Bridging the Research Gap: Making HRI Useful to Individuals with Autism

Elizabeth S. Kim,
Department of Computer Science, Yale University
Rhea Paul, Frederick Shic,
Child Study Center, Yale University
and
Brian Scassellati
Department of Computer Science, Yale University

While there is a rich history of studies involving robots and individuals with autism spectrum disorders (ASD), few of these studies have made substantial impact in the clinical research community. In this paper we first examine how differences in approach, study design, evaluation, and publication practices have hindered uptake of these research results. Based on ten years of collaboration, we suggest a set of design principles that satisfy the needs (both academic and cultural) of both the robotics and clinical autism research communities. Using these principles, we present a study that demonstrates a quantitatively measured improvement in human-human social interaction for children with ASD, effected by interaction with a robot.

Keywords: Human-robot interaction, autism, methods, socially assistive robotics

1. Introduction

For more than a decade, a diverse set of robotics research groups have examined the responses of individuals with autism spectrum disorders to robots (see Scassellati, Admoni, & Matarić, 2012, or Diehl, Schmitt, Villano, & Crowell, 2012 for reviews). These investigations have been driven by the widespread incidence of ASD (estimated to affect one in every 88 children in the U.S. (Autism and Developmental Disabilities Monitoring Network, CDC, 2012), the need for early and sustained intervention, and the high levels of support needed by many individuals throughout their lives (Volkmar, Lord, Bailey, Schultz, & Klin, 2004). Among the core symptoms in ASD are impairments in social interaction and in communication (American Psychiatric Association, 2000). In recognition of these central deficits, researchers have made attempts at using non-human partners to facilitate social interactions, for instance through pet-assisted therapy (Martin & Farnum, 2002; Redefer & Goodman, 1989), computer-assisted therapy (Bosseler & Massaro, 2003; Hetzroni & Tannous, 2004; e.g., Silver & Oakes, 2001) and virtual reality-based approaches (Parsons & Mitchell, 2002; e.g., Strickland, 1997). While these have shown some success, there has been limited investigation of the parameters of the conditions necessary to generalize the benefits to interactions with human partners. Among non-human partners, robots that provide instruction, support, and assistance through social interaction have been seen as a potential

Because of a conflict of interest with one author and the guest editors, this paper was handled by an alternate editor, Sara Kiesler. The decision to accept or reject this paper was made without consulting the guest editor.

Journal of Human-Robot Interaction, Vol. 1, No. 1, 2012, Pages 26-54. DOI 10.5898/JHRI.1.1.Kim

mechanism for supporting therapy and daily living for individuals with ASD (Scassellati, 2005; Tapus, Matarić, & Scassellati, 2007). Robots promise unique practical advantages over other non-human interactive partners. First, they can potentially provide identical delivery of stimuli, establishing a uniquely high level of control in diagnosis and assessment (Scassellati, 2005). Second, they do not require the months or years of training that animal assistants may need, as robots can be designed to allow flexible customization. Third, they offer a potential for tactile interface and the immediacy of embodied agency, which computer programs and virtual realities cannot provide. In cases in which human or animal therapeutic aids may be unavailable or are prohibitively expensive, and where software or virtual reality therapeutic tools cannot provide sufficient embodiment, robots may provide an especially useful addition to therapy.

Studies from within the robotics and human-robot interaction (HRI) communities have shown exciting, but often preliminary, benefits to individuals with ASD, including increased engagement in tasks, increased levels of attention, and novel social behaviors such as joint attention and spontaneous imitation, when robots are part of the interaction (Diehl et al., 2012; Ricks & Colton, 2010; Scassellati et al., 2012). While these studies generate excitement within the robotics community, as well as substantial publicity and attention from news media, the results have gained relatively little attention from the clinical community; clinicians tend to view the application of HRI for autism as a science in its infancy. The reason for this is complex and stems from a series of cultural differences between the research practices of robotics and those used in clinical research for autism, some of which are enumerated by Diehl et al. (2012). While any interdisciplinary effort is likely to face challenges stemming from differences in terminology or lack of familiarity with each other's methods, in the case of robotics and clinical autism research, differences run deep and thus require conscientious negotiation to overcome.

This paper seeks to accomplish three goals: (1) to describe some of these cultural differences that currently hamper the uptake of HRI results into clinical autism research communities; (2) to suggest collaborative solutions, which may produce results that better demonstrate clinical utility; and (3) to present a novel research study on robots and children with autism, which demonstrates a proof of concept and illustrates our own collaborative solutions. Our perspective on these issues has grown from a ten-year collaboration between a robotics research group and a clinical research facility that specializes in the study and treatment of autism spectrum disorders. It is our hope that our suggestions will provide a roadmap to help the field of HRI for autism to outgrow its infancy (proofs of concept), and move into adulthood (clinical acceptance).

2. A Cultural Divide

Any two distinct and mature research fields are likely to have substantially different methodologies and research cultures. In this section, we describe some of the critical differences between the ways the HRI and clinical communities typically plan, carry out, and report experimental studies. For simplicity, we will refer to the *robotics community* to indicate the fields, groups, and venues within which most of the extant findings on robotics and autism have previously been published. This group is primarily represented by roboticists with backgrounds in computer science or engineering. The *clinical community* will refer to the fields and groups who conduct research on the diagnosis and treatment of autism, and the venues within which they share their findings. This group is represented by clinical practitioners, developmental psychologists, and social and behavioral therapists.

We emphasize from the outset that our purpose is not to cast doubt over the methods and practices of either community. Rather, it is our position that to exclusively adopt the methods of one community or another would hinder progress towards the ultimate goal of partnership between these communities: using robots to aid the diagnosis and therapy of individuals with ASD. To conduct research and development according to only one community's standards would render results inaccessible to the majority of the other. Instead, we propose collaborative solutions—ways to negotiate logistical compromises and to design to each community's standards—that address some of the most pressing concerns of each group while making the results at least partially accessible to both. We frame this discussion around the differences in three critical areas: research approach, study design, and publication and dissemination.

2.1 Research Approach

Within this collaborative space, the ultimate aim of both roboticists and clinicians is to determine the parameters within which and the mechanisms by which robots can improve interventions for, or assessment of, individuals with autism. Despite this shared ultimate goal, research approach and motivation differ. While we sometimes loathe admitting it, research in robotics is often driven by the capabilities of our robots rather than the needs of a target user. Funding awards, and their sponsored research endeavors, tend to focus on technological innovation, and the demonstration of feasibility of use. Each time a robot acquires a new capability, a search for applications that can take advantage of that new capability follows. The motivation for this approach is sensible: technological innovation can rapidly open new application areas and make fundamental changes to the kind of services that can be provided. Clinical research, on the other hand, is primarily driven by the specific needs of the target population. Funding and research efforts are directed toward questions that are most likely to reap substantial benefits for individuals with ASD. This fundamental and initiating distinction results in critical differences in the ways in which the two research communities approach collaborative works—as well as the ways funding agencies evaluate results. Clinicians have been hesitant to explore robotics technology in part because a clear case for the utility of robots in this area has not been made. What needs of a child with ASD does the robot fulfill, what support does it provide to the family, or what diagnostic value does offer to a clinician? To date there are no rigorous, controlled, sample-based demonstrations of a robot's improving symptoms, family support, or characterizations of individuals with ASD. Unfortunately, it is often not possible to answer these questions in advance of technology development. On the other hand, because little is known both about how to design human-robot interactions for individuals with autism and about how these individuals will respond and benefit from interactions with a robot, technology cannot be developed strictly in advance of deployment to address specific needs of individuals with ASD. At best, contemporary research collaborations strive to utilize a design process that considers input from diverse stakeholders (including clinicians, families, and other users), and iteratively advances technology to meet needs that are, in turn, iteratively specified by the user community.

Differences in fundamental approach between robotics and clinical research communities lead, in turn, to differences in desired outcomes from studies. At present time, roboticists in clinical collaborations tend to seek proofs of concept, that is, a demonstration of a robot's successful engagement in interactions that are pleasant or socially appropriate, or that resemble an assessment, therapeutic or educational scenario. While engaging interactions are fundamental to effective interventions or assessments, a proof of concept alone will likely be insufficient to motivate clinical use. In clinical studies, research is validated only when a clearly specified benefit to the end user has rigorously been presented. But demonstrations of engaging interactions with a robot do not necessarily show any specific clinical or functional benefit for the end user with ASD. From a pessimistic view a critic might claim that all existing robot-autism studies to date show only the ability for children with ASD to adapt to interactions with a robot, and that effectively training children to engage with robots will have no benefits to their ability to interact with other children or adults.

The study of HRI applications for autism is nascent. Given limited knowledge of the beneficial combinations of robotic form, type of interaction, and characteristics of affected individuals, at present time research efforts necessarily tend to focus on proofs of concept. In addition, efforts to define a clear transition model between human-robot engagements and human-human engagements, plans for moving from dyadic child-robot interactions to triadic child-robot-adult interactions, or other structural mechanisms offer the possibility of moving collaborations toward demonstrated clinical utility.

Collaborative studies can provide data to support investigation of questions uniquely asked by each individual community, as well as questions shared by both. Roboticists seek to improve technologies, in order to better investigate the uses of HRI in autism, and clinicians investigate behavioral or biological markers which may distinguish individuals with ASD from those with typical development, as well as cognitive mechanisms which may be activated during interaction with a robot (Diehl et al., 2012). All these are important questions to answer en route toward demonstrations of the clinical utility of robots. As this interdisciplinary field gains knowledge and data, both technologically- and clinically-focused investigations can be iteratively advanced and refined. Studies can simultaneously acquire clinical interaction data necessary for shaping robotics development while investigating the parameters facilitating clinical utility. For instance, roboticists are interested in fully automating robotic perception of, and response to, human actions. However, better understanding of (and data from) heterogeneous behaviors among individuals with autism is needed, in order to inform and train such designs. In the mean time, robots often operate under secret, manual control which affords the (false) appearance of autonomous robotic behavior (the Wizard of Oz paradigm; Riek, 2012; Scassellati et al., 2012). While using Wizard of Oz-style control, clinicians can make detailed observations of children's responses to robots, roboticists can acquire data which can inform next-generation autonomous perception and action technologies, and both sides of a collaboration can investigate proofs of concept.

2.2 Study Design

A second set of differences exists regarding the typical methodologies employed in each discipline. Evaluations of robotics technology often focus on proof-of-concept, that is, a demonstration of a system's effect for one or more people (often, $n < 5$). These small numbers are typically constrained by the research effort's focus on demonstrating the viability of the design and implementation, and the effort required to construct a reliable, well-engineered device. Focus on the technology may initially cause roboticists to overlook the significant resources required to test with specialized populations, the difficulties associated with accessing the target population, and methodological rigor in user testing. To date, clinical validity and applicability have been difficult to gauge in studies of robotic applications for autism. This is due to insufficient provision of standardized characterizations of participants; or to insufficient control allowing comparison between a robot's effects on individuals with and without autism, or comparison between effects of interaction with a robot and that with an alternative device or person (Diehl et al., 2012). The gold standard for proving efficacy of a medical or behavioral treatment is consistency in findings from multiple, independently conducted, randomized, double-blind clinical trials, each of which requires experimenters blind to knowledge of individual participants' assignments to comparative groups, and participants blind to the parameters of the experiment. Practically, however, double-blinding can be an extremely difficult standard to meet in autism research, because the differences between participants with ASD and controls is often apparent, and the nature and intention of a given task or intervention can be obvious to participants and to the experimenter. For this reason, clinical research in autism frequently uses alternative designs in order to evaluate the efficacy of an intervention (or specificity and sensitivity of an assessment). However it is necessary to approach these designs with appropriate levels of clinical rigor. As discussed by Reichow, Volkmar, and Cicchetti (2008) the clinical autism research community has defined rubrics for evaluating the validity of evidence from experimental interventions, for the purpose of practical dissemination and application. Such standards include using adequately powerful sample sizes for group designs, using appropriate control conditions in both group and single subject designs, and generally obtaining standardized characterizations of participants which can be compared to other research (see Reichow et al., 2008). With respect to study design standards, robotics researchers face a long tradition and deeply ingrained methodology and must adapt to the practices of the clinical community. Clinical standards are also not negotiable within the space of collaboration with roboticists because such standards impact legal, educational, and medical decisions regarding the provision of care to affected individuals (Reichow & Volkmar, 2011). Studies with larger, statistically valid, comparisons are beginning to be conducted (e.g., Feil-Seifer & Matarić, 2011) and the reporting mechanisms for single subject, or case study, designs (which require specific design considerations to have traction within the clinical community; see Kazdin, 2011; Reichow et al., 2008) have also begun to gain acceptance within the robotics community.

In moving to studies that adhere to clinical standards, more standardized mechanisms for participant recruitment, for reporting population statistics, and for the analysis of data with respect to control groups will become necessary. Many current robot-autism studies recruit participants in an ad hoc fashion, as obtaining access to populations for many groups is non-trivial, even within collaborations with clinicians. In addition, clear inclusionary criteria and recruitment procedures are essential to ensure a representative sampling, which is the basis of any statistical conclusion.

Along similar lines, clear characterization, as mentioned above, is fundamental for comparison among disparate research findings. Such comparisons, in turn, make possible definition and refinement of the parameters allowing effective application of HRI in autism treatment or assessment and the investigation into the cognitive mechanisms which such applications might engage (Diehl et al., 2012; Reichow et al., 2008). Participants in existing studies often have been described using a simple diagnostic label (or even just as "autistic"). As the expression of symptoms within ASDs are extremely heterogeneous and the level of impairment ranges from very mild to very severe, these simple labels are typically not sufficient for providing a clear picture of the abilities and selective deficits faced by these individuals (Diehl et al., 2012). In the clinical community rigorous characterizations of socio-cognitive abilities is performed for all study participants, using externally validated protocols (Reichow et al., 2008). For example, for ASDs, assessment tools include the autism diagnostic interview–revised (ADI-R; Lord, Rutter, & Couteur, 1994), the childhood autism rating scale (CARS; Schopler, Reichler, & Renner, 1986), and the autism diagnostic observation schedule (ADOS; Lord et al., 2000). These standardized tools allow for comparison of populations across research studies. These assessments can be lengthy and expensive, as each requires administration by a trained clinician, and each must have been performed close in time to the experimental study, as developmental changes in children with ASD can be substantial over short periods of time. Finally, most proof-of-concept studies from the robotics community focus exclusively on children with ASD and do not provide a comparative sample of typically developing children or children without ASDs having other impairments which are frequently comorbid or symptomatic of ASDs, such as intellectual disabilities or specific language delays. A common objection to existing studies is that many of the effects seen when children with ASD interact with robots (especially increased attention, and high motivation) would be seen in any child when they are given a new robot toy to play with. The use of control groups as described above is standard practice in the clinical community, but has only begun to have more widespread, and increasingly standard, usage in robotics. In these aspects, robotics groups will most likely need to adopt the more standardized reporting mechanisms of the clinical community. However, some flexibility from the clinical community must be offered, as very few research groups have the resources to span the range of assessment, engineering design, and large-scale testing required for a large statistical sample. For those robotics groups lacking access to highly experienced clinicians who have been specifically trained in administering ADOS or ADI-R, CARS may present a slightly more accessible alternative, administrable by physicians, special educators, school pathologists, and speech pathologists who may have little experience with individuals with autism. Another, even more accessible but clinically comparable alternative is the Social Communication Questionnaire (SCQ; Rutter, Bailey, & Lord, 2003), which can be completed by parents or primary caregivers, and which is frequently used in clinical studies to affirm control participants' negative diagnoses.

Also frequently important in clinical research are measures of other kinds of cognitive development, frequently measured with IQ tests such as the Differential Abilities Scale (Elliott, 2007), Wechsler Intelligence Scale for Children (Wechsler, 2003), or with the Mullen Scales of Early Learning (Mullen, 1995) where individuals may be too young for other tests. We advocate for the reporting of standardized IQ assessments, which we expect may be more readily available, given their utility in a broader range of disabilities and our assumption that professionals trained in their administration may be relatively accessible, particularly through schools.

Clearly there is a tradeoff to be made between resources devoted to characterization and comparability and specificity of characterization, and it is for each individual collaboration to negotiate this tradeoff.

2.3 Publication and Dissemination

A final set of cultural differences concerns the timing and location of publication and dissemination of research results. Both the clinical community and the robotics community have their own established publication standards and venues, and the differences between these standards has implications for reporting results, for expectations of young researchers regarding tenure and promotion, and the evaluation of students. High-quality results in robotics typically appear as shorter length papers (6 to 12 pages) in annual conferences, many of which are peer-reviewed, highly competitive venues and result in archival publications. A robotics student might be expected to publish 1-2 such conference papers each year, and a lengthier journal article that covers multiple conference publications appearing every few years. In contrast, the clinical community typically publishes their primary results as longer manuscripts (10 to 30 pages) in monthly or quarterly peer-reviewed, and similarly highly competitive journals. A student in the clinical community might be expected to publish one such paper every few years and to support that publication with the presentation of non-archival posters and talks, at conferences and meetings. These differences are perhaps the most difficult to overcome as they involve the expectations of the entire research communities who evaluate the work of these scientists, not just the researchers involved directly in the collaboration. An approach used in other interdisciplinary fields is to allow each collaborator to publish directly in their own preferred high-quality venue. This can be difficult in this case, as publication in an archival computer science conference proceeding can at times block publication in a high quality clinical journal, which expects all of the data reported to be first run material that does not appear in other archival sources. It is our experience that these issues can be accommodated only by clear communication between the research collaborators about their expectations and needs regarding publication and clear communication of the difficulties involved in these interdisciplinary research issues to reviewers of student performance, tenure and promotion committees, and project reviewers.

3. Suggested Bridges for Collaboration

Methods in each community are valid within each, and funding and other resources reflect—indeed determine—the expectations each community must satisfy in their research. Here we suggest ways to negotiate the cultural differences we've outlined above, to foster collaborations which can further efforts toward demonstrated utility of robotic applications to intervention and assessment of ASD.

Ultimately, to be successfully accepted as a diagnostic or intervention tool, a robot's utility must be demonstrated with statistical significance over a large sample. This standard is generally required in the medical community to establish the evidence basis of any diagnostic tool or treatment's efficacy. Obviously there are personally affective, cultural, and legal implications to establishing any treatment as evidence-based. In the case of communication interventions, few treatments have met this rigorous standard, and typically only over narrowly targeted behaviors (Prelock, Paul, & Allen, 2011). Along the way to this gold standard, there are other effective ways to establish validity within a clinical community. The key here is control. Interventions with broader behavioral targets frequently employ single case experimental designs (for example, changing criterion, reversal, multiple baseline, or alternating-treatment designs; see Kazdin, 2011) to establish non-statistical control over the many other changes developing children with autism may experience at the same time during which they receive treatment. Roboticists facing limited access to clinical resources may wish to consider single subject designs with rigorous control, such as a reversal (ABA) design, in which each participant's behavior is observed (A) before introduction of treatment (e.g., interaction with a robot), (B) just after or while treatment is being applied, and then (A) again, well after treatment has been withdrawn.

With respect to characterization and participant selection, researchers in both fields often face logistical (and funding) limitations on the assessments they can provide, as well as the participants they can recruit. As our understanding of the parameters allowing viable interactions between individuals with autism and robots improves, and as questions of utility become thus more possible to answer, we expect funding to explore specific subpopulations will become increasingly available. In the meantime, often given limited funding, both roboticists and clinicians must collaborate with other ongoing clinical studies having funding which can support expensive assessments. Thus, access to experimental participants is limited to collaboration with existing assessments. Here we suggest a compromise to both communities: that they recognize the intent of most current studies, in the application of robotics to autism, is to establish proof of concept, and that they allow incremental evolution in the specification of viable parameters; that is, that they forgive such proof of concept studies when their experimental samples are broader or slightly different from what in principle may be the ideal population for the application in question. To make such proofs of concept viable and useful, current research should seek as detailed a characterization as possible, to help further both communities' understanding of the technological and clinical parameters that allow individuals with autism to successfully interact with robots. Generally, we suggest that researchers from both communities recruit the largest number of participants that their resources allow, from the subpopulation whom they anticipate will demonstrate the greatest utility of the robotic application. Where n is small, we suggest that researchers design according to well-established single-case methodologies (Kazdin, 2011).

Publication may be the most challenging arena in which to negotiate collaboration. Typically, funding agencies supporting each party will expect first-author publication. How can collaborators split results into two publications without compromising ethics by withholding results from the first publication? There is no perfect solution to this problem. Rather, it is our experience that pre-nuptial agreements can be made (and often require adjustment, depending on results of primary and exploratory analyses), and will often be determined based on funding allocation and who is putting in the most effort and resources. Part of this negotiation can be to identify which research questions are better suited to which community, and then to design experiments and plan analyses according to the planned order of publication. Mechanistic or explanatory analyses tend to require much greater effort, which may be better supported by staff in larger clinical groups. Thus, proof of concept questions, which may require less effort to answer, may be better targets for robotics publications, especially because roboticists may be less interested in some of the finer analyses. Of course there is a lot of overlap, so negotiation is needed.

As technologies and proofs of concept evolve, collaborations between roboticists and clinicians may find greater opportunities to answer questions about the utility of robots in intervention and assessment. We expect that both communities will find it increasingly useful, at this point, to publish such findings in clinical venues, while technical innovations will likely make a greater impact within robotics venues.

4. Case Study in Social Response of Children with ASD to a Robot

To illustrate a collaboration between the robotics and clinical communities, here we present results from a novel experiment over a sizeable sample group of children with ASD, and a smaller but statistically powerful sample of age-matched typical controls. We will discuss the results of this experiment, and the choices, which supported this collaborative study.

4.1 Motivation and Research Questions

Studies showing successful therapy with visual biofeedback from surface muscular sensors for communication disorders (Andrews, Warner, & Stewart, 1986; Gentil, Aucouturier, Delong, & Sambuis, 1994) initially motivated us to consider using other technology-based feedback for therapy-like vocal prosody practice. Atypical prosody has frequently been reported as one of the telltale indicators of odd social behavior in individuals with ASD (Paul, Augustyn, Klin, & Volkmar, 2005). With a goal of determining the viability and utility of incorporating a robot into ongoing experimental interventions and assessments for affective expression in prosody production, we pilot tested a robot interaction in which four school-aged children (two females and twin males, ages ranging from 4.9 to 10.1 years) repeatedly practiced using encouraging prosody to help the Pleo robot (described in greater detail in section 4.1.1) complete a task.

In pilot tests, three participants appeared to exhibit more positive affect during and immediately following interaction with the robot. They also verbally engaged with the robot in repeated trials, producing prosodic and verbal expressions of encouraging affect when interacting with the robot. Two pilot participants also spoke more and engaged in more eye contact with the members of our experimental team, following interaction with the robot. The same two pilot participants also spoke to the robot with heightened variation in prosodic expression of affect. In addition, these participants seemed to make more eye contact and orient themselves to face experimenters more after interaction with the robot. These encouraging social improvements motivated us to examine the statistical stability of such effects, during and immediately after interaction with the robot.

We formulated two hypotheses. First, we expected that children with ASD and those with typical development (TD; that is, a control sample) would equally (a) engage in, and (b) enjoy, interaction with a social robot in a brief, repetitive verbal task. Second, we hypothesized that, more so than controls, children with ASD would show the following improvements in interpersonal social behavior following interaction with the robot: (a) higher levels of participation in pre-scripted one-on-one interviews, and (b) increased time spent facing the interviewer.

Though our pilot studies suggested improvements in eye contact, we did not measure this behavior due to technical limitations: manual eye-tracking was not possible because of insufficient video recording resolution. Initially we planned a third measure of change in interpersonal social behavior, namely that more so than participants with TD, participants with ASD would increase the variety of types of prosodically expressed affect after interacting with the robot. We have not completed analyses of affective prosody. The differences we observed in pilot testing were remarkable but subtler than can be captured by established five-emotion-category coding. We continue to work to establish stable, reliable measurements to describe the subtler affective variations we initially observed.

To address our two hypotheses, we recruited two comparison groups of school-aged children, a group with ASD and a control group with TD. We designed a three-part protocol, beginning with (1) a semi-structured interview to establish individualized baseline social behaviors, followed by (2) interaction with the robot, and ending with (3) a post-robot interview, used to gauge changes in each participant against his or her own baseline.

Primary dependent variables included Likert ratings of affective valence and engagement with the interviewer or task during robot interaction; and total duration, time spent speaking, and time spent orienting to face the interviewer, in the pre- and post-robot interviews. These measurements are described in greater detail in Section 4.2.4.

4.2 Study Design and Methods

Given our long-term aim to explore robots as intervention supplements for atypical prosody, we designed a robot interaction to provide opportunities for participants to practice encouraging prosody. To test our hypotheses regarding immediate effect following robot interaction, pre- and post-robot interviews were designed to balance natural conversation with controlled parallel structure to allow comparison between the two interviews.

Each participant interacted with the socially expressive robot Pleo (Figure 1) for 4 to 8 minutes. Before and after robot interaction, participants completed two brief (3- to 16-minute), parallel, semi-structured interviews.

The interviews and robot interaction all were conducted in a therapy and research examination room, in the presence of an interviewer (a clinically trained research assistant) and another adult who secretly operated the robot (author ESK or another, trained robotics graduate student). Following the final interview, children were offered optional, unstructured time (henceforth, *free play* time) to interact with Pleo.

The interviews and robot interaction were video recorded, and behavioral observations were annotated, following interaction, from these video recordings. When he or she would tolerate it, each participant also wore a lightweight head-mounted boom microphone for analysis of speech prosody production. (Prosody analysis is ongoing and is not presented in this article).

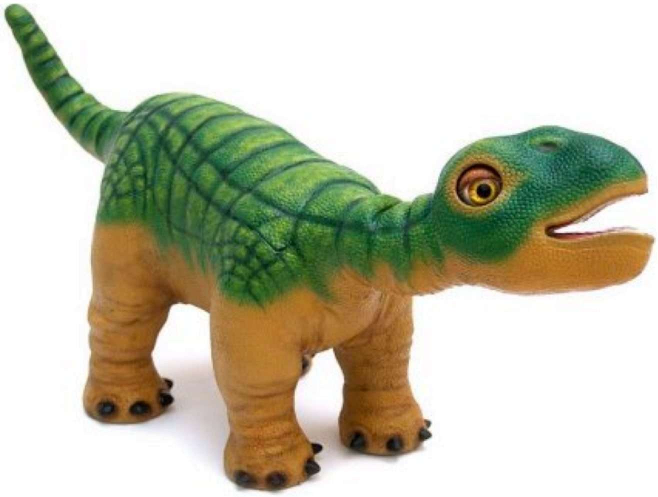


Figure 1. In our human-robot interaction study, participants spoke to Pleo, a small, commercially produced, toy dinosaur robot. Pleo was designed to be expressive of emotions and attention.

4.2.1 Participants

We recruited participants (ages 9 to 14 years) with and without a recent autism spectrum disorders diagnosis, and established two comparative groups of participants, ASD and control, respectively. The ASD group included 18 participants (15 male and three female; ages ranging from 9.1 to 14.97 years, M = 10.9, SD = 1.7). This gender ratio is roughly consistent with reported gender ratios of prevalence of ASD in the United States, of 3.5- or 4.0-to-1, male-to-female (Volkmar et al., 2004). A 19th participant was excluded from analysis because the robot interaction was interrupted by battery malfunction. The control group (ages ranging from 10.0 to 13.7, M = 11.7, SD = 1.3) included 11 participants (five female) with typical development and one (male) participant with specific language delay but no ASD diagnosis.

Diagnoses of children with a previous ASD diagnosis were confirmed (or ruled out in the case of the participant with specific language delay), using Module 3 of the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000), by two experienced psychologists at the Yale Child Studies Center, within one day of participating in the present study. Typical development diagnoses were confirmed using clinical judgment the lifetime Social Communications Questionnaire (SCQ; Rutter, Bailey, & Lord, 2003). All participants with typical development scored 8 or lower on the SCQ.

IQ was evaluated for the ASD group using the Differential Abilities Score (DAS; Elliott, 2007) and the Wechsler Intelligence Scale for Children–4th Edition (WISC-IV; Wechsler, 2003); and for the control group, the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999). The ASD and control groups were well matched on verbal and cognitive abilities, with all participants having Verbal and Performance (or nonverbal) IQ above 70 (ASD VIQ, M = 102.6, SD = 21.4; ASD PIQ, M = 107.8, SD = 19.5; control VIQ, M = 109.7, SD = 17.6; control PIQ, M = 111.7, SD = 14.2).

4.2.2 Robot, Robot Behavior, and Robot Control

The Pleo robot was used in the robot interaction portion of the study. We were motivated to use the Pleo platform by our past observation that adults with typical development spontaneously use intensely affective prosody when instructed to speak to the Pleo robot (Kim, Leyzberg, Tsui, & Scassellati, 2009). Pleo (Figure 1) is an affectively expressive, commercially produced, toy dinosaur robot, recommended for use by children ages 3 and up, and measuring approximately 21 inches long by 6 inches wide by 8 inches high. It was formerly produced and sold by UGOBE Lifeforms. It is untethered, battery-powered, and has 15 degrees of freedom. We extended third-party software to make Pleo controllable by a handheld television remote control, through the built-in infra-red receiver on its snout, allowing us to playback any one of 13 custom recorded, synchronized motor and sound scripts. Pleo plays sounds through a loudspeaker embedded in its mouth.

We pre-programmed Pleo with eight socially expressive and three walking behaviors (forward, left, and right). Each behavior included synchronized motor and nonverbal vocal recordings (performed by author ESK). Social behaviors are listed in Table 1. When Pleo was not executing one of these 11 behaviors, it performed an idling behavior to maintain the appearance of animacy. Pleo's idling behavior included occasional slight hip wiggling, head turning or raising, and subtle tail wagging, all of which were performed randomly in time.

We used Wizard-of-Oz style robot control (Steinfeld, Jenkins, & Scassellati, 2009), allowing participants to believe that Pleo was behaving autonomously, while an investigator secretly manually operated the robot. We chose Wizard-of-Oz style control to fulfill our design objective that the robot should express reliable, contingent social behavior in response to speech. We did not expect that speech recognition technology would be sufficiently reliable with this population to afford highly reliable perception. This is especially the case for the heterogeneous presentations of social behaviors we expected to encounter among children with ASD (Volkmar & Klin, 2005).

In order to obscure the true role of the robot controller to participants, the interviewer instead introduced the robot controller as "Pleo's trainer," who would observe the protocol in order to take care of Pleo and to gauge its progress in overcoming its fear. The robot controller sat in between and about two feet behind the interviewer and the participant. Throughout the protocol, the robot controller sat silently, watching the interview or interaction between the robot and participant, occasionally glancing down at papers on a clipboard. Very infrequently, if the participant or interviewer addressed the robot controller, she or he would respond. During the robot interaction segment, the robot controller used a handheld television remote control, hidden beneath the clipboard, to operate Pleo. The robot controller left her or his seat only to set up or remove Pleo for the robot interaction segment, and freely answered the participant's questions during optional, post-protocol playtime with Pleo.

Table 1: Pleo's eight pre-programmed affectively expressive behaviors. Pleo also was pre-programmed with a forward, left, and right walking behavior, and with an idling behavior to maintain the appearance of animacy.

Affect expressed	**Movements**	**Non-verbal vocalization sounds roughly like…**
Greeting or Affirmative	Tail wags, head raises.	a prolonged, enthusiastic "Hi!"
Fatigue	Legs bend, head lowers, tail lowers.	an extended, relaxed yawn.
Excitement	Tail wags vigorously, head rises high, hips wiggle.	"Woohoo!"
Fear and Surprise	Tail rises rapidly. Then tail lowers, hips quiver rapidly, head lowers.	a high-pitched abrupt "Oh!" followed by a quavering "Ohhh…"
Fear and Uncertainty	Tail raises, then hips and shoulders quiver, and head lowers.	"Eech!"
Boredom	Head and tail lower slightly and loll slowly, side-to-side.	a short, aimless, hummed melody.
Enthusiastic Affirmative	Head raises quickly, tail raises and wags briskly.	"Aye aye!"
Elation	Head rises, tail raises and wags, hips shake, legs bounce.	a victory song.

It is important to note that most children, including children with typical development, entirely or largely ignored the robot controller during the interviews, robot interaction, and optional following playtime. In addition, only one of 31 experimental participants and 5 pilot participants asked whether Pleo was controlled by remote, and neither that participant, nor any other, guessed that Pleo's trainer was in fact controlling the robot.

4.2.3 Experimental Protocol

We designed our robot interaction protocol to provide opportunities for children to speak to the robot using affectively expressive prosody, with the objective of examining effects on affective prosody toward another person following robot interaction. We were also interested in gauging the effects of social interaction with the robot on other social behaviors that are commonly problematic for speaking children with ASD. These include face-to-face orientation to another person; spontaneous production of topically relevant utterances; indication of interest, or relevant response to, a story told by another person; unusual focus on a topic of special interest; and appropriate expression of emotion using vocal prosody. Pre- and post-robot interviews in this protocol were designed to facilitate measurement over these various behaviors. Analysis of these behaviors is ongoing.

Protocol Environment and Instructions. Experimental procedures took place in a clinical testing room roughly identical to rooms (or, in some cases, the very same room) in which the participant completed a battery of other assessments and research protocols preceding this experiment. For the entire protocol, including both interviews and the robot interaction, participants sat facing a long table, with the interviewer seated about two feet to the side of the participant (during the robot interaction, the interviewer also served as a confederate, guiding the participant through the interaction.) The robot controller sat between the two, about two feet behind (farther from the table). Throughout the entire protocol, the tabletop was covered with a six-foot-long play-mat, illustrated with “Dino World,” a green- and brown-colored jungle scene, striped with a series of four blue rivers. The protocol environment can be viewed in Figure 2.

Prior to entering the protocol environment, the interviewer gave participants a brief overview of the protocol’s interview-interaction-interview structure. The interviewer also gave detailed instructions for the robot interaction: “After we talk for a few minutes, Pleo will come out. He is a small dinosaur robot. We are training Pleo to get over his fear of water. He will walk across Dino World. But it has rivers, and he is afraid of them. You can help him when he’s scared, by talking to him in your encouraging voice. Pleo’s trainer will be there, to make sure he’s okay and to see how he does.” When each participant entered the protocol environment, the interviewer introduced him or her to Pleo’s trainer (whose role as the robot controller was kept secret from the participant).

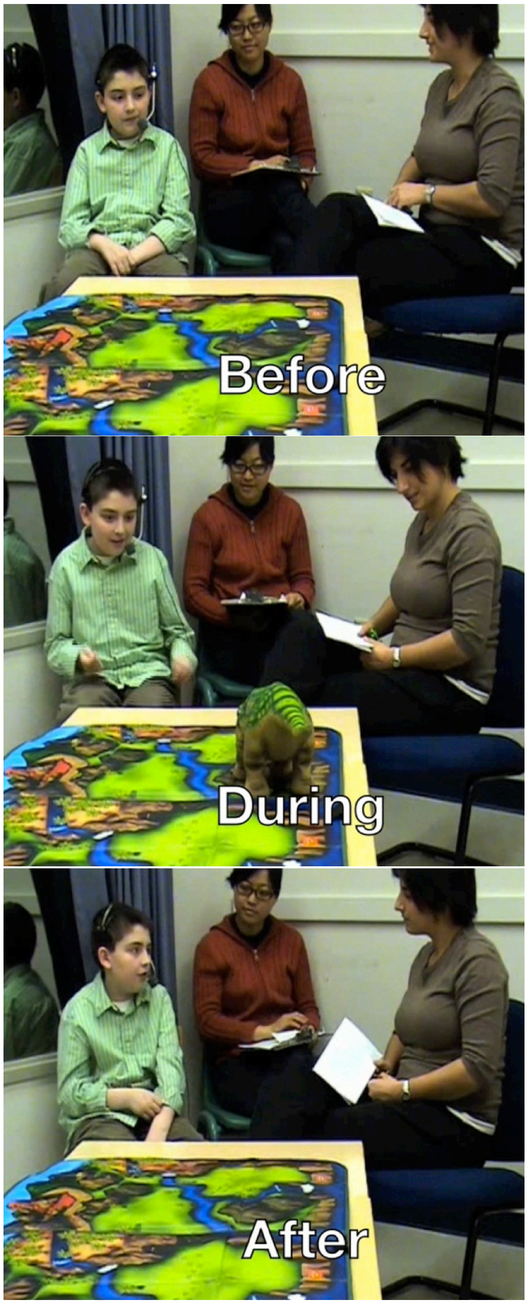
During the interviews, the robot was hidden in an unmarked cardboard carrying case. In pilot testing we observed that the robot’s presence distracted children from listening to instructions, suggesting that it would distract them from engaging in interviews as well. We also kept Pleo hidden to control potential effects of familiarization to the robot’s presence, between pre- and post-robot interview performance. For the post- robot interview, and if participants asked to play with the robot during the pre-robot interview, the interviewer explained, “Pleo is having a nap now.”

Pre- and Post-Robot Interview Protocol. Interviews were conducted in the same setting as the robot interaction, with the participant seated in front of the play-mat used in the robot interaction. The interviewer sat two feet to the left of the participant, and the robot controller sat between and slightly behind the participant and interviewer.

Pre- and post-robot interviews were semi-structured in the sense that the interview was conversational and allowed the participant to introduce topics of his or her own interest. However, the interviewer attempted to limit spontaneous discussion, in order to complete a pre-defined series of conversational objectives. As each objective was completed, the interviewer attempted to redirect the conversation to the next.

We designed the pre- and post-robot interviews to be almost entirely parallel in structure to each other, in order to facilitate comparison between the two and to control for confounding variations between the two. Each interview began with an opportunity for the participant to freely talk (for up to three distinct points of new information) about two of three topics suggested by the interviewer (animals, pets, and hobbies in the pre-robot interview, and previous experiences with robots, dinosaurs, and favorite things to learn about in the post-robot interview); a story told by the interviewer about a time when she needed encouragement; and two opportunities for the participant to spontaneously ask what happened next in the interviewer’s narrative. The interviewer then asked the participant to discuss a hypothetical or remembered episode in which someone helped, or could help, the participant by encouraging him or her. Finally, the interviewer asked the participant to model or recall (produce) an example encouraging utterance that was, or might be, helpful. Abbreviated examples of prompts delivered by the interviewer, in both the pre- and post-robot interviews are provided in Table 2. The scripts illustrate the parallel structure of the interviews, which was designed to control for conversational content and turn-taking balance when comparing pre- and post-robot social behaviors with the interviewer.

Throughout the participant’s conversational turns in the first two interview tasks, the interviewer responded to the participant’s utterances. For example, one participant said, “I’m interested in history,” to which the interviewer responded, “Yeah, you had World War II history books with you yesterday.” To provide opportunities for the participant to show interest in the interviewer’s personal story, the interviewer first paused for three seconds, and if the participant did not comment or ask about the interviewer’s story, the interviewer asked, “Do you want to know what happened?”



(a) Pre-Robot Interview

(b) Robot Interaction

(c) Post-Robot Interview

Figure 2. These three images, captured from a video recording of a participant with ASD, show the (a) pre-robot interview, (b) robot interaction, and (c) post-robot interview, within our clinical testing environment. During robot interaction, the Pleo robot walked across the illustrated play mat, toward the participant. Pictured (from left to right) are a participant, the robot controller, and the interviewer. In the post-robot interview, this participant spent 11% more time facing the interviewer than he did in the pre-robot interview. In the ASD group, we found such the size of such increases to be negatively associated with age.

All interviews were conducted by a research assistant with extensive clinical experience in conducting experimental language and communication protocols with children with ASD. Interviews were 3 to 16 minutes long. The few longer interviews stretched out because of the participant's hesitations to respond or persistent redirection to topics of his or her own interest. A few interviews lasted slightly longer because the participant left his or her seat, at which point the interviewer had to coax the participant to be reseated before resuming the interview.

We controlled for effects of novelty of the first interview and increasing familiarity to the second interview in two ways. First, participants were familiar with the interviewer because our protocol concluded a one- to two-day battery of assessments and experimental protocols, over which she hosted all participants. In addition, the interviewer conducted four of these preceding protocols, including two with experimental protocols featuring brief interview components, the Gray Oral Reading Test (Wiederholt & Bryant, 2001), and an experimental protocol to assess Theory of Mind.

We controlled for novelty and familiarization to the interview structure by designing our semi-structured interviews to roughly parallel the structure of another longer (30- to 40-minute) experimental protocol, the Yale in vivo Pragmatic Protocol (YIPP; (Paul, 2005)), which all participants completed within one day of, and prior to, our study. The YIPP is designed for children ages 9 to 17 years, and like our interviews, provides opportunities for the child to spontaneously expound on a topic of choice, and to indicate interest in the YIPP interviewer's stories about herself. Although YIPP interviews were conducted by another clinician (not our protocol's interviewer), the parallel interview structures were intended to control for novelty and familiarity effects of the semi-structured interview format.

Robot Interaction Protocol. The robot interaction protocol was designed to provide the participant with opportunities to direct affectively and verbally encouraging utterances to the robot. The interviewer mediated the robot interaction by providing instructions to the participant, and by reminding the participant to speak, or clarifying Pleo's affective communications to the participant, if he or she hesitated to speak to Pleo.

At the end of the pre-robot interview, the robot controller or interviewer brought Pleo out from its unmarked cardboard carrying case to the start position on the far end of the play-mat, with its face oriented approximately toward the participant. The robot controller or interviewer then sat again. The robot controller remained silent unless the participant or interviewer directly addressed her or him. Participants rarely addressed the robot controller, and the interviewer typically addressed the robot controller only occasionally, when Pleo's feet became caught on the play-mat. While the robot controller placed Pleo at the start of play-mat, the interviewer briefly reiterated instructions to the participant: "Use your encouraging voice when Pleo gets scared of crossing the rivers." At this point, the interviewer introduced the participant to Pleo.

The robot interaction protocol opened with a brief introductory sequence to familiarize the participant with Pleo's communicative capabilities, followed by Pleo's walking across the play-mat toward the participant. For the familiarization sequence, the interviewer guided the participant through two tasks: a greeting to Pleo and a directive to begin crossing the play-mat. The interviewer first instructed the participant to greet Pleo. If after a second prompt the participant would not do so, the interviewer greeted Pleo: "Hi, Pleo!" Pleo responded to the participant's or interviewer's greeting by expressing the behavior *Greeting or Affirmative* in return, raising its head, nonverbally vocalizing a greeting, and wagging its tail (Table 1 describes Pleo's eight affectively expressive behaviors in detail). The interviewer then instructed the participant to tell Pleo, "Let's get started!" Again, if after a second prompt the participant would not tell Pleo to begin, the interviewer did so instead. Pleo responded to the participant or interviewer's directive by expressing an *Enthusiastic Affirmative*.

At each blue river painted on the play-mat, Pleo stopped walking and expressed *Fear and Surprise*, to elicit robot-directed speech from the participant. At each river crossing, a series of three increasingly restrictive prompts were delivered by Pleo and the interviewer, to encourage the participant to speak to Pleo. These prompts were structured in the style of errorless teaching, such that any speech toward Pleo was accepted. Following Pleo's initial *Fear and Surprise* expression, after a 3-second pause, if the participant did not speak to Pleo, the robot then expressed *Fear and Uncertainty*. If after a 3-second pause, the participant still did not speak to Pleo, the interviewer told the participant, "I think Pleo is scared. You can help him by talking to him in your encouraging voice." Finally, if after a third 3-second pause, the participant still did not speak to Pleo, the interviewer herself encouraged Pleo, for example, "Don't be scared, Pleo. You can cross the water!" Once the participant or interviewer had spoken to Pleo, whether encouraging or not (e.g., one participant expressed disgust at Pleo's hesitation at the fourth river and said, "Come on, Pleo. It's just water."), Pleo expressed an *Enthusiastic Affirmative*, crossed the river, and then expressed *Excitement*. The interviewer then narrated, with a variation of the phrase, "He did it! I think talking to him helped!"

Table 2: Prompts for parallel, semi-structured pre- and post-robot interviews.

Objective or Task	**Pre-Robot Interview Script**	**Post-Robot Interview Script**
Free exposition (two of three topics)	Do you have any pets at home? Do you have a favorite animal? I used to collect hippo toys. Do you collect anything?	Have you played with a robot before? What do you know about dinosaurs? What do you like learning about?
Interest in interviewer 1	When I was younger, I was afraid to learn to swim, and it caused me trouble. *Pause for 3 seconds.*	I used to love playing video games. One time I got in a bit of trouble because of it. *Pause for 3 seconds.*
Interest in interviewer 2pre	My girl scout troupe was planning a canoeing trip, and I was the only one left who hadn't passed the swim test. *Pause for 3 seconds.*	(See 2post below.)
No task (interviewer resolves her story)	My dad was really encouraging. He'd say, "Don't worry! You'll be ok. Just give it a try!" That really helped me.	I stayed up past my bedtime one night to beat the game. My brother stayed up with me and encouraged me. He said, "You can do it! Keep going!" I got in trouble for staying up late, but I was happy I beat the game.
Interest in interviewer 2post	(See 2pre above.)	I don't really play video games anymore. *Pause for 3 seconds.*
No task (interviewer resolves her story)	In the end I got to go canoeing.	My PlayStation broke.
Describe encouraging situation	If you were scared to do something, do you think it would help if someone encouraged you?	Do you think it would help if someone encouraged you with something you had trouble with?
Model encouraging statement	What kinds of things did/would they tell you to help?	What kinds of things did/would they tell you to help?

After crossing the final river, Pleo stepped across a finish line marked with red tape, off the play-mat, and onto the end of the table, inches away from the participant. Pleo expressed *Elation* (a victory song and dance), and the interviewer congratulated the participant on helping Pleo finish his task. The interviewer then explained that Pleo would rest while they spoke (for the post-robot interview), and the robot controller removed Pleo from the table and returned it to its carrying case.

Free Play Protocol. Following the post-robot interview, the interviewer asked each participant if he or she would like to play with the robot. Three participants (all with ASD) were not offered free play, due to time constraints. In addition, all but three participants (two with ASD, and one with TD) who were offered free play accepted, and if parents were available, they were allowed to join the free play interaction. Free play was discontinued when participants or parents chose to stop, or when the interviewer determined that the participant was losing interest.

4.2.4 Social Behavior Measurements

The dependent variables in this experiment are measurements of the quality of participants' social behavior. During the robot interaction portion of the protocol we judged ratings of *affective valence*, and of *engagement* in the robot encouragement task (or engagement with the robot or other people). During the pre- and post-robot interviews we annotated, and summed the durations of, brief episodes during which the participant turns his or her head to face the confederate or the robot controller (*face-to-face orientation*); and measured the interviews' durations themselves (*pre- and post-robot interview durations).*

As part of an exploratory analysis, we also measured the duration of the optional free play session, which followed the post-robot interview (*free play duration*).

Affective Valence During the Robot Interaction. Two raters independently judged the valence of each participant's affect, from video recordings, for 5-second intervals of the robot interaction, judging one out of every four 5-second intervals (or 5 of every 20 seconds). Affective valence was rated on a Likert-type scale from 0 to 5; where 0 and 1 represented intensely negative and negative affect, respectively; 2 and 3, neutral affect with more negative than positive valence, or visa versa, respectively; and 4 and 5, positive and intensely positive affect, respectively. Inter-rater reliability was measured both as percent agreement and as weighted kappa, in both cases, allowing raters to disagree by one point. Agreement was 98%, and kappa was .78.

Engagement During the Robot Interaction. Two raters independently judged video recordings for engagement and compliance of each participant's engagement in the task, or engagement with the robot, the confederate, or the robot controller. Again, these ratings were determined for one out of every four 5-second intervals (i.e., 5 of every 20 seconds) of the robot interaction. Engagement was rated on a Likert-type scale from 0 to 5. Ratings of 0 and 1 represented intense non-compliance and non-compliance, respectively. For example if, during the 5-second interval in question, the participant stood and walked away from the table on which the robot interaction took place, this interval would receive a 0 rating for engagement; or if, in a 5-second interval, the participant hung his head and refused to comply with the interviewer's request to speak to the robot, the interval in question would receive a rating of 1. Ratings of 2 and 3 indicated neither non-compliance nor positive display of interest in the task or with the confederate or robot controller, with more or less reinforcement required on the part of the confederate. For instance, if the participant complied with instructions to speak to the robot, or answered the confederate's questions, but only after several prompts from the confederate, this would warrant a rating of 2; or if the participant required two to three prompts from the confederate before responding or speaking to the robot, even if the reason was interaction with the confederate, this warranted a rating of 3. Ratings of 4 and 5 indicated positive expressions of engagement with the task or other people. For instance, a rating of 4 was given to intervals in which the participant complied immediately following the confederate's request to speak to the robot or answer a question, or in which no request was made while the robot walked, and the participant maintained their gaze on the robot, or looked at the confederate or robot controller without disrupting the progress of the task of speaking to the robot. A rating of 5 was given if the participant spontaneously engaged the confederate or robot (e.g., created encouraging phrases to the robot which had not been offered as examples by the confederate, or spoke to the robot spontaneously and not only when the confederate had instructed the participant to speak), or changed his or her posture (e.g., leaned forward) to nonverbally interact with the robot. Inter-rater reliability was measured both as percent agreement and as weighted kappa, in both cases, allowing raters to disagree by one point. Agreement was 95%, and kappa was .67.

Face-to-Face Orientation During Interviews. We did not plan to explore questions about eye contact because we did not expect participants to tolerate wearing automatic head-mounted eye tracking devices, or to remain stationary enough to facilitate automatic table-mounted eye tracking; and because our video recordings were not of sufficient resolution to facilitate manual eye tracking.

Instead, we explored *face-to-face orientation*, the behavior during which a participant turned his or her head to face the interviewer. (Given limited space, the participant's and interviewer's chairs were typically left facing the play mat, roughly parallel to each other, and never arranged such that an angle formed between their front edges would form an angle smaller than 90 degrees). Face-to-face orientation appears to be a novel metric in autism research. We initially considered face-to-face orientation as a surrogate for eye contact, as these two behaviors overlap in function (it is difficult to make eye contact with someone, without turning to face that person). However, there also appear to be some distinct functions, as well. For example, a listener or speaker may break eye contact in concentration or lower her eyes to reduce the affective intensity of a conversation, but may still orient her face to face the other's. Face-to-face orientation appears to give conversation partners access to one's facial expression, for instance, of affect.

From video recordings, we used VCode software (Hagedorn, Hailpern, & Karahalios, 2008) to mark the beginnings and ends of episodes during which each participant angled his or her head such that he or she was oriented face-to-face with the interviewer. Face-to-face orientation was defined as occurring when the angle between the participant's and the interviewer's faces was smaller than 20 degrees in any direction. More specifically, this angle was defined at the intersection between two vectors, each parallel to the participant or interviewer's line of sight, if the eyes were looking straight ahead. Face-to-face orientation episode markings were verified in VCode, which can synchronize visualizations of episodes and video from which they are annotated. In this paper we analyzed the percent time spent in face-to-face orientation (as a sum over the duration of all episodes, divided by the duration of the video).

Face-to-face orientation was examined for children in the ASD group and a small subset of children in the TD group (n=3) in both the pre- and post-interviews. Fewer children were examined in the TD group because of analysis time-constraints and expectations of a wide separation between the ASD and TD group on this measure.

Interview Durations Before and After Robot Interaction. We annotated the beginning of the pre-robot interview to be the time when the confederate led the participant into the protocol setting (the clinic testing room) and introduced the participant to the robot controller. The post-robot interview began after the robot controller or confederate removed the robot from the table. Both pre- and post-robot interviews ended when the confederate delivered or reminded the participant of instructions for the next segment of the protocol (i.e., robot interaction or free play, following the pre- and post-robot interviews, respectively). The duration of the interview was largely controlled by the participant, and as such provides an easy to calculate surrogate for the child's willingness to continue and elaborate upon the presented interview scenarios. Note that this measure is not without its complexities, a point we will return to in the discussion.

4.3 Results

Like typically developing controls, participants with ASD had no difficulties engaging with the robot as indicated by engagement ratings (TD: M=4.36, SD=.50; ASD: M=4.27, SD=.62; t(27)=.39). Similarly, affective valence during the robot interaction was similar between groups (TD: M=3.68, SD=.63; ASD: M=3.60, SD=.69; t(28)=.33).

There was also no difference in the amount of time children with ASD spent in the pre- or post-robot interview (pre: TD: M=267s, SD=57s; ASD: M=286s, SD=108s; t(27.8)=.65; post: TD: M=312s, SD=72s; ASD: M=395s, SD=199s; t(21.7)=1.6), nor any between-group differences in additional time spent in the post-robot interview as compared to the pre-robot interview (i.e.

$time_{delta}$ = $time_{post}$-$time_{pre}$) ($time_{delta}$: TD: M=45s, SD=85s; ASD: M=86s, SD=166s; t(25.1)=.85). However, within-subject, paired comparisons between the time spent in the pre- and post-robot interviews indicated a significant increase of children with ASD (t(16)=2.13, p=.05) but not for TD children (t(10)=1.7, p=.11). The ASD group also spent significantly longer than the TD group playing with the robot during free play (TD: M=207s, SD=49s; ASD: M=307s, SD=137s; t(20.3)=2.7, p=.02, Cohen's d=0.97).

Children with ASD, as compared to TD children, appeared to face the interviewer less in both the pre- (TD: M=.76, SD=.34; ASD: M=.30, SD=.26, d=1.52) and in the post-robot interviews (TD: M=.85, SD=.14; ASD: M=.31, SD=.23, d=2.84). Because of the small N of the TD group for this measure, the group difference in face-to-face looking ratio was not significant (p=.14); even so the post-robot group difference was highly significant (p<.01). Paired t-tests (that is, repeated measures) analysis among individuals in the ASD group showed no change in face-to-face looking ratios from the pre- to the post-robot interview.

An exploratory analysis of the cognitive and behavioral associations with primary outcome variables in ASD indicated trends such that those participants, with smaller increases in pre- to post-robot interview duration ($time_{delta}$), displayed more negative affect during the robot interaction (affect x $time_{delta}$: r=.48, p=.050), worse language skills (CELF language standard score x $time_{delta}$: r=.50, p=.042), and greater levels of social and behavioral impairments (ADOS total: r=-.49, p=.055). In children with TD, the affect relationship was not observed (affect x $time_{delta}$: r=.14), but the same direction of the language relationship was suggested (CELF language standard score x $time_{delta}$: r=.48, p=.14). No ADOS scores were available for the TD group. For time spent face-to-face in children with ASD, the change from the pre- to post-robot interview was negatively associated with chronological age (r=-.66, p<.01); this relationship with age was not noted for either the pre- (r=.33, p=.25) or post-robot interview face-to-face ratios (r=-.22, p=.45).

5. Discussion

5.1 Summary of Results

The results of this study confirm our first hypothesis, that children with ASD and their typically developing peers engage and enjoy verbal interaction with the robot to similar extents. Our observations only partially support our second set of hypotheses: a) compared to the TD group, children with ASD spent more time in the post-interview process, but b) did not show a greater increase in face-to-face orientation. We will discuss the implications of these findings in turn.

It is clear that children with ASD were motivated to interact with the robot, as indicated by their greater predisposition to spend time with the robot (compared with children in the control group) when given an option to play with the robot freely subsequent to the post-robot interview. Although two of the three participants who opted out of playtime had an ASD, in one case it appears he exhausted himself by talking at great length during the post-interview about the robot with the interviewer. In a second case, the participant was suspected of having comorbid diagnosis with oppositional defiant disorder, and was generally reluctant to participate in all phases of this and other experiments during his visit. In the case of the one child with TD who opted against free play, she simply seemed uninterested in playing with the robot. During free play, some participants in both the control and ASD group showed great ingenuity in understanding the robot's "water"-sensing mechanisms or logic, or in exploring what the robot enjoyed, feared, or understood. For example, participants frequently hypothesized that the robot was programmed to fear the color blue, which it sensed through the camera on its snout. Several participants tested their hypotheses by finding blue objects in the room and holding them up to Pleo's snout.

The results of our study, which confirm our first hypothesis, add to the mounting evidence that robots may be a highly tolerated, and even enjoyable, component of intervention for children with ASD. It is important to note that during the robot interaction phase, children with ASD showed similar levels of affective enjoyment and engagement as typically developing children. By comparison, in natural social situations and under laboratory testing conditions, children with ASD often exhibit limited affective response (Joseph & Tager-Flusberg, 1997; Kasari, Sigman, Mundy, & Yirmiya, 1990; Yirmiya, Kasari, Sigman, & Mundy, 1989).

The observation of an increased change in time spent in the post-robot interview session for children with ASD as compared to TD children (our second hypothesis) suggests that interaction with the robot may lead to greater verbal elaborations or increased verbal participation by children with ASD. However, it is important to note the limitations of this very coarse measure. First, though the variability of the interview duration is largely controlled by the participant, the interview itself is pre-scripted and pre-planned. For this reason, there is a limit to how much leeway each child can be afforded in terms of true "back-and-forth" verbal exchanges with the interviewer. Second, many of the questions can be answered quite succinctly (e.g. "Do you have any pets at home?"); to spend additional time in these phases of the interview may suggest difficulty understanding the question or may result from the expression of incoherent, meandering, or off-topic responses. Third, though the structure of the pre- and post-robot interaction interviews was designed to be parallel, we cannot rule out the possibility that specific tasks may be more accessible to participants in one group versus the other. For example, the "encouraging situation" question (Table 2) may require access to long-term memory in the pre-interview, but in the post-test an example might be readily accessible from the robot-interaction phase. Of course, the design of conducting two parallel interviews sequentially over a short period of time may in of itself bias results, as participants may become more comfortable with increasing interactions to the interviewers. The measure of "total interview duration" thus coarsely suggests the behavioral changes resulting from robot interaction may be found for individuals with ASD, but does not isolate the mechanism, nor provide an unambiguous description of causal relationships or quality of the responses. Further fine-grained analyses of the video recordings will have to be conducted to decipher the underlying structure responsible for increased changes in pre- to post-robot interview duration; an expanded study, consisting of repeated exposures over multiple sessions, would be necessary to gauge the generalizability and repeatability of our observations.

Along these lines, a close examination of the variability associated with the outcome measures examined in this study suggested a wide heterogeneity of responses in the ASD group. In an effort to decode this variability, we examined correlations between clinical features and performance metrics. The results suggest that those children with fewer social and communicative deficits responded to the robot interaction with greater enthusiasm, as reflected by increased time in post-robot interviews and higher affect ratings while interacting with the robot. This suggests that while, as a group, children with ASD behaved similarly to those with TD for most aspects of the robot interaction, responses to the robot interaction were modulated by the degree of socio-cognitive impairments. However, these relationships may also suggest that floor effects could exist in the outcome measure of total interview duration. In other words, the positive relationship between language skills and increases in interview duration post-robot interaction suggest that it is the more verbally capable participants with ASD who may be responsible for the observed between-group (TD vs. ASD) differences in increased interview time; conversely, the children with lower verbal ability may be stretched to their capacity in both the pre- and post-robot interviews. Again, these results highlight the need to carefully examine the relationships among hypotheses, outcome measures, and the individual characteristics of participants in interpreting the results of interactions between robots and children with ASD.

It is also clear, however, from our analysis of face-to-face interactions, that interacting with the robot does not generally result in increased orienting towards the face for our participants with ASD. Though children with ASD exhibit the expected decreased orienting to the face before the robot interaction, the frequency of their decreased looking remains virtually unchanged post-robot interaction. Though from a certain point of view this result is disappointing, from another point of view the result is quite understandable. In the state in which this study was conducted (Connecticut, US), the standard of care for individuals with autism is quite high. In fact, a recent study of community and standard care practices in toddlers with ASD suggests that treatment-as-usual now produces results that are competitive with more specialized intervention programs (Steiner, Goldsmith, Snow, & Chawarska, 2012). It may be optimistic to assume that an extremely brief guided interaction with a robot might be able to effect a change on one of the most highly-targeted behaviors for individuals with ASD: eye-contact and natural conversation. However, examination of the relationships between change in face-to-face looking pre- to post-robot interview showed a prominent negative relationship with age, suggesting that the paradigm as a whole might be more suited, and more effective, for younger children with ASD.

While there are many other aspects of this data that can be examined, especially regarding the degree to which children utilized appropriate prosodic intonation and their production of socially appropriate behavior, here we report only on a subset of possible measures that 1) point towards potential effects that may be due to engagement with the robot, 2) further our understanding of the applicability of the design across the heterogeneity of the autism spectrum, 3) are immediately available and accessible, and 4) illustrate points regarding clinical-HRI partnerships. In the case of this study, the working agreement we have with our clinical partners is to first publish the results of the study here, within a robotics-oriented venue, while preparing for additional analyses that will clarify and solidify our understanding of our data.

5.2 Our Collaborative Strategy

This study illustrates our approach to collaboration, which we hope will help other roboticists and clinical researchers to understand and navigate the cultural differences between their respective fields. We will in turn examine specific points that we have highlighted in our examination of differences in practice, using the presented work as a case study of a collaborative strategy, maximizing the advantages of both fields while eliminating the greatest barriers from collaboration and communication.

In terms of our research approach, we chose to focus on proof of concept, that school-aged children with high functioning and ASD would engage and enjoy a verbal task with an inexpensive, commercially produced robot under seamless interactive control. Although we are interested in automating the robot's perception and behaviors, we chose to focus on the proof of concept by using Wizard of Oz-style control, and to use our investigation of proof of concept to gather data that may support future technological research into automation. We also furthered our clinical agenda by collecting copious speech data and interaction data, which we will continue to analyze, to understand in greater detail the ways that participants interacted with the robot and with people afterward.

Our agreement was to publish results from our study first as a proof-of-concept manuscript in a robotics-oriented publication venue. This was acceptable, and necessary, for several reasons. First, as one of the few larger-N studies of robot-child interaction in autism research, this study highlights the applicability, acceptability, and potential of social robots to effect meaningful change in children with ASD. Publishing this paper sooner, rather than later, enables other researchers to see the advantages of these larger designs and the advantages of detailed clinical characterization in informing our understanding of what works and for whom. Second, it is important that roboticists, who will be on the front line of implementing the technically challenging but critical elements of HRI studies of autism, be given ample information regarding the details and hurdles that will help them design similar studies. Early dissemination of the study protocols and provision of usable (though not ideal) metrics of evaluating change will help these researchers adapt their own platforms and speed up development and evaluation time. Third, and perhaps one of the key issues informing our decision to publish these results here first, is that we estimate that the design, creation, implementation, verification, and evaluation of more detailed measures of interview dynamics, prosody, and semantic content will take approximately 3 months of time at our available level of funding. Factoring in additional statistical analyses and rigorous accounting of individual participant characterization variables, we estimate that the next iteration of these study results will be completed in 5 months.

This decision did not come about haphazardly, but instead reflects our lengthy discussions and a priori agreements well in advance of the start of the study. Of course, research is not a static process, and, when dealing with such a new field such as HRI studies of autism, it is difficult to predict exactly what methods, techniques, protocols, and measures will bear fruit. Here we were guided by clinical insights that informed our study design in advance, and a long-term collaboration built around understanding each party's expectations. We expected that it would be necessary to publish preliminary analyses and proof-of-concept before a final, more detailed examination could fully explore the space of our results. The clinical members of our research team, in turn, expect (and it is our expectation, will receive) our full support in the second iteration of analyses.

Of course, such agreements come also with consequences. First, because we froze the current state of analyses to submit this study, the measures that we employ are necessarily coarse, and to an extent, incomplete. Our study could benefit, for instance, by detailed ratings of affect and engagement during the interviews. Consistent with standards in the field, this would also require a second coder to confirm the accuracy and reproducibility of the more qualitative assessments. Our study could likewise benefit from a careful transcription of verbal exchanges during interviews, complete with timings of utterances. We could then distill from these data sets measures relating to the frequency of verbal production by the children, the semantic content of their speech, and the dynamics of the conversation between the child and the interviewer. Finally, difficulties in obtaining reliable operationalized protocols for evaluating prosodic quality in interviews for children with ASD suggest that standard approaches need to be adopted to capture more subtle prosodic differences between study participants with ASD and the control group.

The second consequence of our decision to publish these results here first are that it may make it more difficult to later publish results regarding the second iteration of analyses in a more clinically themed journal. It was the opinion of the clinical members of our team that though this concern was valid, the more detailed and clinically-oriented second round of analyses and interpretations should make the second manuscript quite distinct from the presentation here. In other words, it was a risk that everyone was willing to take.

5.2.1 Understanding Differences in Approach

At the intersection of robotics and autism research, differences in approach result in a number of potential pitfalls. Researchers in engineering fields typically focus on the development of methodologies, approaches, and processes. By contrast, researchers from clinical fields focus on specific issues relating to clinical populations. While the robotics community tends to focus on novel platforms for delivering treatment, the clinical research community focuses primarily on the treatment itself. A researcher in the robotics community gains greatly from expanding the vision of the possible, and so a successful proof of concept is in many ways a sufficient enterprise in and of itself. Yet, applications tied only to proof-of-concept studies, even though they may provide great benefits to a clinical population, may be left languishing in the land of "potential ideas" for years without a direct translation of those ideas into clinical applicability. This is quite a dangerous position, because without feedback from researchers focusing on clinical utility, the robotics community may drive novel technologies in unproductive directions while neglecting application areas that may have demonstrable clinical impact. Similarly, approaches that focus exclusively on tried-and-true engineering tools and platforms may be left languishing in the equally perilous land of "outdated technology" when more modern and capable technologies provide possible solutions that could not have been considered with more mature technologies. Without attending to the rapidly changing landscape of technical advancement, clinicians face the difficult prospect of struggling to adapt technologies that have already been replaced with more convenient, efficient, or capable solutions.

The study outlined in this manuscript, we believe, illustrates a way in which healthy collaborations between robotics labs and clinical enterprises can be formed. Beyond the typical skills that are necessary for any collaboration to succeed (e.g. mutual respect, open dialogue, rapid feedback), it was necessary for both our groups to understand our respective differences at a much deeper level.

5.2.2 Understanding Differences in Study Design

As mentioned above, the focus on technical novelty and innovation in the robotics community differs from the focus on clinical utility in the clinical research community. This also has implications for the methods that are the standard for each field. With respect to design, we chose to prioritize proof of concept over technological development. For instance, we feel that speech recognition innovations will be required in order to replace Wizard of Oz with automation, but we have decided to justify such an investment first with a demonstration of a highly socially responsive robot, whether automated or manually operated.

Sample Sizes. In the robotics community, a proof-of-concept paper may include 1-6 participants with developmental disabilities. This is sufficient to illustrate the technical advances of the robotics platform, show feasibility, and provide a glimpse at the potential of the advances. However, studies that aim to demonstrate clinical utility involving just a few participants are often regarded by clinicians and developmental researchers as being questionable and insufficiently powered to identify reasonable trends, even if effect sizes are large and results are statistically significant. A recent survey by Diehl et al. (2012) indicated in an extensive review of robotics work in autism that only six studies have involved more than six participants with ASD, and in this context discusses the need for larger and more rigorous studies to better define the role robotics can play in autism research.

In our study, we collected data from nearly 20 participants with ASD and 10 TD controls. This represents the largest group of participants with ASDs in a robotics study to date. We should note that while a large sample size is advantageous for identifying robust positive findings, it is even more valuable in the context of interpreting negative findings. In our study, we found that TD children did not increase in their post-robot interview time as compared to their pre-robot interview time, whereas participants with ASD did. We went so far as to mention that we may be less enthusiastic about the negative result identified in the TD group, given the small sample size. However, we should also point out that this is still far larger than 90% of control groups employed in robotics papers reviewed by Diehl et al. (2012). In this fashion, our perspective on sufficient group sizes has been heavily influenced by the clinical expertise contributing to our work, a perspective that helps us to strive for higher standards in robotics-autism research.

Clear Characterization. One of the most pressing challenges presented to researchers studying ASDs is, as identified by the Interagency Autism Coordinating Committee (2011), the heterogeneity present in the disorder. While the definition of autism spectrum disorders can be neatly summarized by a single reference to the DSM-IV (American Psychiatric Association, 2000), the complexity and heterogeneity of the autism spectrum (e.g., see Happé, Ronald, & Plomin, 2006) is easily overlooked by researchers with limited autism experience. Characteristics of individuals with ASD range from extremely high intelligence and relatively subtle communicative or social difficulties, to no language ability, comorbid and debilitating intellectual disabilities, and almost non-existing social function. Even within a relatively “high-functioning” group of individuals with ASD, behavioral and cognitive characteristics can range widely: verbal communication can be difficult to elicit or flow unceasingly, visual-spatial competency can be average or remarkably superior, adaptive functioning can be well preserved or severely impaired. Understanding the nature of these individual characteristics can often be a nuanced and subtle process, requiring high levels of clinical insight and care to decipher (e.g., see Karmiloff-Smith, 2006). In other words, knowing that the target population has ASD is necessary but not sufficient to understand all of the complexities of an experimental interaction with robots. Ideally, we would know in advance which subpopulations to target, and the expected behaviors of the targeted subpopulations on selected outcome measures. However, given the nascent state of our interdisciplinary field, such knowledge, is often unavailable at the time of experimental design. For this reason, larger-N proofs of concept and exploratory investigations are critical for the understanding of heterogeneity in behavioral responses among individuals with ASD, and thus essential to the advancement of robotics research in autism.

In this study we collaborated with leading experts in speech and language pathology in ASDs, coordinating with a team of expert clinicians and researchers in ASD. Their added insight was extremely valuable, and greatly enhanced the interpretability of our study. For instance, the clear clinical guidelines they provided indicated that the individuals comprising the ASD group indeed were all affected by ASD. We were also able to establish that, despite the several negative findings involving between-group differences of affect, engagement, and pre- and post-robot interview times, these results did not hold up for all individuals with ASD. Instead, we found that the higher-functioning participants with ASD responded more enthusiastically to the study, possibly suggesting that the particular paradigm employed in this study might be most engaging for individuals with PDD-NOS or Asperger syndrome, who typically exhibit less severe autism symptoms than children with autism (Walker et al., 2004).

Rigorous Metrics and Statistical Considerations. Data from HRI studies typically employ a structure that lends itself to standard statistical analyses; participant groups are of equal size, drawn from the same population and tested under equal experimental condition. The standard statistical analyses conducted on these studies (typically, t-tests and ANOVAs) are subject to assumptions based around this standard format. Even within the clinical literature, standardized approaches rely on statistical methods that provide value only when these assumptions hold. As studies at the interface of robotics and clinical research must often depart from these traditional formats, whether due to the heterogeneity and availability of the target population or the experimental and adaptive nature of the technology, analysis and interpretation of even large volumes of data must be done carefully and with respect to these underlying assumptions.

This study, while somewhat elementary in its statistical needs, benefited from a careful examination of assumptions inherent in the selected statistical tests. In addition, the choice to use pre- and post-interview times as surrogates for self-motivated verbal elaborations, in the absence of more refined measures on changes in behavior, was aided by perspectives provided by multiple investigators. As this study matures in its analysis, the benefits and interpretability of the results will be greatly aided by collective emphasis on rigorous statistical modeling and the selection of the most appropriate outcome measures for analysis. Similarly, the lessons learned from this study, in partnership with clinical experts, will help pave the way for the design of future studies aimed at isolating specific properties of robots that are most important to effecting change in children with ASD.

5.2.3 Understanding Perspectives on Publication and Dissemination

At the fundamental level, researchers from the robotics community and clinical researchers have a lot in common. They share the same high levels of inquisitiveness and curiosity, the same desire for rigorous truth, and the same goal of leveraging science to improve our understanding of the world and the lives of others. Yet, despite this, the language and perspectives of robotics researchers and clinical scientists can be very disconnected and a concerted effort to educate our collaborators in both fields must be made regarding publication venues.

First, clinical researchers may not understand the scope and magnitude of a robotics conference paper. To gain that perspective, they sometimes have to be informed that high-impact conferences may have similarly, or even more, competitive submission processes than prestigious journals. Furthermore, it is often not clear to clinical collaborators the great importance that conference publications have for career advancement among junior roboticists, engineers, and computer scientists. For this reason, clinical partners to robotics laboratories may question whether it is worthwhile to devote time and resources toward the development of a well-written conference paper; Patterson, Snyder, and Ullman (1999) provide a succinct discussion of the impact of conference publications on the evaluation of computer scientists.

Second, whereas publishing an abstract in a psychology or other social science conference typically will not hinder publication of a corresponding journal article, in submitting a full-length, archived computer science conference paper that summarizes all clinical results may preclude publication in a peer-reviewed journal. The reason is that many, especially high-profile, journals have extensive requirements for innovation and novelty of work presented; that is, journals tend to actively prohibit the reporting of results which have been detailed in print elsewhere, whether prior to, during, or immediately after submission of the journal manuscript.

There are several options joint robotics-clinical collaborations can choose when deciding where and when to publish. First, they can forego conferences altogether, in favor of waiting to submit results to an appropriate journal. This has the advantage of maximizing the chances that study will be able to be accepted to journals, but runs the risk in fast-paced technology research areas of closing opportunities to be the first group in the field to publish concomitant technological advancements, while waiting for journal publication, which typically take longer than conference papers to submit, review, and publish. In addition, a publication in a journal with a clinical focus may not contribute to evaluations of a robotics researcher, when competing for grants and positions, under evaluation by other computer scientists and engineers; these evaluators may prefer high-quality conference publications in technological fields. A second option is to publish the study in a conference first. This, of course, may raise problems concerning the novelty of work, which will likely impact clinical scientists most. A third collaborative solution, such as the approach that we have taken here, is to publish work-in-progress that can document the sophistication and innovation of the technical aspects of the study, over a preliminary population or analyses in progress; later, a following journal submission can represent results from a larger sample or more extensive analysis, either of which is likely to be considered a significant advance over—and thus a finding distinct from—the initial conference publication. In the case of the study presented in this article, we have chosen to present data most relevant to the robotics community (i.e., findings about gross engagement with the robot and about possible indicators for the most appropriate target population) within this robotics venue, while reserving additional analyses of specific behavioral impact for a later publication in a clinical venue. Note, we do not advocate hiding, or “trimming” of data to achieve this collaborative negotiation; such an approach could present ethical challenges, since scientists are expected to report results as fully as possible. Rather, as we have done, we suggest targeting research questions to robotics and clinical publication venues during experimental planning, and, where necessary and possible, the freezing of analyses while publications are pending.

In all cases, collaborators should establish a clear dialogue early, and should negotiate publication plans in advance, to best avoid conflict and to maximize the mutual benefits of the joint project. Roboticists’ careers, and their relationships with funders and other evaluators, could be injured by surprise decisions, at the conclusion of extensive technological development and data collection, that results cannot be published for as long as a year. Likewise, clinicians who have heavily invested time and resources into a study would face problems with career development and evaluation, if faced with a surprise rejection from a journal due to previous technical conference publication.

5.3 Establishing Common Ground by Minimizing Risk

Collaboration can succeed only if involved parties communicate effectively; this, in turn, requires that each understand the others’ motivations, needs, and resources. A common ground, though perhaps not as noble as one would like, is mutual self-interest: the roboticist very much wants to see his or her platform used; the clinical researcher very much wants to provide new avenues for effecting positive changes in the population that is his or her expertise. It is important to consider that such a pairing poses significant risks to both sides of the collaboration: by pairing with clinicians and developmental experts, the roboticist takes a chance that his or her proof of concept may ultimately advance to a demonstration of non-effectiveness; the clinical expert, by wagering on a new technology, risks spending valuable clinical resources (especially personnel time and access to participants from a small and specialized population) on the exploration of nascent technology, instead of on investigation of better understood, and thus less risky, paradigms.

It is useful to understand the risks each community faces from a financial perspective. Robotics work is design- and development-heavy: much of the costs associated with creating a new robotics platform involve design work, machining, programming, and countless hours of trouble-shooting. Clinical work, especially experimental trials, are delivery-heavy: much of the costs associated with running a successful clinical research enterprise involve careful study design, an extended period of experimental delivery, and rigorous statistical analysis and interpretation. Development time, in the robotics community, is measured in months; experiment delivery time are measured in weeks. In the clinical research community, these time-frames tend to be reversed. This means that in robotics work, time is largely spent in the process of rapid prototyping, deployment, and re-development. On the other hand, clinical partners will spend most of their time conducting the same trial over and over again.

Our collaboration in this study began with considerations to minimize risk to our clinical partners. First, this entailed ensuring that the robot platform was free from glitches, crashes, and other issues that might interfere with the delivery of the experimental protocol. Our debugging and testing phases were far more extensive than would have been usual for a non-collaborative proof-of-concept study. Second, interfaces between the robot and the experimenter controlling the robot were robustly designed; guaranteeing that rapid response to the behaviors of children could be accommodated. Roboticists without extensive clinical experience may overlook the potentially terrific expense required to conduct a rigorous experiment with special populations. Recruitment can be difficult, especially for less prevalent disorders. Access to a specific age-range or a subgroup of individuals with specific characteristics in addition to the disorder itself (e.g. *higher functioning 10- to 12-year-old children with ASD*), which is useful in controlling the experiment from a statistical vantage, can make recruitment even more difficult. Furthermore, clinical characterization requires both tremendous coordination of staff and considerable personnel costs. In other words, even besides study and platform design costs, expenses per participant can be quite high (upwards of several hundred dollars per participant). These costs, in addition to the importance of consistency, make mistakes in this work prohibitively expensive. Third, while roboticists often benefit from demonstrating innovation or proof-of-concept using expensive, one-of-a-kind prototypes, the potential cost for damage to or destruction of these prototypes makes involvement in a clinical environment potentially prohibitively expensive. Our efforts leveraged a commercially available robot platform that could be easily replaced with minimal cost when damaged during a clinical visit. While this risk analysis is particular to our two research groups, a similar analysis of risk can be of great benefit in advancing an initial interdisciplinary conversation to a long-term and viable collaborative effort.

6. Conclusions

In this paper we have discussed barriers that have hindered the ability of the robotics community to elevate their research in developmental disorders (especially autism) to a level that achieves clinical utility. We have discussed these challenges in the context of the novel findings in our recent study of robot interaction with children with ASD, the largest study of interaction between robots and individuals with ASD to date. We illustrate through this study and our collective experiences, that roboticists can overcome these collaborative difficulties through close partnerships and clear lines of dialogue with clinical experts. We have highlighted areas in which roboticists can clearly benefit from clinical expertise, and advantages such partnerships offer toward designing the next generation of experimental robots for therapeutic and evaluative applications in social skills and communication.

7. Acknowledgements

This material is based upon work supported by Microsoft Research and the National Science Foundation under grants No. 0835767, No. 0968538, and No. 1117801, and an NSF Expedition in Computing (award #1139078).. We thank Lauren Berkovits, Dan Leyzberg, Elaine Short, Erica Newland, Eugenia Gisin, Hilary Rose Barr, Yael Fuerst, Holly Pritchett, Elizabeth Schoen Simmons, Rebecca Loomis, and Jonathan A. Kelley, for their capable assistance with robot development, data collection, annotation, and editing. We also thank the editors at the Journal for Human-Robot Interaction, and our three anonymous reviewers, for their insightful feedback.

8. References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. Arlington, VA: Author.

Andrews, S., Warner, J., & Stewart, R. (1986). EMG biofeedback and relaxation in the treatment of hyperfunctional dysphonia. *International Journal of Language & Communication Disorders*, *21*(3), 353–369. doi:10.3109/13682828609019847

Autism and Developmental Disabilities Monitoring Network (ADDM), Centers for Disease Control and Prevention (CDC), U.S. Department of Health and Human Services. (2009). *Prevalence of the autism spectrum disorders (ASDs) in multiple areas of the United States, 2004 and 2006*. Retrieved from http://www.cdc.gov/ncbddd/autism/addm.html

Bosseler, A., & Massaro, D. W. (2003). Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. *Journal of Autism and Developmental Disorders*, *33*(6), 653–672. doi:10.1023/B:JADD.0000006002.82367.4f

Diehl, J. J., Schmitt, L. M., Villano, M., & Crowell, C. R. (2012). The clinical use of robots for individuals with autism spectrum disorders: A critical review. *Research in Autism Spectrum Disorders*, *6*(1), 249–262. doi:10.1016/j.rasd.2011.05.006

Elliott, C. D. (2007). *Differential Ability Scales-II (DAS-II)*. San Antonio, TX: Pearson. Retrieved from http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8338-820

Feil-Seifer, D., & Matarić, M. J. (2011). Automated detection and classification of positive vs. negative robot interactions with children with autism using distance-based features. *Proceedings of the 6th International Conference on Human-Robot Interaction*, HRI ’11 (pp. 323–330). Lausanne, Switzerland: ACM. doi:10.1145/1957656.1957785

Gentil, M., Aucouturier, J.-L., Delong, V., & Sambuis, E. (1994). EMG biofeedback in the treatment of dysarthria. *Folia Phoniatrica et Logopaedica*, *46*(4), 188–192. doi:10.1159/000266312

Hagedorn, J., Hailpern, J., & Karahalios, K. G. (2008). VCode and VData: Illustrating a new framework for supporting the video annotation workflow. *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI ’08 (pp. 317–321). Napoli, Italy: ACM. doi:10.1145/1385569.1385622

Happé, F., Ronald, A., & Plomin, R. (2006). Time to give up on a single explanation for autism. *Nature Neuroscience*, *9*(10), 1218–1220. doi:10.1038/nn1770

Hetzroni, O., & Tannous, J. (2004). Effects of a Computer-Based Intervention Program on the Communicative Functions of Children with Autism. *Journal of Autism and Developmental Disorders*, *34*(2), 95–113. doi:10.1023/B:JADD.0000022602.40506.bf

Interagency Autism Coordinating Committee. (2011). *2011 IACC strategic plan for autism spectrum disorder research*.

Joseph, R. M., & Tager-Flusberg, H. (1997). An investigation of attention and affect in children with autism and Down Syndrome. *Journal of Autism and Developmental Disorders*, *27*(4), 385–396. doi:10.1023/A:1025853321118

Karmiloff-Smith, A. (2006). Atypical epigenesis. *Developmental Science*, *10*(1), 84–88. doi:10.1111/j.1467-7687.2007.00568.x

Kasari, C., Sigman, M., Mundy, P., & Yirmiya, N. (1990). Affective sharing in the context of joint attention interactions of normal, autistic, and mentally retarded children. *Journal of Autism and Developmental Disorders*, *20*(1), 87–100. doi:10.1007/BF02206859

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.

Kim, E. S., Leyzberg, D., Tsui, K. M., & Scassellati, B. (2009). How people talk when teaching a robot. *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, HRI ’09 (pp. 23–30). New York, NY, USA: ACM. doi:10.1145/1514095.1514102

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., et al. (2000). The Autism Diagnostic Observation Schedule—Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. *Journal of Autism and Developmental Disorders*, *30*(3), 205–223. doi:10.1023/A:1005592401947

Lord, C., Rutter, M., & Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, *24*(5), 659–685. doi:10.1007/BF02172145

Martin, F., & Farnum, J. (2002). Animal-assisted therapy for children with pervasive developmental disorders. *Western Journal of Nursing Research*, *24*(6), 657–670. doi:10.1177/019394502320555403

Mullen, E. M. (1995). *Mullen Scales of Early Learning* (AGS.). San Antonio, TX: Pearson.

Parsons, S., & Mitchell, P. (2002). The potential of virtual reality in social skills training for people with autistic spectrum disorders. *Journal of Intellectual Disability Research*, *46*(5), 430–443. doi:10.1046/j.1365-2788.2002.00425.x

Patterson, D., Snyder, L., & Ullman, J. (1999). Best practices memo: Evaluating computer scientists and engineers for promotion and tenure. *Computing Research News*.

Paul, R. (2005). Assessing communication in autism spectrum disorders. In F. R. Volkmar, R. Paul, A. Klin, & D. J. Cohen (Eds.), *Handbook of autism and pervasive developmental disorders* (3rd ed., Vol. 2, pp. 799–816). Hoboken, NJ: John Wiley and Sons.

Paul, R., Augustyn, A., Klin, A., & Volkmar, F. R. (2005). Perception and production of prosody by speakers with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *35*(2), 205–220. doi:10.1007/s10803-004-1999-1

Prelock, P. A., Paul, R., & Allen, E. M. (2011). Evidence-based treatments in communication for children with autism spectrum disorders. In B. Reichow, P. Doehring, D. V. Cicchetti, & F. R. Volkmar (Eds.), *Evidence-based practices and treatments for children with autism* (pp. 93–169). New York, NY: Springer US. Retrieved from http://www.springerlink.com/content/g520308543808731/abstract/

Redefer, L., & Goodman, J. (1989). Brief report: Pet-facilitated therapy with autistic children. *Journal of Autism and Developmental Disorders*, *19*(3), 461–467. doi:10.1007/BF02212943

Reichow, B., Volkmar, F., & Cicchetti, D. V. (2008). Development of the evaluative method for evaluating and determining evidence-based practices in autism. *Journal of Autism and Developmental Disorders*, *38*(7), 1311–1319. doi:10.1007/s10803-007-0517-7

Reichow, B., & Volkmar, F. R. (2011). Evidence-based practices in autism: Where we started. In B. Reichow, P. Doehring, D. V. Cicchetti, & F. R. Volkmar (Eds.), *Evidence-based practices and treatments for children with autism* (pp. 3–24). New York, NY: Springer US. Retrieved from http://www.springerlink.com/content/w34707658g646lg5/abstract/

Ricks, D. J., & Colton, M. B. (2010). Trends and considerations in robot-assisted autism therapy. *2010 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 4354–4359). Anchorage, AK: IEEE. doi:10.1109/ROBOT.2010.5509327

Riek, L. D. (2012). Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, *1*(1), 119-136.

Rutter, M., Bailey, A., & Lord, C. (2003). *Social Communication Questionnaire: SCQ (W-381)*. Los Angeles, CA: Western Psychological Services. Retrieved from http://portal.wpspublish.com/portal/page?_pageid=53,70432&_dad=portal&_schema=PORTAL

Scassellati, B. (2005). Quantitative metrics of social response for autism diagnosis. *IEEE International Workshop on Robot and Human Interactive Communication, ROMAN 2005* (pp. 585 – 590). doi:10.1109/ROMAN.2005.1513843

Scassellati, B., Admoni, H., & Matarić, M. J. (2012). Robots for use in autism research. *Annual Review of Biomedical Engineering*, *14*.

Schopler, E., Reichler, R. J., & Renner, B. R. (1986). *The Childhood Autism Rating Scale (CARS): For diagnostic screening and classification of autism*. New York, NY: Irvington.

Silver, M., & Oakes, P. (2001). Evaluation of a new computer intervention to teach people with autism or asperger syndrome to recognize and predict emotions in others. *Autism*, *5*(3), 299 –316. doi:10.1177/1362361301005003007

Steinfeld, A., Jenkins, O. C., & Scassellati, B. (2009). The oz of wizard: simulating the human for interaction research. *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, HRI '09 (pp. 101–108). San Diego, CA: ACM. doi:10.1145/1514095.1514115

Strickland, D. (1997). Virtual reality for the treatment of autism. In G. Riva (Ed.), *Virtual reality in neuro-psycho-physiology: cognitive, clinical and methodological issues in assessment and rehabilitation* (pp. 81–86). Amsterdam, Netherlands: IOS Press.

Tapus, A., Matarić, M. J., & Scassellati, B. (2007). Socially assistive robotics [Grand challenges of robotics]. *Robotics Automation Magazine, IEEE*, *14*(1), 35 –42. doi:10.1109/MRA.2007.339605

Volkmar, F. R., & Klin, A. (2005). Issues in the classification of autism and related conditions. In F. R. Volkmar, R. Paul, A. Klin, & D. J. Cohen (Eds.), *Handbook of autism and pervasive developmental disorders* (3rd ed., Vols. 1-2, Vol. 1, pp. 335–364). Hoboken, NJ: John Wiley and Sons.

Volkmar, F. R., Lord, C., Bailey, A., Schultz, R. T., & Klin, A. (2004). Autism and pervasive developmental disorders. *Journal of Child Psychology and Psychiatry*, *45*(1), 135–170. doi:10.1046/j.0021-9630.2003.00317.x

Walker, D. R., Thompson, A., Zwaigenbaum, L., Goldberg, J., Bryson, S. E., Mahoney, W. J., Strawbridge, C., et al. (2004). Specifying PDD-NOS: A Comparison of PDD-NOS, Asperger Syndrome, and Autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, *43*(2), 172–180. doi:10.1097/00004583-200402000-00012

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence (WASI)*. San Antonio, TX: Pearson Assessment. Retrieved from http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8981-502

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children (WISC-IV)* (4th ed.). San Antonio, TX: Pearson Assessment. Retrieved from http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8979-044

Wiederholt, J. L., & Bryant, B. R. (2001). *Gray Oral Reading Test (GORT-4)* (4th ed.). San Antonio, TX: Pearson Assessment. Retrieved from http://psychcorp.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8116-577&Mode=summary

Yirmiya, N., Kasari, C., Sigman, M., & Mundy, P. (1989). Facial expressions of affect in autistic, mentally retarded and normal children. *Journal of Child Psychology and Psychiatry*, *30*(5), 725–735. doi:10.1111/j.1469-7610.1989.tb00785.x

Authors' names and contact information: E. S. Kim, Department of Computer Science, Yale University, New Haven, CT, USA. Email: elizabeth.kim@yale.edu; R. Paul and F. Shic, Child Study Center, Yale University; B. Scassellati, Department of Computer Science, Yale University.