9-1-2020

Measuring the Perceived Social Intelligence of Robots

Kimberly A. Barchard
University of Nevada, Las Vegas, kim.barchard@unlv.edu

Leiszle Lapping-Carr
University of Nevada, Las Vegas

R. Shane Westfall
University of Nevada, Las Vegas

Andrea Fink-Armold
Uniiversity of Nevada, Las Vegas

Santosh Balajee Banisetty
University of Nevada, Reno

See next page for additional authors

Measuring the Perceived Social Intelligence of Robots

KIMBERLY A. BARCHARD, LEISZLE LAPPING-CARR, R. SHANE WESTFALL, and ANDREA FINK-ARMOLD, Department of Psychology, University of Nevada Las Vegas, Las Vegas, NV, USA
SANTOSH BALAJEE BANISETTY and DAVID FEIL-SEIFER, Department of Computer Science & Engineering, University of Nevada Reno, Reno, NV, USA

Robotic social intelligence is increasingly important. However, measures of human social intelligence omit basic skills, and robot-specific scales do not focus on social intelligence. We combined human robot interaction concepts of beliefs, desires, and intentions with psychology concepts of behaviors, cognitions, and emotions to create 20 Perceived Social Intelligence (PSI) Scales to comprehensively measure perceptions of robots with a wide range of embodiments and behaviors. Participants rated humanoid and non-humanoid robots interacting with people in five videos. Each scale had one factor and high internal consistency, indicating each measures a coherent construct. Scales capturing perceived social information processing skills (appearing to recognize, adapt to, and predict behaviors, cognitions, and emotions) and scales capturing perceived skills for identifying people (appearing to identify humans, individuals, and groups) correlated strongly with social competence and constituted the Mind and Behavior factors. Social presentation scales (appearing friendly, caring, helpful, trustworthy, and not rude, conceited, or hostile) relate more to Social Response to Robots Scales and Godspeed Indices, form a separate factor, and predict positive feelings about robots and wanting social interaction with them. For a comprehensive measure, researchers can use all PSI 20 scales for free. Alternatively, they can select the most relevant scales for their projects.

CCS Concepts: • **General and reference → Empirical studies**; **Measurement**; • **Human-centered computing → User studies**; *HCI theory, concepts and models*;

Additional Key Words and Phrases: Social intelligence, human-robot interaction, human-computer interaction, socially assistive robotics

ACM Reference format:
Kimberly A. Barchard, Leiszle Lapping-Carr, R. Shane Westfall, Andrea Fink-Armold, Santosh Balajee Banisetty, and David Feil-Seifer. 2020. Measuring the Perceived Social Intelligence of Robots. *ACM Trans. Hum.-Robot Interact.* 9, 4, Article 24 (September 2020), 29 pages.
https://doi.org/10.1145/3415139

The authors thank the Nevada Space Grant Consortium (NNX15AK48A), the National Institute for Health (P20GM103650), and the National Science Foundation (IIS-1719027) for their support of this research.
Authors' addresses: K. A. Barchard and A. Fink-Armold, Department of Psychology, University of Nevada, Las Vegas, 4505 Maryland Parkway, Las Vegas, NV, 89154-5030; emails: {kim.barchard, andrea.fink-armold}@unlv.edu; L. Lapping-Carr, Department of Psychiatry and Behavioral Sciences, Northwestern University Feinberg School of Medicine, 446 E Ontario St, 1576, Chicago, IL 60611; email: leiszle.lapping-carr@northwestern.edu; R. S. Westfall, Psychology Department, Western Wyoming College, 2500 College Drive, Rock Spring, WY, 82901; email: rwestfall@westernwyoming.edu; S. B. Banisetty and D. Feil-Seifer, Department of Computer Science & Engineering, University of Nevada Reno, Mail Stop 0171, 1664 N Virginia St, Reno, NV 89557; emails: santoshbanisetty@nevada.unr.edu, dave@cse.unr.edu.

2573-9522/2020/09-ART24
https://doi.org/10.1145/3415139

1 INTRODUCTION

Social intelligence is defined as the ability to interact effectively with others to accomplish your goals [Ford and Tisak 1983]. Social intelligence is critically important for social robots that interact and communicate with people [Dautenhahn 2007]. For example, a social robot may be expected to teach to, learn from, and collaborate with people. Such tasks are easier if the robot can process the social context and react with appropriate social behavior. Even when robots are deployed in non-social roles or engaged in non-social tasks, social intelligence nevertheless helps robots be effective. For example, if robots are annoying, people sometimes kick, punch, or want to damage them [Brščić et al. 2015; Mutlu and Forlizzi 2008]. Although it will be a long time before robots exhibit true social intelligence, existing robots can nonetheless create the perception of social intelligence, which itself facilitates human robot interaction (HRI).

Many scales have been designed to measure the social intelligence of humans (e.g., Agran et al. [2016], Danielson and Phelps [2003], Frankovsky and Birknerova [2014], Riggio [1989], Silvera et al. [2001], and Tahiroglu et al. [2014]). In theory, these scales could be adapted to rate the social intelligence of robots. However, these scales omit basic skills that are essential for smooth social interactions, because most humans (including children) have these skills. For example, almost all humans understand that other people have thoughts, emotions, and behaviors; can distinguish humans from non-humans; know that people are individuals; and remember their previous interactions with specific people. Most robots do not have these basic skills, and the robots that do can legitimately be considered to have higher social intelligence. Thus, scales designed to measure social intelligence in humans are missing many essential basic skills that are important to evaluate in robots.

Even when a particular skill seems relevant to robots, the particular items that are used to measure that skill in humans are often not usable with robots. For example, the Social Skills Survey [Agran et al. 2016] refers to making objectionable gestures, something that would not pertain to a robot lacking limbs. Given that social intelligence is relevant to many different types of robots, a measure of social intelligence in robots should ideally be applicable to a wide variety of embodiments and behaviors. It should not assume that a robot has any particular type of body or is able to engage in any particular behavior. For example, the items should be applicable both to robots that can move, pick up objects, and speak, and to robots without those capabilities. Moreover, the items should not assume that the robot has any particular cognitive or emotional capacities. For example, the items should be applicable both to robots that have visual, audio, and speech processing abilities, and to those that do not. Because many of the items on human social intelligence measures are irrelevant to targets that lack humanoid bodies or functionalities, it is impossible to adequately adapt the items for use with robots.

Given the importance of social intelligence for successful HRI, several aspects of social intelligence have been included in scales designed to assess robots. These scales include Likability and Perceived Intelligence [Bartneck et al. 2009]; Perceived Sociability, Perceived Adaptivity, Trust, and Anxiety [Heerink et al. 2010]; Welcome, Appealing, and Unobtrusive [Moshkina 2012]; and Pleasure and Warmth [Ho and MacDorman 2010]. However, these scales often include material that goes beyond social intelligence. For example, Trust [Heerink et al.] measures whether the user would trust the robot's advice and also whether the user would follow that advice; Anxiety [Heerink et al.] measures both whether the robot is scary and whether the user would be afraid of breaking the robot; and Perceived Intelligence [Bartneck et al.] includes many abilities beyond social intelligence. In addition, a robot might be considered likable, unobtrusive, and safe for reasons besides its social intelligence. Moreover, existing scales for robots do not capture many

important aspects of social intelligence, such as the ability to identify and recognize humans. Thus, existing HRI scales cannot adequately measure perceptions of robotic social intelligence.

Given the limitations of existing measures to assess the perceived social intelligence of robots, a new measure was needed. Therefore, our goal was to create a comprehensive measure of perceived robot social intelligence, including more basic social intelligence concepts and skills than can be found in the human social intelligence literature, which can be applied to a range of robotic embodiments and behaviors. See the test manual by Barchard et al. [2018] for information about the methods we used to select areas to study and methods we used to draft items. Researchers could use our entire set of 20 scales if they needed a comprehensive measure of robot social intelligence, or they could select the specific scales that are most relevant to their own research questions.

In this article, we explain how our 20 scales conceptually measure various aspects of robot social intelligence. We then evaluate the factor structure and internal consistency reliability of each scale, examine the relation of these new scales to overall social intelligence and to existing robot rating scales (likability, perceived intelligence, etc.), examine the relation of these new scales and the existing robot perception scales to how robots make people feel and whether people want contact with the robots, and compare five different robot videos using our 20 scales. We conclude that the new Perceived Social Intelligence (PSI) Scales provide unique information that is not captured by current robot rating scales and can distinguish between robot videos that suggest different levels of social intelligence.

1.1 Our Conceptualization of Social Intelligence

We created 20 scales (see the Appendix for scale definitions and items) that measure perceived social intelligence in four different ways. First, our scales provide an overall measure of perceived social competence. This overall assessment gauges the extent to which the robot appears to have strong social skills in general. While separate measurement of the components of social competence will often be beneficial, sometimes a measure of overall social intelligence will be sufficient. In addition, an overall assessment can be useful for understanding how the different components impact overall perceptions of robots.

Second, our scales measure whether robots appear able to identify people in three different ways: to detect human presence, to distinguish individuals from each other, and to determine which people are together. Identifying people in these ways could allow robots to have increasingly refined interactions. Detecting human presence could allow robots to keep appropriate social distance and to initiate interactions. For example, a robot might weave through furniture to cross a room, then stop a few feet away from a woman and say hello. Recognizing individuals could allow robots to use previous interactions (such as user preferences and previous conversations) to adapt their behavior. For example, the robot might address the woman by name and ask if she is ready to go to lunch. Determining which people are together could allow robots to avoid disrupting ongoing social interactions (e.g., not walking between her and her friend) and to generalize from one person to another (e.g., asking if her friend would like to accompany them).

Third, the scales measure whether robots appear to have nine information processing abilities related to people. In HRI, there are fundamental concepts of beliefs, desires, and intentions (e.g., Davis and Ramulu [2017] and Georgeff and Lansky [1987]). In psychology, there are somewhat similar concepts of behaviors, cognitions, and emotions (e.g., Weiten [2017]). We believe these two approaches can be integrated. We conceptualize desires as part of emotions, beliefs as part of cognitions, and the ability to infer someone's intentions as predictions of their behaviors. This led us to nine information processing abilities related to people. Thus, we measure the extent to which robots appear able to (1) recognize, (2) adapt to, and (3) predict (a) human behaviors, (b) human cognitions (including beliefs), and (c) human emotions (including desires).

Recognizing, adapting to, and predicting human behaviors are the most fundamental aspects of social intelligence. Imagine a robot is attempting to help a man with arthritis make lunch. First, the robot might attempt to discern what the man is doing: that he has removed two pieces of bread from a bag and placed a jar on the counter. Next, the robot might adapt its behavior based upon its perception of what the man does. If the man says, "Please open that," the robot might reach for the jar and open it. Finally, the robot might attempt to predict what the man will do next. If the robot predicted the man would use the jar it just opened, it might push the jar back towards the man without needing to be asked, leaving the man free to attend to other things. Thus, recognizing, adapting to, and predicting human behaviors could allow robots to work cooperatively and smoothly with people.

Recognizing, adapting to, and predicting human cognition could also aid effective social interaction. If a tutoring robot could recognize why a child got a math problem wrong, it could adapt its feedback appropriately. Similarly, if it could predict concepts that are likely to cause a child confusion, it could forestall that confusion by focusing on common misconceptions.

Recognizing, adapting to, and predicting human emotions can help build and maintain relationships. If a robot could recognize emotions, it could adapt its behavior. For example, if a caregiver robot noticed that a person was crying, it could ask what was wrong. Perhaps the person spilled food on themselves. If the robot could predict that the person will want to change their clothes, it could fetch clean clothes for the person, without the person needing to provide explicit instructions. Thus, if robots can adjust their behavior based upon how people feel and what they want, they can build personal connections and increase the likelihood that the people they are caring for will listen to and respond to their suggestions.

Fourth, our scales measure seven aspects of social presentation, the ability to appear to be a desirable social partner: someone who is friendly, helpful, caring, and trustworthy, and who is not rude, conceited, or hostile. Being trustworthy [Parales-Quenza 2006] and having a prosocial attitude are key components of social intelligence [Marlowe 1986]. Robots are perceived as more friendly if they engage in acts of social intelligence such as mimicry and praise [Kaptein et al. 2011], and friendliness facilitates interpersonal connections [Albrecht 2004]. Moreover, social intelligence predicts cooperative behavior [Kinga and Ibolya 2013]. Thus, if a robot seems to be a desirable social partner, this will likely increase the frequency and duration of HRI and increase human cooperation and compliance, thus assisting the robot in accomplishing its goals in social interactions. Note, however, that we are not simply measuring if a robot is liked. Presenting oneself as a desirable social partner might lead to being liked, but being liked is influenced by many things besides social intelligence, such as appearing somewhat human-like, but not too human-like [Ho and MacDorman 2010].

2 METHOD

2.1 Participants

Participants were recruited using Amazon's online marketplace MTurk for a two-hour study paying $15. Because previous research has sometimes found cultural differences in perceptions of robots (e.g., Kamide and Arai [2017]), for this first study of the PSI Scales, we restricted data collection to a single country: the United States. We used two methods to limit participants to the United States: We used both MTurk (which requires a United States Social Security number to register) and Qualtrics (whose location screening is based upon IP addresses). We took this joint data screening approach because prior research suggests that some MTurk workers provide fraudulent social security numbers to pretend that they are in the United States and because IP screening services are not sufficient to effectively screen for location [Dennis et al. 2019]. Indeed, in our study

10 MTurk workers were given access to our study website by MTurk, but were rejected by Qualtrics because they were located outside the United States.

To further improve data quality, we restricted participants to MTurk workers who had completed at least 500 tasks and had a minimum acceptance rate of 95%, based on Peer et al. [2014] finding that these participants provide significantly better data: They are more likely to pass attention checks, provide data with higher internal consistency reliability, score lower on measures of response bias due to socially desirable responding, better replicate well-established anchoring effects, and are less likely to mark the mid-point of scales. In addition, we assigned each participant a qualification to ensure that they could not complete the study more than once. With these restrictions, 313 people with MTurk Worker IDs accessed the webpage for our study.

To ensure that participants would be able to view the online videos properly, we also screened the devices they used. They had to use a computer (not a cell phone), and they had to prove they could hear audio: They were presented with one of 10 sounds (e.g., chicken, train, dog) and had to identify which sound it was. Four of our participants initially failed this audio check, but then returned to the study, passed the audio check, and completed the study. Finally, we screened out 11 participants who did not complete the ratings for all five videos. Given our multiple screening methods, we are confident in the quality of our final participants. In addition, we screened individual data points thoroughly; see below.

This screening resulted in a final sample of 296 adults (145 female, 150 male). They ranged in age from 19 to 72 (mean 37.39, SD 11.50). When asked to select one of the following six categories, 80.4% identified as White, 7.1% as African-American, 5.1% as Asian, 4.1% as Hispanic, 0.3% as Native American, and 3.0% as other. The vast majority (97.6%) reported speaking English as their first language.

Participants had an average of five robot-related experiences in the last year (including both real and fictional robots), but rated themselves as having relatively little familiarity with real robots (mean 30.51/100, range 28.92–31.17).

2.2 Procedure

Participants completed the study online. After consenting to participate, participants completed questions about their demographics and their background with robots. Next, participants viewed five videos showing interactions between humans and robots. For each video, they completed several measures regarding their impressions of the robot. No time limit was given for the study and participants were allowed to take breaks. Finally, participants were debriefed and compensated for their time.

2.3 Materials

Five videos (1–3 minutes long) depicting HRI were selected to represent a wide range of robot social intelligence. We received permission from the videos' authors to both use and edit these videos for the current study. The videos were shown in the following order. Robovie (a large humanoid robot) asks a woman to lie about seeing an aquarium on her tour of a research lab; the woman agrees and is shown lying to the robot's supervisor [Kahn et al. 2015]. NAO (a small humanoid robot) performs community service for shoplifting batteries, apparently to no effect, as the NAO then steals batteries from a friend's bike light, resulting in a crash [De Greeff et al. 2014]. A robotic ottoman encourages people to put their feet up and cues them that it wants to leave [Sirkin et al. 2015]. PR2 (a large, boxy, humanoid robot) works cooperatively with two different people to stack blocks in specific patterns, adjusting its behavior based upon unexpected human behaviors [Devin et al. 2017]. Dragonbot (a small, fluffy, toy dragon robot) takes turns telling stories with children and creating visual depictions of the stories on an iPad [Kory 2014]. Thus, the videos showed

robots with a variety of body types and physical abilities, acting in a variety of ways in a variety of contexts.

2.4 Measures

2.4.1 Background Information.

2.4.1.1 Demographic variables. Participants completed items requesting their gender, age, first language, length of time they had been speaking English, and comfort with reading, writing, speaking, and listening to English, and asking whether they identified as White, African-American, Asian, Hispanic, Native-American, or other.

2.4.1.2 Familiarity with real robots. Mimicking the procedure used by Sung et al. [2009], we asked participants to rate their familiarity with 12 real robots selected from the robot hall of fame (www.robothalloffame.org). For each robot, we provided the name, the manufacturer, and an image. We asked participants to rate their familiarity with each robot using a five-point scale, 1 = *not at all familiar*, 2 = *slightly familiar*, 3 = *moderately familiar*, 4 = *very familiar*, and 5 = *extremely familiar*.

2.4.1.3 Broad experience with robots. To measure previous experience with robots, we used the five-item measure from MacDorman et al. [2009]. These items ask participants how often they have read or watched robot-related materials, attended robot-related events, had physical contact with robots, and built or programmed robots. Most items include both real robots and fictional ones; for example, they do not distinguish between reading comics and journal articles, or between watching documentaries and science fiction movies.

2.4.2 Ratings of the Robot in Each Video. After watching each video, participants completed several rating scales about that robot.

2.4.2.1 Perceived Social Intelligence (PSI) Scales. The PSI Scales were designed to measure the extent to which robots are perceived as possessing the 20 aspects of social intelligence described above. To facilitate the use of our scales, we wrote items using the International Personality Item Pool (IPIP [Goldberg et al. 2006]) third-person format. The IPIP is a public domain set of 3000+ items, available at http://ipip.ori.org/, that can be used for free and adapted as needed.

While many existing IPIP items are relevant to social intelligence, few could be used to measure the perceived social intelligence of robots. Many IPIP items are not relevant to robots (e.g., referring to friends), many assume social/intellectual skills that few or no current robots have (e.g., having emotions or episodic memories), and many assume certain body types or functions (e.g., being able to hear, speak, or move). Therefore, most of our items were newly created, so that they would apply to a wide range of robots. To ensure adequate content coverage, we drafted dozens of items for each of the 20 concepts. Then, we selected the best six items based upon their apparent relevance to the construct and the clarity of their phrasing. The Appendix provides a complete list of all scales, abbreviations, and definitions; see the Appendix. As shown there, all items use a five-point agreement scale to indicate how well the statement describes a particular robot: 1 = *Strongly Disagree*, 2 = *Disagree*, 3 = *Neutral*, 4 = *Agree*, and 5 = *Strongly Agree*. For detailed information about how to administer and score the PSI Scales, see the test manual [Barchard et al. 2018].

In this current study, to partially reduce possible order effects and carry-over effects in our study, we presented the 120 PSI items in a different order for each of the five robot videos.

2.4.2.2 Desire for social interaction. To assess whether participants desire social interaction with the robot, we designed a five-point agreement scale that participants used to indicate if they would like to meet this robot, be friends with this robot, live with this robot, and work with this robot.

2.4.2.3 Emotional reactions. To assess the emotional reactions of participants to the robots, we used an eight-item measure based upon Lövheim's [2012] adaptation of Tomkin's [1962 1963 1981 1991] theory of affect. We asked participants the extent to which the robot made them feel enjoyment, fear, surprise, shame, anger, distress, interest, and disgust, using a five-point scale where 1 = *Not at all*, 2 = *Slightly*, 3 = *Moderately*, 4 = *Very*, and 5 = *Extremely*.

2.4.2.4 Social Response to Robots Semantic Differential Scales. Participants completed four scales from the Social Response to Robots Semantic Differential Scales [Moshinka 2012]: Welcome, Appeal, Unobtrusiveness, and Naturalness. Each item was presented on a five-point scale, with one adjective anchoring the low end of the scale and another adjective anchoring the high end of the scale, for example, Unwelcome—Welcome, Boring—Interesting, Annoying—Inoffensive, and Fake—Natural. Positive adjectives were consistently placed on the right. Each scale was prefaced with the phrase "This robot is".

2.4.2.5 Godspeed Indices. Participants completed two of the Godspeed Indices [Bartneck et al. 2009]: Likability (first four items) and Perceived Intelligence. We omitted the last item from the Likability Scale because it overlapped with Moshinka's Welcome Scale. Once again, each item was presented on a five-point scale, with one adjective anchoring the low end of the scale and another adjective anchoring the high end of the scale, for example, Unfriendly—Friendly, Unintelligent—Intelligent. Positive adjectives were consistently placed on the right. Each scale was prefaced with the phrase "This robot is."

2.5 Data Screening

Data were carefully screened in SPSS to prevent participants who were univariate or multivariate outliers from having undue influence on the results. To identify univariate outliers for each of the 120 PSI items, we calculated z-scores for each response, as recommended by Tabachnick and Fidell [2019]. We identified 324 of the 178,800 responses (0.18%) as outliers because they had z-scores greater than 3.29, corresponding to a p-value of .001. Tabachnick and Fidell [2019] recommend moving univariate outliers to make them less deviant. Following this advice, we moved each outlier one point closer to the mean for that item. For example, if the mean on the five-point scale was 1.78 and scores of 5 were identified as outliers, those scores were changed to 4.

To identify participant-robot combinations that were multivariate outliers within each of the 20 PSI Scales, we calculated Mahalanobis distances. Across the 20 PSI Scales, between 0.7% and 5.3% of participant-robot combinations were multivariate outliers, as evidenced by Mahalanobis distances that exceeded the chi-squared critical value corresponding to a p-value of .001. As suggested by Tabachnick and Fidell [2019], these participant-robot pairs were deleted from the analyses for those scales.

2.6 Item Selection

The purpose of our study was to select items for the 20 PSI Scales and evaluate the quality of those final scales. For each scale, we selected the best four items. To do this, we used two analyses. First, we conducted exploratory factor analyses within each of the 20 scales to identify items that were strongly associated with the constructs being measured. To determine the number of factors, we conducted a parallel analysis [Cota et al. 1993; Horn 1965]. To implement the parallel analyses, we used the *psych* package [Revelle 2018] in R [R Core Team 2017] with the RStudio interface [RStudio Team 2016]. Seventeen scales had just one factor (as planned). For those scales, we selected items that had strong relations with that factor. Three scales had two correlated factors (i.e., AC, CON, and CAR). For those scales, we selected items that had large salient relations with the largest factor and/or the orthogonalized general factor. Second, we used item response theory to select the

items that provided the most information for discriminating between respondents. Item information curves were calculated using the *mirt* package [Chalmers 2012] in R with Samejima's [1969] graded response model. We selected items that provided the most information, as evidenced by information curves that were tall and wide. When a scale had more than four items with high factor coefficients and high information, we selected items that provided unique content to increase validity, and we included at least one reverse-coded item to reduce the influence of acquiescence response bias.

In addition to selecting the best four items for each of the 20 PSI Scales, we also selected the best single item for each based upon the factor analysis results. These 20 items constitute a short form that can measure all aspects of social intelligence in a concise format.

Recognizing that gender may impact perceptions of social others, we additionally considered whether there exist differences in men and women's responses to the PSI scales. Gender differences appear across a range of topics related to social intelligence, including implicit beliefs about acceptable emotion expression in childhood [Thomassin and Seddon 2019], beliefs about the impact of body appearance on social standing [Wang et al. 2019], and perceptions of threat and attractiveness based on facial expressions [Hester 2019]. Men and women also differ in how they perceive technology. For example, a study in Austria found that women participants have more positive attitudes towards new health or support devices, while men have more positive attitudes towards new communication and entertainment devices [Halmdienst et al. 2019]. Given that men and women sometimes have different perceptions, we compared the variance-covariance matrices for men and women for each of the 20 scales using Box's M. For most scales, there were no significant differences, indicating that the above item selections would apply equally to men and women. For three scales (HST, IH, and CON), there were statistically significant differences between men and women. However, these differences were negligible. HST and IH each had one factor for both men and women, with salient factor coefficients for all items. CON had two factors for both men and women, with the same items on each factor for the two genders. Therefore, for all 20 scales, we recommend the same best four items and the same best single item for men and women.

3 RESULTS

3.1 Evaluating the Quality of the Full PSI Scales

3.1.1 Internal Consistency. To determine if each of the 20 PSI Scales measures a coherent construct, we checked how many constructs underlie each scale using exploratory factor analysis and we measured internal consistency reliability using coefficient alpha. Three separate exploratory factor analysis methods – parallel analysis [Cota et al. 1993; Horn 1965], the minimum average partial test (MAP test [Velicer 1976]), and the scree plot [Cattell 1966]—all showed that each PSI Scale has only a single underlying factor. Moreover, coefficient alpha [Cronbach 1951] showed that each scale has strong internal consistency (range .75–.94, average .86). See Table 1. We conclude that the 20 PSI Scales each measure a single coherent construct.

The two PSI total scores also had strong internal consistency. See the top two rows in Table 1. However, these total scores would not be expected to have just a single factor (see the section below on the exploratory factor analysis).

The four Social Response to Robots Semantic Differential Scales and the two Godspeed Indices also had strong internal consistency (range .89–.96, average .93). See bottom section of Table 1.

3.1.2 Relations with among the PSI Scales. The 20 PSI Scales had meaningful inter-scale correlations. See Table 2. Being perceived as rude, conceited, and hostile had negative correlations with most other scales (1 of these 51 correlations was positive but negligible). The remaining scales consistently had positive inter-correlations. Scales that measure conceptually related content had

Table 1. Internal Consistencies of the Robot Perception Scales

Scale	Coefficient Alpha [95% confidence interval]
PSI Scales	
Short Form Total	.93 [.92, .93]
Long Form Total	.96 [.95, .96]
Social Competence	.92 [.91, .93]
Identifies Humans	.83 [.81, .84]
Identifies Individuals	.94 [.94, .95]
Identifies Social Groups	.85 [.84, .86]
Recognizes Human Behaviors	.84 [.82, .85]
Adapts to Human Behaviors	.89 [.88, .90]
Predicts Human Behaviors	.83 [.81, .84]
Recognizes Human Cognitions	.86 [.85, .87]
Adapts to Human Cognitions	.80 [.78, .82]
Predicts Human Cognitions	.91 [.90, .92]
Recognizes Human Emotions	.90 [.89, .91]
Adapts to Human Emotions	.88 [.87, .89]
Predicts Human Emotions	.88 [.87, .89]
Friendly	.82 [.80, .83]
Helpful	.88 [.87, .89]
Caring	.84 [.83, .86]
Trustworthy	.90 [.89, .90]
Rude	.85 [.83, .86]
Conceited	.75 [.73, .77]
Hostile	.87 [.86, .88]
Social Response to Robots Semantic Differential Scales	
Welcome	.96 [.96, .97]
Appeal	.93 [.93, .94]
Unobtrusiveness	.89 [.88, .90]
Naturalness	.92 [.92, .93]
Godspeed Index	
Likability	.96 [.96, .96]
Perceived Intelligence	.92 [.91, .92]

strong correlations, as expected. However, many inter-scale correlations were small (54 of the 190 correlations were less than .40 in absolute value), demonstrating that these scales are measuring several distinct constructs.

3.1.3 Relations with Existing Robot Perception Scales. The 20 PSI Scales correlated in meaningful ways with the four Social Response to Robots Scales and the two Godspeed Indices. See Table 3. For example, the PSI Helpful and Trustworthy Scales had strong positive correlations with Likability and Welcome. All of the other relations were also in the expected directions, and correlations were usually moderate to strong.

The PSI Long Form total, Short Form total, and Social Competence Scale had their strongest correlations with Perceived Intelligence, Appeal, and Naturalness. This suggests that the PSI Scales are indeed measuring perceived social intelligence, and that appearing socially intelligent makes a

Table 2. Correlations among the 20 PSI Scales

	Socially Competent	Identifies Humans	Identifies Individuals	Identifies Social Groups	Recognizes Human Behaviors	Adapts to Human Behaviors	Predicts Human Behaviors
Socially Competent	1.00						
Identifies Humans	.46**	1.00					
Identifies Individuals	.74**	.43**	1.00				
Identifies Social Groups	.69**	.36**	.78**	1.00			
Recognizes Human Behaviors	.45**	.68**	.43**	.39**	1.00		
Adapts to Human Behaviors	.53**	.54**	.42**	.49**	.78**	1.00	
Predicts Human Behaviors	.69**	.50**	.61**	.65**	.61**	.67**	1.00
Recognizes Human Cognitions	.77**	.39**	.73**	.76**	.55**	.60**	.76**
Adapts to Human Cognitions	.78**	.43**	.67**	.69**	.52**	.65**	.76**
Predicts Human Cognitions	.73**	.27**	.68**	.76**	.38**	.48**	.75**
Recognizes Human Emotions	.83**	.39**	.77**	.78**	.43**	.50**	.73**
Adapts to Human Emotions	.81**	.34**	.74**	.76**	.37**	.48**	.69**
Predicts Human Emotions	.78**	.35**	.72**	.78**	.41**	.51**	.81**
Friendly	.75**	.46**	.56**	.48**	.42**	.52**	.54**
Helpful	.49**	.47**	.31**	.29**	.59**	.67**	.51**
Caring	.74**	.37**	.61**	.59**	.36**	.48**	.59**
Trustworthy	.49**	.29**	.26**	.24**	.38**	.52**	.41**
Rude	−.45**	−.45**	−.26**	−.17**	−.41**	−.50**	−.37**
Conceited	−.27**	−.36**	−.07**	−.00	−.26**	−.36**	−.21**
Hostile	−.24**	−.39**	−.06*	.02	−.29**	−.36**	−.18**

(Continued)

Table 2. Continued

	Recognizes Human Cognitions	Adapts to Human Cognitions	Predicts Human Cognitions	Recognizes Human Emotions	Adapts to Human Emotions	Predicts Human Emotions	
Recognizes Human Cognitions	1.00						
Adapts to Human Cognitions	.80**	1.00					
Predicts Human Cognitions	.83**	.78**	1.00				
Recognizes Human Emotions	.83**	.83**	.83**	1.00			
Adapts to Human Emotions	.81**	.81**	.80**	.90**	1.00		
Predicts Human Emotions	.83**	.82**	.89**	.88**	.84**	1.00	
Friendly	.60**	.67**	.53**	.67**	.65**	.59**	
Helpful	.488**	.58**	.40**	.45**	.44**	.42**	
Caring	.67**	.73**	.65**	.76**	.76**	.69**	
Trustworthy	.44**	.53**	.38**	.44**	.43**	.39**	
Rude	−.31**	−.47**	−.25**	−.35**	−.36**	−.30**	
Conceited	−.15**	−.31**	−.08**	−.19**	−.22**	−.13**	
Hostile	−.10**	−.26**	−.03	−.13**	−.16**	−.08**	
	Friendly	Helpful	Caring	Trustworthy	Rude	Conceited	Hostile
Friendly	1.00						
Helpful	.68**	1.00					
Caring	.78**	.65**	1.00				
Trustworthy	.65**	.78**	.68**	1.00			
Rude	−.61**	−.74**	−.60**	−.70**	1.00		
Conceited	−.46**	−.60**	−.46**	−.62**	.81**	1.00	
Hostile	−.44**	−.60**	−.40**	−.56**	.84**	.80**	1.00

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

robot more desirable and seem more human-like. The PSI Scales that assess apparent information processing abilities and the apparent ability to identify people had their highest correlations with these same three scales, hinting that these skills may be the core of social intelligence.

In contrast, the PSI social presentation scales had strong correlations with Likability, Welcome, and Appeal. This suggests that the social presentation scales successfully measure whether robots are seen as desirable social partners. The social presentation scales also had strong correlations with Perceived Intelligence, suggesting either that having good social presentation skills is recognized as requiring intelligence or, at least, that the robots that seemed to have strong social presentation skills in this study also happened to seem smart.

Table 3. Correlations of the PSI Scales with the Other Robot Perception Scales

PSI Scale	Social Response to Robots Scale				Godspeed Index		R^2
	Welcome	Appeal	Unobtrusiveness	Naturalness	Likability	Perceived Intelligence	
Long Form Total	.59**	.64**	.33**	.60**	.64**	.71**	.62**
Short Form Total	.57**	.63**	.33**	.59**	.63**	.69**	.59**
Overall							
Social Competence	.44**	.59**	.16**	.61**	.51**	.62**	.54**
Identifying People	*.26*	*.40*	*.05*	*.44*	*.29*	*.42*	
Identifies Humans	.30**	.35**	.09**	.28**	.32**	.39**	.19**
Identifies Individuals	.26**	.45**	.04	.53**	.31**	.46**	.37**
Identifies Social Groups	.21**	.39**	.03	.50**	.24**	.42**	.33**
Information Processing Ability	*.40*	*.49*	*.19*	*.52*	*.41*	*.56*	
Recognizes Human Behaviors	.36**	.40**	.26**	.30**	.33**	.50**	.26**
Adapts to Human Behaviors	.49**	.48**	.38**	.38**	.49**	.61**	.39**
Predicts Human Behaviors	.39**	.48**	.18**	.52**	.37**	.57**	.40**
Recognizes Human Cognitions	.37**	.49**	.19**	.59**	.38**	.58**	.45**
Adapts to Human Cognitions	.49**	.56**	.24**	.55**	.53**	.63**	.49**
Predicts Human Cognitions	.32**	.46**	.12**	.56**	.34**	.50**	.39**
Recognizes Human Emotions	.39**	.54**	.12**	.60**	.44**	.55**	.47**
Adapts to Human Emotions	.40**	.53**	.13**	.59**	.45**	.54**	.45**
Predicts Human Emotions	.35**	.48**	.11**	.58**	.37**	.53**	.65**
Social Presentation	*.59*	*.48*	*.42*	*.33*	*.67*	*.54*	
Friendly	.55**	.58**	.25**	.48**	.64**	.57**	.50**
Helpful	.64**	.51**	.49**	.34**	.68**	.61**	.53**
Caring	.56**	.57**	.30**	.53**	.65**	.59**	.52**
Trustworthy	.64**	.51**	.50**	.37**	.68**	.61**	.53**
Rude	−.67**	−.50**	−.49**	−.30**	−.76**	−.59**	.61**
Conceited	−.55**	−.35**	−.48**	−.16**	−.63**	−.42**	.45**
Hostile	−.51**	−.31**	−.45**	−.13**	−.62**	−.39**	.43**
R^2	.54**	.46**	.40**	.43**	.68**	.59**	—

* $p < .05$. ** $p < .001$. The average absolute correlations for each group of scales is presented in the subsection header row. The statistical significance of these averages was not assessed.

All types of PSI Scales had their smallest correlations with Unobtrusiveness, which was the only scale to have non-significant correlations. Thus, being perceived as intrusive is not strongly related to being perceived as socially intelligent. However, Unobtrusiveness did have moderate relations with social presentation: Being seen as intrusive makes a robot less desirable as a social partner.

To determine if the PSI Scales provide new information beyond the existing robot perception scales, we examined the relations between the PSI Scales and other measures of robot perceptions by regressing the PSI Scales and the other scales on each other. First, we used the 20 PSI Scales to predict the Social Response to Robots Scales and Godspeed Indices. Then, we used the Social

Response to Robots Scales and Godspeed Indices to predict the 22 PSI Scales (the 20 original scales and the Long Form and Short Form totals).

The PSI Scales were able to predict the Social Response to Robots Scales and Godspeed Indices. As shown by the last row of Table 3, the PSI Scales predicted large portions of the variance in all of these scales. The smallest portion was for the Unobtrusiveness Scale, but the PSI Scales still accounted for 40% of its variance (corresponding to a multivariate correlation of .62). Impressively, the PSI Scales accounted for roughly two-thirds of the variance in the Likability Index (corresponding to a multivariate correlation of .82). Thus, the PSI Scales cover much of the content that is included in the Social Response Scales and Godspeed Indices, but these latter scales still provide some unique information.

Similarly, the Social Response to Robots Scales and Godspeed indices were able to predict the PSI Scales. As shown by the last column of Table 3, these scales accounted for roughly half of the variance in the PSI social presentation scales and almost half of the variance in several PSI information processing scales. Impressively, these scales predicted almost two-thirds of the variance in the Predicts Human Emotions Scale (corresponding to a multivariate correlation of .79). However, three PSI Scales were not well-predicted: Identifies Humans (R^2 = .19), Identifies Social Groups (R^2 = .33), and Recognizes Human Behaviors (R^2 = .26). Thus, the PSI Scales capture important abilities that are not covered by the Social Response to Robots Scales and Godspeed Indices.

3.1.4 Exploratory Factor Analysis. To better understand the relations among the robot perception scales, we conducted an exploratory factor analysis using all 26 robot perception scales: the 20 PSI Scales, the four Social Response to Robots Scales, and the two Godspeed Indices. We determined that there were three factors based upon parallel analysis, the MAP test, the scree test, and interpretability. Parallel analysis suggested three factors, the MAP test suggested four, and the scree test suggested either two or five factors. We concluded there were probably three or four factors because these three methods are all typically accurate to within one factor, but parallel analysis and the MAP test are usually the most accurate [Velicer et al. 2000]. We chose to examine the three-, four-, and five-factor solutions because conclusions are more likely to be distorted if researchers use too few factors than if they use too many factors [Fabrigar et al. 1999]. For each factor solution, we selected the optimal rotation based upon the number of complex scales (fewer is better), the number of hyperplanar coefficients (more is better), and the extent of correlation among the factors (smaller is better). We rejected the resulting five-factor solution because it included a factor that overlapped substantially with the other factors (this factor had salient coefficients only for variables that also fell on other factors). Of the remaining solutions, the three-factor was the most interpretable from a substantive standpoint and is therefore shown in Table 4.

We named these three factors Behavior, Mind, and Social Presentation. The Behavior factor included the three scales related to behavior, the ability to identify humans, and the tendency to be helpful. We interpret this factor as the tendency of the robot to coordinate its behaviors with humans in a helpful manner. The Mind factor included each of the six scales related to cognitions and emotions. The Social Presentation factor included each of the scales that were designed to measure social presentation skills, as well as all of the Godspeed Indices (Likability and Perceived Intelligence) and three of the Social Response Scales (Welcome, Unobtrusiveness, and Appeal).

Several noteworthy findings occurred. Identifies Humans fell on the same factor as the three behavior scales, rather than the factor where Identifies Individuals and Identifies Groups fell. Helpfulness had a secondary relation with the Behavior factor. Predicts Human Behavior had its largest relation with the Mind factor, rather than the factor where the Recognizes Human Behaviors and Adapts to Human Behaviors Scales fell. Social Competence fell on the Mind factor. Finally, Naturalness fell on the Mind factor. We will explore these findings in the Discussion section.

Table 4. Exploratory Factor Analysis of the 26 Robot Perception Scales

Scale	Factor			h^2
	1	2	3	
PSI Recognizes Human Behaviors	**.85**	.13	.06	.85
PSI Identifies Humans	**.77**	.08	.08	.69
PSI Adapts to Human Behaviors	**.64**	.25	.24	.75
PSI Recognizes Human Emotions	.03	**.92**	.05	.89
PSI Predicts Human Emotions	.06	**.92**	−.03	.87
PSI Predicts Human Cognitions	.01	**.91**	−.04	.81
PSI Adapts to Human Emotions	−.04	**.91**	.10	.85
PSI Identifies Social Groups	.11	**.87**	−.19	.77
PSI Recognizes Human Cognitions	.19	**.83**	−.01	.83
PSI Identifies Individuals	.14	**.81**	−.11	.71
PSI Socially Competent	.08	**.80**	.17	.80
PSI Adapts to Human Cognitions	.17	**.75**	.19	.81
SRRS Naturalness	−.10	**.67**	.19	.50
PSI Caring	−.08	**.67**	**.46**	.78
PSI Predicts Human Behaviors	**.40**	**.65**	.01	.76
PSI Friendly	.07	**.53**	**.47**	.68
GI Likability	−.10	.24	**.84**	.82
PSI Rude	−.18	−.02	**−.83**	.82
PSI Conceited	−.12	.16	**−.83**	.71
PSI Hostile	−.21	.24	**−.81**	.72
SRRS Welcome	−.07	.21	**.79**	.72
PSI Trustworthy	.03	.22	**.74**	.70
SRRS Unobtrusiveness	−.03	−.07	**.70**	.46
PSI Helpful	**.34**	.15	**.65**	.77
SRRS Appeal	−.07	**.47**	**.52**	.59
GI Perceived Intelligence	.11	**.43**	**.52**	.65
Factor 1	1.00	.33	.30	
Factor 2		1.00	.27	
Factor 3			1.00	

Salient pattern matrix coefficients are in boldface. h^2 = communality. PSI = Perceived Social Intelligence. SRRS = Social Response to Robots Scale. GI = Godspeed Indices. Factor 1 = Behavior. Factor 2 = Mind. Factor 3 = Social Presentation.

The four-factor solution was similar to the three-factor solution. The first two factors remained the same: They measured Social Information Processing Abilities and Coordinating Behaviors with Humans. The third factor divided into two: one for the PSI Scales (which used Likert-type items) and one for the Social Response to Robots Scales and Godspeed Indices (which used semantic differential scales)—though the Likability Index was related to both of these factors. These two factors had their highest correlations with each other. Given that the four-factor solution did not make substantive distinctions between the scales (only methodological ones), we prefer the three-factor solution.

3.1.5 How People Feel About the Robots. To determine if the 28 robot perception scales (i.e., the 20 individual PSI Scales, the two PSI total scores, the four Social Response to Robots Scales, and the two Godspeed Indices) are related to how respondents feel about robots, we correlated these scales with desire for social interaction with the robots and with positive and negative feelings about the robots. See Table 5.

Every one of the 26 robot perception scales correlated significantly with Desire for Social Interaction and Positive Feelings. The correlations were highest for the Social Response Scales, Godspeed Indices, and PSI social presentation scales. For example, respondents were more likely to want to interact with the robot and feel positively about the robot if they perceived it as appealing, likable, and caring. The Social Response to Robots Scales and Godspeed Indices tended to have larger correlations with Desire for Social Interaction and Positive Feelings than the PSI scales—even than the PSI social presentation scales.

Most of the robot perception scales also correlated significantly with Negative Feelings about the robot. Once again, the correlations were highest for the Social Response Scales, Godspeed Indices, and PSI social presentation scales; however, the correlations were smaller than they had been for the Desire for Social Interaction and Positive Feelings Scales. Respondents were most likely to have negative feelings if they perceived the robot as hostile, conceited, rude, and not likable. Overall social competence, information processing ability, and the ability to identify people had only small and often non-significant relations with Negative Feelings.

3.1.6 Relation with Overall Social Competence. To determine whether each of the 26 robot perception scales can be considered a measure of perceived social competence, we correlated the PSI Social Competence Scale with each of the remaining PSI Scales and with the four Social Response to Robots Scales and the two Godspeed Indices. See the last column of Table 5. Every one of the 26 robot perception scales had a statistically significant correlation with Social Competence. The PSI information processing abilities and identifying people abilities had the largest correlations. Social Competence was most closely related to the apparent ability to deal effectively with human emotions and cognitions, to identify individuals, and to appear friendly and caring. Interestingly, being rude, conceited, and hostile had negative correlations with Social Competence, as might be expected, but these correlations were smaller than the correlations for the positive social presentation skills (appearing friendly, caring, helpful, and trustworthy). This suggests that rudeness may often be seen as a matter of incompetence rather than poor intentions: It takes both good intentions and skill to be a competent social partner.

The Social Response to Robots Scales and Godspeed Indices typically had moderate correlations with Social Competence, indicating that they are measuring related constructs but not social intelligence itself. The one exception was Unobtrusiveness, which had a low correlation with Social Competence.

3.1.7 Distinguishing Between Robots. The purpose of the PSI Scales is to allow researchers to evaluate the social intelligence of robots, allowing researchers to pinpoint the effects of different robot bodies, behaviors, and contexts. Therefore, to determine if the robot perception scales can distinguish between robots who, based upon their actions, seem to vary in their social intelligence, we used multi-level modeling to compare the average scale scores for the robots in the five videos, after controlling for differences across participants. Specifically, we used the *lme4* package [Bates et al. 2012] in R to complete linear mixed effects analyses with random intercepts for each participant (as recommended by Dixon [2008]), using maximum likelihood estimation (as recommended by Luke [2016]). Both the Short Form and Long Form total scores, every individual PSI Scale, every Social Response to Robots Scale, and both Godspeed Indices were able to make distinctions between the five robot videos. See Table 6.

Table 5. Correlations of the Robot Perception Scales with the Criterion Variables

Robot Perception Scale	Desire for Social Interaction	Positive Feelings	Negative Feelings	Social Competence
PSI Scale				
Long Form Total	.55**	.47**	−.10**	—
Short Form Total	.56**	.46**	−.10**	—
Overall				
Social Competence	.42**	.43**	.01	—
Identifying People	*.27*	*.31*	*.10*	*.63*
Identifies Humans	.24**	.26**	−.10**	.46**
Identifies Individuals	.28**	.33**	.09**	.74**
Identifies Social Groups	.29**	.34**	.12**	.69**
Information Processing Ability	*.41*	*.39*	*.05*	*.71*
Recognizes Human Behaviors	.31**	.28**	−.06*	.45**
Adapts to Human Behaviors	.42**	.35**	−.12**	.53**
Predicts Human Behaviors	.41**	.40**	.02	.69**
Recognizes Human Cognitions	.40**	.39**	.07*	.77**
Adapts to Human Cognitions	.47**	.42**	−.04	.78**
Predicts Human Cognitions	.41**	.38**	.07*	.73**
Recognizes Human Emotions	.42**	.43**	.04	.83**
Adapts to Human Emotions	.43**	.42**	.00	.80**
Predicts Human Emotions	.40**	.41**	.06*	.78**
Social Presentation	*.48*	*.32*	*.27*	*.49*
Friendly	.49**	.41**	−.11**	.75**
Helpful	.55**	.36**	−.22**	.49**
Caring	.55**	.43**	−.11**	.74**
Trustworthy	.56**	.34**	−.24**	.49**
Rude	−.46**	−.29**	.34**	−.45**
Conceited	−.39**	−.19**	.39**	−.27**
Hostile	−.36**	−.20**	.45**	−.24**
Social Response Scale	*.62*	*.45*	*.27*	*.44*
Welcome	.62**	.45**	−.27**	.44**
Appeal	.65**	.61**	−.10**	.59**
Unobtrusiveness	.43**	.22**	−.23**	.16**
Naturalness	.48**	.49**	.10**	.61**
Godspeed Index	*.56*	*.43*	*.23*	*.57*
Likability	.58**	.40**	−.30**	.51**
Perceived Intelligence	.54**	.46**	−.15**	.62**

* $p < .05$. ** $p < .001$. The average absolute correlations for each group of scales is presented in the subsection header row. The statistical significance of these averages was not assessed.

Table 6. Mean Scores on the Robot Perception Scales

Scale	Dragonbot Stories	Robovie Aquarium	PR2 Stacking	NAO Stealing	Ottoman Feet up	$\chi^2(4)$
PSI Scale						
Short Form Total	74.62^{b}	75.09^{b}	63.38^{a}	62.86^{a}	60.90^{a}	420.41**
Long Form Total	298.08^{c}	295.76^{c}	253.45^{b}	250.95^{b}	242.77^{a}	406.54**
Social Competence	15.04^{c}	14.68^{c}	10.36^{a}	12.24^{b}	9.77^{a}	536.88**
Identifies Humans	17.22^{c}	17.17^{c}	16.38^{b}	15.87^{a}	16.18^{ab}	111.60**
Identifies Individuals	14.55^{d}	16.08^{e}	10.32^{b}	13.78^{c}	8.92^{a}	721.75**
Identifies Social Groups	11.51^{b}	13.35^{d}	9.70^{a}	12.47^{c}	9.33^{a}	409.98**
Recognizes Human Behavior	16.04^{c}	15.66^{bc}	16.81^{d}	15.24^{b}	14.49^{a}	157.12**
Adapts to Human Behavior	15.18^{bc}	14.57^{b}	15.78^{c}	13.13^{a}	13.01^{a}	195.85**
Predicts Human Behaviors	13.37^{c}	13.56^{c}	12.29^{b}	12.06^{b}	11.25^{a}	118.55**
Recognizes Human Cognitions	12.63^{c}	13.28^{c}	11.18^{b}	11.65^{b}	8.94^{a}	382.67**
Adapts to Human Cognitions	13.63^{c}	14.01^{c}	11.26^{ab}	11.59^{b}	10.66^{a}	315.63**
Predicts Human Cognitions	11.37^{c}	12.03^{d}	8.85^{a}	10.32^{b}	8.31^{a}	344.53**
Recognizes Human Emotions	13.41^{c}	13.79^{c}	9.18^{a}	11.82^{b}	9.46^{a}	489.90**
Adapts to Human Emotions	12.76^{c}	12.93^{c}	8.81^{a}	11.10^{b}	8.68^{a}	504.16**
Predicts Human Emotions	12.36^{c}	12.70^{c}	9.44^{a}	11.10^{b}	9.02^{a}	339.82**
Friendly	16.24^{d}	14.92^{c}	12.15^{a}	12.72^{b}	12.43^{ab}	465.64**
Helpful	16.50^{d}	14.52^{b}	15.29^{c}	11.68^{a}	13.95^{b}	410.11**
Caring	14.14^{b}	14.30^{b}	10.12^{a}	10.43^{a}	10.45^{a}	512.28**
Trustworthy	15.33^{d}	12.85^{c}	13.25^{c}	9.26^{a}	12.14^{b}	538.02**
Rude	5.96^{a}	6.54^{b}	7.81^{c}	10.77^{e}	8.51^{d}	540.02**
Conceited	6.77^{a}	7.77^{b}	7.97^{bc}	10.97^{d}	8.39^{c}	430.42**
Hostile	5.38^{a}	5.85^{ab}	6.21^{bc}	9.12^{d}	6.52^{c}	424.92**
Social Response to Robots Scale						
Welcome	22.78^{d}	20.36^{c}	19.18^{b}	16.21^{a}	18.77^{b}	316.53**
Appeal	21.66^{c}	19.84^{b}	16.53^{a}	17.36^{a}	17.17^{a}	263.92**
Unobtrusiveness	17.78^{c}	17.08^{c}	19.53^{d}	14.89^{a}	15.85^{b}	212.48**
Naturalness	15.95^{e}	14.88^{d}	11.38^{b}	13.08^{c}	10.03^{a}	389.67**
Godspeed Index						
Likability	18.54^{d}	17.44^{c}	14.81^{b}	12.37^{a}	15.05^{b}	510.06**
Perceived Intelligence	20.08^{c}	19.94^{c}	18.97^{b}	15.33^{a}	15.33^{a}	428.73**

** $p < .001$. Superscripts that are different denote means that are significantly different ($p < .05$).

The two robots that spoke (Robovie and Dragonbot) obtained the best score on most of the robot perception scales. The two remaining humanoid robots (PR2 and NAO) obtained the next best scores, with the robotic ottoman obtaining the worst scores on most scales. There were a few exceptions to this general pattern. First, the PR2 robot that cooperatively stacked blocks obtained the highest scores for recognizing and adapting to human behavior and for being unobtrusive. Second, the robotic ottoman was rated as more helpful, trustworthy, welcome, likable, and unobtrusive than the NAO that shoplifted and stole from a friend.

3.2 Evaluating the Quality of the PSI Short Form

To evaluate the quality of the 20-item PSI Short Form, we completed four analyses. First, to evaluate whether the Short Form is an adequate substitute for total scores from the complete set of PSI

Scales, we correlated total scores from the Short Form with total scores over all 80 items. It had a nearly perfect correlation ($r = .98$, $p < .001$). Second, to evaluate whether the Short Form total scores can be interpreted as an overall measure of social intelligence, we correlated its total scores with social competence (calculated as the sum of the three Social Competence items that were not included on the Short Form). This correlation was very strong ($r = .82$, $p < .001$). Third, to evaluate internal consistency reliability, we calculated coefficient alpha. It was excellent (coefficient alpha = .93). Finally, to determine if the Short Form could distinguish robots that appear more socially intelligent from robots that appear less socially intelligent, we compared the Short Form scores across the five videos included in this study. Mirroring the results for the Long Form, the dishonest Robovie and the story-telling Dragonbot had the highest perceived social intelligence, the block-stacking PR2 and the shoplifting NAO had the next highest, and the robotic ottoman had the lowest perceived social intelligence.

4 DISCUSSION

4.1 The Current Study

4.1.1 Scale Design. Measuring social intelligence in robots is critical to research on HRI and thus to the integration of robots in public, private, and work settings. However, no robot social intelligence measures exist. Existing measures of people's perceptions of robot abilities do not target social intelligence specifically. Measures of human social intelligence examine skills that are too advanced to be relevant to current or near-future robots and assume that the target being rated has humanoid features and functions. Therefore, we created 20 new scales to measure a wide range of social intelligence abilities in robots.

We sought to create cohesive scales that could reliably discriminate the perceived social ability of different robots. For each scale, we therefore drafted items that would be applicable to many different kinds of robots in many different contexts. We selected items that had high factor analysis coefficients (indicating that they are measuring a coherent construct) and good information (indicating that they can distinguish raters who perceive a particular robot as socially intelligent from raters who perceive a particular robot as not socially intelligent). Each of the resulting 20 four-item PSI Scales has good internal consistency and a single factor, thus indicating that each scale measures a single coherent concept.

The 20 PSI Scales were related to each other in meaningful ways. Scales that measured conceptually related content had strong positive correlations, and being perceived as rude, conceited, and hostile had negative correlations with most other scales. However, many inter-scale correlations were small, demonstrating that these scales are measuring several distinct constructs.

4.1.2 Similarities Between the Robot Perception Scales. This study included five robots that acted in a variety of ways in a variety of contexts to demonstrate a wide range of social intelligence. Importantly, the PSI Scales, Social Response to Robot Scales, and Godspeed Indices tended to rank the robots in the five videos in the same way: The robots that spoke obtained the highest scores and the robotic ottoman obtained the lowest scores. The exceptions to this general trend reinforce the validity of all these scales. For example, the PR2 robot had the highest scores on the Adapting to Human Behavior Scale, which makes sense given that it changed the blocks it was using after its human partner left to talk on the phone, and the NAO robot had the highest scores on the Rude, Conceited, and Hostile Scales, and the lowest scores on the Welcome and Likability Scales, which makes sense given that it was the only robot to engage in criminal activity. Thus, respondents appear to be making fine distinctions between the robots in the videos based upon the specific qualities they are being asked to evaluate, rather than responding to all items based upon their global impressions of each robot. These findings support the quality and usefulness of all the scales.

The PSI Scales, Social Response to Robots Scales, and Godspeed Indices all had significant relations with overall social intelligence and were able to predict positive and negative feelings about the robots and desire for social interaction with the robot. The convergence of these results supports the quality of all three sets of scales. Moreover, these findings tell us that people like socially intelligent robots and want to spend more time with them.

Furthermore, each of the PSI Scales could be predicted from the Social Response to Robots Scales and Godspeed Indices, and each of the Social Response to Robots Scales and Godspeed Indices could be predicted from the PSI Scales. Some of these relations were very strong. For example, the PSI Scales predicted roughly two-thirds of the variance in the Likability Index (corresponding to a multivariate correlation of .82, which represents a large portion of its reliable variance). The PSI social presentation scales, in particular, had substantial overlap with the Social Response to Robots Scales and Godspeed Indices. For example, robots were considered more likable and welcome if they were less rude and more helpful. Furthermore, the PSI social presentation scales, Godspeed Indices, and almost all of the Social Response to Robots Scales grouped together on a single factor. We labeled this factor Social Presentation and consider it to be an important part of social intelligence in robots. Social presentation is the ability of the robot to appear to be a desirable social partner, a valuable skill that is assessed by all three sets of scales.

The overlap with the Social Response to Robots Scales and Godspeed Indices was not limited to the PSI social presentation scales. The Social Response to Robots Scales and Godspeed Indices were able to predict roughly two-thirds of the variance in the Predicts Human Emotions Scale (corresponding to a multivariate correlation of .79, which accounts for a large portion of its reliable variance). The largest correlations with Predicts Human Emotions were for the Naturalness Scale and Perceived Intelligence Index. Thus, robots that appear to predict human emotions seem the most human-like and the most intelligent. Of the skills we assessed in this study, predicting emotions is likely perceived as one of the most challenging and thus may be the one that most closely marks whether a robot was acting like a human. In summary, the large meaningful relations between the PSI Scales, Social Response to Robots Scales, and Godspeed Indices support the validity and usefulness of all these scales.

4.1.3 Differences Between the Robot Perception Scales. Despite their similarities, each of the three sets of scales also provide unique information. The PSI social information processing scales and identifying people scales capture the core of social intelligence. These scales had large correlations with overall social competence and had only small to moderate correlations with positive feelings and desire for social interaction, indicating that they are measuring perceived *skills*, not just the tendency to make a good impression. These scales formed two factors. The Behavior factor focuses on the recognition of, adaptation to, and prediction of behavior, the ability to identify humans, and the tendency to be helpful. We interpret this factor as the robot's tendency to appear to helpfully coordinate its behaviors with humans. The Mind factor focuses on social information processing skills related to cognitions and emotions, the ability to identify individuals and groups, and overall social competence. We interpret this factor as measuring the robot's apparent ability to understand and interact with people's minds. These two factors, Behavior and Mind, are clearly in line with our original definition of social intelligence: the ability to interact effectively with others to accomplish one's goals [Ford and Tisak 1983].

Two PSI social presentation scales—Friendly and Caring—acted similarly to the PSI social information processing scales. They had some of the highest correlations with overall social competence and fell on the Mind factor. Perhaps social intelligence is a prerequisite for being perceived as friendly and caring: If a robot has poor social intelligence, people cannot tell if it has good intentions.

The remaining PSI social presentation scales, the Social Response to Robots Scales, and the Godspeed Indices appear to measure constructs that are related to social intelligence, but not social *intelligence* itself. They formed a separate factor (labelled Social Presentation) and had only moderate correlations with social competence. Thus, being welcoming and likable are somewhat separate from interacting effectively with people. These scales had high correlations with positive feelings about the robot and desire for social interaction, suggesting that social presentation may be critical to adoption and use of robots. However, these high correlations also suggest that these scales are strongly influenced by overall positive impressions. This conclusion is in line with Ho and MacDorman's [2010] analysis of the Godspeed indices (of which Likability and Perceived Intelligence were used in our study): The Likability Index correlated highly with the Anthropomorphism (.73), Animacy (.74), and Perceived Intelligence Indices (.71), suggesting that these indices do not distinguish between the specific constructs of interest and overall positive appraisal.

The 26 robot perception scales were not able to predict negative feelings as well as they were able to predict positive feelings and desire for social interaction. Most of the robot perception scales assess positive qualities that make people enjoy the robots and want to spend time with them. Only the PSI Rude, Conceited, and Hostile Scales focus on negative qualities that make people dislike robots. The ability to measure these negative characteristics is thus a unique strength of the PSI Scales.

The factor analysis results demonstrate that the PSI Scales, Social Response to Robots Scales, and Godspeed Indices each provide unique information. This conclusion is reinforced by the fact that the PSI Scales were often able to predict less than 50% of the variance in the Social Response to Robots Scales and Godspeed Indices, which in turn were able to account for less than one-third of the variance of some of the PSI Scales (such as the Identifies Humans, Identifies Social Groups, and Recognizes Human Behaviors Scales).

The PSI Scales were least able to predict the Social Response to Robots Unobtrusiveness Scale. In addition, Unobtrusiveness had only a small correlation with overall social competence ($r = .16$) and was the only scale to have non-significant (and near zero) correlations with some PSI Scales. This scale is measuring something different from the other scales we examined.

Unobtrusiveness may have been hard to predict because of our study design. In our five videos, the robots almost always interacted with humans in a cooperative manner. There was very little intrusive behavior that might be perceived as annoying and unhelpful. However, in natural settings, intrusive and annoying behavior may be much more common. If so, the Unobtrusiveness Scale may have a stronger relation with overall social intelligence. Furthermore, robots with poor social skills may be perceived as interrupting people and interfering with their work. This may be particularly true of task-focused robots working around humans if they do not adjust their behavior for different workflows and physical environments (e.g., Mutlu and Forlizzi [2008]). The Unobtrusiveness Scale is currently the only scale that measures this key perception, making this an important, unique contribution of the Social Response to Robots Scale.

In summary, the PSI Scales, Social Response to Robots Scales, and Godspeed Indices have strong relations with each other and overall social competence, but each set of scales also captures unique perceptions about robots. Had we included all of the scales from the Social Response to Robots Scales and Godspeed Indices, further unique contributions would doubtless have been found. The area of perceptions of robots is broad, and all three sets of scales are helpful in capturing some parts of this domain. Researchers should use whichever scales best capture the parts of the domain in which they are most interested.

4.1.4 What We Learned About the Social Intelligence of Robots. Examining the factor analytic relations between the 26 robot perception scales provides five insights into the nature of robot

social intelligence. First, the Behavior factor included a secondary relation for the Helpfulness Scale. Thus, when the five robots in this study coordinated their behaviors with humans, this coordination was primarily helpful. This makes sense. While it is possible for a robot (or human) to carefully time and execute its behaviors to hurt others (e.g., sticking out a foot to trip someone as they walk by), most behavioral coordination assists in the completion of cooperative endeavors (e.g., stacking blocks together, sharing stories with each other, or just passing each other in a hallway without colliding).

Second, the Predicts Human Behaviors Scale had its largest relation with the Mind factor, not the Behavior factor. There is a maxim that the best predictor of future behavior is past behavior [Meehl 1986]. However, this maxim only appears to hold if the future behavior is the same as the past behavior: Through a sort of inertia (or momentum), people keep doing the same things they used to do. To predict that someone will do something *different* from what they are currently doing, we need to understand their thoughts and feelings, their goals and desires.

Third, the Identifies Humans, Identifies Individuals, and Identifies Groups Scales did not all fall on the same factor. The Identifies Humans Scale fell on the Behavior factor. This suggests that the best way to identify humans is based on how they act. People walk, talk, open and close doors, and type on computers. Most office chairs do not. This explains why the Turing [1950] test is so difficult: It is people's behaviors (rather than their ideas) that primarily mark them as human. In contrast, the Identifies Individuals Scale fell on the Mind factor, suggesting that the best way to identify *specific* people is based on their feelings and cognitions: While most people talk, different people say different things depending upon what they think and feel. Similarly, the Identifies Groups Scale fell on the Mind factor, suggesting that the best way to identify whether people are part of a social group is their cognitions and emotions: What usually joins people into a group is not that they all walk or all talk or all type on a computer, but that they share a political philosophy or like the same music.

Fourth, the Social Response to Robots Naturalness Scale was the only non-PSI scale that did not have its strongest relation with the Social Presentation factor. Robots can be likable without being natural. Instead, this scale fell on the Mind factor. Participants rated the robots as more natural, human-like, conscious, lifelike, and animate if the robot seemed to be able to understand people's thoughts and emotions.

Fifth, the PSI Social Competence Scale fell on the Mind factor. This suggests that overall social competence is most closely related to the ability to interact smoothly with people's minds: knowing and responding to what people think and feel. The Social Competence Scale did not fall on the Behavior factor and thus is less related to coordinating ones' physical actions with others' (which might be related more to a robot's physical grace than its intellectual abilities). Finally, the Social Competence Scale did not fall on the Social Presentation factor, which suggests that a robot can be socially competent without being a particularly desirable social partner. For example, consider the shoplifting NAO robot: It was rated as more socially competent than two of the other robots, but less trustworthy and more rude, conceited, and hostile than any other robot.

4.1.5 Evaluating the PSI Short Form. In addition to creating four-item scales to measure each of the 20 PSI constructs, we also created a short-form that includes the single best item from each of those 20 scales. This PSI Short Form had excellent psychometric properties. It had a nearly perfect correlation with the full-length form, a strong correlation with overall social competence, and excellent internal consistency. It also successfully distinguished between the overall social intelligence of the robots in the five videos, ordering the robots from highest to lowest in the identical order as the Long Form. We conclude that the 20-item Short Form provides an efficient and inclusive snapshot of overall social intelligence.

4.1.6 Limitations. Because of possible cultural differences in perceptions of social intelligence, this study was deliberately limited to participants from a single country, the United States. This limits our ability to generalize from the present findings. For example, Kamide and Arai [2017] found that participants in Japan are more comfortable with conventional (i.e., non-anthropomorphic) robots, while participants in the United States are more comfortable with robots that act like humans [Kamide and Arai 2017]. Analogous differences may occur for perceived social intelligence. For example, in our study, robots were considered more desirable social partners (and aroused more positive feelings) if they seemed more natural and human-like, but this finding may not hold in all other countries. Therefore, future research should explore how perceptions of robots vary between and within groups and contexts.

Additionally, there were five limitations with our study design that could usefully be corrected in future work. First, each video showed a unique robot in a unique context engaged in a unique narrative. This created a wide range of perceived social intelligence, but does not allow us to determine if it was the robots' appearance, context, or behavior that most influenced its rated social intelligence. To separate the effects of robot appearance, context, and behavior, future research would need to vary these independently. Second, the five videos were shown in a fixed order for all participants. If comparisons among stimuli are the primary interest, future research should randomize the order of the videos to eliminate potential order effects. Third, the PSI items were presented in a unique order for each of the five videos. If comparisons among items or scales are of primary interest, future research should ideally randomize the order of the items for each participant to eliminate order effects, or if comparisons of items and scales are not of particular interest, future research can keep the order the same for all participants, so that order effects are the same for every stimuli being rated. Fourth, in our study, the robot perception scales were not able to predict negative feelings as well as they were able to predict positive feelings; this may have been because the robot videos selected for our study did not create many negative feelings among participants. If future researchers are particularly interested in negative feelings, they should ensure that their stimuli create a wide range of negative feelings in their participants. Finally, asking participants about their demographics (e.g., sex and age) at the beginning of the study could have primed participants to answer according to stereotypes related to their demographics. Future research may benefit from asking demographic questions at the end of the study.

4.2 Future Research

Researchers can examine a wide variety of research questions using our scales. Researchers can examine how robot behaviors, robot bodies, and context influence perceived social intelligence. For example, how close should a robot approach a person with Alzheimer's disease to be considered friendly and trustworthy? If they act the same, is a fuzzy purple dragon perceived as more trustworthy than a shiny silver dog? When a hospital delivery robot announces its arrival, how do workflow and physical environment influence perceived rudeness [Mutlu and Forlizzi 2008]? Moreover, how do all of the above factors interact? Perhaps verbal announcements are considered ruder than screen displays in noisy environments, but friendlier in quite environments.

Future research could also explore how people make their judgments. For example, if a robot greets a person by name, perhaps people will usually assume the robot recognizes them, remembers their previous interactions, and likes them. Similarly, if a robot uses appropriate eye contact, perhaps people will assume the robot is friendly and caring. Thus, certain key robot behaviors might lead people to attribute a wide range of skills and characteristics to the robot. One method of determining how people make their judgments would be to ask them directly. Another method would be to vary robot behaviors or bodies (or the information about robot behaviors and

bodies that people have access to) to determine how perceptions of social intelligence change. For example, when a NAO robot demonstrated that it understood that an experimenter had a false belief (justifying improved scores on Recognizing Human Cognition and Social Competence), participants also rated it higher on Adapts to Human Cognitions, Predicts Human Behavior, Predicts Human Behavior, and Identifies Individuals [Sturgeon et al. 2019].

Researchers can also examine relations with other variables. How does perceived social intelligence influence goal completion? (E.g., Does robot trustworthiness help a child learn vocabulary?) How is perceived social intelligence related to other cognitive and personality variables? (E.g., If a self-driving car is perceived (perhaps incorrectly) as being able to recognize and predict human behaviors, will it be considered morally responsible for any accidents that occur?) Do these relations vary across countries or groups?

Finally, perceived social intelligence may be equally important for artificial intelligence programs (e.g., personal assistants, navigation systems, and voice-activated searches). Fortunately, we designed our items so that they do not assume the target has any particular form (e.g., a head, arms, or legs) or functions (e.g., hearing, speech, or motion detection). Therefore, researchers and developers are encouraged to consider whether these concepts and scales would be helpful for targets besides robots.

5 CONCLUSION

Robots are increasingly present in public, private, and occupational settings, and thus need to be able to interact smoothly with people. True robotic social intelligence is still a long way off. Nonetheless, to study and enhance HRI, roboticists need a method of measuring the extent to which robots are perceived as socially intelligent. The PSI Scales provide just such a measurement tool. For example, the PSI Scales have been used to evaluate robots executing social behaviors pertaining to navigation. A new socially aware navigation planner has been created (SAN [Banisetty et al. 2019]) that takes into consideration interpersonal distance with humans and social norms such as going to the end of a line of people and forming a circle when joining a group. Prior work on SAN used performance metrics such as path efficiency and distance travelled [Sebastian et al. 2017], trajectory differences [Ramírez et al. 2016], and number of proxemic intrusions [Helbing et al. 2002]. However, these performance metrics only address the motion that a robot takes: People's perceptions of the robot's navigation behavior were not examined. The PSI Scales bridge that gap. A preliminary study using the PSI Scales shows that a robot with traditional navigation is perceived as less socially intelligent than a robot with SAN in *waiting in a queue* and *group formation* scenarios [Honour et al. 2019].

Although the PSI Scales need to be examined in other cultures and with different stimuli, this study provides strong initial evidence that each scale measures a coherent construct and that the 20 scales are distinct from each other. If researchers need a comprehensive measure of perceived social intelligence and also need to examine individual aspects of social intelligence, they can use the full set of 20 four-item scales. Alternatively, researchers can select individual scales or use the 20-item short form to get brief, comprehensive assessment of people's perceptions. The PSI items were designed to be versatile. They apply to a wide range of robot roles, behaviors, and contexts. They also apply to a wide range of robotic embodiments and could be applied to non-embodied artificial intelligence.

Given the variety of research questions and study designs for which the PSI items could be used, we encourage researchers to use whichever PSI items are relevant to their own research goals. Most researchers will need only a few scales for any particular project. Moreover, they might only need a few items from those scales. To allow researchers to be explicit about which items and scales they used in their research, we hereby give test users permission to reproduce

items from the PSI Scales in scientific publications and other venues. The PSI Scales are located on the International Personality Item Pool website (https://ipip.ori.org/newMultipleconstructs.htm) and in the Appendix. These items can be freely used and adapted as needed.

A APPENDIX

A.1 Perceived Social Intelligence Scales

Each PSI Scale contains four items. If multiple scales are being used, intermix the items. To calculate the total score for one of these 20 PSI Scales, first reverse score any items that have an R after the item number (e.g., reverse score = 6—original score), then average the four item scores.

For each of the 20 PSI Scales, the single best item is given first. The PSI Short Form consists of these 20 items. To calculate the total score, first reverse score the items for the Rude, Conceited, and Hostile scales, then sum the 20 item scores.

Social Information Processing

Social Competence (SOC) – The robot appears to have strong social skills.
This robot…

1 is socially competent
2 is socially aware
3R is socially clueless
4 has strong social skills

Identifies Humans (IH) – The robot appears to detect human presence.
This robot…

1 notices human presence
2R mistakes humans for inanimate objects
3 knows when a human is nearby
4R fails to notice when humans are around

Identifies Individuals (II) – The robot appears to identify and recognize people as individuals.
This robot…

1 recognizes individual people
2 remembers who people are
3R cannot tell people apart
4 remembers its shared history with each person

Identifies Social Groups (IG) – The robot appears to discern which people are with each other.
This robot…

1 knows if someone is part of a social group
2 knows which people are together
3R ignores the fact that people are together
4 figures out which people know each other

Recognizes Human Behaviors (RB) – The robot appears to detect people's behaviors.
This robot…

1 notices when people do things
2 detects human movement
3 can figure out what people are doing
4 notices when people try to interact with it

Adapts to Human Behaviors (AB) – The robot appears to adapt its behavior appropriately based upon people's behaviors.
This robot…
1 adapts effectively to different things people do
2 appropriately changes what it is doing based on what others around it are doing
3 knows how to react to what people do
4 adapts its behavior based upon what others do

Predicts Human Behaviors (PB) – The robot appears to anticipate people's behavior.
This robot…
1 anticipates people's behavior
2 predicts human movements accurately
3R has no idea what people are going to do
4 knows how people will react to things it does

Recognizes Human Cognitions (RC) – The robot appears to detect people's thoughts and beliefs.
This robot…
1 can figure out what people think
2 knows when people are missing information
3 can figure out what people can see
4 understands others' perspectives

Adapts to Human Cognitions (AC) – The robot appears to adapt its behavior appropriately based upon people's thoughts and beliefs.
This robot…
1 adapts its behavior based upon what people around it know
2R ignores what people are thinking
3 selects appropriate actions once it knows what others think
4 knows what to do when people are confused

Predicts Human Cognitions (PC) – The robot appears to anticipate people's thoughts and beliefs.
This robot…
1 anticipates others' beliefs
2 figures out what people will believe in the future
3 knows ahead of time what people will think about certain situations
4 anticipates what people will think

Recognizes Human Emotions (RE) – The robot appears to detect people's emotions.
This robot…
1 recognizes human emotions
2R has trouble understanding what people are feeling
3 notices people's emotional reactions
4 knows what people like

Adapts to Human Emotions (AE) – The robot appears to adapt its behavior appropriately based upon people's emotions.
This robot…
1 responds appropriately to human emotion
2 knows what to do when a person is emotional
3R acts the same regardless of how people feel
4 is good at responding to emotional people

Predicts Human Emotions (PE) – The robot appears to anticipate people's emotions.
This robot…
1 anticipates others' emotions
2R has no idea how people will feel in different situations
3 knows ahead of time how people will feel about its actions
4 knows what people are going to want in different situations

Social Presentation
Friendly (FRD) – The robot appears to enjoy social interactions.
This robot…
1 enjoys meeting people
2 likes spending time with people
3 is sociable
4R prefers being alone

Helpful (HLP) – The robot appears willing to assist in tasks.
This robot…
1 tries to be helpful
2 is cooperative
3 values cooperation over competition
4 wants to help people

Caring (CAR) – The robot appears to care about the well-being of others.
This robot…
1 cares about others
2 is compassionate
3 feels concern for people who are in distress
4R feels little concern for others

Trustworthy (TRU) – The robot appears deserving of trust.
This robot…
1 is trustworthy
2 is honest
3 is sincere
4 is ethical

Rude (RUD) R – The robot appears rude and disrespectful.
This robot…
1 is impolite
2 is rude
3R is respectful
4R is nice to people

Conceited (CON) R – The robot appears overly proud of itself or its abilities.
This robot…
1 thinks it is better than everyone else
2 is self-centered
3 is condescending
4R is modest

Hostile (HST) R – The robot appears antagonistic and violent.
This robot…
1 tries to hurt people
2 is violent
3R is peaceful
4 is mean to people

REFERENCES

M. Agran, C. Hughes, C. A. Thoma, and L. A. Scott. 2016. *Social Skills Survey*. PsycTests. DOI: http://dx.doi.org/10.1037/t50828-000

K. Albrecht. 2004. *Social Intelligence: The New Science of Success*. Pfeiffer, San Francisco, CA.

K. A. Barchard, L. Lapping-Carr, R. S. Westfall, S. B. Banisetty, and D. Feil-Seifer 2018. *Perceived Social Intelligence (PSI) Scales test manual*. Unpublished psychological test and test manual. Observer report of 20 aspects of social intelligence of robots, with four items per scale. Retrieved from https://ipip.ori.org/newMultipleconstructs.htm.

S. B. Banisetty, S. Forer, L. Yliniemi, M. Nicolescu, and D. Feil-Seifer. 2019. Socially-aware navigation: A non-linear multi-objective optimization approach. arXiv preprint arXiv:1911.04037.

C. Bartneck, D. Kulic, E. Croft, and S. Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robots* 1 (2009), 71–81. DOI: http://dx.doi.org/10.1007/s12369-008-0001-3

D. M. Bates, M. Maechler, and B. Bolker. 2012. lme4: Linear mixed-effects models using S4 classes. *R package version* 0.999999-0. Retrieved from https://cran.r-project.org/package=lme4.

D. Brščić, H. Kidokoro, Y. Suehiro, and T. Kanda. 2015. Escaping from children's abuse of social robots. In *Proceedings of ACM/IEEE International Conference on Human-robot Interaction*. 59–66. DOI: https://doi.org/10.1145/2696454.2696468

R. B. Cattell. 1966. The meaning and strategic use of factor analysis. In *Handbook of Multivariate Experimental Psychology*, R.B. Cattell (Ed.). Rand McNally, Chicago, IL, 174–243. DOI: http://dx.doi.org/10.1007/978-1-4613-0893-5_4

A. A. Cota, R. S. Longman, R. R. Holden, and G. C. Rekken. 1993. Comparing different methods for implementing parallel analysis: A practical index of accuracy. *Educ. Psychol. Meas.* 53 (1993), 865–875. DOI: http://dx.doi.org/10.1177/0013164493053004001

R. P. Chalmers. 2012. mirt: A multidimensional item response theory package for the R environment. *J. Stat. Softw.* 48, 6 (2012), 1–29. DOI: http://dx.doi.org/10.18637/jss.v048.i06

L. J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16 (1951), 297–334. DOI: http://dx.doi.org/10.1007/BF02310555

C. K. Danielson and C. R. Phelps. 2003. The assessment of children's social skills through self-report: A potential screening instrument for classroom use. *Meas. Eval. Couns. Dev.* 35 (2003), 218–229.

D. N. Davis and S. K. Ramulu. 2017. Reasoning with BDI robots: From simulation to physical environment—Implementations and limitations. *Palad. J. Behav. Robot.* 8 (2017), 39–57. DOI: http://dx.doi.org/10.1515/pjbr-2017-0003

K. Dautenhahn. 2007. Socially intelligent robots: Dimensions of human-robot interaction. *Philos. Translat. Roy. Society B: Biol. Sci.* 362 (2007), 679–704. DOI: http://dx.doi.org/10.1098/rstb.2006.2004

J. de Greeff, O. Blanson-Henkemans, A. Fraaije, L. Solms, N. Wigdor, and B. Bierman. 2014. Child-robot interaction in the wild: Field testing activities of the ALIZ-E project. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*. 148–149. DOI: http://dx.doi.org/10.1145/2559636.2559804

S. A. Dennis, B. M. Goodson, and C. Pearson. 2019. Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behav. Res. Account.* DOI: http://dx.doi.org/10.2139/ssrn.3233954

S. Devin, A. Clodic, and R. Alami. 2017. About decisions during human-robot shared plan achievement: Who should act and how? In *Proceedings of the 9th International Conference on Social Robotics*, A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, and H. He. (Eds.). Lecture Notes in Computer Science, 10652. Springer, Cham, Switzerland. DOI: http://dx.doi.org/10.1007/978-3-319-70022-9_45

P. Dixon. 2008. Models of accuracy in repeated-measures designs. *J. Mem. Lang.* 59, 4 (2008), 447–456. DOI: http://dx.doi.org/10.1016/j.jml.2007.11

L. R. Fabrigar, R. C. MacCallum, D. T. Wegener, and E. J. Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Meth.* 4 (1999), 272–299. DOI: http://dx.doi.org/10.1037/1082-989X.4.3.272

M. E. Ford and M. S. Tisak. 1983. A further search for social intelligence. *J. Educ. Psychol.* 75, 2 (1983), 196–206. DOI: http://dx.doi.org/10.1037/0022-0663.75.2.196

M. Frankovsky and Z. Birknerova. 2014. Measuring social intelligence—The MESI Methodology. *Asian Soc. Sci.* 10, 6 (2014). DOI: http://dx.doi.org/10.5539/ass.v10n6p90

M. P. Georgeff. and A. L. Lansky. 1987. Reactive reasoning and planning. In *Proceedings of the 6th National Conference on Artificial Intelligence (AAAI'87)*. 2 (1987), 677–682.

L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. C. Gough. 2006. The International Personality Item Pool and the future of public-domain personality measures. *J. Res. Personal.* 40 (2006), 84–96. DOI: http://dx.doi.org/10.1016/j.jrp.2005.08.007

N. Halmdienst, M. Radhuber, and R. Winter-Ebmer. 2019. Attitudes of elderly Austrians towards new technologies: Communication and entertainment versus health and support use. *Europ. J. Age.* 16, 4 (2019), 513–523. DOI: http://dx.doi.org/10.1007/s10433-019-00508-y

M. Heerink, B. Krose, V. Evers, and B. Wilinga. 2010. Assessing acceptance of assistive social agent technology by older adults: The Almere model. *Int. J. Soc. Robots* 2 (2010), 361–375. DOI: http://dx.doi.org/10.1007/s12369-010-0068-5

D. Helbing, I. J. Farkas, P. Molnar, and T. Vicsek. 2002. Simulation of pedestrian crowds in normal and evacuation situations. *Pedest. Evac. Dyn.* 21, 2 (2002), 21–58.

N. Hester. 2019. Perceived negative emotion in neutral faces: Gender-dependent effects on attractiveness and threat. *Emotion* 19, 8 (2019), 1490–1494. DOI: http://dx.doi.org/10.1037/emo0000525.supp

C. C. Ho and K. F. MacDorman. 2010. Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Comput. Hum. Behav.* 26 (2010), 1508–1518. DOI: http://dx.doi.org/10.1016/j.chb.2010.05.015

A. Honour, S. B. Banisetty, and D. Feil-Seifer. 2019. The perceived social intelligence of robots with socially-aware navigation. Retrieved from https://scholarworks.unr.edu/bitstream/handle/11714/6509/HonourAnaLisa_76233992_Honour_AnaLisa_2019_acc.pdf?sequence=1& isAllowed=y.

J. L. Horn. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika* 30 (1965), 179–185. DOI: http://dx.doi.org/10.1007/bf02289447

P. H. Kahn, T. Kanda, H. Ishiguro, B. T. Gill, S. Shen, H. E. Gary, and J. H. Ruckert. 2015. Will people keep the secret of a humanoid robot?—Psychological intimacy in HRI. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*. IEEE Computer Society, 173–180. DOI: http://dx.doi.org/10.1145/2696454.2696486

H. Kamide and T. Arai. 2017. Perceived comfortableness of anthropomorphized robots in U.S. and Japan. *Int. J. Soc. Robo.* 9 (2017), 537–543. DOI: http://dx.doi.org/10.1007/s12369-017-0409-8

M. Kaptein, P. Markopoulos, B. Ruyter, and E. Aarts. 2011. Two acts of social intelligence: The effects of mimicry and social praise on the evaluation of an artificial agent. *AI Soc.* 26, 3 (2011), 261–273. DOI: http://dx.doi.org/10.1007/s00146-010-0304-4

S. Kinga and K. Ibolya. 2013. The predictive value of social intelligence for cooperative behavior in a task-oriented interaction paradigm: A pilot study. *Erdé. Pszichol. Szemle* 14, 2 (2013), 255–274.

J. Kory. 2014. *Storytelling with Robots: Effects of Robot Language Level on Children's Language Learning*. Master's Thesis, Media Arts and Sciences, Massachusetts Institute of Technology, Cambridge, MA.

H. Lövheim. 2012. A new three-dimensional model for emotions and monoamine neurotransmitters. *Med. Hypoth.* 78, 2 (2012), 341–348. DOI: http://dx.doi.org/10.1016/j.mehy.2011.11.016

S. G. Luke. 2016. Evaluating significance in linear mixed-effects models in R. *Behav. Res. Meth.* 49, 4 (2016), 1494–1502. DOI: http://dx.doi.org/10.3758/s13428-016-0809-y

K. F. MacDorman, S. K. Vasudevan, and C.-C. Ho. 2009. Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. *AI Soc.* 23 (2009), 485–510. DOI: http://dx.doi.org/10.1007/s00146-008-0181-2

H. A. Marlowe. 1986. Social intelligence: Evidence for multidimensionality and construct independence. *J. Educ. Psychol.* 78, 1 (1986), 52–58. DOI: http://dx.doi.org/10.1037/0022-0663.78.1.52

P. E. Meehl. 1986. Law and the fireside Inductions (with postscript): Some reflections of a clinical psychologist. *Behav. Sci. Law* 7 (1986), 521–550. DOI: http://dx.doi.org/10.1002/bsl.2370070408

L. Moshkina. 2012. Reusable semantic differential scales for measuring social response to robots. In *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*. 89–94. DOI: http://dx.doi.org/10.1145/2393091.2393110

B. Mutlu and J. Forlizzi. 2008. Robots in organizations: The role of workflow, social, and environmental factors in human-robot interaction. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*. 287–294. DOI: http://dx.doi.org/10.1145/1349822.1349860

C. J. Parales-Quenza. 2006. Astuteness, trust, and social intelligence. *J. Theor. Soc. Behav.* 36, 1 (2006), 39–56. DOI: http://dx.doi.org/10.1111/j.1468-5914.2006.00295.x

E. Peer, J. Vosgerau, and A. Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav. Res. Meth.* 46, 4 (2014), 1023–1031. DOI: http://dx.doi.org/10.3758/s13428-013-0434-y

O. A. I. Ramírez, H. Khambhaita, R. Chatila, M. Chetouani, and R. Alami. 2016. Robots learning how and where to approach people. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'16)*. 347–353.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.r-project.org/.

W. Revelle. 2018. psych: Procedures for Personality and Psychological Research. Version = 1.8.12. Retrieved from https://CRAN.R-project.org/package=psych.

R. E. Riggio. 1989. *Social Skills Inventory Manual: Research Edition*. Consulting Psychologists Press, Palo Alto, CA.

RStudio Team. 2016. *RStudio: Integrated development for R*. RStudio, Inc., Boston, MA. Retrieved from http://www.rstudio.com/.

F. Samejima. 1969. Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph No. 17. Psychometric Society, Richmond, VA. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf.

M. Sebastian, S. B. Banisetty, and D. Feil-Seifer. 2017. Socially-aware navigation planner using models of human-human interaction. In *Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'17)*. 405–410.

D. H. Silvera, M. Martinussen, and T. I. Dahl. 2001. The Tromsø Social Intelligence Scale, a self-report measure of social intelligence. *Scand. J. Psychol.* 42 (2001), 313–319. DOI: http://dx.doi.org/10.1111/1467-9450.00242

D. Sirkin, B. Mok, S. Yang, and W. Ju. 2015. Mechanical ottoman: How robotic furniture offers and withdraws support. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*.

S. Sturgeon, A. Palmer, J. Blankenburg, and D. Feil-Seifer. 2019. Perception of social intelligence in robots performing false-belief tasks. In *Proceedings of the International Symposium on Robot and Human Interactive Communication (RO-MAN'19)*. DOI: http://dx.doi.org10.1109/RO-MAN46459.2019.8956467

J. Sung, H. I. Christensen, and R. Grinter. 2009. Robots in the wild: Understanding long-term use. In *Proceedings of the 4th ACI/IEEE International Conference on Human-robot Interaction*. 45–52.

B. G. Tabachnick and L. S. Fidell. 2019. *Using multivariate statistics* (7th ed.). Upper Saddle River, NJ: Pearson.

D. Tahiroglu, L. J. Moses, S. M. Carlson, C. E. V. Mahy, E. L. Olofson, and M. A. Sabbagh. 2014. The Children's Social Understanding Scale: Construction and validation of a parent-report measure for assessing individual differences in children's theories of mind. *Dev. Psychol.* 50, 11 (2014), 2485–2497. DOI: http://dx.doi.org/10.1037/a0037914

K. Thomassin and J. A. Seddon. 2019. Implicit attitudes about gender and emotion are associated with mothers' but not fathers' emotion socialization. *Canad. J. Behav. Sci./Rev. Canad. Sci. Comport.* 51, 4 (2019), 254–260. DOI: http://dx.doi.org/10.1037/cbs0000142

S. S. Tomkins. 1962. *Affect, Imagery, Consciousness, Volume I: The Positive Affects*. Springer, New York, NY. DOI: http://dx.doi.org/10.1037/14351-009

S. S. Tomkins. 1963. *Affect, Imagery, Consciousness, Volume II: The Negative Affects*. Springer, New York, NY. DOI: http://dx.doi.org/10.1037/14351-009

S. S. Tomkins. 1981. The quest for primary motives: Biography and autobiography of an idea. *J. Personal. Soc. Psychol.* 41, 2 (1981), 306–329. DOI: http://dx.doi.org/10.1037/0022-3514.41.2.306

S. S. Tomkins. 1991. *Affect, Imagery, Consciousness, Volume III: The Negative Affects: Anger and fear*. Springer, New York, NY. DOI: http://dx.doi.org/10.1016/0191-8869(93)90053-6

A. M. Turing. 1950. Computing machinery and intelligence. *Mind* 49 (1950), 433_460. DOI: http://dx.doi.org/10.1093/mind/LIX.236.433

W. F. Velicer. 1976. Determining the number of components from the matrix of partial correlations. *Psychometrika* 41 (1976), 321–327. DOI: http://dx.doi.org/10.1007/bf02293557

W. F. Velicer, C. A. Eaton, and J. L. Fava. 2000. Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In *Problems and Solutions in Human Assessment: Honoring Douglas N. Jackson at Seventy*, R. D. Goffin and E. Helmes (Eds.). Kluwer, Boston, MA, 41–71).

X. Wang, F. Teng, Z. Chen, and K.-T. Poon. 2019. Control my appearance, control my social standing: Appearance control beliefs influence American women's (not men's) social mobility perception. *Personal. Indiv. Diff.* 155. DOI: http://dx.doi.org/10.1016/j.paid.2019.109629

W. Weiten. 2017. *Psychology: Themes and Variations* (10th ed.). Cengage Learning, Boston, MA.

Received July 2019; revised June 2020; accepted July 2020