Effects of Different Interaction Contexts when Evaluating Gaze Models in HRI

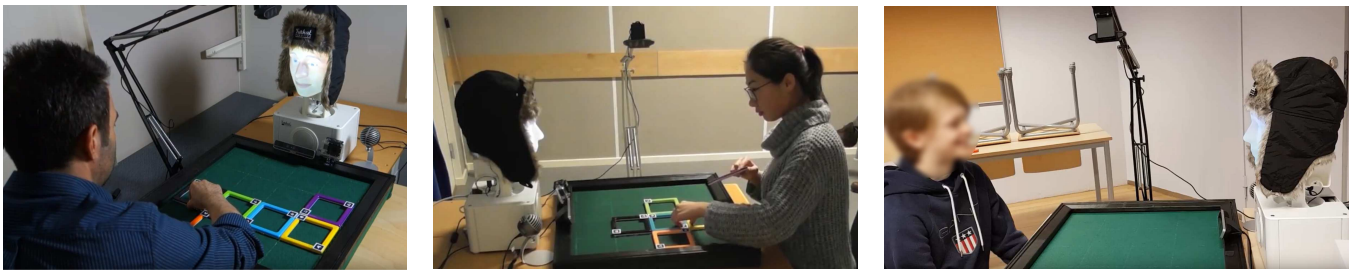
André Pereira
atap@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Catharine Oertel
TU Delft
Delft, Netherlands

Leonor Fermoselle
TNO
Den Haag, Netherlands

Joseph Mendelson
Furhat Robotics
Stockholm, Sweden

Joakim Gustafson
KTH Royal Institute of Technology
Stockholm, Sweden



(a) External Observers (b) Adults in the Lab (c) Children in School

Figure 1: The three different contexts that were studied within the same task.

ABSTRACT

We previously introduced a responsive joint attention system that uses multimodal information from users engaged in a spatial reasoning task with a robot and communicates joint attention via the robot's gaze behavior [25]. An initial evaluation of our system with adults showed it to improve users' perceptions of the robot's social presence. To investigate the repeatability of our prior findings across settings and populations, here we conducted two further studies employing the same gaze system with the same robot and task but in different contexts: evaluation of the system with external observers and evaluation with children. The external observer study suggests that third-person perspectives over videos of gaze manipulations can be used either as a manipulation check before committing to costly real-time experiments or to further establish previous findings. However, the replication of our original adults study with children in school did not confirm the effectiveness of our gaze manipulation, suggesting that different interaction contexts can affect the generalizability of results in human-robot interaction gaze studies.

Permission to make digital or hard copies of all or part 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 components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HRI '20, March 23–26, 2020, Cambridge, United Kingdom
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6746-2/20/03…$15.00
https://doi.org/10.1145/3319502.3374810

CCS CONCEPTS

• **Human-centered computing** → **User studies**.

KEYWORDS

Joint attention, mutual gaze, social robots, social presence

ACM Reference Format:
André Pereira, Catharine Oertel, Leonor Fermoselle, Joseph Mendelson, and Joakim Gustafson. 2020. Effects of Different Interaction Contexts when Evaluating Gaze Models in HRI. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20), March 23–26, 2020, Cambridge, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3319502.3374810

1 INTRODUCTION

While there are some exceptions, most studies in human-robot interaction that propose new advances in robot perception, cognition or behavior generation, neglect to test their technology in different scenarios, user groups or contexts. This affects the validity of the generalization of the results for each technology manipulation as these changes profoundly affect how users evaluate the robots and their own experience in the interaction.

In [25], we contributed to human-robot interaction research in joint attention by proposing a responsive gaze system within a case study where a social robot collaborates with humans in finding the solution of a physical spatial reasoning task. We evaluated our gaze system within this task by performing a user study with adults in a laboratory setting that tested the following hypothesis:

- A robot's use of responsive joint attention gaze cues will increase perceptions of the robot's social presence.

This work revealed that the proposed gaze system caused positive social presence effects on the perception of the robot. In the present work, we attempted two reproductions of our initial evaluation. Using the same method (same robot, task, gaze system, and hypothesis), we investigated the reproducibility of our prior findings in two further contexts: (1) external observers watch videos of a human-robot interaction (manipulation of the perspective to address generalizability from a situated first-person perspective to an observer third-person perspective) and (2) children, at school, interact with the robot (manipulation of the participant population to address generalizability from adult findings to children). We expected that this demographic would be more engaged with the task and that repeating our initial study with children would result in clearer findings in support of our hypothesis. To minimize the impact of differing levels of cognitive development in younger children (e.g., potential differences in theory of mind between ages 5 to 10 [20]), we limited our recruitment in Study 2 to typically developed children of ages 10+.

2 RELATED WORK

If social interaction partners are mutually aware that they are attending to a common target, that phenomenon is generally described as joint attention. Joint attention can drive believable effective interactions between humans and robots [6], can make robots appear more transparent and competent [14], can make object referral appear more natural, pleasant and efficient [18] and improve task performance [23]. Joint attention can have several phases, beginning with mutual gaze to establish attention, proceeding to referential gaze to draw attention to the object of interest, and cycling back to mutual gaze to ensure that the experience is shared [1]. Favorable feelings for robots are further enhanced when eye contact or mutual gaze is used in combination with joint attention [37]. Social eye gaze behavior generation is not only significant for improving collaboration between people and robots [2] but also for creating the illusion of human-like appearance and behavior. In [22], Mutlu et al., implemented gaze behavior in a humanoid robot (Honda's ASIMO) to create natural human-like behavior for storytelling. They found that participants performed significantly better in recalling the robot's story when the robot looked at them more often. Their results also yielded demographic differences by showing that men and women evaluated the robot differently based on the frequency of gaze they received from the robot. In this paper, we attempt to further establish the benefits of a system that uses information from multimodal perception to model responsive joint attention and mutual gaze mechanisms in a social robot [25]. We demonstrate and reflect on how different contexts and demographics can affect robots' perception results when testing different gaze models. This can help strengthen the validity of our previous results and make them more generalizable.

While there are some examples of HRI research that test gaze systems in different experimental conditions, including different demographics, studies that replicate others' are still scarce. In fact, HRI findings are still hardly ever reproduced by other research groups [8] and also not often replicated for confirmation by the same group. In [15], the authors discuss how the replication crisis in psychology [7], where it was shown that out of 100 psychology studies only 39 could be replicated, might also propagate to the sub-field of HRI. Some contributors to the HRI community have alternatively performed multiple studies to improve the reliability of, and to further establish, their research. In [35], Wang et al. showed that robots that are described as having social functions are rated higher for emotional capabilities but not for cognition when compared to robots that are described as bearing an economic value. In their contribution, the authors replicated the same study three times with small variations to confirm this effect. Several studies have also been performed to support Mori's hypothesis [21] on the uncanny valley with robots [30–32]. These studies not only support that people rate highly human-like robots as eerie or uncanny, but also that people exhibit greater avoidance behaviors when encountering human-like robots [31]. In children, the uncanny valley effect was also confirmed, but in this age group, the effect did not translate to avoidance behavior [30]. In [13], the authors survey several studies that show the benefits of physically embodied robots when compared to virtually embodied screen characters and attempt to replicate some of their previous subjective and objective findings. However, the authors found that most of the results are difficult to replicate, as different robots, virtual agents, scenarios and interaction contexts are used in each study.

3 SOCIAL PRESENCE

In this paper, the main measure used to evaluate our responsive joint attention system is social presence. We posit that social presence is one of the most relevant measures that account for the perceived nature of social robots. Social presence is defined as the feeling of "being together with another" [5]. It has been positively correlated with greater enjoyment [12], performance, satisfaction [28, 33], trust [29] and a greater ability to persuade the user [16]. We are still far from having artificially intelligent entities that exhibit the same level of social presence we experience in human-human interactions [13]. However, when that feeling more closely matches the human counterpart and users believe that they are interacting with another social being, they will be engaged and entertained with the artificial entities for longer periods [26] and will consequently have higher intention to continue interacting with the artificial entities in the future [12]. Social presence is an extremely important measure for human-robot interaction. If we were to attribute perfect social presence (the same as a person) to a robot, we could conceive that many social psychology studies in human-human interaction would yield similar results in robots. We advocate using social presence measures to evaluate joint attention systems in robots as the intrinsic bidirectional nature of this measure makes it a sound choice for evaluating the effects of each type of joint attention individually [19]: initiating joint attention (IJA) and responding to others' joint attention (RJA). As such, in this paper, we continue to explore the relationship between the perception of social presence and joint attention in HRI.

4 SPATIAL REASONING TASK

In this section, we describe the task used to perform the studies described in this paper, the robot used in our experiments, its perception capabilities, its responsive gaze system and detail the semi-wizarded interaction.

Table 1: Examples of the most common dialog acts in MagPuzzle. Each dialog act contains 3 to 18 behavior implementations. For illustration purposes, a single behavior line is presented for each dialog act.

Dialog Act	Content	Target
Remind Objective	Remember {user_name}, we will need to build 4 walls a roof and a floor.	Player
Well Played	I could have not thought <gesture(happy)> of a better move myself.	Player
Poorly Played	Are you sure <gesture(skeptical)> you do not want to rethink that move?	Player
Probe Help	I could offer you my opinion if you'd like?	Player
Hint	I think putting a piece here might help.	Square
Good Square	I think this move makes a lot of sense.	Square
Wrong Piece	I would think of a different place for this piece.	Piece
Wrong Piece Explicit	This {piece_color} piece might not be optimally placed at this location.	Piece

4.1 Task Content

MagPuzzle[1] is a spatial reasoning cooperative task where a life-like social robot interacts with human participants. The participants' goal is to visualize a three-dimensional cube and reconstruct it in a two-dimensional space by placing and manipulating six different colored puzzle pieces in a board with a 4 by 4 grid. The task contains three different puzzles, with varying levels of difficulty, where the maximum amount of pieces in a line (row or column) allowed for the solution varies. Most participants have no trouble building the geometry net with a maximum of 4 pieces in a line but have more difficulties when they are limited to a maximum of 3 or 2 pieces in a line.

4.2 Robot's Role

The social robot introduces both itself and the task verbally (utilizing commercial quality synthetic speech) and cooperates in a peer-like manner with participants to arrive at solutions in all three different puzzles. The robot guides users towards a solution by using non-deterministic speech and deictic gaze behavior. The task was purposefully designed to stimulate the establishment of joint attention and so that the robot's guidance is more successful if participants are able to perceive the gaze cues expressed by the robot. However, the participant can voluntarily ignore the robot as the robot's guidance is not essential for the completion of the task.

4.3 Human-like robot embodiment

In our scenario, we use a back-projected robot head [3] placed next to a table that contains the puzzle board (see Figure 1). We decided to use a back-projected robotic head because they have been shown to be capable platforms for allowing gaze reading by humans as they allow for very fine-grained deictic gestures [9]. Our robot is capable of digitally animating quick eye gaze shifts and also utilizes physical servos in its neck to simulate head movements. Believable neck and eye gaze behavior are essential for a task such as ours, where we want users to assume the visual perspective of the robot. The robot can precisely look at different parts of the board because it stands in a fixed position and gaze shifts towards board positions were manually calibrated.

[1]Code and demo videos available on https://github.com/andre-pereira/MagPuzzle

4.4 Multimodal Perception

In front of the robot, we placed an Intel Realsense SR300 RGB-D camera that tracks the user's gaze direction and position. GazeSense[2] is used for gaze tracking and to estimate participants' visual focus of attention. For object detection, we use an open-source augmented reality toolkit[3] that can, in real-time, robustly recognize fiducial markers placed on the puzzle pieces used in our scenario. The augmented reality toolkit relies on a dedicated RGB camera placed on top of the board. For speech activity detection and speech recognition, a high-quality amplified microphone is placed on the side of the robot. Speech recognition and activity results are obtained from the real-time streaming Windows 10 speech recognition cloud service.

4.5 Responsive Gaze System

The robot's responsive gaze system is briefly described in this subsection and fully described in [25]. We previously suggested dividing gaze responsibilities between two concurrent parallel layers.

The proactive gaze layer is focused on generating proactive gaze behavior and a responsive layer sensitive to multimodal environmental cues. The proactive gaze layer is implemented both by a Wizard-of-Oz interface that allows an experimenter to trigger dialog acts and by an idle gaze module that controls the robot's gaze in moments of inactivity. Idle gaze shifts occur when the speech content associated with a dialog act finishes, or the timer for any previous gaze shift ends. The possible targets for idle gaze behavior in *MagPuzzle* are: looking at a player; looking at one of the four quadrants in the board; or looking at a random point in the environment, making the robot seem as if it is either distracted or averting its gaze.

The parallel responsive gaze layer is based on real-time multimodal perception data. This minimizes the robot's idle time and increases the usage of more meaningful responsive joint attention behavior. Users influence our robot's gaze system through speech, by changing their head orientation and by moving pieces on the board. When speech activity from a user is detected, the robot often looks at the user to simulate that it is paying attention. By using the gaze information obtained from our gaze tracker, the robot can look

[2]https://eyeware.tech/gazesense
[3]http://www.artoolkitx.org

at an approximate position of the board that the user is looking at and can look back at the player if the player shifts the gaze towards the robot. While engaged in mutual gaze, the robot also uses information from our gaze tracker to follow the user in real-time at 30 frames per second. Finally, when a participant moves, places or removes a puzzle piece from the board, the robot often assumes that the piece is the current focus of attention and gazes at that location.

4.6 Semi-Wizarded Interaction

The robot's gaze behavior is autonomously operated by the system described above. However, the robot's dialog flow is controlled in a wizard-of-oz manner where an experimenter uses a dedicated interface to trigger dialog acts. The experimenter has access to an interface that shows information about the state of the board, shows where the robot and the user are looking at and presents speech recognition results. We developed a helper artificial intelligence search algorithm (brute force, depth-first) that allows the wizard to know: the correct squares for providing hints to the user; if the board is in a proper or incorrect board state; or if any rule was broken. We also introduced simple rules that make interface elements appear when relevant and disappear when no longer necessary. When experimenters are faced with complex behavior arbitration problems, increased cognitive load can prove to be detrimental [8]. In order to mitigate this and reduce the wizard's cognitive load, it was important to introduce these interface optimizations and automated gaze behaviors.

4.7 Dialog Acts

Dialog acts in MagPuzzle are used to describe and manage the rules of the task, provide simple responses to users' questions, offer advice on what to play next, provide feedback on a move that was just played and motivate or compliment the user. The specific implementation of each dialog act allows for a mix of verbal and non-verbal content in a format that is compact for authoring (see Table 1). When a dialog act is selected, one of its multiple content entries is chosen and parsed by a behavior planner, similar to [27], that outputs the correct multimodal instructions at the correct time to the robot. The planner starts by vocalizing and lip-syncing the text but also supports additional commands that can be synchronized in the middle of the content such as gaze changes, gestures, pauses, or pitch and volume changes. A simple notation for variable replacement allows the authored content to contain dynamic interaction information such as the user's name or the current referred-to piece color. The final property of dialog acts in *MagPuzzle* is that they are associated with a gaze target. The robot first shifts its focus of attention towards the intended target and only then executes the associated content. If selected dialog acts are associated with a square on the board or a quadrant, the robot shifts its current focus of attention to the associated place on the board. In dialog acts not associated with the board, the robot changes its current attention target to the participant. The set of dialog acts used in the task were authored to reflect a friendly cooperative social robot that often smiles, displays facial expressions, calls participants by their names and can refer to puzzle pieces by their color.

5 STUDY CONDITIONS

Similarly to [25], we created two distinct conditions to evaluate our responsive joint attention system. Our experimental condition (**full**) uses the complete joint attention system described in section 4.5. Our **control** condition does not use responsive gaze behaviors based on perceived multimodal information. Instead, the control condition is exclusively controlled by the wizard interface and by the robot's idle behavior. As detailed in [25], we carefully designed the idle module to exhibit believable gaze behavior with the goal of making a fair comparison between both conditions. This manipulation remains constant in both studies presented in the following two sections.

6 STUDY 1 - EXTERNAL OBSERVERS

In the initial study presented in this paper, we test whether third-party observers perceive the robot equipped with our responsive joint attention system as more socially present. Testing our hypothesis within this context had the main objective of showing that our manipulation could also be perceived by external observers.

6.1 Participants

We created a crowd-sourcing task using the figure eight data annotation platform[4]. We assigned our task to 40 native English participants from Canada and the United States.

6.2 Procedure

In this study, we adapt to the context and limitations of a within-subjects crowdsourcing task. We created two videos, that show the same actor from a top shoulder perspective (see Figure 1(a)). The videos focus mainly on the robot and the task but the actor's actions are also clearly visible. In one video, the robot was running the full condition, and in the other, the control condition. The actor was responding similarly to the robot in both videos. The videos were approximately 1:50 minutes in length and displayed the exact same verbal behavior from both the human and the robot side. The only variation was the robot's manipulated non-verbal behavior. Crowd-workers viewed each video once but could re-watch the videos as many times as necessary to reply to a set of questions described below.

6.3 Measures

For the increasingly typical context of using crowd workers to evaluate human-robot interactions based on videos, we proposed a simpler questionnaire that can be used to measure social presence within gaze research. We attempt to make the questionnaire and consequent study design appropriate to this context because the user is not directly interacting with the robot (has no *perception of self*). Participants were only presented with six questions that mirrored the *perception of the robot* items of the social presence measure that we use throughout our studies [11]. Social presence focuses on the degree to which an individual feels interconnected with another entity [5]. To achieve this feeling of interconnection, participants should believe that both the robot's behavioral and psychological engagement is connected to theirs and that their

[4] https://www.figure-eight.com

Table 2: Percentage of selections in the external observers experiment.

	control	full	$\chi^2(1)$	p
Which robot is more aware of the presence of the human? (Co-Presence)	27.5%	72.5%	8.10	**=.004**
Which robot is more attentive to the human? (Attentional Allocation)	37.5%	62,5%	2.5	=.114
Which robot is understanding the human better? (Message Understanding)	25.0%	75,0%	10.00	**=.002**
Which robots behaviour is more dependent on what the human is doing? (Behavioral Int.) [a]	57.5%	37.5%	1.68	=.192
Which robot is perceiving more of what the human is feeling? (Affective Understanding)	40.0%	60,0%	1.60	=.206
Which robot is more emotionally involved with the human? (Affective Int.)	20.0%	80,0%	14.40	**<.001**

[a] 5% did not reply to this item.

behavioral and psychological engagement is connected to the robot. However, an external observer does not actively participate in the experiment and as such, we only look into how the crowd-workers evaluated the robot's behavioral and psychological engagement with the actor. We also decided to represent each dimension of the questionnaire by single forced-choice items (see 2) given that crowd workers did not have the same nuanced experience of real users.

6.4 Results

A Chi-Square goodness of fit test was calculated comparing the occurrence of participants that chose the full condition in each dimension with the expected occurrences (20). Significant deviation from the expected values was found ($\chi^2(5) = 19.55$, p=.002). This test combined with the trend in the results (summarized in Table 2) confirms our hypothesis that third party observers do perceive differences between our conditions and consider the full condition as more socially present.

7 STUDY 2 - REPLICATING WITH CHILDREN

In this study, we attempt to replicate the results of our experiment with adults in a laboratory-interaction context in a new context with children in a school. Here, we assess whether children engaged in real-time interactions with our robot similarly perceive our full manipulation as more socially present.

7.1 Participants

In the previously reported experiment with adults, twenty-nine participants took part in a between-subjects experiment in a laboratory setting at KTH (see Figure 1(b)). In that context, participants were 18 to 34 years old and were rewarded with a cinema ticket. In the new replication study, thirty-nine children from the Nacka International English school in Stockholm, Sweden, between the fifth and seventh grade, participated in the experiment. Their ages ranged from 10 to 12 years-old. Due to technical concerns, the study with adults was reduced to 22 participants (11 full, 11 control) and the children study consisted of 27 participants (16 full, 11 control). Regarding gender distribution in the final pool of participants, the adults study was composed of 11 female and 11 male users whereas the children study was composed of 15 female and 17 male participants.

7.2 Procedure

In the adults in the laboratory experiment, participants were guided to a soundproof room separated from the wizard setup and were asked to sign a consent form and fill in a computer form with demographic information. Next, participants were briefed by the experimenter and instructed to sit down in front of the robot. At the beginning of each interaction, the robot explained the MagPuzzle task in detail and participants completed the set of three puzzles in a time span that lasted from 5 to 10 minutes depending on the participant. Finally, participants filled in a final questionnaire and the experiment was concluded.

In the children study, the procedure was identical with some exceptions. We first obtained consent from the children's guardians before they were allowed to participate in the experiment. Children were assigned 30 minutes of their class time to participate in the experiment and were guided to a room near their classroom (see Figure 1 (c)). When entering the room, children were instructed by the experimenter to sit down and interact with Furhat. The total duration of the interaction was on average longer than the experiment with adults and varied between 4 to 14 minutes.

7.3 Measures

For both studies, in the final questionnaire, we used the updated full social presence measure described in [11] and similarly to [25], we divide the social presence questionnaire in two directions: *perception of self* and *perception of the robot*. However, in this paper, to better compare our results with the observer study and the significant results from the original study, we focus primarily on the *perception of the robot* direction of social presence. Table 3 presents the 18 5-point Likert scale *perception of the robot* items grouped by the social presence dimensions. In addition to the social presence questionnaire, we included customized items (see Figure 3) that measured the participants' perceived task difficulty, the robot's helpfulness, the intention of future usage, the likability of the robot and trust in the robot's judgments.

We also video recorded the interactions and logged all the information collected by the joint attention system, including all gaze shifts from the user and robot, the moves performed in the board and the dialog acts triggered by the wizard. We use this information to present objective results on gaze synchrony and hint success rate. The percentage of mutual gaze (eye contact between the robot and the user), joint attention (both looking at the same location

Table 3: Items used for measuring the perception of the robot's social presence in each direction and dimension [11]. Chronbach's alpha values are reported for both the adults and children studies. Items marked with (R) were reversed in our analysis.

	Items	α_{adults}	$\alpha_{children}$
Co-Presence	The robot noticed me. My presence was obvious to the robot. I caught the robot's attention.	.80	.37
Attentional Allocation	(R) The robot was easily distracted from me when other things were going on. The robot remained focused on me throughout our interaction. (R) I did not receive the robot's full attention.	.73	.15
Message Understanding	My thoughts were clear to the robot. The robot found it easy to understand me. (R) The robot had difficulty understanding me.	.77	.77
Behavioral Interdependence	The behavior of the robot was often in direct response to my behavior. The robot reciprocated my actions. The robot's behavior was closely tied to my behavior.	.61	.65
Affective Understanding	The robot could tell how I felt. (R) My emotions were not clear to the robot. The robot could describe my feelings accurately.	.83	.70
Affective Interdependence	The robot was sometimes influenced by my moods. My feelings influenced the mood of our interaction. My attitudes influenced how the robot felt.	.63	.66
All Items (Total)		**.87**	**.84**

on the board) and mismatch (all other situations) are computed by using the log data from the robot's gaze behavior and user gaze tracker. The hints provided by the robot during the task depend on the participant's ability to perceive the game from the robot's perspective and follow the robot's suggestions within a limited period. We used the log data to calculate the percentage of correct responses within 10 seconds after providing hints.

7.4 Results

7.4.1 Social Presence. For comparing the social presence dimensions between our conditions, similarly to the original study, we used non-parametric Mann Whitney-U tests. Previously, in the adults study, a Mann Whitney-U test supported our hypothesis by indicating that the total measure of *perception of the robot* was significantly higher in the full condition than in the control (U = 27.5, p=.015, r = .46). Message understanding (U = 29.0, p=.018, r = .45) was the dimension most responsible for this effect (see Figure 2). However, the children study revealed no significant differences in the *perception of the robot* social presence direction. For both studies, between our gaze conditions, no significant results were found in any of the dimensions in the direction of *perception of self* and in our custom end-of-experiment questionnaire.

7.4.2 Hints and Difficulty. In the adults study, the hints and feedback provided by the robot were most often correctly assessed. The hints provided by the robot about the next move for the puzzle are given by head pose direction and eye gaze combined with an indirect verbal reference (e.g. "I think putting a piece *here* might help"). Although this type of cue was difficult to perceive, given the resolution of the board with pieces placed close together, the success rate of these hints was 65%. This suggests that participants had a good understanding of the robot's viewpoint of the board. Nonetheless, the success rate of combined verbal and non-verbal cues (e.g., I would arrange this yellow piece differently) reached a 93% success rate.

Similarly to adults, children can successfully take the robot's visual perspective. The success rate of the most common hints provided by the robot was 72%. Explicit cues that combined verbal with non-verbal behaviors achieved a success rate of 94%. However, children faced increased difficulties in completing this task, when compared to adults. This is reflected by a number of key factors. When comparing children to adults, on average, they: spent more time (in seconds) to finish the task (M=464, SD=183.70) vs (M=307, SD=80.01); played a higher average number of incorrect moves (M=69.22, SD=74.60) vs (M=28.71, SD=14.44); and required more hints (M=7.11, SD=4.77) than adults (M=2.57, SD=2.36) to complete the task.

7.4.3 Behavioral Synchrony. The first two rows of Table 4 illustrate the behavior synchrony between our robot and the adult participants from the original study. As expected, given our manipulation, the average time that both the participant and the robot were engaged in joint attention significantly increased in the full condition (M=22.51, SD=15.44) compared to control (M=7.58, SD=6.48); t(11)=-2.415, p<.05, d=-1.38.

The last two rows of Table 4 show the behavioral synchrony results for children. With children, the full condition had significantly less mismatch of attention (M=28.76, SD=9.58) when compared

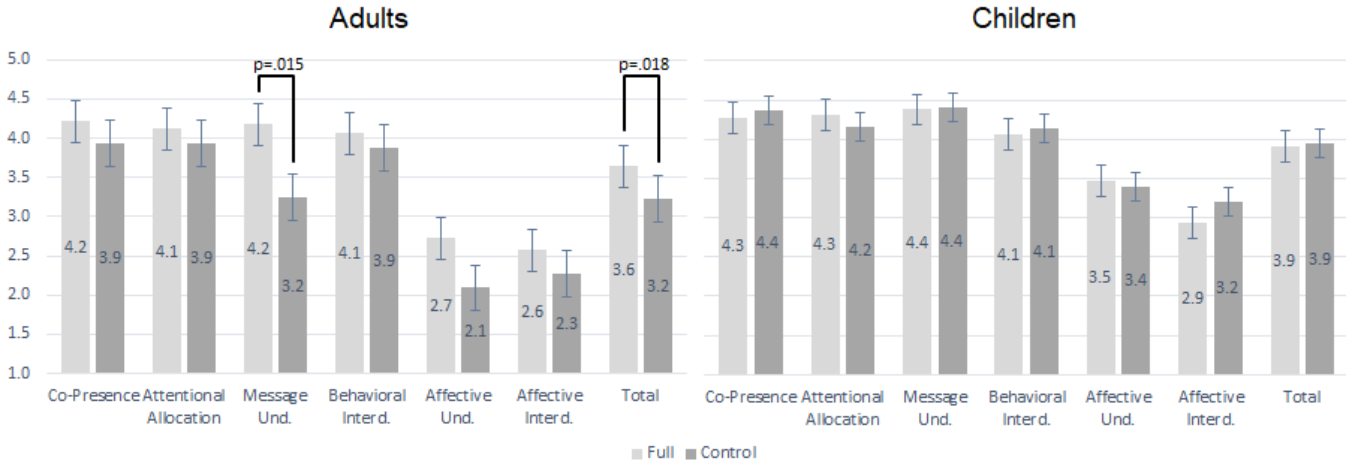


Figure 2: Mean values for the total *perception of the robot* and each of its six social presence dimensions in both real-time experiments.

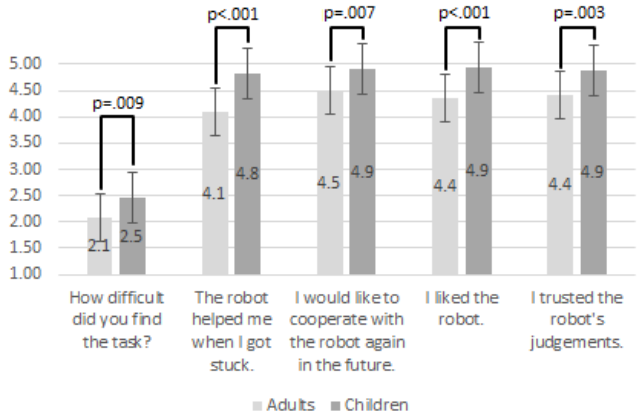


Figure 3: Mean values of the additional end-of-experiment questions in both real-time experiments. Effect size values (from left to right) are .32, .51, .33, .46, .38.

to the control condition (M=42.83, SD=13.01), t(25)=3.24, p<.005, d=1.27. The difference of joint attention between conditions in this study was less pronounced when compared to the study with adults. These differences can be explained by the fact that children asked for more hints and spent more time looking at the board correcting an incorrect move or thinking about their next move. This increased the robot's chances to engage in responsive joint attention but also diluted the differences between conditions.

8 DISCUSSION

The results from the two replication studies reported in this paper suggest that our initial hypothesis was supported with external observers, but not with children. In this section, we further establish the relationship between responsive joint attention and social presence by discussing the implications and limitations of our replication studies.

Table 4: Average percentage of time of mutual gaze, joint attention and mismatched attention in adults vs children.

	Condition	Mutual Gaze	Joint Attention	Mismatch
Adults	Control	44.99	7.58	47.46
	Full	41.69	22.51	35.8
Children	Control	29.49	27.68	42.83
	Full	30.90	40.34	28.76

8.1 Study 1 - External observers

This study suggested that external observers consider the robot in the full condition as more aware of its interaction partner, more capable of understanding messages, and more emotionally involved with the user in the task. In addition to providing further evidence that supports our initial findings with adults, the results from Study 1 afforded us the confidence to perform additional real-time studies. We suggest that third-person observation (e.g., video-based studies) may be used to perform a manipulation check before committing to performing lengthy and costly real-time HRI experiments. Third-person viewpoint studies can be extremely useful for gaze research as (1) external observers can focus their attention on the robot's gaze behavior (2) they are often less resource-intensive and (3) can provide more control/consistency in exposure (i.e., more identical interaction across participants). These studies should not fully replace situated real-time studies but can be considered as an alternative to performing initial human-human studies as proposed in [15]. Before performing their own study with robots, Irfan et al., created several scenarios that aimed to reproduce the social facilitation effect widely reported in human-human interaction [10, 38]. However, after several human-human interaction trials in their target scenarios, the authors could not reproduce the desired effect and decided not to follow through with HRI experiments on the scenarios they designed. Although the measures in external observer studies can be more limiting due to the participants not

being situated in the interaction, these studies can have the advantage of being designed such that users are already visualizing the target robot with the target behaviors which could provide more accurate preliminary evaluations.

8.2 Study 2 - Replicating with Children

Contrary to the initial study with adults, our hypothesis was not supported for the children in school context. Our replication study did not yield the results that we predicted as we did not find any significant difference between our manipulated gaze conditions. However, it can often be more relevant to investigate situations where people did not treat robots manipulated with social behavior as social [8] and understand why certain studies did not yield the authors' desired results, as illustrated in [34]. Below, we discuss the main trends in our data that may explain the lack of perceived differences and limitations of the children study.

8.2.1 Task difficulty differences. Compared to adults, children perceived the MagPuzzle task as more difficult. This is confirmed by our direct question that asks participants how difficult they found the task and in several of our objective measures that point to younger participants having more difficulties in completing the task. In the adults sample, almost none of the participants experienced significant difficulties in the task. The extra cognitive effort experienced by children made them more focused on the task and, as such, exhibited less interest in establishing mutual gaze with the robot to fully appreciate its improved gaze capabilities in the experimental condition (see Table 4). The lack of mutual gaze established with the robot in the children study played a part in the lack of results between our conditions. It has also been previously reported that when participants are working on difficult tasks [36] or are more focused on task performance [17], they tend to ignore embodied social behavior, or even consider it distracting. Researchers performing gaze research in HRI should take into consideration the selection of an appropriate task and interaction context that provides significant opportunities to observe the robot's behavior. While we initially thought that our task would be more engaging for children and thus more appropriate to evaluate our gaze model, it appears that external observers and participants that established increased levels of eye contact or mutual gaze with the robot were more successful at detecting improvements in our manipulated behavior. To better understand the effects of task difficulty in gaze perception among different age groups, future work should include task difficulty as an additional manipulation factor.

8.2.2 Children's higher scores. Despite not reporting differences between our gaze conditions, children, when compared to adults, experienced a higher sense of social presence, found the robot more helpful, would be willing to cooperate more with the robot in the future and liked and trusted the robot to a greater extent. However, these results should be interpreted carefully as they might also be explained by the children's tendency of upper bound scoring found within our data. In both conditions, either because the task or the robot was more compelling for children or because of their innate tendency to please the experimenter, most of the children's scores approached the upper bound of our Likert scales (see Figures 2 and 3). Children may be using extreme responses to second guess the desirable response, as it has been extensively reported that children have a preference for pleasing the experimenter rather than to answer truthfully to subjective questionnaires [4]. Children may also be more affected by the novelty effect of interacting with a robot when compared to the adults that participated in our study. This could have potentially been mitigated by using a task where participants would interact with our system throughout a longer period, possibly within multiple interactions.

8.2.3 Robot's hinting behavior. Regarding hints, it appears that both adults and children can adopt the visual perspective of the robot as they successfully perceive the robot's gaze cues. However, children were better at understanding the robot's cues plausibly because they committed more errors in the task and therefore had more opportunities for developing a better understanding of the robot's hinting behavior. The robot's hinting behavior also appeared to be extremely relevant to the robot's positive overall evaluation. When participants experienced difficulties in the task, the robot was successful at providing useful hints at the right moment.

8.2.4 Developmental age differences. We tried to limit the possible effect of non-uniform childhood development by using participants over the age of ten as identified in the literature. However, behavioral joint attention abilities can still be subject to developmental change during childhood and adolescence [24].

8.3 Social presence

We reiterate the usefulness of using social presence as a measure for performing gaze research in HRI. The relationship between the perception of social presence and gaze research still requires further exploration. Our contribution demonstrates within multiple studies how social presence measures can be used to evaluate the believability of gaze systems in social robots. Given our results, we advocate for future gaze research in HRI to consider using the "perception of the robot" items presented in Tables 2 and 3 to further validate their utility in the area.

9 CONCLUSIONS

In this paper, we investigated gaze models in different HRI contexts. We have conducted two replication experiments with a case study where a robot collaborates with human participants in a spatial reasoning problem-solving task. The first experiment compared the full joint attention system with our control condition and suggested that external observers do perceive the robot in the full condition as more socially present. However, contrary to a previously published experiment where adult participants interacting in a laboratory environment similarly perceived the robot as more socially present, this effect was not confirmed in a replication study with children in a school. Children had greater difficulty with the task and it appears that this heightened difficulty reduced the attention to the robot and increased the attention to the task, which in turn might have reduced their ability to perceive the robot's responsive joint attention behaviors. These results show the importance of performing multiple studies in different interaction contexts for understanding the advantages and limitations of enabling robots with believable gaze mechanisms.

REFERENCES

[1] Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63.

[2] Henny Admoni, Thomas Weng, Bradley Hayes, and Brian Scassellati. 2016. Robot nonverbal behavior improves task performance in difficult collaborations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 51–58.

[3] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems*. Springer, 114–130.

[4] Tony Belpaeme, Paul Baxter, Joachim De Greeff, James Kennedy, Robin Read, Rosemarijn Looije, Mark Neerincx, Ilaria Baroni, and Mattia Coti Zelati. 2013. Child-robot interaction: Perspectives and challenges. In *International Conference on Social Robotics*. Springer, 452–459.

[5] Frank Biocca, Chad Harms, and Jenn Gregg. 2001. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *4th annual international workshop on presence, Philadelphia, PA*. 1–9.

[6] Dan Bohus and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. ACM, 5.

[7] Open Science Collaboration et al. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[8] Kerstin Dautenhahn. 2007. Methodology & themes of human-robot interaction: A growing research field. *International Journal of Advanced Robotic Systems* 4, 1 (2007), 15.

[9] Frédéric Delaunay, Joachim de Greeff, and Tony Belpaeme. 2010. A study of a retro-projected robotic face and its effectiveness for gaze reading by humans. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 39–44.

[10] Joshua M Feinberg and John R Aiello. 2010. The effect of challenge and threat appraisals under evaluative presence. *Journal of Applied Social Psychology* 40, 8 (2010), 2071–2104.

[11] Chad Harms and Frank Biocca. 2004. Internal consistency and reliability of the networked minds measure of social presence. (2004).

[12] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. 2008. The influence of social presence on acceptance of a companion robot by older people. (2008).

[13] Laura Hoffmann and Nicole C Krämer. 2013. Investigating the effects of physical and virtual embodiment in task-oriented and conversational contexts. *International Journal of Human-Computer Studies* 71, 7-8 (2013), 763–774.

[14] Chien-Ming Huang and Andrea Lockerd Thomaz. 2010. Joint Attention in Human-Robot Interaction.. In *AAAI Fall Symposium: Dialog with Robots*.

[15] Bahar Irfan, James Kennedy, Séverin Lemaignan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2018. Social psychology and human-robot interaction: An uneasy marriage. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 13–20.

[16] Sara Kiesler and Jonathon N Cummings. 2002. What do we know about proximity and distance in work groups? A legacy of research. *Distributed work* 1 (2002), 57–80.

[17] Dimosthenis Kontogiorgos, Andre Pereira, Olle Andersson, Marco Koivisto, Elena Gonzalez Rabal, Ville Vartiainen, and Joakim Gustafson. 2019. The effects of anthropomorphism and non-verbal social behaviour in virtual assistants. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. ACM, 133–140.

[18] Gregor Mehlmann, Markus Häring, Kathrin Janowski, Tobias Baur, Patrick Gebhard, and Elisabeth André. 2014. Exploring a model of gaze for grounding in multimodal HRI. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 247–254.

[19] James N Meindl and Helen I Cannella-Malone. 2011. Initiating and responding to joint attention bids in children with autism: A review of the literature. *Research in developmental disabilities* 32, 5 (2011), 1441–1454.

[20] Carol A Miller. 2006. Developmental relationships between language and theory of mind. *American Journal of Speech-Language Pathology* (2006).

[21] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

[22] Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. 2006. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*. Citeseer, 518–523.

[23] Bilge Mutlu, Allison Terrell, and Chien-Ming Huang. 2013. Coordination mechanisms in human-robot collaboration. In *Proceedings of the Workshop on Collaborative Manipulation, 8th ACM/IEEE International Conference on Human-Robot Interaction*. Citeseer, 1–6.

[24] Eileen Oberwelland, Leonhard Schilbach, Iva Barisic, Sstsh C Krall, Kai Vogeley, Gereon R Fink, Beate Herpertz-Dahlmann, Kerstin Konrad, and Martin Schulte-Rüther. 2016. Look into my eyes: Investigating joint attention using interactive eye-tracking and fMRI in a developmental sample. *NeuroImage* 130 (2016), 248–260.

[25] Andre Pereira, Catharine Oertel, Leonor Fermoselle, Joe Mendelson, and Joakim Gustafson. 2019. Responsive joint attention in human-robot interaction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.

[26] André Pereira, Rui Prada, and Ana Paiva. 2014. Improving social presence in human-agent interaction. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 1449–1458.

[27] Tiago Ribeiro, André Pereira, Eugenio Di Tullio, Patrıcia Alves-Oliveira, and Ana Paiva. 2014. From Thalamus to Skene: High-level behaviour planning and managing for mixed-reality characters. In *Proceedings of the IVA 2014 Workshop on Architectures and Standards for IVAs*.

[28] Jennifer Richardson and Karen Swan. 2003. Examing social presence in online courses in relation to students' perceived learning and satisfaction. (2003).

[29] David Herrick Spencer. 2002. *A field study of use of synchronous computer-mediated communication in asynchronous learning networks*. Rutgers The State University of New Jersey-Newark.

[30] Megan Strait, Heather L Urry, and Paul Muentener. 2019. Children's Responding to Humanlike Agents Reflects an Uncanny Valley. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 506–515.

[31] Megan Strait, Lara Vujovic, Victoria Floerke, Matthias Scheutz, and Heather Urry. 2015. Too much humanness for human-robot interaction: exposure to highly humanlike robots elicits aversive responding in observers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 3593–3602.

[32] Megan K Strait, Victoria A Floerke, Wendy Ju, Keith Maddox, Jessica D Remedios, Malte F Jung, and Heather L Urry. 2017. Understanding the uncanny: both atypical features and category ambiguity provoke aversion toward humanlike robots. *Frontiers in psychology* 8 (2017), 1366.

[33] Chih-Hsiung Tu and Marina McIsaac. 2002. The relationship of social presence and interaction in online classes. *The American journal of distance education* 16, 3 (2002), 131–150.

[34] Paul Vogt, Rianne van den Berghe, Mirjam de Haas, Laura Hoffman, Junko Kanero, Ezgi Mamus, Jean-Marc Montanier, Cansu Oranç, Ora Oudgenoeg-Paz, Daniel Hernández García, et al. 2019. Second Language Tutoring using Social Robots: A Large-Scale Study. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Ieee, 497–505.

[35] Xijing Wang and Eva G Krumhuber. 2018. Mind perception of robots varies with their economic versus social function. *Frontiers in psychology* 9 (2018), 1230.

[36] Ina Wechsung, Patrick Ehrenbrink, Robert Schleicher, and Sebastian Möller. 2014. Investigating the Social Facilitation Effect in Human–Robot Interaction. In *Natural Interaction with Robots, Knowbots and Smartphones*. Springer, 167–177.

[37] Tomoko Yonezawa, Hirotake Yamazoe, Akira Utsumi, and Shinji Abe. 2007. Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking. In *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 140–145.

[38] Robert B Zajonc. 1965. Social facilitation. *Science* 149, 3681 (1965), 269–274.