Journal of Autism and Developmental Disorders (2020) 50:199–211
https://doi.org/10.1007/s10803-019-04246-z

ORIGINAL PAPER

Assessing Social Communication and Collaboration in Autism Spectrum Disorder Using Intelligent Collaborative Virtual Environments 利用智能协作虚拟环境评估自闭症障碍患者的社会沟通与协作

Lian Zhang[1] · Amy S. Weitlauf[2,3] · Ashwaq Zaini Amat[1] · Amy Swanson[3] · Zachary E. Warren[2,3,4,5] · Nilanjan Sarkar[1,6]

Published online: 3 October 2019

Abstract

我们开发并试验了一种新的智能协作虚拟环境代理（CRETA），用于评估系统和对等交互中的社会通信和协作

Existing literature regarding social communication outcomes of interventions in autism spectrum disorder (ASD) depends upon human raters, with limited generalizability to real world settings. Technological innovation, particularly virtual reality (VR) and collaborative virtual environments (CVE), could offer a replicable, low cost measurement platform when endowed with intelligent agent technology and peer-based interactions. We developed and piloted a novel collaborative virtual environment and intelligent agent (CRETA) for the assessment of social communication and collaboration within system and peer interactions. The system classified user statements with moderate to high accuracies. We found moderate to high agreement in displayed communication and collaboration skills between human–human and human–agent interactions. CRETA offers a promising avenue for future development of autonomous measurement systems for ASD research.

系统对用户语句进行中高精度分类

我们发现，人与人和人与机器交互之间在显示的沟通和协作技能方面有中到高的一致性

CRETA为ASD自主测量系统的未来发展提供了一条很有前途的途径。

Keywords Autism · Technology · Virtual reality · Measurement · Collaboration · Communication

Although a growing body of literature indicates that behavioral and pharmacological interventions have some benefits for many children with ASD (McPheeters et al. 2011; Veenstra-VanderWeele and Warren 2015; Weitlauf et al. 2014), the generalizability and impact of this work has been hampered by a lack of pragmatic outcome measurement strategies for indexing meaningful change in social communication skills. Ideally, social communication measurement strategies could be implemented in a consistent fashion, would be sensitive to reductions in the severity of specific symptoms, and would reflect reductions in functional impairments across settings. Technological approaches offer promise regarding consistency of implementation but, to date, most measures have been confined by methodologies that severely restrict use in broader settings or studies pushing for quick (pass/fail) trial outcomes (Bekele et al. 2013, 2014; Goodwin 2008; Josman et al. 2008; Kandalaft et al. 2013). In this context, developing advanced technological systems capable of both eliciting, autonomously detecting, and quantifying key social communication processing skills in environments highly relevant to deploying these complex skills would represent a powerful advancement of intervention science.

Recent published reviews (Anagnostou et al. 2015; Scahill et al. 2015) evaluating common measurement approaches for indexing change in core ASD areas of impairment have consistently documented several critical challenges that have limited progress in terms of intervention trials. Specifically, available measures are often: (1) reliant on caregiver or other reporting strategies and subject to aspects of reporting bias and imprecision in measurement of core behaviors; (2) confined in coverage to limited aspect of core deficits of the disorder; (3) limited in application to specific ranges of the ASD spectrum; (4) insensitive to change given broad

✉ Amy S. Weitlauf
amy.s.weitlauf@vumc.org

1 Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA

2 Department of Pediatrics, Vanderbilt University Medical Center, Nashville, USA

3 Vanderbilt Kennedy Center, Treatment and Research Institute of Autism Spectrum Disorders, Vanderbilt University Medical Center, 230 Appleton Pl., PMB 74, Nashville, TN 37212, USA

4 Department of Special Education, Vanderbilt University, Nashville, TN, USA

5 Department of Psychiatry, Vanderbilt University Medical Center, Nashville, TN, USA

6 Department of Mechanical Engineering, Vanderbilt University, Nashville, TN, USA

coverage of functional symptom and diagnostic domains; (5) infeasible given extensive specialized training and/or coding procedures for use; and/or (6) laboratory based and limited in ecological validity. Further, a major challenge of appropriately indexing social communication impairments relates to the very nature of this core symptom domain itself. As opposed to restricted and repetitive behaviors, which represent behavioral excesses and more readily observable atypicalities, social communication skills represent areas of impairment related to deficit/absence (i.e., appropriate skill is not deployed in situations when it is expected). As such, paradigms for measuring symptom presence and impact must be capable of eliciting these skills in order to appropriately measure their absence and associated impairment.

Given recent rapid developments in technology, it has been argued that specific computer and virtual reality (VR) based applications could be harnessed to provide effective and innovative clinical tools for meaningfully measuring, and perhaps intervening, regarding social communicative impairments in individuals with ASD. VR technology possesses several strengths in terms of potential application to ASD measure development in that it is capable of realizing platforms that mimic real-world social communication tasks while simultaneously providing quantitative, objective, and reliable measures of performance and processing far beyond any reporter or clinical observational strategy. VR can also depict various scenarios that may not be feasible in a "real world" therapeutic setting, given naturalistic social constraints and resource challenges. As such, VR appears well-suited for creating interactive platforms of meaning for assessing specific aspects of social communication skills. However, many studies have developed and utilized technologies that require rate limiting confederate participation or control (i.e., human's controlling operation of VR or technological systems). Further, most studies of the existing autonomous VR and technological environments applied to ASD have been primarily designed (1) to index learning via aspects of performance alone (i.e., correct or incorrect) within simplistic interaction paradigms or (2) visual processing during social tasks (Lahiri et al. 2013; Mitchell et al. 2007; Parsons and Mitchell 2002). At present, it is difficult to determine to what degree these approaches reflect measurable aspects of real-world interactions. Technological and VR systems that not only gauge performance in specified tasks within the virtual environment, but can also automatically detect, respond to and adjust task characteristics based on communicative and coordinated aspects of interaction, may represent more powerful metrics of social communication development and change.

A promising avenue for designing technological VR systems capable of indexing social communication skills is the collaborative virtual environment (CVE). A collaborative virtual environment (CVE), which is a computer-based, distributed, virtual space for multiplayers to interact with one another and/or with virtual items (Benford et al. 2001). CVE technology offers a flexible alternative to conventional modalities of both in vivo (e.g., social skill groups, peer-mediated programs) and technological intervention (e.g., confederate controlled virtual reality, computerized skill programs) where multiple individuals can share and interact in a virtual space using network-based communication. CVEs preserve the advantages of traditional computer-based intervention systems but also facilitate real-time interactions between real users across distance. In particular, the characteristics of this environment are highly controllable and can be adapted and structured in ways that mimic aspects of real-world interactions. These characteristics can tangibly impact the very nature of the collaborative interaction itself.

Although CVEs provide a promising platform for interactions between real users, CVE-based interventions lack reliable and easy-to-use methods for measuring (1) social communication within these systems and (2) impacts of these systems on children with ASD. The majority of CVEs in this area evaluated system impacts based on self-report questionnaires or users' task-performance. For example, Wallace et al. designed a CVE-system to teach greeting behaviors to children with ASD in a virtual gallery (Wallace et al. 2015). They evaluated the system impacts using a self-report questionnaire, and found that children with ASD, compared to their Typically Developing (TD) peers, were less sensitive to a negative greeting. Millen et al. applied CVEs to promote collaboration among children with ASD, and the results of a self-report questionnaire showed improved engagement of children with ASD in the CVEs (Millen et al. 2011). Cheng et al. designed a CVE-based virtual restaurant to understand empathy of children with ASD (Cheng et al. 2010). They found that these children could appropriately answer more empathic questions after the intervention. Although these methods could gather essential information for system evaluation, they could not be used to understand and analyze users' conversation, which is an essential component during user-to-user interactions in most CVE-based interventions.

In some instances, domain experts have been involved to observe and code not only task-performance but also verbal communication of users within CVEs using a human coding methodology. iSocial is a 3D-CVE aimed at understanding and improving social competency development of children with ASD (Schmidt et al. 2011). In iSocial, children's social behaviors, such as gesture, initiation of conversation, response to others' conversation, and turn-taking in conversation, were manually coded by domain experts for system evaluation using a video coding method (Schmidt et al. 2012). However, manually coding users' behaviors, especially verbal communication, needs significant time and efforts. In addition, the CVE-based intervention systems,

一个有前途的设计技术VR系统的途径，可以索引社会交际技能是协作虚拟环境（CVE）

which utilized this time-consuming method for system evaluation, could not provide real-time feedback to the users.

Although they offer promise for promotion of social interaction and communication skills, two fundamental challenges limit the use of CVE systems as novel measures of social communication. First, the dynamic social interactions with CVE systems are partner dependent in open-ended form. That is, interactions within the CVE change based on specific peer input and as such fundamentally limit consistent, replicable measurement of the interactions themselves. Given that it is not possible to deploy one invariant partner for multiple sessions for multiple subjects, the change in partners across interactions impacts the social communication transaction and any associated metrics. Second, while open-ended CVE systems pose no restriction in verbal communication between partners, subsequent manual coding of both sides of interactions is necessary to understand patterns of communication, creating a resource burden fundamentally limiting realistic scale-up of the paradigm.

One way to address these challenges in measuring social communication within CVEs is to use an intelligent agent that can automatically gauge user performance within the system itself. Intelligent agent technology has been explored as a measure task performance and conversation behaviors of TD individuals in collaborative learning environments (Kumar et al. 2007; Nabeth et al. 2005; Scheuer et al. 2010; Walker et al. 2014). Note that although designing an intelligent agent that cannot be distinguished from a human for unrestricted naturalistic conversation is a challenge yet to be solved (i.e., the Turing test), designing paradigms for controlling, indexing, and altering aspects of interactions within a specific domain may represent an extremely valuable and much more viable methodology (Cauell et al. 2000; Kopp et al. 2005). Researchers in the collaborative learning area have developed intelligent agents to, first, measure important aspects of the collaborative learning interactions, such as topic change (Van Rosmalen et al. 2005), learner understanding (Linton et al. 2003), quality of arguments (Scheuer et al. 2010), and learner motivation (Desmarais and Baker 2012) and then provide feedbacks to help these users based on the measurements. Although these systems were not designed for ASD intervention, they provided useful information about applying intelligent agent technology to measure the behaviors of the children with ASD in CVEs.

Motivated by this body of work, we designed an intelligent agent that can play collaborative games with children with ASD within a CVE while providing verbal prompts and responses. Simultaneously, the agent can generate meaningful measurement outcomes (henceforth termed "features") to gauge users' communication and collaboration skills, such as the types of statements that they made, their ability to move pieces in synchrony with the system, and so on. Collaborative games were selected as interactive tasks in the CVE because they have the potential to prompt users to work and communicate with one another (Benford et al. 2001; Leman 2015). These games also offer a controllable environment that can be embedded with carefully designed strategies which could facilitate and track collaborative interaction between users (Curtis and Lawson 2001; Zancanaro et al. 2007). Previous work supports the utilization of collaborative game protocols; Battocchi and colleagues designed collaborative puzzle games with an enforced collaboration rule, which required two users to take actions simultaneously to encourage them to work together (Battocchi et al. 2009). They evaluated the effect of these games on users' collaborations by measuring their task-performance, such as task completion time and number of moved puzzle pieces. They found that games equipped with the enforced collaboration rule have more positive effects on children with ASD, compared to these games without these types of rules.

In an attempt to combine the benefits of peer-based collaborative games and intelligent agent interaction that does not require heavy burden of human coding, we developed the current CVE system with an intelligent agent called Collaborative viRtual Environment and inTelligent Agent (CRETA). This work had two primary aims. First, in light of existing literature and the need for consistent, replicable, low-resource measurement strategies, we designed an intelligent agent embedded within a CVE that could measure communication as well as collaboration skills of children with ASD (human–agent interactions; HAIs). Second, we then conducted a feasibility study using 20 pairs of ASD and TD children to evaluate whether those measurements reflected important aspects of user behaviors (human–human interactions; HHIs) within the same virtual environment.

In what follows, we first present the CRETA system in which an intelligent agent monitors and indexes the communication and collaboration skills of children with ASD in a CVE. We then describe using the intelligent agent to interact with a person with ASD within the CVE and a feasibility study of its use as a measurement tool. Finally, we present a data analysis framework explaining more thoroughly how the system measured these skills and how we evaluated the system measurements.

Methods

Participants

We conducted a feasibility study with 20 age- and sex-matched pairs (40 participants), in order to evaluate whether the intelligent agent could measure both communication and collaboration skills of children in the CVE. The study was approved by the university Institutional Review Board

(IRB) with developmentally appropriate consent and assent obtained from participants and their primary caregivers.

Participants with ASD were recruited from an existing clinical research registry (see Table 1). They had existing ASD diagnoses from licensed clinical psychologists based upon DSM-5 criteria as well as Autism Diagnostic Observation Schedule scores (Lord et al. 2000). Where available, ADOS Calibrated Severity Scores as well as IQ scores from the original diagnostic visits of participants (mean of 7.29 years prior to current study participation participation) are provided in Table 1. Additional inclusion criteria included the use of spontaneous phrase speech and average intelligence. Participants in the TD group were recruited through an electronic recruitment portal accessible to community families. To index initial autism symptoms and screen for autism risk among the TD participants, parents of all participants completed the Social Responsiveness Scale, Second Edition (SRS-2) (Constantino and Gruber 2002) and the Social Communication Questionnaire Lifetime (SCQ) (Rutter et al. 2003).

System Development

Overarching System Design

We designed a CVE system (CRETA) wherein children with ASD played a highly controlled series of games with both an AI agent as well as a TD partner. These games were created explicitly to require effective social communication and collaboration for successful completion. During these games, the system utilized a real-time measurement strategy wherein specific features hypothesized as relevant to effective interactions were automatically measured by the system. We also used expert human coders to index the quality of interactions relative to the system measurements. This allowed us to analyze the ability of the AI interactions and automatic measurement system to predict the quality of social communication as rated by a human, both with the agent itself as well as dynamic human peers. We hypothesized that this data would allow us to develop powerful predictive models of social communication quality using a within-system quantification and controlled analogue interactions.

As described in detail below, the measurement system (CRETA) was composed of CVE and an intelligent agent. The CVE component was designed for two users to converse and play collaborative games with each other. The collaborative games were equipped with strategies to elicit both communication and collaboration between the two users. The intelligent agent component was designed to interact with users as well as generate meaningful features to measure their communication and collaboration skills.

CRETA offers two interactions modes: (i) human–human interactions (HHI) where it allows for humans to interact with one another, and (ii) human–agent interactions (HAI) where participants interact with the agent. During HHIs, participants interact with a peer while the agent monitors and assesses their collaboration and communication skills. During HAIs, the intelligent agent acts as a partner to perform consistent, controlled, and replicable interactions with the participant, simultaneously monitoring and assessing the participant's communication and collaboration skills within the HAI itself. In this way, CRETA allows us to compare how participants interact with other humans within dynamic, variable contexts, to how they interact with the highly controlled system, providing information about behaviors across contexts as well as validation data regarding the intelligent agent as a measurement tool.

System Components

A critical component of CRETA is the CVE, which allows two participants in different locations to interact (within this study, to complete puzzle games) in a shared virtual environment. Within the CVE, although users cannot see each other's faces, they can interact in two ways: conversing via audio chat functionality and playing collaborative puzzle games. The CVE can also record aspects of participant game performance, such as successful game completion and amount of time spent working together to move puzzle pieces. More information about CVE design and implementation can be found in [reference redacted for blinded review].

To elicit communicative and collaborative behaviors, we designed nine trials involving puzzle games in which we varied characteristics of turn-taking, color view, and target

Table 1 Participant characteristics

	Age (years)	Female/male	SRS-2 T-score mean (SD)	SCQ Total mean (SD)	ADOS calibrated severity score mean (SD)	IQ score mean (SD)
ASD (N = 20)	13.39 (2.07)	4/16	78.33 (9.50)	22.65 (8.63)	8.28 (1.41)	85 (22.26)
TD (N = 20)	13.50 (2.30)	4/16	47.33 (8.36)	2.86 (4.03)	n/a	n/a

ASD Autism Spectrum Disorder, *TD* typically developing, *SD* standard deviation, *SRS-2* social responsiveness scale-second edition, *SCQ* social communication questionnaire (current), *ADOS* autism diagnostic observation schedule, *IQ* Intelligence Quotient

movement to ensure that games were both challenging as well as capable of eliciting meaningful measurement targets. For example, in some trials, only one user could see the color of the pieces; in another, users were required to move pieces simultaneously for successful placement. These variations required participants to talk with each other and work together for successful game completion.

Intelligent Agent

The intelligent agent is a computer program developed using machine learning and natural language processing technologies. This agent serves two functions that allow it to act as a measurement tool. First, it can converse and play games with a participant in a controllable, consistent way. Second, it can generate meaningful "features" related to core participant communication and collaboration skills. These features are then utilized to assess target participant skills.

Dialogue Act Classification

To "understand" what participants are saying, the intelligent agent first transcribes participants' speech to text in real time using speech recognition software (Google Cloud Speech APIs; https://cloud.google.com/speech/). Then, it "understands" the human language using a dialogue act classification. "Dialogue acts" are defined as the specialized performative function that an utterance plays in a language (Stolcke et al. 2000). Dialogue act features, such as requests for information (McManus and Aiken 1995), providing information (Gogoulou et al. 2008), and acknowledging other people's actions (Vieira et al. 2004), have been shown to be useful in understanding group discussion behaviors of children with ASD and TD children. As seen in Table 2, we defined dialogue acts in our system based on previous work, in which we recorded conversations between ASD and TD users and discerned five different classes of collaboration-related communication: *request color*, *provide*, *direct movement*, *acknowledge*, and *request object* (reference redacted for blinded review).

The dialogue act class of each sentence was computed using a machine learning method, which classified each sentence into one of the predefined dialogue acts. After "understanding" the human language, the intelligent agent generated responses based on some rules. For example, if a user asked, "What color is this puzzle piece?", the intelligent agent understands this sentence as a *request color* dialogue act. Then, it responds to the user and says, "It is red".

Feature Creation

We then imposed a classification methodology on to aspects of verbal-communication and task-performance which were automatically indexed by the system. Verbal- and task-related performance metrics were grouped into "features," or clearly defined categories reflecting aspects of spoken language and within-game actions. This feature classification allowed for examination of patterns of usage as well as change over time and across gaming partner (i.e., human vs. agent).

Verbal-communication feature categories were derived from existing literature into collaborative communication for individuals with ASD as well as expert consensus of involved research staff who reviewed written transcripts in pilot work (Bauminger-Zviely et al. 2013). As seen in Table 2, verbal-communication features relevant to target outcomes were analyzed based on frequency per minute and included *word frequency* (total words uttered), *request- color*

Table 2 System-generated features and their descriptions

Feature	Description	Example
Verbal communication (per minute)		
Word frequency	How many words a user speaks	–
Request color frequency	How many times a user asks for color information	What color is this piece?
Provide frequency	How many times a user provides game information	That piece is red.
Direct movement frequency	How many times a user directs movements	Move piece number three.
Acknowledge frequency	How many sentences confirm receipt of information	Got it!
Request object frequency	How many times a user asks for objects	Which one do you want to move?
Sentence frequency	How many sentences a user speaks	–
Task performance (per minute)		
Success frequency	Number of puzzle pieces successfully placed (per game)	–
Failure frequency	How many times a user fails to successfully place a puzzle piece	–
Collaboration time	How much time two users spend simultaneously moving puzzle pieces	–
Dragging time	The total time a user spends dragging puzzle pieces	–
Collaborative movement ratio	The ratio of collaboration time to dragging time	–

(number of times user ask for puzzle color), *provide* (number of times user provide information related to game), *direct movement* (number of times user give direction to move puzzle), *acknowledgement* (number of times user acknowledges receipt of information), *-request object* (number of times user requests an object), and *sentence frequency* (total number of sentences) (see Table 2). The system also included 5 task-performance features. These include how many puzzle pieces were successfully moved (*success frequency*), how many times a participant failed to accurately place puzzle pieces (*failure frequency*), how long a participant dragged puzzle pieces (*dragging time*), and how often two users simultaneously moved puzzle pieces together (*collaboration time*).

As it monitored dialogue and game actions, the intelligent agent combined what the participants said as well as game play information to make sense of participant behavior. Finally, based on this understanding, the intelligent agent generated verbal responses or prompts as well as game-related responses using a finite state machine (FSM) (Clarke et al. 1986). A FSM is a computational framework which displays the process by which a system transitions between states (e.g., moving puzzles pieces versus giving directions) based upon external input (e.g., what a participant says or does). Only one state can be active at a time. In other words, using a FSM allowed the agent to ask the participant questions, respond to what was said, and move puzzle pieces according to the participant's spoken instructions, in ways similar to how another human participant would respond.

In summary, CRETA consisted of a virtual environment in which two users, seated in different locations, could play games either with each other or with the system's intelligent agent. These puzzle games were designed to withhold information or require simultaneous movement in order to promote collaboration and communication between players. The intelligent agent monitored not only whether and how quickly puzzles were successfully completed, but also what types of dialogue acts users offered and how frequently (e.g., how often a child with ASD asked about the color of a puzzle piece). It did this utilizing machine learning methods based upon types of dialogue derived from our previous work. These features were monitored both for human–human interactions between two users as well as for human–agent interactions, in which participants played games with the system itself.

Procedure

Experimental Procedure

Each ASD/TD pair completed a one-visit experiment. First, participants were provided with text and audio instructions about how to play games in the CVE. Next, the participants played nine collaborative puzzle games in a random order. Each game was played in a HHI mode followed by a HAI mode. In the HHI mode, a participant with ASD played a game with their TD partner child for 1–3 min, as determined by preliminary work. In a HAI mode, each child played the same game with the intelligent agent. Between each game, there was a 10-s break. Each experiment lasted approximately 40 min. Participants and their computer screens were audio and videotaped during the experiments.

Measurement Procedure

We evaluated the potential of the system to measure both communication and collaboration skills of children with ASD. Specifically, the system automatically generated features representing verbal-communication and task-performance outcomes to represent participant behaviors in HAIs as well as HHIs. We evaluated not only the ability of the system to generating these features, but also whether these system-generated features could measure communication and collaboration skills relative to human coders, who analyzed transcripts offline according to a standardized scoring rubric and provided the "ground truth" for target skills.

Ground Truth of the Communication and Collaboration Skills

Obtaining human ratings of target features was important to establishing CRETA as a potential replacement for traditional time-consuming human-coding methodologies. Two human raters, trained research assistants familiar with ASD and technological intervention, watched videos of the experiments, and rated both communication and collaboration skills of the participants in order to provide the ground truth of these skills. These two human raters utilized a continuous rating scale (−4 to 4, with higher scores reflecting a higher level of collaboration or communication). They made these judgments based on independent observation and clinical impression of participants after each game was completed. The primary human rater, who was blinded to the study, rated all the experiments, and her ratings were used as the ground truth of the skills. The secondary human rater, who was not blinded, rated 25% of the all the experiments for the purpose of establishing reliability. Inter-rater reliability (Spearman's rho) using this continuous scale was moderate for communication ($r_s = 0.42$, $p < .001$) and collaboration ($r_s = 0.54$, $p < .001$).

Results

We evaluated the preliminary functioning of our system using three steps. First, we compared human transcriptions and ratings to system generated classifications in order to

determine the *feature accuracy*, or system classification error rates. We dropped features which were not accurately assessed from further analyses. Next, we computed *correlations between human ratings and system ratings of communication and collaboration skills*. This allowed us to evaluate whether the system was capable of automating a human coder's perceptions of target skill areas. Finally, *we compared participant performance in HAIs to performance in HHIs*. This allowed us to evaluate whether the participant's interactions with the agent accurately reflected how they interacted with another human user, whose behaviors were less predictable.

Feature Accuracy

We analyzed the ability of CRETA to accurately generate verbal-communication features. We did this by evaluating the performance of the speech recognition software using its word error rate (Klakow and Peters 2002). Then, we evaluated the performance of the dialogue act classification model by showing its confusion matrix, which included true positives, true negatives, false positives, and false negatives of each dialogue act class (Srinivasan and Petkovic 2000). Finally, we computed accuracies and error rates of system-generated verbal-communication features relative to human coders, for all participants in both HAIs and HHIs conditions, and retained those with acceptable accuracies.

The five-class dialogue act classification revealed accuracies much higher than random (20%, given the five possible classifications) across both conditions (HAIs: 70.27% out of 1337 sentences; HHIs: 68.78% out of 868 sentences). The overall error rate of the speech recognition was 18.01% in HAIs and 23.16% in HHIs. We then assessed the accuracy of feature recognition by dividing the number of system-classified incidents by the number of human-coded incidents. That is, we calculated how many times an offline human coder classified a dialogue act in one way, and the system classified the same act in a different way. In HAIs, relatively high accuracies were found for *provide* (83%) and *acknowledge* (82%). Accuracy was more modest for *direct movement* (48%), which was confused approximately 50% of the time with *provide* in HAIs. In contrast, in HHIs, *direct movement* had the highest accuracy (90%), with *provide* (61%), and *acknowledge* (66%) yielding more modest results. Provide was confused with *direct movement* (28% of instances) and *acknowledge* (10% of instances), whereas *acknowledge* was confused with *provide* (22% of instances) and *direct movement* (22% of instances). *Request color* and *request object* frequencies were very low (< 3% of total dialogue acts across both conditions) and, as such, they were dropped from further analyses.

Correlational Analyses

Description of Approach

We first computed correlations between the system-generated features and *continuous human skill ratings* using Spearman's rank correlation, a non-parametric measure of rank correlation between two variables (Krishnaiah 1980), because these features did not follow a normal distribution. A machine learning model was built to measure communication skills using the system-generated features and ratings of communication skills; while another machine learning model was built to measure collaboration skills using these features and rating of the collaboration skills. In addition, we trained two models to classify these skills based on balanced training data. The balanced training data were generated by randomly under-sampling the majority class, which is a commonly used resampling techniques to improve classification performance in unbalanced datasets. The performance of these models in measuring these skills was evaluated using their classification accuracies, which were computed using a sixfold cross-validation method.

Continuous Correlation of Human–Agent Interaction

As seen in Table 3, we correlated human ratings of communication and collaboration with participant performance in HAI. Across both ASD and TD groups, we found a strong negative correlation between the system generated *failure frequency* feature and human ratings of collaboration (ASD: $r_s = -0.60$, $p < 0.001$, TD: $r_s = -0.23$, $p < 0.001$) and human ratings of communication (ASD: $r_s = -.49$, $p < .001$, TD: $r_s = -0.24$, $p < .05$). That is, the lower the continuous human ratings, the more frequently the system registered *failure* during an HAI session. More collaborative participants were also more likely to provide information (ASD: $r_s = 0.35$, $p < 0.001$, TD: $r_s = 0.31$, $p < 0.05$).

Differences emerged across groups, as well, with associated means and t-tests presented in Table 4. TD participants were more likely to complete a task successfully, as shown by significantly higher rates of *success frequency* (ASD: $m = 82.32$, $sd = 34.47$; TD: $m = 110.77$, $sd = 39.20$; $t(21) = 4.56$, $p < .001$). Some human ratings of communication and collaboration skills were significantly correlated with features for participants with ASD, but not for TD participants. In general, human coders rated TD participants as more communicative ($t(21) = 4.13$, $p < .001$) and collaborative ($t(21) = 5.21$, $p < .001$) than participants with ASD. Additionally, participants with ASD who were rated as more communicative and collaborative were more likely to have success during tasks, whereas human ratings of TD participants did not seem to relate to task success (Communication: ASD: $r_s = 0.38$, $p < .001$, TD: $r_s = 0.03$,

Table 3 Spearman's Rank correlations between features and continuous human ratings in Human–Agent Interactions

System-generated feature	Human ratings for ASD		Human ratings for TD	
	Communication	Collaboration	Communication	Collaboration
Word frequency	0.33**	0.12	0.59**	0.26*
Provide frequency	0.45**	0.35**	0.34*	0.31*
Direct movement frequency	0.22	0.03	0.24	−0.03
Acknowledge frequency	−0.00	0.09	−0.20	−0.07
Sentence frequency	0.47**	0.38**	0.28**	0.18
Success frequency	0.38**	0.54**	0.03	0.18
Failure frequency	−0.49**	−0.60**	−0.24*	−0.23**
Collaboration time	0.30	0.43**	−0.04	0.06
Dragging time	−0.37**	−0.33**	0.02	0.08
Collaborative movement ratio	0.10	0.17	−0.17	-0.03

ASD autism spectrum disorder, *TD* typical development

**p < .001; *p < .05

Table 4 Means, standard deviations, and *t* tests for differences across groups for features and continuous human ratings in Human–Agent Interaction

	Human agent interactions		
	ASD Mean (sd)	TD Mean (sd)	*t* test *t*(21)
Word frequency	337.82 (174.02)	324.82 (168.43)	0.35
Provide frequency	43.68 (32.47)	59.05 (25.67)	2.44
Direct movement frequency	29.27 (19.15)	22.59 (16.11)	1.42
Acknowledge frequency	33.18 (31.80)	42.73 (25.08)	1.23
Sentence frequency	140.73 (48.00)	161.09 (50.78)	3.37***
Success frequency	82.32 (38.47)	110.77 (39.20)	4.56***
Failure frequency	1.77 (1.69)	0.42 (0.80)	3.95***
Collaboration time	40.08 (20.84)	47.59 (15.98)	2.15
Dragging time	575.30 (218.32)	436.53 (127.87)	2.89**
Collaborative movement ratio	0.08 (0.04)	0.11 (0.03)	4.17***
Communication skills ratings	6.57 (9.23)	15.41 (8.77)	4.13***
Collaboration skills ratings	6.64 (9.43)	16.09 (9.24)	5.21***

ASD Autism Spectrum Disorder, *TD* Typical Development

***$p < .001$; **$p < .01$

p = ns, Collaboration: ASD: $r_s = 0.54$, p < .001, TD: $r_s = 0.18$, p = ns). Participants with ASD also spent more time working with the intelligent agent to move pieces ($r_s = 0.43$, p < 0.001), including significantly more time spent dragging pieces (indicating less efficient collaborative movement; $r_s = -0.33$, p < 0.001). As seen in Table 4, participants with ASD spent significantly more time dragging pieces and had more variability in time spent than TD participants (ASD: $m = 575.30$, $sd = 218.32$; TD: $m = 436.63$, $sd = 127.87$; $t(21) = 2.89$, p < .01).

Correlations Between Features of HHIs and Features of HAIs

In order to evaluate whether participant behaviors in HAIs could reflect important aspects of their behaviors in HHIs, we used Spearman's Rank correlation to compare the system-generated features of HAIs to the system-generated features of HHIs. As shown in Table 5, there were strong, positive correlations between *word frequency* (ASD: $r_s = 0.67$, p < 0.001 TD: $r_s = 0.56$, p < .001), *provide frequency* (ASD: $r_s = 0.54$, p < 0.001 TD: $r_s = 0.31$, p < 0.05), *sentence frequency* (ASD: $r_s = 0.69$, p < 0.001 TD: $r_s = 0.51$, p < 0.01), *collaboration time* (ASD: $r_s = 0.59$, p < 0.001 TD: $r_s = 0.61$, p < 0.01) and *dragging time* (ASD: $r_s = 0.31$, p < 0.05 TD: $r_s = 0.61$, p < .001) for both participants with ASD and TD in HHIs and HAIs. The correlation for *direct movement* was significant for ASD ($r_s = 0.42$, p < 0.001) but not for TD ($r_s = 0.26$, p = ns). The correlations for *acknowledge frequency, success frequency, failure frequency*, and *collaborative movement ratio* were not significant.

Table 5 Spearman's Rank Correlations between features in Human–Human and Human–Agent Interactions

System generated features	**ASD**	**TD**
Word frequency	0.67**	0.56**
Provide frequency	0.54**	0.31*
Direct movement frequency	0.42**	0.26
Acknowledge frequency	0.13	0.20
Sentence frequency	0.69**	0.51**
Success frequency	0.07	0.03
Failure frequency	0.10	0.24
Collaboration time	0.59**	0.61**
Dragging time	0.31*	0.61**
Collaborative movement ratio	0.06	0.20

ASD Autism Spectrum Disorder, *TD* typical development

**$p < .001$; *$p < .05$

Discussion

To address important deficits in existing measurement tools for quantifying social-communicative behaviors, we developed and applied a novel intelligent agent that could elicit, detect, and quantify social communication and collaboration within a CVE. We examined its accuracy by comparing its ratings to those of human observers. We also analyzed within-system objective measurements of participant success related to task performance and execution across time and conditions. Our results indicated that (i) the system accurately generated features that reflected several skills of participants with ASD and their TD peers while they played games with the intelligent agent, (ii) interactions with the intelligent agent reflected important aspects of peer-mediated interactions in the CVE, and (iii) the system performed similarly on most measured features across participants with and without ASD, capturing some differences in performance across groups. These findings offer important preliminary support for the use of technological systems to independently elicit and assess behaviors that, rather than occurring in isolation within a virtual environment, may reflect skills shown in interactions with other people.

Most existing CVE intervention systems attempt to measure aspects of peer-mediated interactions by generating task-performance features only. Given the importance of verbal-communication behaviors to social interaction, however (Owen-DeSchryver et al. 2008; Schmidt et al. 2012), and the time-consuming nature of human coding methodologies, we endowed our system with the capacity to generate task-performance and verbal-communication skills within the virtual environment. We did this utilizing not only interactions with an intelligent agent that was trained via our own machine learning algorithms, but also by validating findings using interactions with typically developing peers. CRETA accurately classified statements as fitting within dialogue acts at much higher rates than chance (> 68% across both conditions), with modest error rates in speech recognition (18–24%). It robustly assessed not only word and sentence frequency, but also pre-defined dialogue acts which reflect important components of communication and collaboration. These results offer promise for future iterations of the machine learning methodology.

Our system automatically generated meaningful verbal-communication and task-performance features to measure both communication and collaboration skills of the participants within the virtual environment. We discovered that aspects of participant behavior captured by the system directly correlated with human ratings of overall skills, suggesting that the system captured data with potential for clinical or measurement-related significance in ways that could systematically automate human coders' evaluation. Although human ratings of communication and collaboration did not relate to the within-system success of TD individuals, they did relate to the success for individuals with ASD. Interestingly, however, low ratings of communication and collaboration were associated with task failures across both groups. Participants with ASD also spent more time dragging pieces with the computer mouse. Future work should examine the potential contributions of fine motor deficits (Fulceri et al. 2018; Kaur et al. 2018) or visual attention, planning, or coordination (Dowd et al. 2012; Samad et al. 2017) on CVE performance, especially for individuals with ASD. It will also be important for future investigations to examine what independent factors may contribute to task failure as opposed to success, given group differences in how these features related to collaboration and communication skills.

We found that ways in which participants interacted with our agent reflected important aspects of how they interacted with other human users, who in this study were typically developing peers. As compared to interactions with the agent, our system was able to capture how much users were talking, how often they provided information, the time they spent moving pieces, and the time that they spent working together at similar rates found to their interactions with the agent. These communicative and interactive behaviors have support within the existing literature as being important for peer-mediated interactions, including word and sentence frequency (Hourcade et al. 2013), success frequency (Buaminger-Zviely et al. 2013), and collaboration time (Wilson and Russell 2007). Therefore, CRETA has the potential to save time, effort, and costs associated with a human coding methodology when assessing aspects of communication and collaboration skills for children with ASD and their peers.

In addition to validating system capacity and functionality, our findings may have implications for clinical intervention as well as assessment. Many existing interventions to promote the social use of language focus on young children

rather than adolescents (Parsons et al. 2017); social skills interventions for adolescents have a well-defined literature (Vernon et al. 2018; Laugeson et al. 2012, 2014) but rely upon the opinion of an expert human therapist to judge treatment goals and progress toward such. Additionally, measurement strategies for tracking social language may focus on questionnaires or tests administered at defined points in time (Parsons et al. 2018) rather than in an ongoing, adaptive fashion. Our findings offer preliminary support for the ability of multi-user online tasks to assess for communication and collaboration skills as part of structured, semi-naturalistic interactions, not only with other humans, but also with an artificially intelligent agent that can assess areas of need, measure progress, and offer practice.

Although the results of this preliminary validation study are promising, it has several limitations which warrant discussion. First, the sample size was relatively small, and not well characterized regarding current levels of cognitive functioning, autism symptoms, and language skills. It will be important for future work to examine system functionality across diagnostic subgroups, ages, and tasks to better understand not only potential for immediate impact, but also change and appropriateness over time. Additionally, the experimental design consisted of only one session. We succeeded in designing a measurement system that automatically measured important skills within the CVE, and results indicate that it has the potential to automatically measure both communication and collaboration skills of children with ASD. However, this system is in the very early stages of design and implementation. In future work, we will utilize CRETA for real-time measurements with more participants, a counter-balanced experimental design, more robust interrater reliability, and a longer intervention duration to further assess potential system impact. Importantly, a more stringent interrater reliability process will give us increased confidence that future system iterations reflect consensus ratings of communication and collaboration skills across clinicians. We will also use this data to further train our machine learning algorithms to classify skills, particularly communication skills, some of which showed higher error rates due to their very low frequency.

Second, as part of these future iterations, we will test CRETA across a broader variety of interactions. The games utilized in this work were carefully designed to maximize potential for collaborative and communicative behaviors related to achieving specific goals (e.g., placing a certain colored puzzle piece in the correct spot). Introducing more flexibility or different goals, perhaps related to communication (e.g., finding out your partner's favorite food), would allow CRETA to expand beyond goal-focused communication and performance skills, laying the groundwork for expanding the dialogue act classification system. Training of existing and new classifications will require additional conversational and performance data acquired through additional work. Future studies would also benefit from a companion assessment in which the ASD-TD pairs interact face-to-face to gather comparative data regarding communication and collaboration outside of a technological system.

Third, the small number of system-generated features was limited by the preliminary nature of the work. For example, the features *request color* and *request object* were not frequently uttered by the participants, which has implications for future iterations of game design. Additionally, other human behaviors relay important communication information that may influence collaborative efforts in a partner, including eye gaze, body language, and facial expression. Future versions of this system will utilize separate algorithms to capture information related to eye gaze, gesture usage, and emotion recognition. Understanding the role of non-verbal communication and developing intelligent technologies for capturing such within controlled, scaffolded virtual interactions holds additional promise for measurement tools in the next iteration of this work. Another important avenue of work relates to the content of spoken language. Natural language processing to examine speech content was beyond the scope of this work and, indeed, presents a technological hurdle not fully surpassed within the literature (Amirhosseini and Kazemian 2019). However, devising systems that can recognize core content or frequencies of speech/language targets may allow for more systematic measurement of skills and change in such, whether in response to different tasks, across groups, or over time. Promoting social communication and collaboration may be key to optimizing a variety of lifespan outcomes for individuals, such as employment, educational success, and relationships, areas in which young adults with ASD show well-documented vulnerabilities (Bennett et al. 2018; Connor et al. 2019; Ohl et al. 2017). CRETA and associated systems may hold potential as training tools, capable of utilization either within broader systems of care or by individuals themselves, offering feedback regarding areas of need as well as outlets for real-time interaction with peers.

Although they limit the generalizability of our findings, these limitations also offer exciting opportunities for additional exploration of how CVEs, intelligent agents, and machine learning methodologies can be harnessed to capture important information about communication and collaboration skills for children on the autism spectrum. To our knowledge, CRETA is the first system designed to assess collaboration and communication using a virtual agent as well as peer-to-peer interactions. Our findings suggest that intelligent technologies hold the potential for measurement of core skills related not only to autism spectrum disorder, but other developmental diagnoses which may impact how individuals convey information, work together, and accomplish goals.

Author Contributions LZ, AW, ZW, and NS conceived of the study and crafted the experimental design. LZ, NS, and ZW helped design the CVE procedure. LZ, ZW, and AS provided oversight of study implementation across Vanderbilt sites. LZ, AA, and AS assisted with data collection and analysis for manuscript preparation. LZ, AW, AA, ZW, and NS significantly participated in drafting the article, revising it critically, and providing final approval of the manuscript. All authors are in agreement with accountability for all aspects of the work.

Funding This project was funded by the National Institutes of Health (1R21MH111548-01).

Compliance with Ethical Standards

Conflicts of interest The authors declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human subjects were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Amirhosseini, M. H., & Kazemian, H. (2019). Automating the process of identifying the preferred representational system in neuro linguistic programming using natural language processing. *Cognitive Processing*. https://doi.org/10.1007/s10339-019-00912-3.

Anagnostou, E., Jones, N., Huerta, M., Halladay, A. K., Wang, P., Scahill, L., … Dawson, G. (2015). Measuring social communication behaviors as a treatment endpoint in individuals with autism spectrum disorder. *Autism, 19*(5), 622–636. https://doi.org/10.1177/1362361314542955.

Battocchi, A., Pianesi, F., Tomasini, D., Zancanaro, M., Esposito, G., Venuti, P., … Weiss, P. L. (2009). *Collaborative Puzzle Game: A tabletop interactive game for fostering collaboration in children with Autism Spectrum Disorders (ASD).* Paper presented at the Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces.

Bauminger-Zviely, N., Eden, S., Zancanaro, M., Weiss, P. L., & Gal, E. (2013). Increasing social engagement in children with high-functioning autism spectrum disorder using collaborative technologies in the school environment. *Autism, 17*(3), 317–339. https://doi.org/10.1177/1362361312472989.

Bekele, E., Crittendon, J., Zheng, Z., Swanson, A., Weitlauf, A., Warren, Z., et al. (2014). Assessing the utility of a virtual environment for enhancing facial affect recognition in adolescents with autism. *Journal of Autism and Developmental Disorders, 44*(7), 1641–1650. https://doi.org/10.1007/s10803-014-2035-8.

Bekele, E., Zheng, Z., Swanson, A., Crittendon, J., Warren, Z., & Sarkar, N. (2013). Understanding how adolescents with autism respond to facial expressions in virtual reality environments. *IEEE Transactions on Visualization and Computer Graphics, 19*(4), 711–720. https://doi.org/10.1109/TVCG.2013.42.

Benford, S., Greenhalgh, C., Rodden, T., & Pycock, J. (2001). Collaborative virtual environments. *Communications of the ACM, 44*(7), 79–85.

Bennett, A. E., Miller, J. S., Stollon, N., Prasad, R., & Blum, N. J. (2018). Autism spectrum disorder and transition-aged youth. *Current Psychiatry Reports, 20,* 103.

Cauell, J., Bickmore, T., Campbell, L., & Vilhjálmsson, H. (2000). Designing embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, & E. Churchi (Eds.), *Embodied conversational agents* (pp. 29–63). Cambridge: The MIT Press.

Cheng, Y., Chiang, H.-C., Ye, J., & Cheng, L.-H. (2010). Enhancing empathy instruction using a collaborative virtual learning environment for children with autistic spectrum conditions. *Computers & Education, 55*(4), 1449–1458.

Clarke, E. M., Emerson, E. A., & Sistla, A. P. (1986). Automatic verification of finite-state concurrent systems using temporal logic specifications. *ACM Transactions on Programming Languages and Systems, 8*(2), 244–263. https://doi.org/10.1145/5397.5399.

Connor, A., Sung, C., Strain, Z., Zeng, S., & Fabrizi, S. (2019). Building skills, confidence, and wellness: Psychosocial effects of soft skills training for young adults with autism. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-019-03962-w.

Constantino, J. N., & Gruber, C. P. (2002). *The social responsiveness scale*. Los Angeles: Western Psychological Services.

Curtis, D. D., & Lawson, M. J. (2001). Exploring collaborative online learning. *Journal of Asynchronous Learning Networks, 5*(1), 21–34.

Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction, 22*(1–2), 9–38.

Dowd, A. M., McGinley, J. L., Taffe, J. R., & Rinehart, N. J. (2012). Do planning and visual integration difficulties underpin motor dysfunction in autism? A kinematic study of young children with autism. *Journal of Autism and Developmental Disorders, 42*(8), 1539–1548.

Fulceri, F., Grossi, E., Contaldo, A., Narzisi, A., Apicella, F., Parrini, I.,… & Muratori, F. (2018). Motor skills as moderators of core symptoms in Autism Spectrum Disorders: Preliminary data from an exploratory analysis with Artificial Neural Networks. *Frontiers in Psychology, 9*, 2683.

Gogoulou, A., Gouli, E., & Grigoriadou, M. (2008). Adapting and personalizing the communication in a synchronous communication tool. *Journal of Computer Assisted Learning, 24*(3), 203–216.

Goodwin, M. S. (2008). Enhancing and accelerating the pace of autism research and treatment the promise of developing innovative technology. *Focus on Autism and Other Developmental Disabilities, 23*(2), 125–128. https://doi.org/10.1177/1088357608316678.

Hourcade, J. P., Williams, S. R., Miller, E. A., Huebner, K. E., & Liang, L. J. (2013). Evaluation of tablet apps to encourage social interaction in children with autism spectrum disorders. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3197–3206).

Josman, N., Ben-Chaim, H. M., Friedrich, S., & Weiss, P. L. (2008). Effectiveness of virtual reality for teaching street-crossing skills to children and adolescents with autism. *International Journal on Disability and Human Development, 7*(1), 49–56.

Kandalaft, M. R., Didehbani, N., Krawczyk, D. C., Allen, T. T., & Chapman, S. B. (2013). Virtual reality social cognition training for young adults with high-functioning autism. *Journal of Autism and Developmental Disorders, 43*(1), 34–44. https://doi.org/10.1007/s10803-012-1544-6.

Kaur, M., Srinivasan, S. M., & Bhat, A. N. (2018). Comparing motor performance, praxis, coordination, and interpersonal synchrony between children with and without Autism Spectrum Disorder (ASD). *Research in Developmental Disabilities, 72,* 79–95.

Klakow, D., & Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication, 38*(1–2), 19–28.

Kopp, S., Gesellensetter, L., Krämer, N. C., & Wachsmuth, I. (2005). *A conversational agent as museum guide–design and evaluation of a real-world application.* Paper presented at the International Workshop on Intelligent Virtual Agents.

Krishnaiah, P. R. (1980). *Handbook of statistics* (Vol. 31). New Delhi: Motilal Banarsidass Publishe.

Kumar, R., Rosé, C. P., Wang, Y.-C., Joshi, M., & Robinson, A. (2007). Tutorial dialogue as adaptive collaborative learning support. *Frontiers in Artificial Intelligence and Applications, 158,* 383.

Lahiri, U., Bekele, E., Dohrmann, E., Warren, Z., & Sarkar, N. (2013). Design of a virtual reality based adaptive response technology for children with autism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 21*(1), 55–64. https://doi.org/10.1109/TNSRE.2012.2218618.

Laugeson, E. A., Ellingsen, R., Sanderson, J., Tucci, L., & Bates, S. (2014). The ABC's of teaching social skills to adolescents with autism spectrum disorder in the classroom: The UCLA PEERS® program. *Journal of Autism and Developmental Disorders, 44*(9), 2244–2256.

Laugeson, E. A., Frankel, F., Gantman, A., Dillon, A. R., & Mogil, C. (2012). Evidence-based social skills training for adolescents with autism spectrum disorders: The UCLA PEERS program. *Journal of Autism and Developmental Disorders, 42*(6), 1025–1036.

Leman, P. J. (2015). How do groups work? Age differences in performance and the social outcomes of peer collaboration. *Cognitive science, 39*(4), 804–820.

Linton, F., Goodman, B., Gaimari, R., Zarrella, J., & Ross, H. (2003). *Student modeling for an intelligent agent in a collaborative learning environment.* Paper presented at the International Conference on User Modeling.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., … Rutter, M. (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. https://doi.org/10.1023/a:1005592401947.

McManus, M. M., & Aiken, R. M. (1995). Monitoring computer-based collaborative problem solving. *Journal of Interactive Learning Research, 6*(4), 307.

McPheeters, M. L., Warren, Z., Sathe, N., Bruzek, J. L., Krishnaswami, S., Jerome, R. N., et al. (2011). A systematic review of medical treatments for children with autism spectrum disorders. *Pediatrics, 127*(5), e1312–e1321. https://doi.org/10.1542/peds.2011-0427.

Millen, L., Hawkins, T., Cobb, S., Zancanaro, M., Glover, T., Weiss, P. L., & Gal, E. (2011). *Collaborative technologies for children with autism.* Paper presented at the Proceedings of the 10th International Conference on Interaction Design and Children.

Mitchell, P., Parsons, S., & Leonard, A. (2007). Using virtual environments for teaching social understanding to 6 adolescents with autistic spectrum disorders. *Journal of Autism and Developmental Disorders, 37*(3), 589–600. https://doi.org/10.1007/s10803-006-0189-8.

Nabeth, T., Razmerita, L., Angehrn, A., & Roda, C. (2005). InCA: A cognitive multi-agents architecture for designing intelligent & adaptive learning systems. *Computer Science and Information Systems, 2*(2), 99–114.

Ohl, A., Grice Sheff, M., Small, S., Nguyen, J., Paskor, K., & Zanjirian, A. (2017). Predictors of employment status among adults with Autism Spectrum Disorder. *Work, 56*(2), 345–355.

Owen-DeSchryver, J. S., Carr, E. G., Cale, S. I., & Blakeley-Smith, A. (2008). Promoting social interactions between students with autism spectrum disorders and their peers in inclusive school settings. *Focus on Autism and Other Developmental Disabilities, 23*(1), 15–28.

Parsons, L., Cordier, R., Munro, N., & Joosten, A. (2018). The feasibility and appropriateness of a peer-to-peer, play-based intervention for improving pragmatic language in children with autism spectrum disorder. *International Journal of Speech-Language Pathology, 2,* 1–13.

Parsons, L., Cordier, R., Munro, N., Joosten, A., & Speyer, R. (2017). A systematic review of pragmatic language interventions for children with autism spectrum disorder. *PLoS ONE, 12*(4), e0172242.

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*(Pt 5), 430–443.

Rutter, M., Bailey, A., & Lord, C. (2003). *The social communication questionnaire: Manual*. Los Angeles: Western Psychological Services.

Samad, M. D., Diawara, N., Bobzien, J. L., Harrington, J. W., Witherow, M. A., & Iftekharuddin, K. M. (2017). A feasibility study of autism behavioral markers in spontaneous facial, visual, and hand movement response data. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26*(2), 353–361.

Scahill, L., Aman, M. G., Lecavalier, L., Halladay, A. K., Bishop, S. L., Bodfish, J. W., … Dawson, G. (2015). Measuring repetitive behaviors as a treatment endpoint in youth with autism spectrum disorder. *Autism, 19*(1), 38–52. https://doi.org/10.1177/1362361313510069.

Scheuer, O., Loll, F., Pinkwart, N., & McLaren, B. M. (2010). Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning, 5*(1), 43–102.

Schmidt, M., Laffey, J. M., Schmidt, C. T., Wang, X., & Stichter, J. (2012). Developing methods for understanding social behavior in a 3D virtual learning environment. *Computers in Human Behavior, 28*(2), 405–413.

Schmidt, M., Laffey, J., & Stichter, J. (2011). *Virtual social competence instruction for individuals with autism spectrum disorders: Beyond the single-user experience.* Paper presented at the Proceedings of CSCL.

Srinivasan, S., & Petkovic, D. (2000). *Phonetic confusion matrix based spoken document retrieval.* Paper presented at the Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., … Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics, 26*(3), 339-373.

Van Rosmalen, P., Brouns, F., Tattersall, C., Vogten, H., Bruggen, J., Sloep, P., et al. (2005). Towards an open framework for adaptive, agent-supported e-learning. *International Journal of Continuing Engineering Education and Life Long Learning, 15*(3–6), 261–275.

Veenstra-VanderWeele, J., & Warren, Z. (2015). Intervention in the context of development: Pathways toward new treatments. *Neuropsychopharmacology, 40*(1), 225–237. https://doi.org/10.1038/npp.2014.232.

Vernon, T. W., Miller, A. R., Ko, J. A., Barrett, A. C., & McGarry, E. S. (2018). A randomized controlled trial of the Social Tools And Rules for Teens (START) program: An immersive socialization intervention for adolescents with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 48*(3), 892–904.

Vieira, A. C., Teixeira, L., Timóteo, A., Tedesco, P., & Barros, F. (2004). *Analyzing on-line collaborative dialogues: The oxentchê–chat.* Paper presented at the International Conference on Intelligent Tutoring Systems.

Walker, E., Rummel, N., & Koedinger, K. R. (2014). Adaptive intelligent support to improve peer tutoring in algebra. *International Journal of Artificial Intelligence in Education, 24*(1), 33–61.

Wallace, S., Parsons, S., & Bailey, A. (2015). Self-reported sense of presence and responses to social stimuli by adolescents with ASD in a collaborative virtual reality environment. *Journal of Intellectual and Developmental Disability, 42*(2), 131–141.

Weitlauf, A. S., McPheeters, M. L., Peters, B., Sathe, N., Travis, R., Aiello, R., … Warren, Z. (2014). *Therapies for Children With Autism Spectrum Disorder: Behavioral Interventions Update*. Rockville: Agency for Healthcare Research and Quality.

Wilson, G. F., & Russell, C. A. (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human Factors*, *49*(6), 1005–1018.

Zancanaro, M., Pianesi, F., Stock, O., Venuti, P., Cappelletti, A., Iandolo, G., … Rossi, F. (2007). Children in the museum: an environment for collaborative storytelling. In O. Stock, M. Zancanaro (Eds.), *PEACH-intelligent interfaces for museum visits* (pp. 165–184). Berlin: Springer.