subject's ToM is impacted during robot interactions. There is also a lack of objective data in ToM research that relates the facilitator type (human vs. robot) to task performance. Understanding this multimodal data will provide valuable insight about the effectiveness of SARs as facilitators of ToM interactions.

This paper focuses on using a robot facilitator to deliver a ToM scenario in order to answer the following research questions: (1) Does performance change when ToM tasks are delivered by humans vs. robots? (2) How does participants' self-perceived ToM differ from their measured ToM? (3) How does participants' perception of the robot delivering a task change the participant's performance within the experiment? (4) How is the participants' response time affected by human or robot facilitators?

2 Related Work

Theory of Mind has been thoroughly researched over the past four decades after it was first introduced by Premack and Woodruff [3]. In more recent years, many ToM studies are focused on child development, specifically from ages 3 to 5 [6]. Wellman et al. [7] found that children under the age of 3 continuously failed the false-belief task due to their lack of conceptual understanding of the world. The same study found that children aged 4 and up were able to regularly pass false belief tasks they were given. Bernstein et al. [8] did find that in certain cases, middle and older aged adults have more errors on false belief tasks than children.

Theory of Mind has been extensively studied in children, but there has also been a few studies about Theory of Mind in adults. Much of this ToM research in adults has been geared towards understanding the declination of ToM capabilities over one's lifetime. There has been a debate on this topic which Pardini and Nichelli addressed in "Age-Related Decline in Mentalizing Skills Across Adult Life Span" [9]. In their work, they selected four groups of participants based on ages (range: 25–75) to complete ToM tasks. Their results "found that performance of the early middle-aged people was not significantly different from that of the young group, whereas both had better scores if compared with the other two groups. We also found that the old middle-aged group presented a deficit in this domain if compared with the two younger groups but was less impaired if compared with the elderly subjects." These results suggest that there is a peak in ToM capabilities in early to middle adulthood.

However, the debate continues as other studies come to the conclusion that ToM capabilities do deteriorate with age. In their study: "Social Understanding: How does it Fare with Advancing Years?" where they tested older adults ages (60–82 years old) and younger adults ages (20–46) Sullivan and Ruffman found that "The younger group performed significantly better than the older group on the theory of mind stories ... There was a trend for the younger group to do better on the control stories as well, although the effect failed to reach significance" [10].

The research presented in this paper analyzes ToM in early adulthood. According to previous research, the results found by this study are unlikely to

be marred by developing or declining ToM in participants. This sort of change in the participants' abilities to understand others' mental states would pose issues if one participant were to be further along in the development process than another, indicating higher ToM scores, when in reality, they were just more mature in that regard. This is not an issue for the results of this paper as the subject pool was all in the young adult category from Pardini and Nichelli's work [9].

2.1 False-Belief Tasks

The Theory of Mind is commonly tested through false belief tasks, which were introduced by Wimmer and Perner in 1983 [11]. The first false-belief task was administered to children between three and nine years old; each given two pictures. The first picture showed a main character placing an object in 'x' place. The second picture allowed the children to witness the object being moved to 'y' place in the absence of the main character. The children were then asked where the main character would look for the object. This is a false-belief task because the child participants know where the object actually is while the main character in the story does not.

There are two types of false-belief tasks: first- and second-order. A first-order false-belief task measures "An individual's ability to comprehend false beliefs of others" [12]. A commonly used first order false belief task is called the "Smarties." A child is asked to predict what another child thinks is in the Smarties box. In reality, there is a pencil in the box [13]. A second-order false-belief tasks is "what an individual thinks about another person's thoughts" [12]. A commonly used second-order false-belief task is the Perner and Wimmer (1985) "ice-cream van story." Mary is the only character who knows the actual location of the ice cream truck and is asked to predict where John thinks the ice cream truck is located. John knows the former location of the truck but not the current location. Though Mary knows where the truck truly is, she must consider if what John knows is different from what she knows [11]. A variation of the original false belief task is the "Sally - Anne Task" [14]. This task expands the original task by using props to visually demonstrate the scenario, instead of showing story images. Most recent works are variations of this task, depicted in Fig. 6.

2.2 Theory of Mind Inventory

The Theory of Mind Inventory survey (ToMI:SR-A) is a 60 question survey designed to establish a ToM score for the test taker [15]. The results are described as 'self-assessed' as each participant provides answers to subjective questions about themselves. The answers are recorded on a 20-unit continuum similar to the traditional Likert-type scale. The subject is asked to rank how well they relate to each question on this scale. The answers are recorded for each question and averaged over all 60 questions, providing each subject a score of 0–20, with

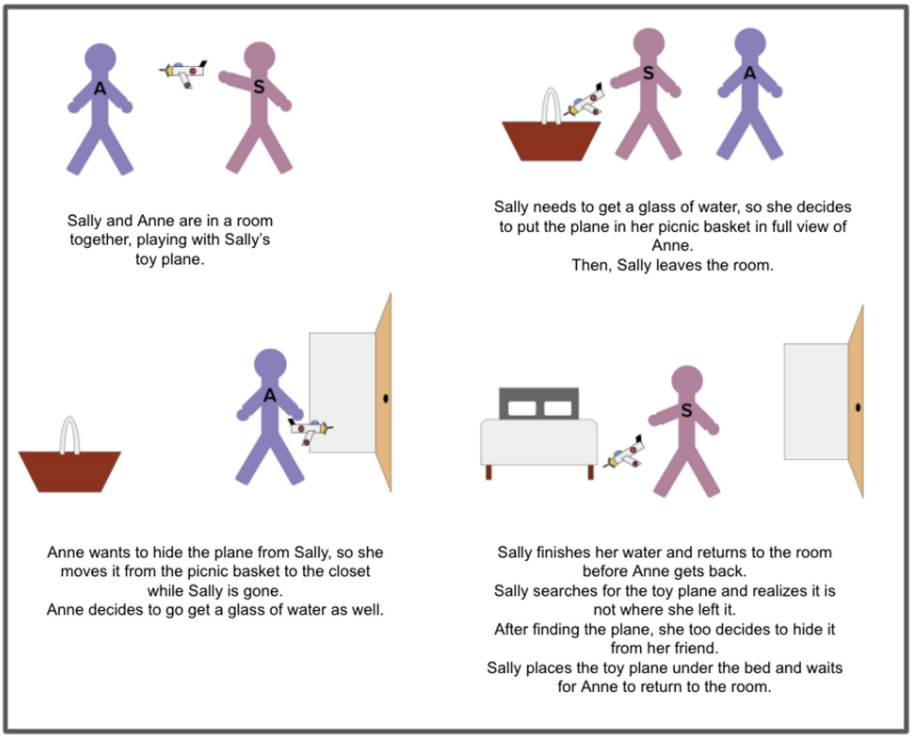

Fig. 1. Photos that participants were shown in video tasks with human and robot facilitators. This is the "Sally-Anne" task. The object being moved is the toy plane. It is first moved into the basket, then it is moved twice with the absence of one of the characters. This gives enough movement of the object to ask the participants feasible questions for the false-belief tasks.

accuracy to one decimal place (e.g. 17.4). Neurotypical subjects with well developed ToM are expected to score at or above a 15 out of 20 on the ToMI:SR-A survey [15].

2.3 Testing Robot Theory of Mind Capabilities

Recently, there has been some ToM research involving robots. One study used robots as the facilitators in their scenarios [4], while other studies have used robots as subjects who are supposed to complete first-order false belief tasks [16–18]. Another study expanded the previous research by using robot subjects who are intended to complete higher-order false belief tasks [4].

While these previous studies use robots, few of them directly investigate the way that humans react to robot facilitators in terms of subject performance and perception of the robot facilitator. The previously mentioned studies also do not compare the performance differences between human and robot delivered tasks.

To the best of our knowledge, very little prior work has been done to identify the differences between human and robot facilitators on false-belief tasks. It has been suggested that robot facilitators could create less subject/facilitator bias when it comes to asking questions for false-belief tasks. One study found that when children were asked questions pertaining to a false-belief tasks, they answered incorrectly because they felt they were being questioned about their knowledge instead of being asked about the beliefs of the character in the story. Using a robot facilitator eliminated this feeling within the children participants [19]. There have, however, been some studies on how humans react to robot facilitators in scenarios and interactions outside of Theory of Mind [20]. The presented study analyzes audio data of subjects when answering questions from robot facilitators vs human facilitators, which allows researchers to directly observe the impact the robot facilitator has on subject performance.

The Godspeed Robot Perception Questionnaire assesses how people perceive the presence of humanoid robots in their environments using 5 overall categories: Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety [21]. One study had subjects interact with a lifelike humanoid robot and then complete the Godspeed Robot Perception Questionnaire. The results indicated that the more lifelike a robot was, the lower subject perceptions were. [20].

3 Methodology

False belief tasks are often used to assess an individual's ToM, where a participant is asked to identify a character's beliefs in a story when the participant knows those beliefs to be false [12]. The presented within-subjects experiment used such a task and manipulated facilitator types (human or robot).

3.1 Experiment Protocols

The research was approved by the Rochester Institute of Technology Institutional Review Board (ethics review). Various measures were built into the experiment to protect participant confidentiality and the experimental validity. All subjects completed an informed consent form prior to their participation. Subject and facilitator names were also changed before video recording began to help deidentify the data, due to the virtual nature of the experiment.

Thirty human subjects (mean age: 22.1 years; standard deviation: 4.0) completed four false-belief task scenarios. Each task was delivered virtually via Zoom by a recorded human or a NAO robot facilitator (see Fig. 2) and were accompanied by story-line pictures. After the story was told, the participants were asked questions about the characters' beliefs. The first two tasks were considered easy,

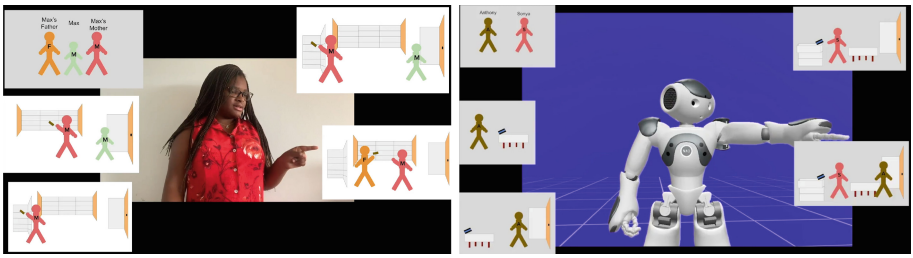

Fig. 2. Human and humanoid robot facilitators presenting false-belief scenarios to subjects who are tasked with identifying characters' beliefs in the story.

while the last two were considered difficult. The experiment manipulated the facilitator type (human or robot), where participants were randomly assigned a facilitator ordering. Each participant completed one easy and one hard task with a robot and the remaining tasks with a human. A standard ToM instrument survey [15] was administered prior to scenario completion to rate an individual's ToM. The participant's task performance (correctly answered questions), response duration times, and speech rate (number of words divided by response duration) were analyzed for each task. The subjects also rated their perceptions of the humanoid robot at the end using the Godspeed Questionnaire [21].

In each of the four tasks, pictures appeared on the screen as facilitators reached certain parts within the story. The images stayed on the screen for the duration for the task, including while the questions were being asked. This helps prevent the experiment from becoming memory based, as the participants are able to refer to the images so that they do not forget the order of the scenes in the story. The easy tasks were designed with two questions that were fairly straightforward for the subject to answer correctly. The harder tasks were designed with 5 questions that required the subject to think deeper about what they had heard in the video and how it compared to what they knew themselves.

3.2 NAO Robot Facilitator

The study used a virtual NAO robot. Figure 3 depicts Choregraphe, the graphical programming interface for the NAO. The drag and drop programming function was used to control the robots arm and head motions in order to have the NAO closely mimic the human facilitators' movements.

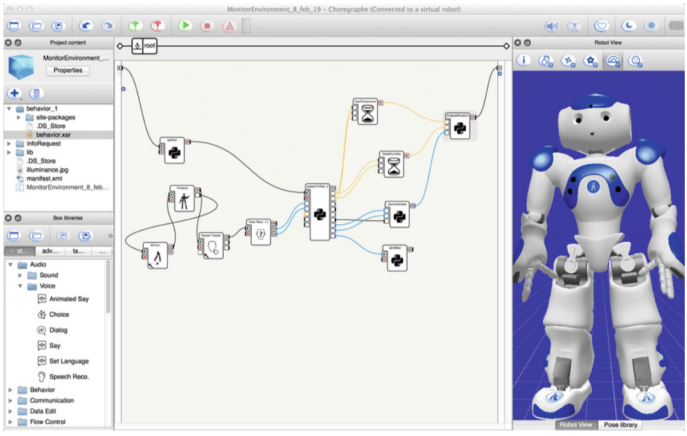

Fig. 3. Image of drag and drop program used to program the NAO.

4 Results

The first research question was: ***Does performance change when ToM tasks are delivered by humans vs. robots?*** This was addressed by comparing the number of correct answers each participant got on each robot facilitated task to the number of correct answers of each participant on the human facilitated task for each task. Comparing the number of correct answers a participant gave as the tasks were facilitated by a human or by a robot allowed for an analysis of changes in performance as the facilitator type changed. A Mann-Whitney U-test (see Table 1) found no significant difference for task performance between the robot and human facilitators for individual tasks. In addition, there was no significant difference between the percentage of correct answers for human- (median = 86%) and robot-facilitated tasks (median = 86%). These findings indicate that a robot-facilitated ToM study yields comparable task performance results to the same study using human facilitators.

Table 1. Mann-Whitney U-test results for robot vs. human.

Task	p-value	Median robot	Median human	w-value
1	0.39	1.5	2.0	227.0
2	0.65	2.0	2.0	221.0
3	0.80	5.0	5.0	226.0
4	0.92	4.0	4.0	229.5

Figure 4 shows the comparison between the number of correct answers on robot and human facilitated tasks for task 1. Comparing the performance on both types of facilitators allows the difference to be visible.

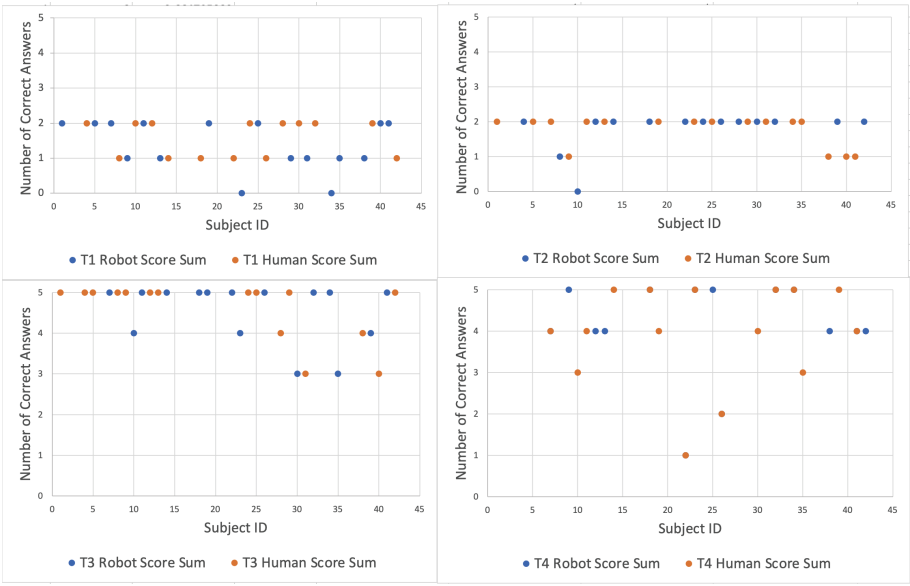

Fig. 4. Comparison of the number of correct answers per subject when the task was facilitated by a human vs. by a robot. Participants answered two questions on tasks 1 and 2 and five questions on tasks 3 and 4.

On task 1, 16 participants were shown a robot facilitator and 14 participants were shown a human facilitator. 50% of subjects correctly answered both questions from task 1 with the robot facilitator. 57% of subjects correctly answered both questions from task 1 with the human facilitator. For the robot facilitated task $M = 1.4$ and $SD = 0.74$ and for the human facilitated task $M = 1.57$ and $SD = 0.51$. Subjects scored, on average, higher on task 1 with human facilitators than with robot facilitators. This means that the participants were more likely to get one or even two questions correct for the first task when it was administered by a human. Although this preliminary data suggests that human facilitated tasks yield better results than robot facilitated tasks, this is likely not the case. A Mann-Whitney U-test was performed on this data and found that there was no difference between the type of facilitator (human vs. robot) and the scores on the task questions, as shown in Table 1.

On task 2, 14 participants were shown a robot facilitator and 16 participants were shown a human facilitator. 86% of subjects correctly answered both questions from task 2 with the robot facilitator. 75% of subjects correctly answered both questions from task 2 with the human facilitator. For the robot facilitated task M = 1.8 and the SD = 0.73 and for human facilitated task M = 1.75 and SD = 0.44. 16% more subjects correctly answered both questions for task 2 when shown with a robot facilitator.

On task 3, 15 participants were shown a robot facilitator and 15 participants were shown a human facilitator. 67% of subjects correctly answered all five questions from task 3 with the robot facilitator. 73% of subjects correctly answered all five questions from task 3 with the human facilitator. For the robot facilitated task M = 4.5 and SD = 0.76 and for human facilitated task M = 4.6 and SD = 0.74. More subjects correctly answered all five questions when given the human facilitator vs the robot facilitator.

On task 4, 15 participants were shown a robot facilitator and 15 participants were shown a human facilitator. 33% of subjects correctly answered all five questions from task 4 with the robot facilitator, whereas 40% of subjects correctly answered all five questions from task 4 with the human facilitator. For the robot facilitated task M = 4 and SD = 0.92 and for the human facilitated task M = 3.93 and SD = 1.2. More subjects correctly answered all five questions when given the robot facilitator vs. the human facilitator.

The second research question was: ***How does participants' self-perceived ToM differ from their measured ToM?*** This was addressed by comparing the percentage of correct answers each subject got on the robot facilitated tasks to their composite score on the Theory of Mind Self-Assessment. The same was also done to compare the percentage of correct answers each subject got on the human facilitated tasks to their composite score on the Theory of Mind Self-Assessment. This comparison was used because if a subject passed the false-belief task, that showed they were able to comprehend the mental state of another person. The more questions a subject gets correct, the more likely they are to have the Theory of Mind capabilities [22].

Figure 5 shows the comparison between the combined percentage of correct answers each subject got on the robot facilitated tasks vs. their composite score on the Theory of Mind Self Assessment. All participants were shown 2/4 task videos with robot facilitators. The Theory of Mind 60 question survey is a self-assessment. Therefore, there is always the risk of subject bias. Comparing that score to how well they do on the false-belief task will provide a more accurate prediction of a subject's Theory of Mind level.

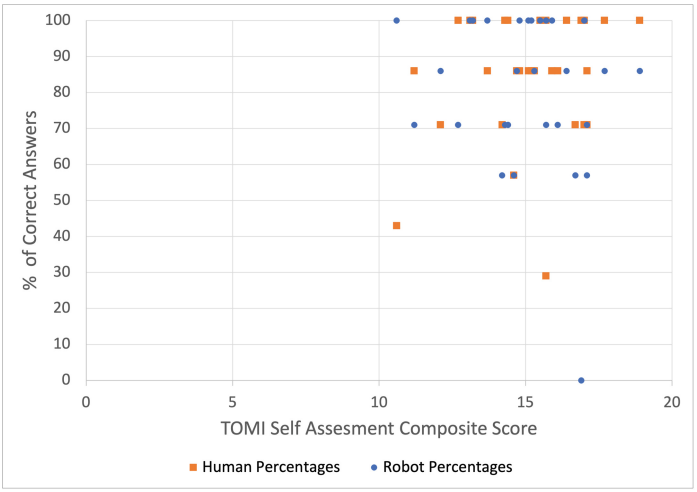

Fig. 5. Comparison of the percentage of correct answers per subject on two of four tasks with a robot facilitator, versus on two of four tasks with a human facilitator. The self-assessed composite score does not appear to have an effect on the percentage of correct answers and vice versa. The lowest percentage of correct answers (20%) does not correlate to the lowest composite score (11.19).

A Spearman's correlation test determined the relationship between task performance and the ToM inventory scores. No significant correlation was found for the robot-facilitated or human-facilitated sessions (see Table 2). These findings suggest that the use of a robot facilitator was comparable to a human facilitator.

Our third research question was: ***How does participants' perception of the robot delivering a task change the participants' performance within the experiment?*** This was addressed by comparing the subject composite score on the Theory of Mind Self-Assessment to the average rating they gave the humanoid robot based on the Godspeed Robot Perception Questionnaire. We hypothesized that how a subject viewed the humanoid robot in the experiment could affect their Theory of Mind performance in the given tasks. We also hypothesized that how a subject perceived the humanoid robot could positively or negative affect their performance on false-belief tasks given a robot facilitator.

Table 2. Spearman's rank correlation analysis comparing the percentage of correct answers on robot/human facilitated tasks with subjects' ToM Composite score.

Spearman values	Robot facilitated tasks vs. ToM composite score	Human facilitated tasks vs. ToM composite score
r - value	−0.12	0.12
p - value	0.53	0.51

Figure 6 shows the comparison between the subjects' Theory of Mind composite score and their average rating of the humanoid robot. It was thought that a higher rating of the humanoid robot would indicate a higher Theory of Mind composite score. For the average robot perception rating M = 3.45 and SD = 0.57. For the Theory of Mind composite score, for males M = 15.14 and SD = 1.86, for females, M = 15.16 and SD = 2.16. A Spearman's rank correlation was done with these two variables and found p = 0.46 and r = 0.14. These findings suggest no significant correlation exists between the average robot perception ratings and the Theory of Mind composite scores.

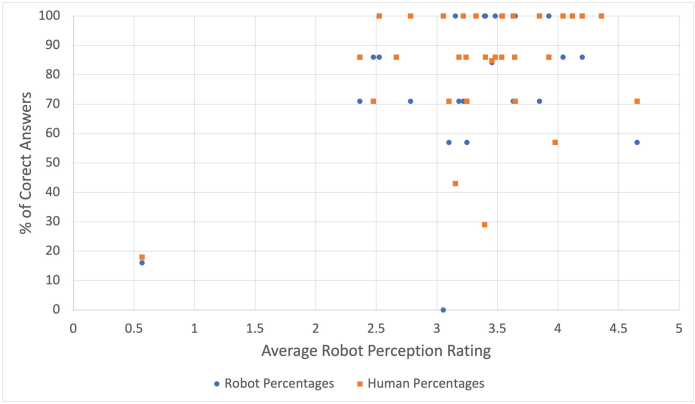

Fig. 6. Comparison between combined percentage of correct answers each subject got on the two out of four tasks they were shown which were robot facilitated and their average rating of the humanoid robot.

The self-assessed composite score does not appear to have an effect on the average robot perception rating and vice versa. The lowest robot rating of 2.36 does not correlate to the lowest composite score (11.19). 40% of subjects correctly answered all questions for robot facilitated videos. For the average robot perception rating, M = 3.45 and SD = 0.57. For percentage of correctly answered questions for robot facilitators M = 84.2 and SD = 16, These findings suggest no significant correlation between the average robot perception rating and the Theory of Mind composite score. The self-assessed composite score does not appear to have an effect on the percent of correct answers on robot facilitated tasks. The lowest percentage of correct answers on the robot facilitated tasks 57% does not correlate to the lowest composite score (2.36)

The fourth research question was: ***How is participants' response time affected by human or robot facilitators?*** By assigning timestamps to the beginning and end of each response, we were able to obtain the average response duration. A Mann-Whitney U-Test found no significant difference (w = 895, p = 0.77) between the participants' average response duration for robot (median = 18.35 s) and human (median = 19.30 s) facilitators. No significant correlation

was found between the response duration and task performance for robot (r= −0.03, p = 0.82) and human (r= −0.01, p = 0.94) facilitators.

Figure 7 shows average response times for each task as they were delivered by a human or a robot. Over the span of four tasks, response time decreased. Average response times for task 1 were higher than those for task 2, etc. Originally, we thought that the response time for tasks 3 and 4 would be greater than response times for tasks 1 and 2 because they were designed to be more difficult, but this was not the case, likely due to the participants becoming acquainted with the process and thus answering quicker.

No significant difference was found (w = 971, p = 0.41) between the participants' speech rates for robot (median = 0.87 words/sec) and human facilitated tasks (median = 0.74 words/sec). There was also no significant correlation between speech rate and task performance for robot (r = −0.03 , p = 0.88) and human facilitated tasks (r = 0.12, p =0.53). These findings show comparable results between human- and robot-facilitated tasks.

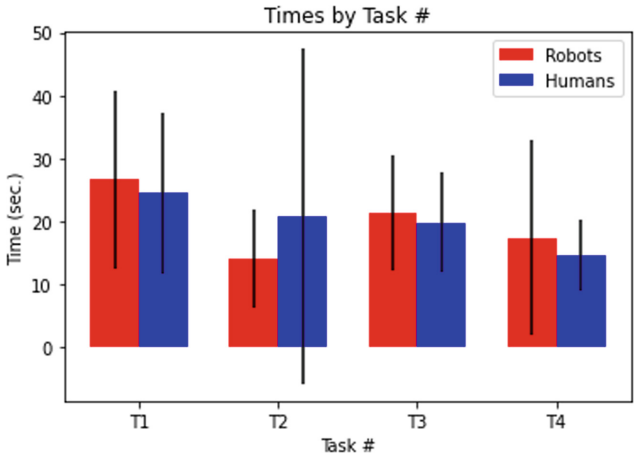

Fig. 7. The difference between response times on human and robot facilitated tasks. The black error bars are meant to illustrate the range of answers. One outlier was disregarded in the timestamp analysis as they had an average response time on task 2 of 104 s, which is over a minute longer than the mean.

5 Discussion

This study sought to investigate four research questions concerning a SAR facilitating first- and second-order ToM false-belief tasks. The first research question focused on determining if a NAO robot facilitator yields similar results to a human facilitator. Results showed that participant performances on the false-belief tasks were comparable between the two types of facilitators. This result

suggests that robots can be effectively used as facilitators in future human subject research on ToM and potentially other similar research topics. Using robot facilitators can increase consistency among facilitation as robot facilitators are able to recite the same script in exactly the same manner between trials, unlike a human facilitator whose delivery may naturally vary. Recognition bias is also likely to decrease with the use of robot facilitators similar to the NAO robot because of its lack of human facial features [19]. Often, robot facilitators are easy to access and require less human work hours to set up than it would take for a human facilitator to actually run the experiment. While these robots are unlikely to replace human facilitators, they can still be used effectively to augment current approaches to human subject research.

Another important question is whether the participant's self-perceived ToM, as measured by means of a subjective questionnaire, differs from their measured ToM (performance on false-belief tasks) for robot and human facilitators. We expected that a significant correlation would exist for both facilitators; however, no significant correlation was found. This result indicates that there may be a disassociation between a person's self-perceived ToM and their actual ToM for both human and robot facilitators. This lack of correlation yields little to no additional information on the effectiveness of a robot facilitator.

The third research question investigated whether a participant's perception of a humanoid robot is related to their perceived ToM. No significant relationship was found, which indicates that a person's perception of a humanoid robot does not negatively impact their perceived ToM. This result further demonstrates that robots can deliver false-belief tasks as long as the human has a positive perception of the robot. Non-humanoid robots may not produce similar results.

The last question determined how question response times and participant speech rates are impacted by facilitator type, as the two objective metrics may indicate how facilitator type impacts a participant's objective response. No significant differences existed for either metric, which again further demonstrates that robots can deliver false-belief tasks with similar results to humans.

These results are preliminary. The size of the subject pool was limited to 30 people. This is partially due to the time frame of the experiment and limited resources for participant compensation. With more subjects, the results would be more precise. Future work should address this problem. Additionally, the average Theory of Mind composite score did not yield the expected results. According to the official Theory of Mind Instrument Documentation for the survey administered, there were some scores for both male and female participants that were below the cutoff score (2 male, 1 female for a cutoff of 13/20). The cutoff score is defined as the minimum score to be achieved such that the subject still possesses ToM. Three participants below the cut-off score may be attributed to the experiment's virtual design. Subjects tended to lose interest and pay less attention to the 60-question survey, which could be attributed to the virtual administration. An experimenter had to share their screen with the survey to the participant, who filled out their answers on a virtual form.

The results from the Godspeed Robot Perception Questionnaire were unexpected. None of the responses had an average below two out of five, which means that, on average, participants were not made uncomfortable by any aspect of the robot and its actions. Previous work suggests that when a human robot becomes more lifelike, people's perception rating of it will decrease. This is called the uncanny valley theory [20,23]. The concept suggests that as the appearance of the robot becomes more human they seem more familiar until a point when the response from an observer quickly drops from positive into strong revulsion. In many instances, things that are close to lifelike or closely mimic humans are expected to have human qualities but the robots are often not capable of this [24]. The robot in this experiment (NAO) is not lifelike enough to seem to mimic human reality. As a result, the subjects in this study did not have a negative reaction to the robot. If another robot had been used, it is likely that the participants would have reacted differently. Typically the NAO robot scores highly on the Godspeed Robot Perception Questionnaire [25,26].

Another reason why the ratings of the robot were high is the subjects' focal points. After completing the false-belief tasks, many subjects remarked that they had been so focused on following along with the videos that they paid less attention to the robot. This would change their conscious perception of the robot; though it may not have changed their responses or subconscious reactions. The balance of attention (looking at the robot vs. watching the videos) is something which should be addressed and mitigated in future work.

The NAO robot likely will not fully replace human researchers, but there is evidence supporting the effectiveness of its usage in augmenting human-subject studies. Participants reacted positively to the robot, and they performed comparably when false-belief tasks were delivered by human or robot facilitators. The robot-facilitated tasks did not yield longer response times or different speech rates. There was a significant difference in measured and surveyed ToM performances, but this did not correspond to the robot facilitator as subjects performed similarly on the ToMI survey regardless of the facilitator they were later shown. Overall, it is evident that robot facilitators yield the same results as human facilitators in human subject research on ToM.

6 Conclusion

This paper analyzed whether human subjects' performance on false belief tasks changed for human and robot facilitators. The research further examined how results from a standard ToM survey [15] correlated to true/false belief tasks performance. Additionally, speech data was analyzed to compare response duration and speech rates between human- and robot-facilitated tasks. Overall, no significant difference was found between the human and robot facilitator. The study's result supports the feasibility of using a robot facilitator for ToM tasks, which allows robots to be used more widely in related application scenarios, such as psychological/pedagogical tests and interventions.

Acknowledgement. This material is based upon work supported by the National Science Foundation under Award No. IIS-1851591. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Tironi, A., Mainetti, R., Pezzera, M., Borghese, N.A.: An empathic virtual caregiver for assistance in exer-game-based rehabilitation therapies. In: 2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH), pp. 1–6. IEEE (2019)
2. Ko, S., et al.: The effects of robot appearances, voice types, and emotions on emotion perception accuracy and subjective perception on robots. In: Stephanidis, C., Kurosu, M., Degen, H., Reinerman-Jones, L. (eds.) HCII 2020. LNCS, vol. 12424, pp. 174–193. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60117-1_13
3. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behav. Brain Sci. **1**(4), 515–526 (1978)
4. Thellman, S., Silvervarg, A., Ziemke, T.: Some adults fail the false-belief task when the believer is a robot. In: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, 479–481 (2020)
5. Dissing, L., Bolander, T.: Implementing theory of mind on a robot using dynamic epistemic logic. IJCA **I**, 1615–1621 (2020)
6. Cadinu, M.R., Kiesner, J.: Children's development of a theory of mind. Eur. J. Psychol. Educ. **15**(2), 93–111 (2000)
7. Wellman, H.M., Cross, D., Watson, J.: Meta-analysis of theory-of-mind development: the truth about false belief. Child Dev. **72**(3), 655–684 (2001)
8. Bernstein, D.M., Thornton, W.L., Sommerville, J.A.: Theory of mind through the ages: older and middle-aged adults exhibit more errors than do younger adults on a continuous false belief task. Exp. Aging Res. **37**(5), 481–502 (2011)
9. Pardini, M., Nichelli, P.: Age-related decline in mentalizing skills across adult life span. Exp. Aging Res. **35**, pp. 98–106 (2009)
10. Sullivan, S., Ruffman, T.: Social understanding: how does it fare with advancing years? Br. J. Psychol. **95**(1), 1–18 (2004)
11. Wimmer, H., Perner, J.: Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition **13**(1), 103–128 (1983)
12. Ward, T., et al.: False-belief task. In: Encyclopedia of Autism Spectrum Disorders, pp. 1249–1249 . Springer, New York (2013). https://doi.org/10.1007/978-1-4419-1698-3_91
13. Gopnik, A., Astington, J.W.: Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. Child Development, pp. 26–37 (1988)
14. Baron-Cohen, S., Leslie, A.M., Frith, U.: Does the autistic child have a "theory of mind"? Cognition **21**(1), 37–46 (1985)
15. Hutchins, T.L., Prelock, P.A., Lewis, L.: Technical Manual of the Theory of Mind Inventory: Self Report-Adult (ToMI:SR-A). Unpublished Manuscript. Available at theoryofmindinventory.com (2019)

16. Arkoudas, K., Bringsjord, S.: Toward formalizing common-sense psychology: an analysis of the false-belief task. In: Pacific Rim International Conference on Artificial Intelligence. Springer (2008) 17–29. https://doi.org/10.1007/978-3-540-89197-0_6
17. Breazeal, C., Gray, J., Berlin, M.: An embodied cognition approach to mindreading skills for socially intelligent robots. Int. J. Robot. Res. **28**(5), 656–680 (2009)
18. Sindlar, M.P., Dastani, M.M., Dignum, F., Meyer, J.-J.C.: Mental state abduction of BDI-based agents. In: Baldoni, M., Son, T.C., van Riemsdijk, M.B., Winikoff, M. (eds.) DALT 2008. LNCS (LNAI), vol. 5397, pp. 161–178. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-93920-7_11
19. Baratgin, J., Dubois-Sage, M., Jacquet, B., Stilgenbauer, J.L., Jamet, F.: Pragmatics in the false-belief task: let the robot ask the question! Front. Psychol. **11**, 3234 (2020)
20. Haring, K.S., Matsumoto, Y., Watanabe, K.: How do people perceive and trust a lifelike robot. In: Proceedings of the World Congress on Engineering and Computer Science, vol. 1, Citeseer (2013)
21. Bartneck, C., Kulić, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int. J. Soc. Robot. **1**(1), 71–81 (2009)
22. Campbell, D. et al.: Theory of mind. In: Encyclopedia of Autism Spectrum Disorders, pp. 3111–3115. Springer, New York (2013). https://doi.org/10.1007/978-1-4419-1698-3
23. Mori, M., MacDorman, K.F., Kageki, N.: The uncanny valley [from the field]. IEEE Robot. Autom. Mag. **19**(2), 98–100 (2012)
24. Ho, C.C., MacDorman, K.F.: Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. Comput. Hum. Behav. **26**(6), 1508–1518 (2010)
25. Lehmann, H., Rojik, A., Hoffmann, M.: Should a small robot have a small personal space? Investigating personal spatial zones and proxemic behavior in human-robot interaction (2020)
26. Thunberg, S.: Investigating the social influence of different humanoid robots. Undergraduate thesis, Linköping University (2017)