Measuring the Efficacy of Robots in Autism Therapy: How Informative are Standard HRI Metrics?

Momotaz Begum
Computer Science
University of Massachusetts Lowell
mbegum@cs.uml.edu

Richard W. Serna
Psychology
University of Massachusetts Lowell
Richard_Serna@uml.edu

David Konktak
Crotched Mountain Rehabilitation Center, Greenfield, NH
David.Kontak@crotchedmountain.org

Jordan Allspaw
Computer Science
University of Massachusetts Lowell
allspaw.j@gmail.com

James Kuczynski
Computer Science
University of Massachusetts Lowell
james.perl12@gmail.com

Holly A. Yanco
Computer Science
University of Massachusetts Lowell
holly@cs.uml.edu

ABSTRACT

A significant amount of robotics research over the past decade has shown that many children with autism spectrum disorders (ASD) have a strong interest in robots and robot toys, concluding that robots are potential tools for the therapy of individuals with ASD. However, clinicians, who have the authority to approve robots in ASD therapy, are not convinced about the potential of robots. One major reason is that the research in this domain does not have a strong focus on the efficacy of robots. Robots in ASD therapy are end-user oriented technologies, the success of which depends on their demonstrated efficacy in real settings. This paper focuses on measuring the efficacy of robots in ASD therapy and, based on the data from a feasibility study, shows that the human-robot interaction (HRI) metrics commonly used in this research domain might not be sufficient.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: [Performance evaluation (efficiency and effectiveness)]; J.4 [**Social and Behavioral Sciences**]: [Psychology]

Keywords

Autism spectrum disorders (ASD), efficacy, robot, human-robot interaction (HRI) metrics

1. INTRODUCTION

Robotics research has demonstrated that many individuals with ASD (IwASD) express elevated enthusiasm (e.g. increase in attention [18], imitation ability [11], verbal utterances [17], social activities [29], etc.) while interacting with robots. A comprehensive survey on this research is available in [6, 27]. A long line of research is dedicated to the design of robots with appropriate physical features [24], control architectures [9], evaluation metrics [26], and HRI algorithms [10] that can be used in ASD diagnosis and therapy. Despite these efforts, the targeted end-users of this technology (IwASDs, their caregivers, and clinicians) are neither aware nor convinced of the role of robots in ASD therapy [8]. Recently, a number of systematic reviews and a meta-analysis of the technology-based interventions for IwASDs (which reviewed research articles published before December 2011) have concluded that the robot based studies with IwASDs fail to meet a set of criteria commonly observed to assess the outcome of an ASD therapy [13, 23]. The problem lies in the fact that the vast majority of robotics research in this domain shows the 'likability' of robots but fails to demonstrate a robot's efficacy in therapy and/or diagnosis of ASD [8, 18]. Accordingly, we are observing a recent paradigm shift where more research is focusing on investigating the efficacy of robots in ASD therapy through improved research design [4, 12, 16, 22, 28], investigation of robots' features and abilities [25], and increasing robots' autonomy [5]. Defining appropriate efficacy metrics is one of the fundamental components in investigating the efficacy of robots in ASD therapy. There is, however, no report in the existing literature on efficacy metrics for robots in ASD therapy.

Efficacy metrics indicate how well a robot is producing the intended therapeutic effect in an IwASD. A vast majority of robotics research in this domain uses some common HRI metrics (e.g. gaze direction, verbal/non-verbal communication cues, affective responses, etc.) as a measure of a robot's impact on an IwASD. We have recently conducted a study to investigate whether it is clinically feasible to teach a new skill to IwASDs through a robot [20]. During the study, a robot was used to teach a basic skill of greeting someone in a socially acceptable manner to a group of low-functioning IwASDs. We defined two metrics, `Skill execution` and `Prompt dependency`, which measured the efficacy of the robot-mediated therapy in achieving the intended therapeutic goal. An analysis of the study out-

HRI '15 March 02 - 05 2015, Portland, OR, USA
ACMÂĚ978-1-4503-2883-8/15/03$15.00
http://dx.doi.org/10.1145/2696454.2696480.

comes shows that a set of common HRI metrics (namely, `Gaze`, `Communication`, and `Affect`) have no correlation with these two efficacy metrics. This lack of correlation indicates that the common HRI metrics may not be able to gauge the efficacy of a robot in ASD therapy. To the best of our knowledge, this is the first work that presents efficacy metrics for robots in ASD therapy and provides a comparative analysis of efficacy metrics with some common HRI metrics.

2. EFFICACY AND HRI METRICS

Any ASD therapy, in a clinical domain, operates with some fundamental goals. In general, a therapy: 1) aims to teach an IwASD a new behavior/skill or to eliminate a problem behavior [30], 2) aims to improve the quality of life, level of independence, health, and well-being of an IwASD and help him/her integrate in the society in a better way [1], and 3) when successful, should produce outcomes that last [1]. It is expected that an IwASD will be able to execute a behavior learned through the therapy independently, outside of the therapy setting, and with the people in his/her everyday life. Clinicians use different metrics, on a case-by-case basis, to monitor the progress made by an IwASD and measure the overall outcome of a therapy. To establish the efficacy of robots in this domain, a robot-mediated ASD therapy should also operate with the same goals, and metrics should be defined to clearly indicate how the use of a robot is linked to the positive outcome of a therapy.

As the core deficits of ASD are related to social behaviors and communication abilities [3], robotics researchers historically focused on a set of social and communicative cues directed by IwASDs toward robots to assess the quality of IwASD-robot interaction. The most common HRI metrics used in this case are `Gaze` (the duration or the number of times an IwASD looked at the robot) [5, 18, 19, 26, 28, 29], `Communication` (number of verbal/non-verbal communication with the robot, total number of words exchanged with the robot, etc.) [17, 26, 29], `Affect` (being in an affective state or showing affective responses to the robot) [7, 9, 18, 24], `Attention` (focusing on the robot) [18, 19], `Imitation` (imitating a robot's action or speech) [11, 19, 24], and `Proxemics` (being in a close proximity of the robot) [10]. Although these metrics perform well to demonstrate the general enthusiasm expressed by an IwASD when (s)he is around a robot, none of these studies provides a clear indication of whether the participants (IwASDs) were actually able to learn the target behavior from the robot and executed it independently, outside of the study setting.

Our recent study investigated whether a robot-mediated ASD therapy can achieve some of the clinical goals of a standard ASD therapy. We propose two efficacy metrics to monitor the progress of our participants and the overall outcome of the robot-mediated therapy: `Skill execution` and `Prompt dependency`.

1. `Skill execution, SE`: A measure of the ability of a participant to execute a target social skill (with the robot and with other people, within and outside of the therapy setting).
2. `Prompt dependency, PD`: A measure of a participant's ability to execute a target skill without the help of the robot.

These two metrics, when analyzed together, indicate the efficacy of a robot in teaching a new skill to an IwASD. The goal of the robot-mediated therapy is to maximize `Skill execution` (ideally, `SE` = 100%) while minimizing `Prompt dependency` (ideally, `PD` = 0). The trends of `SE` and `PD` over the duration of a therapy provide important information about the efficacy of a therapy. This paper shows that some of the common HRI metrics listed in this section do not show any meaningful correlation with these two efficacy metrics.

3. A FEASIBILITY STUDY ON TEACHING THROUGH A ROBOT

Single-subject research design has a unique value in autism research [15]. Recent literature suggests that group-based design, although considered as the gold-standard in clinical research, might not be the only choice to prove the efficacy of an ASD therapy [21]. We conducted a single-subject study at the Crotched Mountain Rehabilitation Center (CMRC), a special education school in New Hampshire, USA, to investigate the feasibility of teaching a new skill to IwASDs through a robot. The study was approved by the Institutional Review Boards (IRB) of the University of Massachusetts Lowell and CMRC. Informed consent from the parents or guardians of the participants was collected before the study began.

3.1 Study Design

Based on a discussion with the clinicians at CMRC, we identified the lack of social greetings as a common deficit among many IwASDs attending the school. Accordingly, the goal of the study was to teach the basic skill. A therapy was designed to teach saying 'Hi' or 'Hello' in response to a social greeting. The study was a single-subject, multiple baseline type and followed the basic guidelines for single-subject research design outlined in [15].

3.1.1 Participants

Two inclusion criteria were decided for the participants.

1. An individual diagnosed with any form of ASD.
2. An individual with ASD who, according to his/her current therapists at the CMRC, has one or more of the prerequisite abilities to initiate/respond to social greetings but does not generally do so. The ability to imitate, verbal ability, the physical ability of waving hands, etc. were considered a few of the prerequisites to learn the skill of initiating/responding to social greetings.

Five students from the school matched these inclusion criteria and were recruited for the study. Due to space limitation, results from three participants will be reported in this paper. Participant demographics and diagnostic information are presented in Table 1. All of the participants are male which is consistent with the fact that ASD is more common among boys than girls [2]. Note that our participants are mostly toward the lower end of the spectrum (generally, IQ$<$ 70), a less investigated population in robot-based ASD research. Due to compromised cognitive abilities, all of our participants, according to a school regulation, were required to be accompanied by a professional caregiver at all times. Accordingly, the caregiver was included in the therapy design. Informed consent from the caregivers was collected before the study began.

Table 1: Participants' Information

ID	Sex	Age (Yrs)	Diagnostic information
P1	M	19	General diagnosis: Autism Assessment tool: AAMR Adaptive Behavior Scale - School Score:Very poor - Average, age equivalence range 3 Yrs - 4 Yrs
P2	M	14	General diagnosis: Classic Autism, Mood disorder Assessment tool: Adaptive Behavior Scale - School (ABS-S:2) Score: Age equivalence range 3 Yrs - 6 Yrs 3 Mos
P3	M	13	General diagnosis: Autism Assessment tool: Leiter International Performance Scale, Verbal Behavior Milestones Assessment and Placement Program IQ: 40

3.1.2 Variables

A required feature of single-subject research design, among many others, is that the dependent and independent variables should have operational definitions that clearly indicate the performance pattern of participants and allow valid interpretation of results [15]. The independent variable in our study is the robot: whether the therapy delivered through a robot can teach a target skill to the participants. There are two dependent variables: *Responsiveness to the command* (i.e. the frequency of positive response to social greetings) and *Prompt* (i.e. the number of prompts provided by the robot to generate a positive response). The dependent variables are measured repeatedly before the robot-mediated therapy begins (i.e. the baseline measurement phase), during the therapy, and outside of the therapy setting in order to identify the effect of the robot-mediated therapy on the participants. The two efficacy metrics, `Skill execution` and `Prompt dependency`, are directly linked to these two dependent variables but provide a generalized way to measure the outcome of other ASD interventions of similar nature. The paper hypothesizes that these two metrics provide clear indications about the efficacy of a robot in teaching a skill to IwASDs and that the HRI metrics are not sufficient to convey information about therapeutic effects. The dependent variables and the efficacy metrics will be further discussed in Section 3.2.1.

3.1.3 Baseline measurement

Baseline measurement was performed on each participant before his therapy began and continued until a stable pattern was observed, for a minimum of three days. Each participant went through a different duration of the baseline phase in order to conform with the guidelines of multiple baseline design [15]. During each day of the baseline phase for a participant, one person familiar to the participant and one unfamiliar person greeted him by saying "Hi [*Name*]" at two different locations within the school at two different times of the day. Waving hands and/or saying "Hi" or "Hello" (with or without making eye-contact with the greeter) or any other within-context verbal responses were considered as a correct response. Ignoring the greeter or gazing at the greeter with no further verbal/non-verbal expressions was considered as an incorrect response. The mean of the two responses in one day makes one data point while analyzing the results. Data from the baseline phase are shown in Fig. 4 and will be further discussed in Section 4. Table 2 shows the baseline duration of different participants.

Table 2: Overall Description of the Study

Participant ID	Baseline duration (days)	No. of therapy sessions	Duration of the therapy (days)	Generalization training
P1	3	19	9	No
P2	6	24	10	Yes
P3	9	10	6	No

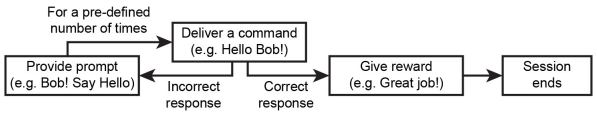

Figure 1: The structure of the ABA-based therapy for teaching the behavior of social greetings

3.1.4 The therapy

The therapy was designed in collaboration with a behavioral scientist (the second author of this paper) who is an expert in autism research. The therapy follows the basic structure of applied behavior analysis (ABA), a widely accepted method for behavioral intervention [14]. Any ABA-based training follows a basic structure: Command or Discriminative stimulus (S^D) $\Longrightarrow$ Prompts (P) (if necessary) $\Longrightarrow$ Reinforcement/Reward (R). Our therapy of teaching social greetings follows the structure shown in Fig. 1. Two different types of prompts are used to evoke the correct response in the participants. The first type of prompt is *modeling*, where the robot models the target behavior with the caregiver of the participant. The second type is *verbal instruction*, where the robot verbally informs the participant the correct way to respond to the command. If a participant correctly responds to a command, no prompt is delivered and the robot provides reward for his correct response. Prompts are delivered only in cases of no response or incorrect response. The *verbal instruction* prompt is delivered only if the *modeling* prompt fails to elicit the correct response.

To facilitate learning, the process 'Command or Discriminative stimulus (S^D) $\Longrightarrow$ Prompts (P) (if necessary) $\Longrightarrow$ Reinforcement/Reward (R)' is iterated for 3 to 5 times per session, irrespective of correct or incorrect responses from the participant. The robot starts each iteration by saying "Let us practice saying 'Hi' again" in order to create a context. The command and prompts are chosen to be simple and easily understandable by the target population. An example of a command is the robot saying 'Hi [name of the participant]' as soon as he enters the therapy room. An example of a verbal instruction prompt is the robot saying '[name of the participant], say 'Hi' to me'. The robot uses similar commands with the caregiver during a modeling prompt (e.g. the robot greets the caregiver by saying 'Hi [name of the caregiver]', and the caregiver responds to this greeting by waving hands and saying 'Hi Blue', as shown in Figs 2(c) - 2(e)). The basic form of reinforcement/reward is the robot appreciating the IwASD for his correct response.

A generalization phase is designed for participants who respond well to the robot-mediated therapy. A two-step proce-

(a) (b) (c) (d) (e)

Figure 2: (a) The humanoid robot 'Blue' (Aldebaran Robotics Inc.) (b) The robot-control interface (c-e) The modeling prompt is helping the participant P1 (on the left) to gradually execute the correct behavior

dure is followed for generalization training. At the first step, a triadic interaction is designed where the caregiver greets the participant immediately after the participant's positive response to a robot's greeting. This triadic interaction design (Robot → Participant ← Caregiver) tries to exploit a participant's 'intent to interact' with the robot to generalize the skill with the human (in this case, the caregiver). When a participant performs well in this triadic interaction, the second step of generalization is introduced where the robot is removed from the scene. In this case, a dyadic interaction is designed where the participant interacts with the robot in the therapy room without the presence of the caregiver. The caregiver waits outside of the therapy room and greets the participant as soon as he comes out of the room.

3.1.5 Study protocol

The therapy was delivered through a humanoid robot from Aldebaran Robotics (we named the robot 'Blue') (Fig. 2(a)). The user-interface shown in Fig. 2(b) was designed to deliver the therapy through the robot in a Wizard-of-Oz manner. The study observed the following protocol:

1. For each participant, the therapy starts after he completes his baseline measurement.
2. Prior to the therapy beginning, each participant goes through a 5 minute habituation session where the robot interacts with the participant while showing its different motor and sensing abilities.
3. Each therapy session lasts for 2 to 4 minutes and follows the procedure described in Section 3.1.4.
4. Only those participants who consistently perform well with the robot in therapy sessions (high `Skill execution` with no `Prompt dependency` for at least six sessions in three consecutive days) are considered for the generalization training.
5. Participants who do not show any sign of improvement after 10 therapy sessions (e.g. no steady pattern in `SE` and `PD` values, no increase in `SE` or decrease in `PD` values, etc.) are discontinued from the study.

A brief summary of the study is provided in Table 2.

3.2 Data Collection and Coding

We collected video, audio, and images from all therapy sessions. Two sets of behaviors of the participants were coded from the video data: behaviors related to the two efficacy metrics (`SE` and `PD`) and behaviors related to three common HRI metrics that are relevant to this study.

3.2.1 Behaviors related to efficacy metrics

The two dependent variables of the study are considered as the behaviors in this category and are defined as follows.

- *Prompt*: The participant is unable to correctly respond to the command and the robot delivered one or both of the two prompts to teach him the correct behavior. This behavior is coded quantitatively by counting the total number of prompts delivered in a session. *Prompt*, inherently, is a component of a therapy and is not provided or measured during the baseline phase.
- *Responsiveness to the command*: The participant waved hand(s) or said "Hi" or "Hello" or any other words within the context of social greetings to the robot or human greeter. Gazing at the greeter without any other verbal or non-verbal expressions is not considered a correct response. This behavior is coded quantitatively by counting the number of commands with appropriate response. *Responsiveness to the command* is also measured during the baseline phase.

The two efficacy metrics are calculated from these two behaviors as follows:

1. `Skill execution, SE`: The *Responsiveness to the command* expressed as a percentage of the total number of commands delivered in a session.

$$\texttt{SE} = \frac{\textit{Responsiveness to the command}}{\text{Total Commands}} X 100\% \quad (1)$$

2. `Prompt dependency, PD`: The average number of prompts required to generate one appropriate response in a therapy session.

$$\texttt{PD} = \frac{\textit{Prompt}}{\textit{Responsiveness to the command}} \quad (2)$$

In the case *Responsiveness to the command*$= 0$ (which implies $\texttt{SE} = 0\%$),

$$\texttt{PD} = \textit{Prompt} \quad (3)$$

ASD therapies generally are time consuming processes where achieving a therapeutic goal might require several months. These two metrics provide a simple way to track the progress of a participant during any ABA-based robot-mediated therapy including the one presented in this paper. For example,

- Increasing trend of `SE` accompanied by a decreasing trend of `PD` indicate a potential improvement of the participant over time and a possible success of the therapy.

Table 3: Cohen's Kappa for Coded Behaviors

Behavior	κ (chance corrected)	κ (chance not corrected)
Prompt	0.82	0.94
Responsiveness to the command	0.92	0.97
Gaze	0.80	0.90
Communication	0.76	0.91
Affect	0.64	0.80

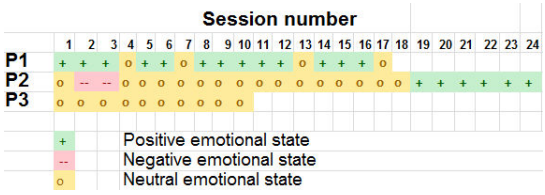

Figure 3: Emotional states of the participants

- Increasing trend of `SE` accompanied by an increasing or constant `PD` indicate the requirement of adopting a strategy to fade out the prompt while keeping the skill intact.
- Decreasing trend of `SE` accompanied by a decreasing or increasing trend of `PD` indicate that the current therapy is not making any positive effect on the participant.

Note that, any other confounding variables that are active during a therapy might alter the interpretations of `SE` and `PD` and should be investigated carefully before drawing any conclusion about the outcome of a robot-mediated therapy.

3.2.2 *Behaviors related to common HRI metrics*

There are three behaviors in this category which are generally considered as standard HRI metrics.

- `Gaze`: Participants look at the robot (`Gaze` at the robot) or at the human caregiver (`Gaze` at the caregiver) during a session. A participant's interaction with his caregiver (e.g. talking or playing) is considered as `Gaze` at the caregiver. The behavior is coded quantitatively by measuring its duration and is expressed as a percentage of the total duration of a session.
- `Communication`: The participant uses words to respond to the robot or initiates a non-verbal communication (e.g. touching the robot, pointing to the robot, etc.) with the robot. Any words directed toward the robot is considered as a form of communication. This behavior is coded quantitatively by counting the number of occurrences during a session and is expressed as occurrences per minute.
- `Affect`: Overall emotional state of the participant while interacting with the robot during a therapy session. The coders reported three emotional states qualitatively: positive, negative, and neutral. Example signs of positive emotional state are the participant seemed happy, was smiling, singing, joyfully talking to the robot or the caregiver, etc. during a session. Examples signs of negative emotional state are the participant seemed stressed, angry (e.g. screaming), sad, non-focused, etc. during a session. Neutral emotional state indicates a participant with no signs of positive or negative emotion.

Based on these definitions, two people coded the behaviors related to efficacy metrics and two people coded the behaviors related to HRI metrics, all independently. Cohen's Kappa was used to ensure inter-coder agreement for each behavior and is listed in Table 3.

4. RESULTS AND ANALYSIS

This section provides an analysis of the coded behaviors to 1) highlight the role of the two efficacy metrics in reporting the efficacy of the robot-mediated therapy, and 2) shed light on the relationship between the two efficacy metrics and the three common HRI metrics (`Gaze`, `Communication`, and `Affect`). *Visual analysis*, a standard tool to present results from single subject multiple baseline study [15], is used to visually compare the responses of the participants (in terms of `SE` and `PD`) during the baseline phase and throughout the course of the robot-mediated therapy. Fig. 4 presents the *visual analysis* of the study outcome. Each participant's baseline phase is staggered with his therapy phase, to conform with the standard of *visual analysis* in single subject design. As a result, session numbers in Fig. 4 are marked as 1 to 20 (3 baseline sessions and 17 human-robot sessions) for P1, 1 to 30 (6 baseline sessions and 24 human-robot sessions) for P2, and 1 to 10 (9 baseline sessions and 10 human-robot sessions) for P3. Figs. 3, 5, 6, and 7 present the HRI metrics of the participants during the human-robot therapy sessions (note that the session numbers in these figures correspond to the human-robot session only and hence, different from that shown in Fig. 4). Pearson's Correlation Coefficient (PCC) is used to measure the statistical correlation between the HRI metrics and the efficacy metrics. Table. 4 lists the PCCs between the two efficacy metrics (`PD` and `SE`) and the two quantitatively coded HRI metrics (`Gaze` and `Communication`). The following sections will provide an individual analysis of results for each participant.

4.1 Participant 1 (P1)

P1 received the robot-mediated therapy in 19 sessions. Video and audio were not recorded in the first two sessions due to technical problems. These two sessions are excluded from the analysis (resulting in 17 robot-mediated sessions). Fig. 5 shows the variation in `Gaze` and `Communication` HRI metrics of P1 during the therapy sessions. P1 consistently expressed more gazing preferences to the robot (on average, 49.9% of the time in a session) than to the caregiver (on average, 11.8% of the time in a session) and made some efforts to communicate with the robot in 94% of the sessions. In 76% of the sessions P1 was in a positive emotional state while in the rest of the sessions he expressed neutral emotions (as shown in Fig. 3). According to the common metrics of HRI, these are considered as signs of positive impact of a robot on an IwASD. However, in spite of all of these positive behaviors, P1 failed to meet the clinical goal of the study: he could not learn to execute the target behavior of responding socially to a greeting without the help of the robot.

As shown in Fig. 4, P1 generated some correct responses in 33% of sessions during the baseline phase (baseline phase did not involve prompting). But during the therapy, in 76%

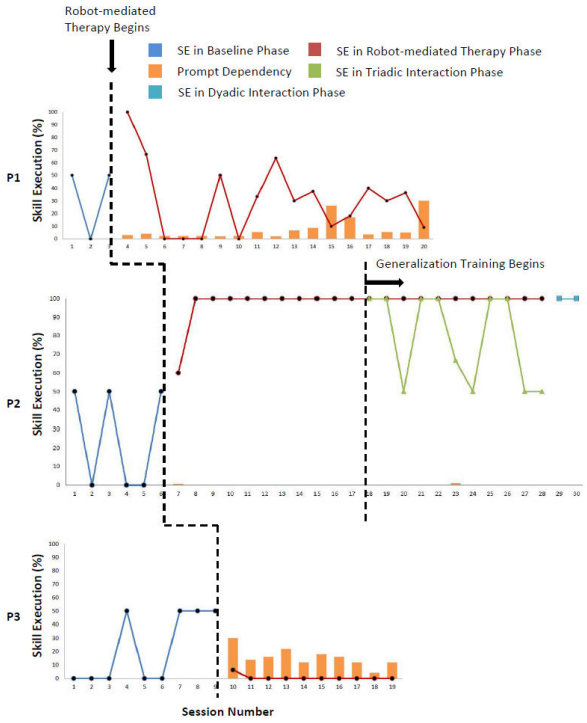

Figure 4: Outcome of the single subject multiple baseline study with three participants

of the sessions P1 needed at least one prompt to appropriately respond to one command while in sessions 6, 7, 8 and 10 several prompts could not yield even one correct response. Overall, the `Skill execution` of P1 did not follow any consistent pattern during the therapy. In addition to that, P1's `Prompt dependency` did not show any decreasing trend over time. The `SE` and `PD` values of P1 clearly show that he did not make any progress with respect to learning to execute the target skill without any help (i.e. the prompts) from the robot. In other words, although P1 shows strong signs, according to the three HRI metrics, that he is interested in the robot, the robot was not able to teach him the target skill through the 17 therapy sessions. This potential mismatch is also revealed in the lack of correlation between the HRI metrics and the efficacy metrics.

The `Gaze` (at the robot) metric has poor correlation (PCC =0.23) with the `SE` and negative correlation with the `PD` (PCC=−0.55). The `Communication` metric shows strong correlation with the `SE` (PCC = 0.87). That is mostly because most of P1's `Communication` with the robot was related to responding to the greeting commands of the robot. The `Communication` metric, however, shows negative correlation with the `PD` (PCC = −0.13).

Overall, analysis of Figs. 4, 3, and 5, and Table 4 reveals the fact that the three HRI metrics can not be used as indicators of the robot's efficacy in P1's therapy.

4.2 Participant 2 (P2)

P2 is the only participant in this study who went through the generalization training. P2 received the therapy in 24 sessions. Analysis of P2's coded behaviors demonstrates another case of potential mismatch between the information conveyed through efficacy metrics and HRI metrics. Fig. 6 shows the variation in `Gaze` and `Communication` HRI metrics of P2 during the therapy. Similar to P1, P2 also showed gazing preferences more toward the robot (on average 59% of the time of a session) than to the caregiver (average is 2.2%) and made efforts to communicate with the robot in every session. Unlike P1, P2 was in positive emotional state in only 25% of the sessions and in 8% of the sessions he expressed negative emotions. P2, however, as opposed to P1, was able to achieve the clinical goal of this study by learning to execute the social behavior outside of the therapy setting. As shown in Fig. 4, P2 was not responding consistently during the baseline phase (`SE`= 50% in 50% of the sessions). But P2's `Skill execution` quickly rose to 100% after the first therapy session and he maintained that high score without any `Prompt dependency` for 11 consecutive therapy sessions. The generalization training (triadic interaction) started from session 18 (human-robot session 12) as shown in Fig. 4 (note that therapy sessions are staggered with the baseline phase). Last two sessions were dedicated to the dyadic interaction. During the generalization training P2 needed only one prompt in Session 23 (human-robot session 17) to execute the target skill. Steady high values of `SE` and low values of `PD` clearly indicate P2's success in mastering the target skill from the robot. The HRI metrics, however, do not show any meaningful correlation with the efficacy metrics. As indicated in Table 4, the `Gaze` (at the robot) metric has very poor correlation with both of the two efficacy metrics while the `Communication` metric shows negative correlation with the two efficacy metrics. Overall, although P2 learned to execute the skill from the robot and P1 failed to do that, the HRI metrics do not convey any information that can differentiate P2 from P1 with respect to achieving the therapeutic goal of the study.

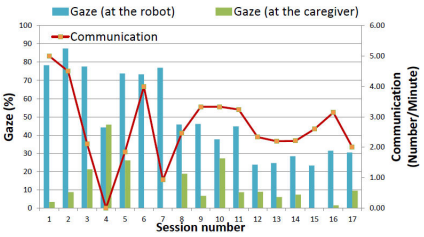

Figure 5: HRI metrics of P1 in human-robot sessions

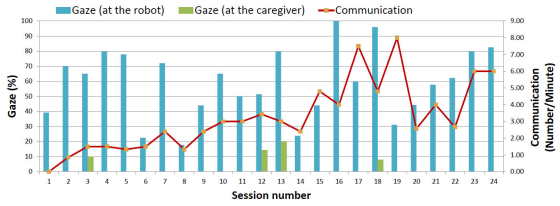

Figure 6: HRI metrics of P2 in human-robot sessions

4.3 Participant 3 (P3)

P3 received the robot-mediated therapy in 10 sessions during which he showed very high gazing preferences for the

Table 4: Correlation Between HRI Metrics and Efficacy Metrics: Pearson's Correlation Coefficient

	SE			PD		
	P1	P2	P3	P1	P2	P3
`Gaze (at Robot)`	0.23	0.04	0.11	-0.55	0.12	0.11
`Communication`	0.87	-0.01	0.64	-0.13	-0.21	0.63

robot (on average 71% of the time) than for the caregiver (only 1.9% of the time), as shown in Fig. 7. P3's gazing preference to the robot is much higher than that of both P1 and P2. P3 also maintained a neutral emotional state throughout the therapy sessions. P3, however, showed minimal effort to communicate with the robot. Overall, HRI metrics represent P3 as a participant who might have moderate to no interest in the robot but certainly did not dislike the robot. Efficacy metrics, however, represents P3 as a participant completely unsuitable for robot-mediated interventions. As shown in Fig. 4, during the baseline phase P3 exhibited some correct responses (`SE` = 40% in 44% of the baseline measurements) but as soon as the therapy started his `Skill execution` drastically dropped to zero. Despite several prompts P3 was completely unable to execute the correct behavior in 90% of the sessions. Accordingly, we concluded that P3 is the kind of participant who might not be suitable for robot-mediated interventions. HRI metrics convey a very little information for drawing such a conclusion.

5. DISCUSSION AND CONCLUSION

Efficacy of robots is a factor which will determine whether robots can actually be employed in clinical settings for ASD therapy. Clinicians require measures to quickly identify how well a robot is performing to achieve a planned therapeutic goal. In this paper, we defined two such efficacy metrics, `Skill execution` and `Prompt dependency`, as a measure of a robot's efficacy to teach a basic skill of social greetings to three low-functioning IwASDs. Both of these metrics are easy to calculate and involve monitoring of two easily observable behaviors of the participants. Results presented in Section 4 show that these two metrics, when analyzed together, created a highly informative picture of how well different participants were performing over the course of the study.

We also analyzed all therapy sessions with respect to three common HRI metrics (`Gaze, Communication,` and `Affect`). Our analysis shows that the HRI metrics performed well in reporting the 'likability' of the robot to the participants. For example, all of the participants showed stronger gazing preference for the robot than for the caregiver and such preferences did not die out over time, enabling us to safely rule out the contribution of the novelty effect in this case. All participants, within their limited verbal and cognitive abilities, made efforts to communicate with the robot, and finally, the robot encountered negative emotions in only 3.9% of the total 51 therapy sessions with three participants. All of these indicate that the participants were not uncomfortable around the robot and might have enjoyed the robot's presence in many cases. This finding is completely in line with the findings of previous research in this domain [6, 27]. However, a comparative analysis of the two sets of metrics (HRI and efficacy) shows that HRI metrics were not able to convey information about the efficacy of the robot in achieving the goal of the therapy. For example, between P1 and P2, HRI metrics indicated P1 was more attracted to the robot than P2 (both of them have approximately similar scores for the `Gaze` and `Communication` metrics, but P1 expressed significantly more positive `Affect` around the robot than P1). In reality, P2 was able to reach the clinical goal of the study while P1 could not master the skill from the robot. The efficacy metrics successfully conveyed this information. These are significant observations and indicate the need of defining appropriate efficacy metrics for robot-mediated ASD therapy. We understand that the sample size of this study was small ($n = 3$) but single subject, multiple baseline design mitigates the negative effect of small sample size.

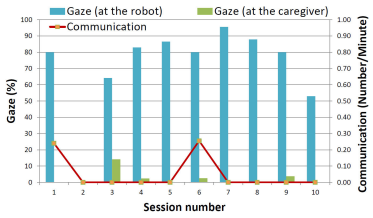

Figure 7: HRI metrics of P3 in human-robot sessions

Another important consideration is the diagnostic condition of the participants. All of the participants in this study were toward the lower end of the spectrum, a population generally not discussed in the majority of robot-mediated ASD research. This diagnostic condition might act as a confounding variable to influence the uncorrelated patterns of the HRI metrics and efficacy metrics that emerged from this study. Further studies are required to investigate this variable and is considered as a future work.

6. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation (IIS-0905228 and IIS-1111125). The authors acknowledge the support of Jonathan Knittle and staff at CMRC to conduct the study.

7. ADDITIONAL AUTHOR

Additional author: Jacob Suarez (Computer Science, University of Massachusetts Lowell email: `Jacob_Suarez@ student.uml.edu`)

References

[1] Autism spectrum disorders: Guide to evidence based interventions. *Missouri Autism Guidelines Initiative*, 2012.

[2] Prevalence of autism spectrum disorder among children aged 8 years: Autism and developmental disabilities monitoring network, 11 sites. *Center for Disease Control and Prevention*, 2014.

[3] C. A. Barker. The triad of impairment in autism revisited. *Child and adolescent psychiatric nursing*, 22(4):189–193, 2009.

[4] E. Bekele, J. A. Crittendon, A. Swanson, N. Sarkar, and Z. E. Warren. Pilot clinical application of an adaptive robotic system for young children with autism. *Autism*, 18:598–608, 2014.

[5] E. T. Bekele, U. Lahiri, A. R. Swanson, J. A. Crittendon, Z. E. Warren, and N. Sarkar. A step towards developing adaptive robot-mediated intervention architecture (ARIA) for children with autism. *Neural Syst. and Rehab. Eng.*, 21:289 – 299, 2013.

[6] J.-J. Cabibihan, H. Javed, M. A. Jr., and S. M. Aljunied. Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism. *Social Robotics*, 5(4):593–618, 2013.

[7] K. Dautehhahn and I. Werry. Towards interactive robotics in autism therapy. *Pragmat Cogn*, 12:1–35, 2007.

[8] J. Diehl, L. Schmitt, C. R. Crowell, and M. Villano. The clinical use of robots for children with autism spectrum disorders: A critical review. *Research in Autism Spectrum Disorders*, 6(1):249–262, 2012.

[9] J. Feil-Seifer and M. J. Mataric. B3IA: A control architecture for autonomous robot-assisted behavior intervention for children with autism spectrum disorders. In *ACM/IEEE Intl. Conf. on Human-Robot Interaction*, pages 328 – 333, 2008.

[10] J. Feil-Seifer and M. J. Mataric. Using proxemics to evaluate human-robot interaction. In *ACM/IEEE Intl. Conf. on Human-Robot Interaction*, pages 143–144, 2010.

[11] I. Fujimoto, T. Matsumoto, P. R. S. D. Silva, M. Kobayashi, and M. Higashi. Mimicking and evaluating human motion to improve the imitation skill of children with autism through a robot. *Social Robotics*, 3:349–357, 2011.

[12] M. A. Goodrich, M. Colton, B. Brinton, M. Fujiki, J. A. Atherton, D. Ricks, M. H. Maxfield, and A. Acerson. Incorporating a robot into an autism therapy team. *IEEE Intelligent Systems Magazine*, 27(2):52–59, 2012.

[13] O. Grynszpan, P. L. Weiss, F. Perez-Diaz, and E. Gal. Innovative technology-based interventions for autism spectrum disorders: A meta-analysis. *Autism*, 18:346–361, 2014.

[14] S. L. Harris and L. Delmolino. Applied behavior analysis: Its application in the treatment of autism and related disorders in young children. *Infants and Young Children*, 14:11–18, 2002.

[15] R. H. Horner, E. G. Carr, J. Halle, G. McGee, S. Odom, and M. Wolery. The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, 71:165 – 179, 2005.

[16] B. Huskens, R. Verschuur, J. Gillesen, R. Didden, and E. Barakova. Promoting question-asking in school-aged children with autism spectrum disorders: Effectiveness of a robot intervention compared to a human-trainer intervention. *Developmental Neurorehabilitation*, 16:345–356, 2013.

[17] E. S. Kim, L. D. Berkovits, E. P. Bernier, D. Leyzberg, F. Shic, R. Paul, and B. Scassellati. Social robots as embedded reinforcers of social behavior in children with autism. *Autism and Developmental Disorders*, 43:1038 – 1049, 2013.

[18] E. S. Kim and B. Scassellati. Bridging the research gap: Making HRI useful to individuals with autism. *Human Robot Interaction*, 1(1):26–54, 2012.

[19] H. Kozima, C. Nakagawa, and Y. Yasuda. Children-robot interaction: A pilot study in autism therapy. *Prog Brain Res*, 164:385–400, 2007.

[20] M.Begum, R. Serna, D. Kontak, and H. Yanco. Robots for therapy of individuals with ASD: Are we there yet? In *Intl. Conf. on Intelligent Robots and Systems (Workshop on Rehabilitation Robotics)*, 2014.

[21] G. Mesibov and V. Shea. Evidence-based practices and autism. *Autism*, 15:114–133, 2011.

[22] C. A. Pop, R. E. Simut, S. Pintea, J. Saldien, A. S. Rusu, J. Vanderfaeillie, D. O. David, D. Lefeber, and B. Vanderborght. Social robots vs. computer display: Does the way social stories are delivered make a difference for their effectiveness on ASD children? *Educational Computing Research*, 49(3):381–401, 2013.

[23] F. D. Reed, S. R. Hyman, and J. M. Hirst. Applications of technology to teach social skills to children with autism. *Developmental Neurorehabilitation*, 5:1003–1010, 2011.

[24] B. Robins, K. Dautenhahn, and J. Dubowski. Does appearance matter in the interaction of children with autism with a humanoid robot? *Interaction Studies*, 7:479–512, 2006.

[25] B. Robins, E. Ferrari, K. Dautenhahn, G. Kronreif, B. Prazak-Aram, G. Gelderblom, B. Tanja, F. Caprino, E. Laudanna, and P. Marti. Human-centered design methods: developing scenarios for robot assisted play informed by user panels and field trials. *Human-Computer Studies*, 68(12):873–898, 2010.

[26] B. Scassellati. Quantitative metrics of social response for autism diagnosis. In *Intl. Workshop on Robots and Human Interactive Communication*, pages 585– 590, 2005.

[27] B. Scassellati, H. Admoni, and M. Mataric. Robots for use in autism research. *Annual Review of Biomedical Engineering*, 14:275–294, 2012.

[28] A. Tapus, A. Peca, A. Aly, C. Pop, L. Jisa, S. Pintea, A. S. Rusu, and D. O. David. Children with autism social engagement in interaction with nao, an imitative robot. *Interaction Studies*, 13(3):315–347, 2012.

[29] J. Wainer, B. Robins, F. Amirabdollahian, and K. Dautenhahn. Using the humanoid robot kaspar to autonomously play triadic games and facilitate collaborative play among children with autism. *Autonomous Mental Development*, 6(3):183 – 199, 2014.

[30] Z. Warren, J. Veenstra-VanderWeele, W. Stone, J. L. Bruzek, A. S. Nahmias, J. H. Foss-Feig, and R. N. J. et al. Therapies for children with autism spectrum disorders. *Vanderbilt Evidence-based Practice Center: Agency for Healthcare Research and Quality (US) Tech. Rep. 11-EHC029-EF*, April 2011.