Journal of Autism and Developmental Disorders
https://doi.org/10.1007/s10803-023-05901-2

ORIGINAL PAPER

The Social Shapes Test as a Self-Administered, Online Measure of Social Intelligence: Two Studies with Typically Developing Adults and Adults with Autism Spectrum Disorder

Matt I. Brown[1,2] · **Patrick R. Heck**[1] · **Christopher F. Chabris**[1]

Accepted: 11 January 2023

Abstract
The Social Shapes Test (SST) is a measure of social intelligence which does not use human faces or rely on extensive verbal ability. The SST has shown promising validity among adults without autism spectrum disorder (ASD), but it is uncertain whether it is suitable for adults with ASD. We find measurement invariance between adults with ($n = 229$) or without ASD ($n = 1{,}049$) on the 23-item SST. We also find that adults without ASD score higher on the SST than adults with ASD ($d = 0.21$). We also provide two, 14-item versions which demonstrated good parallel test-retest reliability and are positively related to scores on the Frith-Happé task. The SST is suitable for remote, online research studies.

Keywords Autism spectrum disorder · Social cognition · Social intelligence · Theory of mind · Online testing

Social and emotional skills are important for adaptive functioning in everyday life (Soto et al., 2020). Clinical researchers have developed an array of psychological assessments to measure these skills (e.g., Abrahams et al., 2019) and to explain difficulties in social responsiveness or behavior observed in various psychological conditions including autism spectrum disorder (ASD; Morrison et al., 2019) and schizophrenia (Pinkham et al., 2018). Cognitive neuroscience researchers have also used these tools to identify brain areas associated with social functioning (Schaafsma et al., 2015; Schurz et al., 2021). Due in part to the COVID-19 pandemic, however, many clinical and research practices have needed to shift from administering in-person assessments to using measures that can be administered remotely during telemedicine visits or online studies (Türközer & Öngür, 2020). This shift to online, non-proctored assessment has already become the predominant mode in employment testing (e.g., Tippins 2015) prior to the pandemic. In contrast, many of the assessments used in clinical or developmental research are designed for in-person, proctored administration. Few studies to date have documented the psychometric properties of social intelligence measures in clinical and typically developing samples (e.g., Gourlay et al., 2020; Pinkham et al., 2018), and there has been little research to determine whether any of these instruments are suitable for remote, online administration.

Although recent work has helped establish new, web-based cognitive ability assessments (e.g., Biagianti et al., 2019; Liu et al., 2020; Wright, 2020), few studies have focused on designing remote measures of social intelligence. This is important given that social intelligence is generally affected by ASD or similar, developmental disorders (Velikonja et al., 2019). Presently, many of the social intelligence measures that can be administered online rely on self- or observer-reports such as the Social Responsiveness Scale (SRS; Constantino et al., 2003), the Autism Quotient (AQ; Baron-Cohen et al., 2001b), or the Broad Autism Phenotype Questionnaire (BAPQ; Hurley et al., 2007). In contrast, many of the validated, performance-based social intelligence tests are designed to be completed in-person and administered by a proctor or trained clinician. This makes them ill-suited for remote administration and presents a challenge given the growing need for remote, web-based assessments. To this end, the Social Shapes Test (SST; Brown et al., 2019) was developed as a simple, self-administered social intelligence

✉ Matt I. Brown
mibrown9015@gmail.com

[1] Geisinger Health System, Lewisburg, PA, USA

[2] Human Resources Research Organization, 66 Canal Center Plaza, Suite 700, 22314 Alexandria, VA, USA

Published online: 09 February 2023

test based on the animated shape task created by Heider and Simmel (1944). To date, however, the SST has only been validated for use with adults without ASD. Therefore, we conducted the present study to examine whether the SST is appropriate for use as a remote, performance-based social intelligence test for adults with ASD.

We consider the SST, along with other existing animated shape tasks, to be measures of social intelligence (SI). We define SI as the ability to perceive and decode the internal states, motives, and behaviors of others (Mayer & Salovey, 1993; Lievens & Chan, 2010). This operational definition overlaps with those for constructs commonly studied in autism research like mentalizing and Theory of Mind (ToM; Luyten et al., 2020). Some scholars have recently expressed concern regarding the accumulation of narrowly defined social and emotional abilities and the potential for jingle and jangle fallacies (Olderbak & Wilhelm, 2020; Quesque & Rossetti, 2020). An example of this concern is that a task in which individuals are asked to identify mental states from pictures of human faces (e.g., the Reading the Mind in the Eyes test; Baron-Cohen et al., 2001) has been variously characterized as a measure of Theory of Mind, mentalizing ability, empathic accuracy, face processing, and emotion recognition across different studies (Oakley et al., 2016). Therefore, we use the more inclusive term SI given its long history in psychological research and its broader use across research fields relative to other terms (e.g., Theory of Mind which is more specific to developmental research, or mentalizing, which is more specific to social cognitive neuroscience).

Measuring Social Intelligence Using Animated Shape Tasks

The original animated shape task was developed by Heider & Simmel (1944) who famously observed that research participants often described the movements of simple animated, geometric shapes in human psychological terms. This pioneering work inspired several streams of research where scholars sought to identify individual differences in Theory of Mind or mentalizing ability using the original film or newly created shape animations. Klin (2000) used the original Heider and Simmel film to create the Social Attribution Task (SAT). In this task, individuals were shown the film and asked to provide written responses to 17 questions about the events in the film (e.g., "What happened to the big triangle"). Each question was asked after participants viewed specific segments of the film. These responses were scored by human raters based on the use of specific kinds of terms indicating concepts such as emotions, mental states, or behaviors. Klin reported that individuals with autism or Asperger's disorder made fewer social attributions compared to individuals without ASD as indicated by using fewer mental or emotional state terms, mentioning fewer personality features of the shapes, and difficulty identifying the social meaning of the shapes' movements. Likewise, SAT scores have been found to predict the severity of ASD-related social symptoms among a sample of children with ASD but average general intelligence (Altschuler et al., 2018). Researchers have observed modest test-retest reliability for SAT scores. Most recently, Altschuler and Faja (2022) reported stronger reliability for spontaneous ToM and cognitive ToM scores but slightly weaker reliability for affective ToM scores. A modified version of the SAT has also been used in neuroimaging research to identify differences in activation of brain regions related to social information processing between individuals with and without ASD (e.g., Vandewouw et al., 2021).

The Frith-Happé animation task is similar to the Social Attribution Task but consists of 12 short films with each featuring two animated triangles (Abell et al., 2000). In each film, the movements of the triangles are meant to depict interactions involving mental states, only physical, goal-directed interaction, or purposeless movement. A recent meta-analysis studies which have used the Frith-Happé animations ($k = 33$ studies) found that individuals with ASD are less able to correctly categorize animations which are designed to depict mentalizing compared to animations containing only goal-directed or random movement (Wilson, 2021). In addition, the Frith-Happé animations have also been used to identify similar difficulties in social attribution among adult patients with schizophrenia (Martinez et al., 2019).

Although most studies have focused on mental state attributions from written responses to these animated shape stimuli, some scholars have adapted these tasks into a multiple-choice test format. A 19-item, multiple choice version of the Social Attribution Task (SAT-MC) was designed by Bell and colleagues (2010). This test uses the same film as the SAT but replaces the narrative responses with targeted multiple-choice questions which are scored as either correct or incorrect. Performance on the SAT-MC has been found to be positively related to performance on other social cognition tasks including the Bell-Lysaker Emotion Recognition Task and the Mayer-Salovey-Caruso Emotional Intelligence Test (Bell et al., 2010). Adults with schizophrenia have also been found to perform significantly worse on the SAT-MC compared to a group of healthy controls (Johannesen et al., 2018; Pinkham et al., 2018). SAT-MC scores were also found to be positively related with social skills as assessed by a standardized role-playing task. In addition, the test has also displayed promising validity in autism research where a recent pilot study by Burger-Caplan et al. (2016) found

that children with an ASD diagnosis scored 0.87 standard deviations worse on the test compared to healthy controls.

Similar to the SAT-MC, White and colleagues (2011) designed a multiple-choice task using the Frith-Happé animations. In this version of the task, individuals are asked to correctly categorize each film as demonstrating either theory of mind, physical interaction, or random movement. Performance on this task has been found to positively correlated with performance on other social intelligence tasks while also displaying modest group score differences favoring IQ-matched, typically developing adults (Brewer et al., 2017). This task has also been administered as an online task completed by adults with and without ASD diagnoses in recent research (Livingstone et al., 2021). However, the multiple-choice version of the Frith-Happé animation task has only been used in seven of the 33 studies identified by Wilson (2021).

One benefit of these various animated shape tasks is that they rely less on reading skill or verbal knowledge and comprehension compared to other measures of SI. For example, tasks like the Faux Pas (Baron-Cohen et al., 1999) or the Hinting task (Corcoran et al., 1995) require reading and interpreting written descriptions of social interactions. Other tasks like the Reading in the Mind in the Eyes Test (RMET; Baron-Cohen et al., 2001a) require knowledge of words used to describe emotional or mental states which are not commonly used in everyday language (Kittel et al., 2022; Peterson & Miller, 2012). These tasks may confound social intelligence with verbal ability where some individuals could use their verbal skills to compensate for low SI (Livingston & Happé, 2017). Another advantage to animated shape tasks is that they are abstract and do not include any obvious cultural or gender cues. These cues, such as those present in emotion recognition tasks which only use faces of White or Caucasian individuals, can result in mean test score differences due to race or ethnicity among clinical and nonclinical populations (Dodell-Feder et al., 2020; Pinkham et al., 2017). In contrast, animated shape tasks have displayed little if any racial or ethnic group differences in past research which makes them potentially suitable for studies involving international samples (Brown et al., 2019, 2022; Lee et al., 2018). However, several of the existing animated shape tasks are not well-suited for remote testing. For example, the original SAT and Frith-Happé animations, require a clinician or administrator to ask questions and to record and score verbal responses from participants. Not only could this introduce confounding effects of verbal ability or rater bias, but it also increases administration time and financial costs (Livingston et al., 2019). Although more recent versions of these tasks use a fixed set of multiple choice questions (e.g., SAT-MC), an administrator is still needed in order to operate the specific video segments for each question. This is problematic because it prevents participants from being able to complete these tasks remotely which likely prevents some researchers from using these tasks as studies increasingly shift from being conducted in-person to online. Other alternative versions of these tasks are primarily designed for neuroimaging studies and are also not well-suited for brief, online assessment (Ludwig et al., 2020).

The Social Shapes Test (SST)

The SST is a 23-item multiple choice test designed to measure individual differences in social intelligence among neurotypical adults. Each SST item consists of a short, 13–23 s animated video which includes a standard set of colored, geometric shapes. Each video features a different social plot where the shapes display a variety of behaviors including bullying, helping, comforting, deceiving, and playing. Some animations were designed to mimic the bullying behavior which appears in the original Heider and Simmel video. Others were designed to represent false belief tasks. In sum, these animations have been found to elicit a similar degree of social attributions in written descriptions compared to those reported by Klin in a prior study (Ratajska et al., 2020). In this study, Ratajska et al. (2020) scored narrative descriptions for each of the SST videos using Klin's (2000) Theory of Mind indices and found that the range of scores for SST items overlapped with those reported for the original Heider and Simmel film. All videos are controlled by the participant and can be viewed as many times as desired. Before starting the SST, participants are given the following instructions:

> "In this task, you will see a series of short, silent, animated videos. The shapes in these videos can be interpreted as people interacting with each other.
> First, please watch each video carefully and completely. After watching the video, select the best answer to the multiple choice question listed below the video. Make sure to answer all of the questions to the best of your ability.
> Next is a practice item. Please watch the video and try your best to answer the question. Note that you are allowed to replay a video as many times as you want while answering the question. Please do not expand the videos to full screen."

Next, all participants are given a sample item followed by feedback indicating the correct response (Fig. 1). All 23 items are subsequently administered in the same order for all participants.

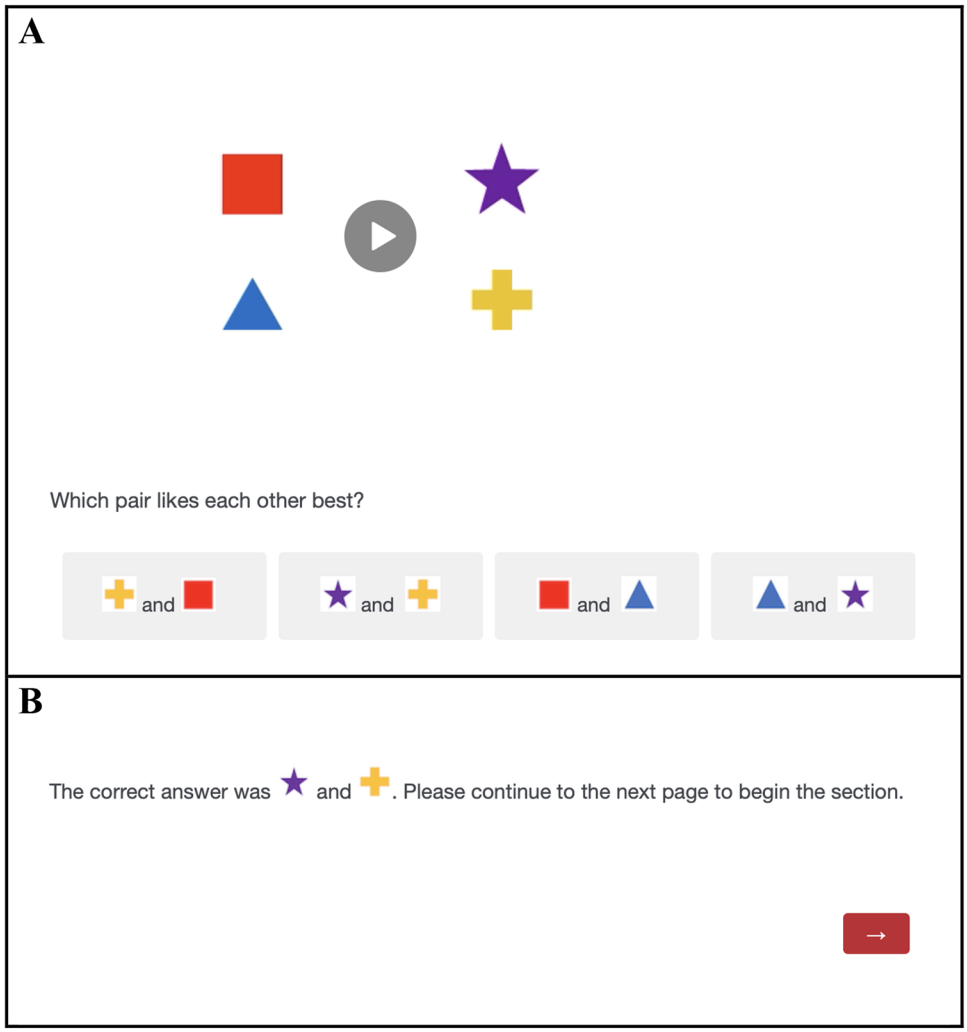

Fig. 1 Practice SST Item. All participants were given a practice item (A) before starting the 23-item SST. After responding to the practice item, participants received feedback which identified the correct response (B)

Unlike other SI tasks, the SST was explicitly designed to be completely self-administered online, as was done in initial validation studies (Brown et al., 2019). All questions are scored using an objective scoring key which helps prevent potential rater bias when scoring open responses which are used in other animated shape tasks (White et al., 2011). Like the SST, an updated version of the Frith-Happé animations was developed for remote, online administration (Livingston et al., 2021). Although versions of the Frith-Happé animations have been found to detect differences in social intelligence between neurotypical adults and adults with ASD, they have rarely been administered to large samples of typically developing adults. Lastly, all SST questions and video files are freely available for research use and can be accessed via the Open Science Framework (https://osf.io/sqxy6). Researchers are free to use the SST videos to administer the test as part of an online survey or to adapt the videos in order to suit their own individual studies. This makes the SST more easily accessible for researchers, especially compared to other video-based social intelligence tests owned or distributed by commercial test publishers (e.g., The Awareness of Social Inference Test – TASIT; McDonald, 2012). The video content in the SST is also relatively short (each animation ranges between 13 and 23 s in length) which helps minimize administration time compared to other, video-based measures of social intelligence (e.g., Movie for the Assessment of Social Cognition; Dziobek et al., 2006).

The SST is also unique in that that it was originally developed and validated using samples of undergraduate college students and crowdsourced participants from Amazon Mechanical Turk (MTurk) who were not selected for prior history or diagnosis of ASD. In these studies, the SST has demonstrated modest internal consistency ($\alpha > 0.65$) and promising convergent validity with other performance measures of social intelligence. Among MTurk workers, SST scores were found to be positively related to emotion recognition ability as assessed by the RMET ($r = .47$). Individuals who scored higher on the SST were also more effective at identifying the correct emotion or mental state based on written scenarios in the Situational Test of Emotional

Understanding ($r = .48$; Brown et al., 2019). In a subsequent study of undergraduate psychology students, those who scored higher on the SST were better at identifying the best behavioral solutions to interpersonal workplace situations in a situational judgment task ($r = .40$; Brown et al., 2022). These relationships remained even after controlling for differences in more general cognitive abilities (e.g., verbal or spatial abilities) and educational attainment. Despite these promising results, however, it is uncertain whether the SST can adequately assess differences in social intelligence among adults with ASD or other developmental disorders.

Present Study

We designed the present study to investigate whether the SST is appropriate to be self-administered remotely to measure social intelligence among adults with ASD. Our first aim is to test for measurement invariance of the SST between adults with or without ASD. Our second aim is to collect further validity evidence for the SST by testing whether unaffected, typically developing adults score higher on the test compared to adults who have been diagnosed with ASD. Based on similarities in test content (e.g., use of similar, geometric shape animations) and existing convergent validity evidence from typically developing adult samples, we expect that the SST and other animated shape tasks measure a similar, underlying social intelligence construct. Therefore, we expect that adults with ASD should score lower on the SST compared to adults without ASD as observed in prior research using similar animated shapes tasks like the Frith-Happé animations or the SAT-MC (Burger-Caplan et al., 2016; Livingstone et al., 2021; Wilson, 2021). We also conducted a second study to gather further reliability and validity evidence for two, alternate 14-item forms of the SST and to compare performance on the SST with scores on an existing animated shape task (the Frith-Happé animation task; White et al., 2011).

Study 1

Methods

Participants

Participants in Study 1 included of a variety of adults with or without a prior diagnosis of autism spectrum disorder (ASD). We recruited 261 participants who self-reported a diagnosis of ASD, autistic disorder, or Asperger's disorder from the Simons Foundation Powering Autism Research for Knowledge (SPARK; SPARK Consortium, 2018). This cohort consists of individuals with ASD and their first-degree relatives. All of these individuals who were recruited for this study presently live independently and did not have a record of cognitive impairment when they joined the SPARK cohort. A broader description of adults in the SPARK cohort was recently reported by Fombonne et al. (2020). All SPARK participants were given a $10 Amazon gift card for completing the study. Although diagnosis history was collected using either self- or parent-reports, rather than direct clinical evaluation, past research has found that this method yields reliable accounts of autism diagnoses in other research registries (Daniels et al., 2012).

To account for the lack of clinical data for the independent adult ASD sample recruited via SPARK, we also recruited a second sample of 25 adults who had previously received a clinical diagnosis of ASD from a neurodevelopmental clinic. All of these 25 individuals had sought clinical services in the Northeastern U.S. and had consented to be contacted for ongoing research studies. Due to the smaller pool of eligible participants compared to the SPARK cohort, these participants were given a larger reward of a $35 Amazon gift card for completing this study. All participants were recruited online and completed the SST without a proctor or administrator. Participants ranged from 18 to 34 years of age (mean age = 20.8, SD = 3.9). Most participants identified as male (20/25; 80%) and as White, non-Hispanic (22/25; 88%). Based on assessment scores obtained from the electronic medical record, the average full-scale IQ score among the ASD group was 86.1 (SD = 22.4). T-scores from the Social Responsiveness Scale (SRS; Constantino et al., 2003) indicated an elevated level of autistic symptoms among most of the participants in clinical ASD group as well (M = 74.2, SD = 12.4).

We also recruited adults without ASD for this study. One group of adults without ASD was recruited from SPARK; these individuals were parents of one or more children with an ASD diagnosis who themselves had never received a diagnosis (SPARK parent; $n = 217$). Although these adults did not report any history of ASD, they may be at a greater genetic risk for ASD compared to the general population. Therefore, we also relied on data collected from adult participants in two prior studies (Brown et al., 2019, 2022) for a comparison group of adults without ASD. Unlike the SPARK parents, we assume that the adult participants from prior studies were not likely to share a potential genetic predisposition to ASD or to other developmental disorders given the relatively low rate of ASD in the general population. There was also no focus on ASD or developmental disorders in either of the prior studies from which these participants were recruited. A total of 829 participants were recruited from undergraduate psychology courses at a public university in the Midwestern U.S. and from Amazon's

Table 1 Study 1 Participant Demographics

Category	ASD ($n = 229$)	SPARK Parents ($n = 217$)	Without ASD ($n = 829$)
Sex			
Male	102 (45%)	99 (45%)	343 (41%)
Female	127 (55%)	119 (55%)	490 (59%)
Race/Ethnicity			
White, Non-Hispanic	191 (83%)	172 (79%)	466 (56%)
Age (in years)			
Mean	33.61	41.76	29.69
Standard Deviation	12.42	8.88	11.47
Educational Attainment			
Less than high school	6 (3%)	4 (2%)	2 (> 1%)
High school degree	44 (19%)	31 (14%)	95 (11%)
Some college	102 (45%)	71 (32%)	515 (62%)
4-year college degree	48 (21%)	51 (23%)	170 (20%)
More than a four-year degree	29 (13%)	61 (28%)	51 (22%)

Note: ASD = participants reporting a diagnosis of autism spectrum disorder, autism disorder, or Asperger's syndrome; Without ASD = adults without ASD from prior studies which were sampled from Amazon's Mechanical Turk or undergraduate psychology students (Brown et al., 2019, 2022); Race/Ethnicity (1 = White/non-Hispanic, 0 = all other racial or ethnic groups); Sex (1 = Female, 0 = Male); Education was self-reported; Group differences in age $F(2,1273) = 100.63$, $p < .001$; Group differences in sex χ^2 (2) = 1.74, $p = .42$; Group differences in race/ethnicity χ^2 (2) = 82.37, $p < .001$; Group differences in educational attainment $F(2,1277) = 20.53$, $p < .001$.

Mechanical Turk. Most of these participants identified as female (59%) and White, non-Hispanic (56%). All prior study adult participants had completed the SST as part of a self-administered, online Qualtrics survey.

Data Cleaning

Prior to data analysis, we removed participants who had a median response time of less than 10 s per SST item. This response time threshold was chosen to increase the likelihood that participants watched the entire video for each item and to remove potential cases of non-purposeful responding (all 23 SST videos were 13 s in length or longer). We observed that a greater proportion of the adults without ASD were removed based on our response time threshold (21%) compared to SPARK participants with ASD (3%). A total of five participants were removed from the clinical sample. We also removed participants who failed to respond correctly to all four attention-check items. In each such item, participants were asked to watch a different shape animation and to identify which of four shapes did not appear in the video. The attention-check items depend only on basic cognitive processes (e.g., vision, attention, memory) and should not require social intelligence to solve. More than 90% of participants with or without ASD were able to correctly identify the missing shape in all four items. This left us with a total sample of $n = 1{,}275$ participants (ASD $n = 229$; SPARK parent $n = 217$; without ASD $n = 829$). We provide a full summary of the key demographic variables for each group in Table 1.

Procedure

All study materials were presented in an online survey which was accessed via a link sent using email. Participants were given a brief set of instructions and a practice item before beginning the SST. Afterwards, participants completed several demographic items regarding their geographical location, educational attainment, self-identified race/ethnicity, and approximate annual income. Participant sex and age for participants recruited via SPARK was provided by the SPARK consortium. All SST scores were calculated as the simple sum of correct responses across the 23 items.

Statistical Analysis

All analyses were performed using R version 3.6.3. We tested for measurement invariance using confirmatory factor analysis models estimated using the *lavaan* package (Rosel, 2012). We also used multiple linear regression in order to statistically control for demographic differences between the three diagnosis groups. We report the standardized mean difference (Cohen's *d*) when interpreting differences in SST scores based on ASD group where negative values indicate lower scores compared to adults without ASD.

Results

In order to investigate our first aim and determine whether the SST functions similarly for adults with or without ASD, we tested for measurement invariance between participants with or without ASD. It is important to rule out measurement invariance in order to determine whether the observed score differences between groups are due to true differences in the construct of interest and not a result of differences in the test's measurement properties (Vandenberg & Lance, 2000). We focused on metric invariance which tests whether the primary factor loadings for test items are equal across different groups. This is tested by first specifying a single-factor, confirmatory factor analysis model where all 23 SST items load on to a single factor. Factor loadings were then estimated for all adults without ASD (participants from prior

Table 2 SST Descriptive Statistics by Diagnosis Group in Study 1

	ASD (n = 229)	SPARK parent (n = 217)	Without ASD (n = 829)
Mean (SD)	15.16 (3.73)	15.58 (3.71)	15.89 (3.42)
Cohen's d [95% CI]	–0.21 [–0.36, –0.06]	–0.09 [–0.24, 0.05]	
α [95% CI]	0.72 [0.67, 0.77]	0.73 [0.68, 0.78]	0.67 [0.64, 0.70]
Mean CITC (SD)	0.27 (0.10)	0.29 (0.12)	0.24 (0.09)

Note. Cohen's d = standardized mean difference relative to adults without ASD; ASD = participants reporting a diagnosis of autism spectrum disorder, autism disorder, or Asperger's syndrome; CITC = corrected item-total correlation

Table 3 Differences in SST Scores between Diagnosis Groups in Study 1

	B(SE)	β	p
Model 1			
ASD	–0.73 (0.26)	–0.08	0.006
SPARK Parent	–0.31 (0.27)	–0.03	0.25
Model 2			
ASD	–0.94 (0.27)	–0.10	< 0.001
SPARK Parent	–0.69 (0.29)	–0.07	0.02
Age	–0.00 (0.01)	–0.01	0.46
Race/Ethnicity	0.78 (0.22)	0.10	< 0.001
Education	0.61 (0.12)	0.15	< 0.001

Note. β = standardized regression coefficient; ASD = autism spectrum disorder; Race/Ethnicity (1 = White/non-Hispanic, 0 = all other participants); Race/Ethnicity (1 = identified as White, 0 = all other participants); Model 1 $R^2 = 0.01$, $F(2, 1272) = 3.99$, $p = .02$; Model 2 $R^2 = 0.04$, $F(6, 1269) = 9.70$, $p < .001$; Sex was not included as a covariate because there were no differences in sex between groups but we also observed no sex difference in SST scores ($d = 0.02$, $t = 0.32$, $p = .74$)

studies and SPARK parent groups combined). Next, this model is estimated using response data from the ASD group while constraining each item loading to be equal to the estimate from the group without a prior diagnosis. Constraining these factor loadings to be equivalent did not significantly reduce overall model fit, $\Delta\chi^2$ (22) = 30.05, $p = .12$. These results were further supported by comparable estimates for internal consistency of the SST separately for participants with ASD (α = 0.72) and without ASD (α = 0.67). Item difficulties (percent correct) within the SST were also highly consistent between groups ($r = .97$, $p < .001$). This indicated that the items which were most difficult for adults with ASD were also most difficult for adults without ASD. Based on these results, we conclude that the SST provides an equivalent measure of social intelligence for adults regardless of ASD diagnosis. We report the psychometric properties and descriptive statistics for the SST within each group in Table 2. Therefore, any subsequent differences in SST scores between groups can be attributed to differences in social intelligence ability and not differences in test functioning between the two groups.

For our second aim, we tested for differences in SST scores between participants with and without ASD using linear regression. Given the heritability of ASD, we also tested for differences between adults recruited for prior SST studies and parents of an affected child (SPARK parents). We first regressed SST scores on dummy-coded diagnosis variables representing adults with ASD and SPARK parents (Model 1 in Table 3). A statistically significant regression coefficient for either dummy-coded variable indicates a meaningful difference in SST scores compared adults without ASD from prior studies. Participants with ASD did score significantly lower on the SST relative to adults without ASD from prior studies (β = –0.08, $p = .006$, $d = -0.21$, 95% CI = [–0.35, –0.06]). We provide a histogram illustrating this difference in SST scores in Fig. 2. SPARK parents also scored lower on the SST compared to adults without ASD but this difference was not statistically significant (β = –0.03, $p = .22$, $d = -0.09$, 95% CI = [–0.25, 0.06]). However, we observed several differences in the demographic makeup across our three comparison groups which may affect these observed test scores (Table 1). In particular, SPARK parents reported greater educational attainment and were older than the other two groups on average. Adults without ASD from prior studies were younger on average, reported greater educational attainment compared to adults with ASD, and were less likely to identify as White, non-Hispanic. Therefore, we also tested for differences in SST score after statistically controlling for participant age, educational attainment, and race/ethnicity. Participants with ASD still scored lower on the SST after controlling for these demographic differences (β = –0.10, $p < .001$; $d = -0.27$). In addition, we observed a significant difference between adults without ASD and SPARK parents when holding age, race/ethnicity, and education constant (β = –0.07, $p < .001$; $d = -0.20$). Among the demographic control variables, educational attainment was positively related to SST scores (β = 0.15, $p < .001$) and participants who identified as White scored higher on the SST, compared to all others (β = 0.10, $p < .001$). Age was not a significant predictor of SST scores in this regression model.

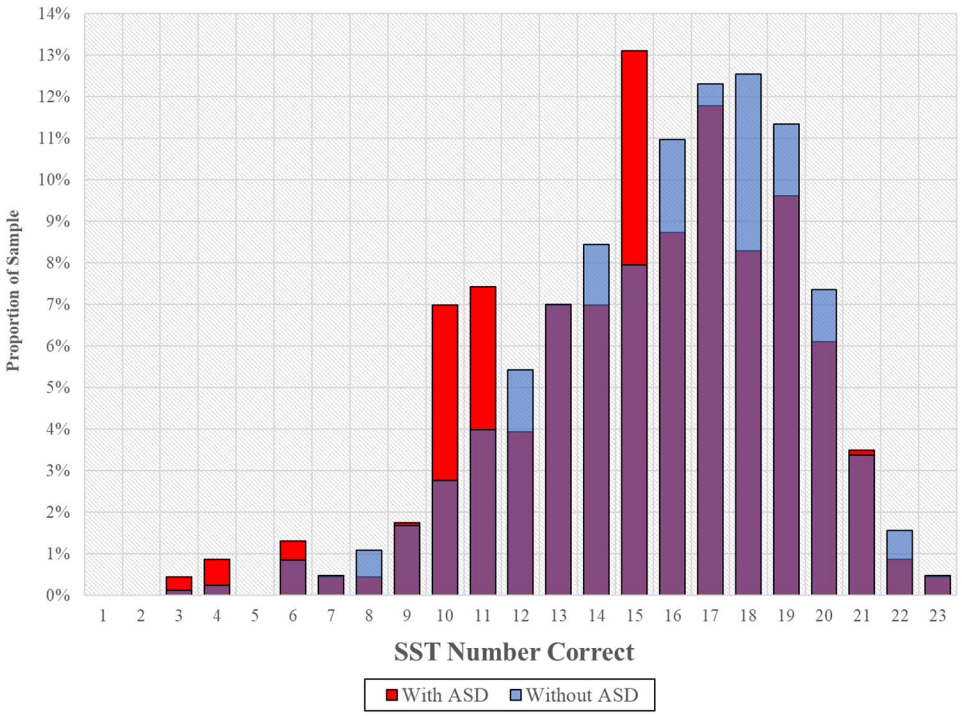

Fig. 2 Relative distribution of SST Scores for participants with ASD ($n = 229$) and without ASD ($n = 829$). Adults with ASD are displayed in red. Adults without ASD are displayed in blue. The y-axis represents the proportion of participants within each group. The x-axis represents the number of correct responses to the 23-item SST.

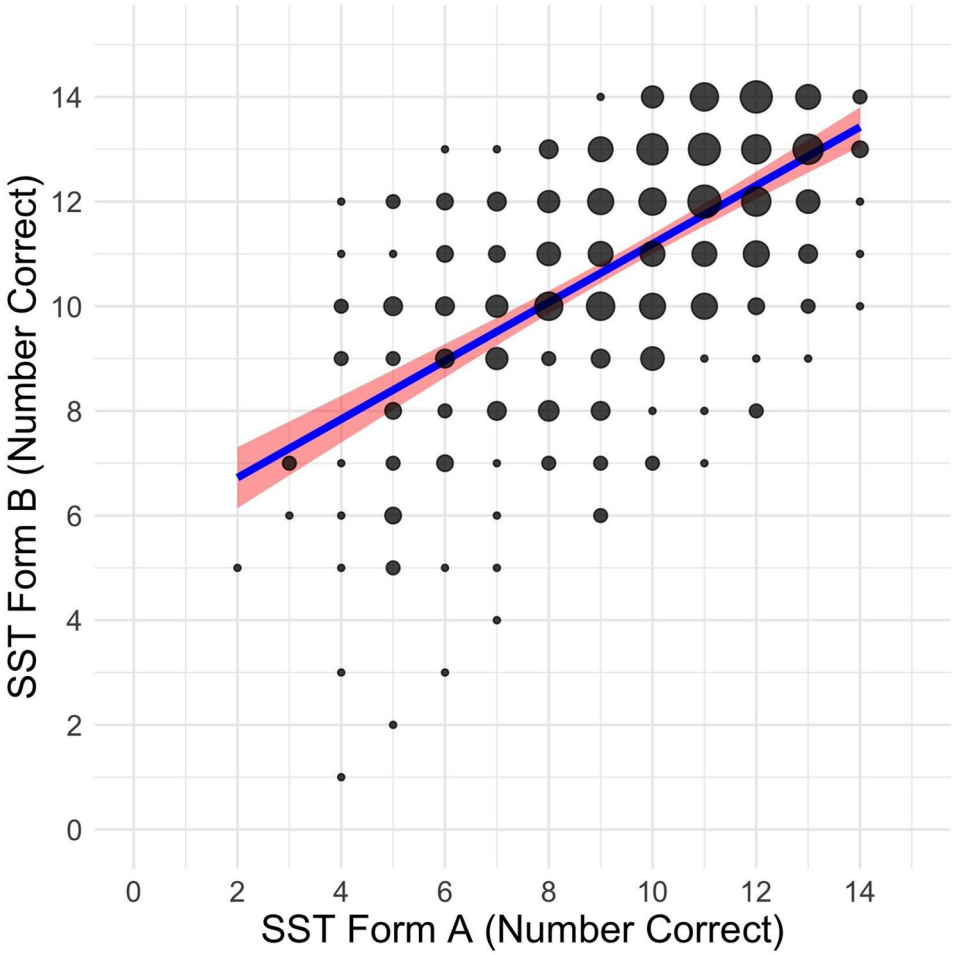

Fig. 3 Correlation between alternate 14-item SST forms in Study 2

Study 2

Although the 23-item SST typically takes roughly 15 min to complete, this may be too long for some studies involving a series of different assessments and measures. Prior studies have developed shorter versions of commonly used SI tests (e.g., short-form Reading the Mind in the Eyes Test; Olderbak et al., 2015) to accommodate researchers who wish to include SI performance measures without making the length of the study discouraging or prohibitive to prospective participants. Therefore, we conducted Study 2 to develop an abbreviated form of the SST which could be used in situations where a shorter administration time is needed while retaining the psychometric properties of the full 23-item version. We also conducted this study in order to estimate the test-retest reliability of this shorter version of the SST. The 23-item version has also only displayed modest internal consistency with estimates ranging between $\alpha = 0.67$ (Brown et al., 2019) to 0.72 among adults with ASD in Study 1. However, some scholars have argued that internal consistency underestimates the reliability of tests when item content is heterogeneous (Neubauer & Hofer, 2022). We aim to estimate the reliability of the SST as the test-retest reliability between two alternate forms as demonstrated for the SAT-MC by Pinkham et al. (2017).

A second goal of Study 2 is to gather further validation evidence for the SST by observing convergent validity with another animated shape task. As noted earlier, the Frith-Happé animation task was recently evaluated for remote, online use in research studies (Livingstone et al., 2021). Although this task uses similar shape animations as featured in the SST, there are some differences between the two measures. A prior version of the Frith-Happé task included questions where participants needed to identify mental or emotional states of specific shapes (White et al., 2011), but most of the research using the tool has only administered questions where participants only categorize the content of each video. In contrast, the SST features more focal shapes in each video (four or five shapes compared to only two in the Frith Happé task) and all SST videos involve social interactions. This potentially allows for greater granularity in measuring differences in social intelligence relative to the four Theory of Mind items in the Frith-Happé task. We include both measures in Study 2 in order to estimate the correlation between the two tasks and to observe whether each task can predict incremental variance in performance on a separate, video-based social intelligence test after controlling for individual differences in general intelligence.

Table 4 Study 2 Demographics

Category	Overall ($n = 387$)
Sex	
Male	209 (54%)
Female	178 (46%)
Race/Ethnicity	
White, Non-Hispanic	308 (80%)
Age (in years)	
Mean	41.63
Standard Deviation	12.08
Educational Attainment	
Less than high school	3 (1%)
High school degree	37 (10%)
Some college	114 (29%)
4-year college degree	158 (41%)
More than a four-year degree	75 (19%)

Methods

Participants

We gathered a sample of 387 U.S.-based adults from Amazon's Mechanical Turk using CloudResearch (formerly Turk Prime; Litman et al., 2017). Participants were paid $5.50 for completing the first survey and were paid $7.00 for completing a second survey one week later. Among the initial 504 participants who completed the first survey, we removed 24 participants who failed either the attention check items from each SST form or did not meet our median response time criteria. Of the remaining 480 participants, 387 returned one week later to complete the second survey (81%). Most participants identified as White (80%) and male (56%). The average participant age was 41.63 years old (SD = 12.08). We report a full summary of participant demographics in Table 4. When comparing participants who did or did not complete the second, follow-up survey, we found no differences in self-reported gender ($\chi^2(1) = 3.66$, $p = .06$) or educational attainment ($t = 0.69$, $p = .49$). Participants who completed both surveys were slightly older ($d = 0.24$, $t = 2{,}29$, $p = .02$) and were more likely to identify as White (80% versus 72%; $\chi^2(1) = 4.45$, $p = .03$) than those who completed only the first survey.

Procedure

All measures were administered using an online survey hosted by Qualtrics. All tasks were completed by participants in the same fixed order. Participants were also recruited to complete an alternate SST form one week after completing the initial form. On average, participants completed the second survey six days after the first administration (ranging from 6 to 10 days between administrations). After the

alternate SST form, participants completed the four Theory of Mind Frith-Happé animations along with the eight feelings questions reported by White et al. (2011), the Social Norm Questionnaire (Kramer et al., 2014), and an 18-item situational judgment test of interpersonal skills.

Measures

We created two 14-item SST forms based on an item analysis of the full 23-item version using item-level data reported by Brown et al. (2019) and supplemental data which was not featured in the published article but is publicly available on the Open Science Framework (https://osf.io/sqxy6/). Several newly written items were created based on existing animation files and were initially evaluated as part of a separate study. Each form featured the same 14 shape animation files but paired each with a different multiple-choice question. Each form also included a single, attention check item from the original 23-item version. All participants were randomly assigned to complete either Form A or Form B in the first survey and completed the alternate form when participating in the second survey.

After completing the SST in the first survey, participants completed the 12-item Frith-Happé animation task (Livingstone et al., 2021; White et al., 2011). In this task, participants viewed short film clips featuring two animated triangles and are asked to categorize each film as demonstrating random movement, physical or goal-directed movement, or mentalizing. Lastly, participants completed the 16-item ICAR cognitive ability test ($\alpha = 0.77$; Revelle et al., 2020).

In the second survey, participants completed the alternate SST form and the objective, eight-item Frith-Happé feelings task (White et al., 2011). To assess knowledge of social norms, we next administered the 22-item Social Norm Questionnaire (SNQ; Kramer et al., 2014) and an 18-item situational judgment test. The SNQ measures knowledge of social norms ($\alpha = 0.68$) and has been observed to correlate positively with other social intelligence ability tests in past research (Baksh et al., 2021). The situational judgment test (SJT) was designed to evaluate understanding of effective behavior in interpersonal interactions in a variety of everyday settings. Each item presents a short, written scenario about a social interaction. We asked participants to identify the most effective and least effective responses to each scenario among five behavioral response options ($\alpha = 0.72$). This methodology is widely used in research and practice to assess interpersonal skills in adults and children (Murano et al., 2020; Webster et al., 2020).

Results

We tested for practice effects and differences in difficulty between the two alternate forms using repeated measures ANOVA. There was no evidence for a practice effect between SST administrations in the first and second surveys, $F(1,385) = 2.07$, $p = .15$. Form A was significantly more difficult compared to Form B, $F(1,385) = 158.20$, $p < .001$, Cohen's $d = 0.55$. This difference in scores between forms was consistent regardless of the order in which the forms were completed. Although participants provided more correct responses to Form B relative to Form A, we found modest test-retest reliability between the alternate forms ($r = .61$, $p < .001$; ICC = 0.52, 95% CI = [0.25, 0.68]; Fig. 3). The test-retest correlation did not vary based on the order in which the alternate forms were completed. We also found similar estimates for internal consistency for each (Form A $\alpha = 0.65$; Form B $\alpha = 0.64$). There was no statistically significant score differences based on participant age, gender, or race/ethnicity for either SST form. These results provide support for the test-retest reliability for the SST across alternate forms plus comparable internal consistency for each shortened form with the original 23-item version in Study 1.

Next, we correlated SST scores with performance on the Frith-Happé animation task. All correlations between study tasks are reported in Table 5. We first calculated separate categorization scores for the Theory of Mind (ToM), goal-directed (GD), and random video trials (e.g., Livingstone et al., 2021; White et al., 2011). Among these three subscales, only the random videos provided adequate internal consistency ($\alpha = 0.64$). In contrast, internal consistency was very weak for the goal-directed ($\alpha = 0.33$) and ToM videos ($\alpha = 0.09$). Two of the ToM videos were incorrectly categorized as representing a physical interaction by a majority of participants ("Coaxing" = 85% of participants, "Seducing" = 70% of participants). The corrected item-total correlations for the four ToM videos were also very weak, ranging between –0.19 and 0.14. Likewise, two of the goal-directed videos were also incorrectly categorized as representing a mental interaction (ToM) by most participants ("Chase" = 59% of participants, "Leading" = 62% of participants). Due to these weak reliability estimates, we use overall performance scores on the Frith-Happé categorization items in our regression analyses (12-item $\alpha = 0.48$). Despite these poor measurement properties, SST scores were positively correlated with categorization of all Frith-Happé videos ($r_{\text{Form A}} = 0.36$, 95% CI = [0.27, 0.44], $p < .001$, $r_{\text{Form B}} = 0.42$, 95% CI = [0.34, 0.50], $p < .001$).

In contrast to the categorization items, we observed stronger measurement properties for the Frith-Happé feelings items. For each of the four ToM animations, participants were asked to identify the correct mental state of the

Table 5 Study 2 Correlation Matrix

	M	*SD*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1. SST A	9.53	2.57									
2. SST B	10.93	2.36	0.61								
3. F-H ToM	0.45	0.22	0.20	0.24							
4. F-H GD	0.46	0.26	0.15	0.19	0.27						
5. F-H Random	0.89	0.21	0.39	0.44	0.18	0.05					
6. F-H Total	0.60	0.16	0.36	0.42	0.71	0.72	0.56				
7. F-H Feelings	5.76	1.78	0.48	0.44	0.21	0.19	0.34	0.36			
8. SNQ	19.20	2.49	0.45	0.45	0.14	0.07	0.37	0.27	0.43		
9. SJT	17.25	5.91	0.39	0.43	0.08	0.06	0.33	0.22	0.38	0.53	
10. ICAR	8.27	3.41	0.47	0.43	0.12	0.19	0.28	0.29	0.33	0.24	0.33

Note. $n = 387$; SST = Social Shapes Test; F-H = Frith-Happé; ToM = Theory of Mind; GD = goal directed; SNQ = Social Norm Questionnaire; SJT = situational judgment test of interpersonal skills; ICAR = International Cognitive Ability Resource

All correlations > $r = .15$ are statistically significant at $p < .001$; correlations > $r = .11$ are statistically significant at $p < .05$

Table 6 Incremental Prediction of Social Judgment and Understanding of Social Norms

	DV = Social Norm Questionnaire			DV = Interpersonal SJT		
	B(SE)	β	*p*	*B*(SE)	β	*p*
Model 1						
ICAR	0.17 (0.04)	0.23	< 0.001	0.57 (0.08)	0.33	< 0.001
Model 2						
ICAR	–0.01 (0.04)	–0.01	0.78	0.27 (0.09)	0.12	0.002
F-H Total	1.09 (0.78)	0.07	0.16	0.88 (1.92)	0.02	0.64
F-H Feelings	0.37 (0.07)	0.26	< 0.001	0.78 (0.18)	0.23	< 0.001
SST	0.29 (0.05)	0.30	< 0.001	0.42 (0.13)	0.18	0.001

Note. β = standardized regression coefficient; ICAR = International Cognitive Ability Resource; F-H = Frith-Happé; Estimates for the SST are calculated using Form A scores; Similar results were found when using Form B scores to predict SNQ (β = 0.32, $p < .001$) and SJT scores (β = 0.28, $p < .001$); SNQ Model 1 $R^2 = 0.05$, $F(1, 385) = 22.03$, $p < .001$; SNQ Model 2 $R^2 = 0.25$, $F(4,382) = 33.86$, $p < .001$; SJT Model 1 $R^2 = 0.11$, $F(1, 385) = 46.31$, $p < .001$; SJT Model 2 $R^2 = 0.21$, $F(4,382) = 26.29$, $p < .001$;

small and large triangle from five response options (see White et al., 2011 for the individual questions). These eight-items displayed better internal consistency (α = 0.61) compared to the categorization task and had corrected item-total correlations ranging between 0.12 and 0.45. Performance on this task was positively related to scores on SST forms A ($r = .48$, 95% CI = [0.40, 0.55], $p < .001$) and B ($r = .44$, 95% CI = [0.35, 0.51], $p < .001$). These correlations were stronger than the observed correlation between overall performance on the Frith-Happé categorization items and the feelings items ($r = .36$, 95% CI = [0.27, 0.44], $p < .001$). Scores from each SST form also accounted for 14% of incremental variance in Frith-Happé feelings scores beyond what could be explained by general intelligence task performance ($\Delta R^2 = 0.14$, $F = 52.44$, $p < .001$). SST scores also accounted for incremental variance in overall performance on the Frith-Happé categorization task beyond the effects of general intelligence ($\Delta R^2 = 0.12$, $F = 28.64$, $p < .001$).

Lastly, we examined whether scores on the SST and Frith-Happé tasks accounted for incremental variance in social knowledge after controlling for individual differences in general intelligence (Table 6). Both SST and Frith-Happé feelings scores were unique predictors of social norm knowledge ($\Delta R^2 = 0.20$, $F = 35.81$, $p < .001$). Likewise, both tasks were also unique predictors of interpersonal skill as measured by the SJT ($\Delta R^2 = 0.10$, $F = 17.61$, $p < .001$). Frith-Happé categorization task scores were not found to be a statistically significant predictor in either model. Although we only report the models when using SST Form A scores in Table 6, we observed the same pattern of results when using scores on Form B. These results provide further support for the validity of the shorter, 14-item SST forms as a correlate of individual differences in social norm understanding and knowledge of effective interpersonal behavior.

General Discussion

Our study is one of the first to explore how adults with ASD perform on a self-administered, online SI test compared to adults without ASD (e.g., Livingstone et al., 2021). Regarding our first study aim, our data in Study 1 provide evidence for measurement invariance for the SST between adults with ASD and a large normative sample of 1,049 participants

without ASD. In support of our second aim, we observed modest group mean SST score differences between adults without and with an ASD diagnosis ($d = 0.21$). We provide a histogram of SST scores for participants with and without ASD in Study 1 (Fig. 2). These results suggest that the SST holds promise as a valid, online, remote assessment of SI for adults in either clinical or subclinical populations. Unlike self- or observer-reported measures of autistic traits which have been found to correlate with personality traits in non-clinical samples (Ingersoll et al., 2011; Schwartzman et al., 2016), past research also indicates that SST scores are practically unrelated to self-reported personality or trait emotional intelligence scores (Brown et al., 2019). We further explored our second aim in Study 2 by observing that both the SST and the Frith-Happé feelings task were unique predictors of understanding of social norms and knowledge of effective behavior in social situations, even after controlling for general intelligence scores. In addition, scores on the SST forms were positively related to performance on the Frith-Happé feelings task and demonstrated better internal consistency relative to the Frith-Happé categorization task. These findings suggest that the SST may be useful as a complement to many of the popular existing self- or observer-reported measures.

These findings are especially promising given the growing need for valid, online, self-administered assessments. Although past research has documented the development of web-based general cognitive ability or intelligence tests (e.g., Brown & Grossenbacher 2017; Liu et al., 2020; Sliwinski et al., 2018; Wright, 2020), few performance-based SI tests besides the RMET have been used online without a proctor. Even though our participants completed the SST outside of a clinical setting and using their own device (e.g., tablet, laptop, or desktop computer), we did not detect any degradation in measurement precision or item validity. Based on these results, the SST appears useful for assessing SI while allowing participants to complete the test remotely without having to travel to a clinic or research site. We also designed alternate, 14-item short forms of the SST which can be used while retaining modest reliability and demonstrating similar validity evidence to what has been reported for the full 23-item version. These forms also displayed convergent validity with other ability measures of social intelligence and knowledge of socially acceptable behavior. This may help researchers recruit larger samples or may make participating in research studies more accessible to potential participants. Based on findings from recent research, the use of animation in the SST may also create a more engaging and enjoyable experience for participants relative to text-based assessments (Karakolidis et al., 2021).

Even though the SST was not explicitly designed to detect ASD or other developmental disorders or to quantify traits related to ASD, the test does appear to be somewhat sensitive to differences in SI between groups of participants with and without ASD. After controlling for demographic differences, we also found that adults without ASD also scored higher on the SST compared to adults without ASD but who are parents of children with ASD. These effect sizes were lower than the difference between patients with schizophrenia and controls reported for the SAT-MC ($d = 0.64$; Pinkham et al., 2018) and for the Frith-Happé ToM task ($d = 0.58$; Wilson, 2021). However, the SST displayed several potential advantages compared to other existing animated shape tasks. The 14-item alternate SST forms displayed slightly stronger test-retest reliability compared to estimates reported for the SAT-MC in prior research ($r = .55$ for controls and $r = .57$ for patients in Pinkham et al., 2017). Both SST forms displayed better internal consistency relative to the Frith-Happé categorization task. The reliability estimates that we observed for the Frith-Happé categorization task were substantially worse than those reported by Livingstone et al. (2021) and suggest that continued research is needed to determine whether these items can adequately assess social intelligence when self-administered online. Similar weaknesses were recently documented by Andersen et al. (2022) who also reported weak reliability for the categorization task in a large sample of adolescents. We argue that these results indicate that the SST may provide a more reliable measure of social intelligence in studies involving adults with and without ASD. Still, further test development work may help improve the sensitivity and further optimize the SST for assessing ability differences within clinical populations. We recommend that future researchers use the 14-item versions of the SST reported in Study 2 given that these forms provided good test-retest reliability and similar internal consistency and convergent validity as previously reported for the 23-item version. We provide the item order and text for both 14-item forms along with all of the video files on the Open Science Framework (https://osf.io/sqxy6).

Implications and Directions for Future Research

We hope that our findings help provide future researchers with the tools to further explore novel ways of assessing social intelligence or similar, more narrowly defined abilities. Researchers have long struggled to develop measures of social intelligence which are empirically distinct from general mental ability or intelligence (Lievens & Chan, 2010). Despite some recent attempts to explain how social intelligence fits within a broader framework of human abilities (e.g., Bryan & Mayer 2021; MacCann et al., 2014), much of the research on social intelligence has been siloed within different subfields where construct labels and measurement methods are often inconsistent (Olderbak & Wilhelm,

2020). This makes it challenging for researchers to integrate findings from different fields and to replicate results from different populations or research settings.

Another important avenue for future research is to determine the boundary conditions for administering the SST online. In our samples, adults with ASD were able to complete the SST outside of a controlled research or clinic setting. However, many of these adults appear to be relatively high functioning, based on their self-reported educational attainment. Future studies should seek to identify criteria which would help researchers determine whether a participant could be expected to provide valid responses in a self-administered, online assessment. Likewise, our samples only included participants who were 18 years of age or older even though animated shape tasks have been used to measure SI among children and adolescents (Altschuler et al., 2018; Burger-Caplan et al., 2016; Salter et al., 2008). Given that the SST items were designed to require as little reading as possible and thus be more independent of verbal ability or language skills, we expect that the test can be used in younger populations, but this has yet to be explored empirically. Thus, further research is needed to observe how this test functions when administered to younger participants in clinical or nonclinical populations.

Future research is also needed to observe the heritability and genetic predictors of SST scores. Twin studies have estimated genetic contributions to individual differences in social cognition (Isaksson et al., 2019) and measures of social functioning (Constantino & Todd, 2003). More specifically, a recent study reported a heritability estimate of 28% for performance on the RMET (Warrier et al., 2018). We would expect to find similar heritability estimates for the SST and similar animated shape tasks based on the convergent validity evidence with the RMET, but this has yet to be empirically observed. This work could potentially determine whether different measures of SI share common genetic influences and how distinct those influences may be from those which predict performance on more general intelligence or cognitive tests.

Study Limitations

There are some limitations on the results we report in this paper. We observed that the SST provided adequate, but not ideal, internal consistency for adults with or without ASD ($0.60 < \alpha < 0.80$). Based on these results, the 23-item and 14-item forms of the SST are best suited for research purposes where even modest reliability may be sufficient for detecting true effects (Schmitt, 1996). These forms are not reliable enough for high-stakes, diagnostic use, where Nunnally (1978) suggests a threshold of $\alpha \geq 0.90$. The modest level of reliability for the SST may have attenuated our observed mean differences between adults with or without ASD in Study 1. Another limitation is that we did not obtain a consistent measure of cognitive ability or intelligence for all participants in Study 1. Therefore, we were only able to control for coarse-grained educational attainment as a proxy for differences in cognitive functioning between adults with or without ASD. Our results in Study 2 indicate that performance on the SST is positively correlated with performance on a general intelligence task (Form A $r = .47$, $p < .001$; Form B $r = .43$, $p < .001$). However, we also found evidence that SST scores do correlate with performance on other social intelligence tasks even after controlling for differences in general intelligence.

Conclusion

Across two studies, we detected differences in social intelligence between adults with and without ASD using a remotely administered, freely available online test. Not only did we find support for measurement invariance between adults with or without an ASD diagnosis, but we also detected modest group mean differences where adults without ASD achieved higher SST scores compared to those with ASD. This effect was still present even after controlling for demographic differences between these two groups. We also designed a shorter, 14-item version of the test in Study 2. These forms provided good test-retest reliability and greater internal consistency compared to the Frith-Happé tasks. We also found that SST scores were related to knowledge of social norms and effective interpersonal behavior even after controlling for differences in general intelligence. These results indicate that the SST is a promising tool for measuring SI, especially in situations where in-person, on-site assessments are either impractical or not possible. Although future research is needed to further optimize the SST and boost its reliability for clinical purposes, this tool may help researchers obtain a quantitative measure of SI while avoiding some of the practical or psychometric limitations of other existing instruments.

Acknowledgements We thank Lauren Walsh for her research assistance and Antoinette DiCriscio, Cora Taylor, and Vanessa Troiani for their helpful feedback on our manuscript. We also thank Jamie Toroney, Kiely Law, and the SPARK Consortium for their assistance in recruiting participants for this study. Additional thanks to Sarah White for giving us permission to use the Frith-Happé animation task and to Michelle Martin-Raugh for providing the situational judgment test administered in Study 2. MIB is now at Human Resources Research Organization (HumRRO) and PRH is now at the Consumer Financial Protection Bureau.

Author's Contribution All authors contributed to the study conception and design. Data preparation, analyses, and interpretation of results were performed by MIB with support from PRH and supervision from

CFC. All authors contributed to, provided feedback on, and approved the research, analyses, and manuscript.

Funding This work was supported by funds from the National Institutes of Health (grant U01MH119705).

Declarations

Conflict of Interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institution and followed the guidelines of the 1964 Helsinki Declaration. Both studies were approved by the Institutional Review Board at Geisinger. Study 1 was also approved by the SPARK Foundation.

Informed Consent All participants gave informed consent to participate in each study.

References

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*, 1–16. https://doi.org/10.1016/S0885-2014(00)00014-9.

Abrahams, L., Pancorbo, G., Primi, R., Santos, D., Kyllonen, P., John, O. P., & De Fruyt, F. (2019). Social-emotional skill assessment in children and adolescents: advances and challenges in personality, clinical, and educational contexts. *Psychological Assessment*, *31*, 460–473. https://doi.org/10.1037/pas0000591.

Altschuler, M., & Faja, S. (2022). Brief report: test-retest reliability of cognitive, affective, and spontaneous theory of mind tasks among school-aged children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *52*, 1890–1895. https://doi.org/10.1007/s10803-021-05040-6.

Altschuler, M., Sideridis, G., Kala, S., Warshawsky, M., Gilbert, R., Carroll, D., Burger-Caplan, R., & Faja, S. (2018). Measuring individual differences in cognitive, affective, and spontaneous theory of mind among school-aged children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *48*, 3945–3957. https://doi.org/10.1007/s10803-018-3663-1.

Andersen, N. K., Rimvall, M. K., Jeppesen, P., Bentz, M., Jepsen, J. R. M., Clemmensen, L., & Olsen, E. M. (2022). A psychometric investigation of the multiple-choice version of animated Triangles Task to measure theory of mind in adolescence. *Plos One*, *17*, e0264319. https://doi.org/10.1371/journal.pone.0264319.

Baksh, R. A., Abrahams, S., Bertlich, M., Cameron, R., Jany, S., Dorrian, T., Baron-Cohen, S., Allison, C., Smith, P., MacPherson, S. E., & Auyeung, B. (2021). Social cognition in adults with autism spectrum disorders: validation of the Edinburgh Social Cognition Test (ESCoT). *The Clinical Neuropsychologist*, *35*, 1275–1293. https://doi.org/10.1080/13854046.2020.1737236.

Baron-Cohen, S., O'Riordan, M., Jones, R., Stone, V. E., & Plaisted, K. (1999). A new test of social sensitivity: detection of faux pas in normal children and children with Asperger syndrome. *Journal of Autism and Developmental Disorders*, *29*, 407–418. https://doi.org/10.1023/A:1023035012436.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001a). The "Reading the mind in the Eyes" test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, *42*, 241–251. https://doi.org/10.1111/1469-7610.00715.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001b). The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*, 5–17. https://doi.org/10.1023/A:1005653411471.

Bell, M. D., Fiszdon, J. M., Greig, T. C., & Wexler, B. E. (2010). Social attribution test – multiple choice (SAT-MC) in schizophrenia: comparison with community sample and relationship to neurocognitive, social cognitive and symptom measures. *Schizophrenia Research*, *122*, 164–171. https://doi.org/10.1016/j.schres.2010.03.024.

Biagianti, B., Fisher, M., Brandrett, B., Schlosser, D., Loewy, R., Nahum, M., & Vinogradov, S. (2019). Development and testing of a web-based battery to remotely assess cognitive health in individuals with schizophrenia. *Schizophrenia Research*, *208*, 250–257. https://doi.org/10.1016/j.schres.2019.01.047.

Brewer, N., Young, R. L., & Barnett, E. (2017). Measuring theory of mind in adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *47*, 1927–1941. https://doi.org/10.1007/s10803-017-3080-x.

Brown, M. I., & Grossenbacher, M. A. (2017). Can you test me now? Equivalence of GMA tests on mobile and non-mobile devices. *International Journal of Selection and Assessment*, *25*, 61–71. https://doi.org/10.1111/ijsa.12160.

Brown, M. I., Ratajska, A., Hughes, S. L., Fishman, J. B., Huerta, E., & Chabris, C. F. (2019). The social shapes test: a new measure of social intelligence, mentalizing, and theory of mind. *Personality and Individual Differences*, *143*, 107–117. https://doi.org/10.1016/j.paid.2019.01.035.

Brown, M. I., Speer, A. B., Tenbrink, A. P., & Chabris, C. F. (2022). Using game-like animations of geometric shapes to simulate social interactions: an evaluation of group score differences. *International Journal of Selection and Assessment*, *30*, 167–181. https://doi.org/10.1111/ijsa.12375.

Bryan, V. M., & Mayer, J. D. (2021). Are people-centered intelligences psychometrically distinct from thing-centered intelligences? A meta-analysis. *Journal of Intelligence*, *9*, 48. https://doi.org/10.3390/jintelligence9040048.

Burger-Caplan, R., Saulnier, C., Jones, W., & Klin, A. (2016). Predicting social and communicative ability in school-age children with autism spectrum disorder: a pilot study of the Social Attribution Task, multiple choice. *Autism*, *20*, 952–962. https://doi.org/10.1177/1362361315617589.

Constantino, J. N., Davis, S. A., Todd, R. D., Schindler, M. K., Gross, M. M., Brophy, S. L., & Reich, W. (2003). Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. *Journal of Autism and Developmental Disorders*, *33*, 427–433. https://doi.org/10.1023/A:1025014929212.

Constantino, J. N., & Todd, R. D. (2003). Autistic traits in the general population. *Archives of General Psychiatry*, *60*, 524–530. https://doi.org/10.1001/archpsyc.60.5.524.

Corcoran, R., Mercer, G., & Frith, C. D. (1995). Schizophrenia, symptomatology and social influence: investigating "theory of mind" in people with schizophrenia. *Schizophrenia Research*, *17*, 5–13. https://doi.org/10.1016/0920-9964(95)00024-G.

Daniels, A. M., Rosenberg, R. E., Anderson, C., Law, J. K., Marvin, A. R., & Law, P. A. (2012). Verification of parent-report of child autism spectrum disorder diagnosis to a web-based autism registry. *Journal of Autism and Developmental Disorders*, *42*, 257–265. https://doi.org/10.1007/s10803-011-1236-7.

Dodell-Feder, D., Ressler, K. J., & Germine, L. T. (2020). Social cognition or social class and culture? On the interpretation of differences in social cognitive performance. *Psychological Medicine*, *50*, 133–145. https://doi.org/10.1017/S003329171800404X.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T., & Convit, A. (2006). Introducing MASC: A movie for the Assessment of Social Cognition. *Journal of Autism and Developmental Disorders*, *36*, 623–636. https://doi.org/10.1007/s10803-006-0107-0.

Fombonne, E., Green Snyder, L., Daniels, A., Feliciano, P., Chung, W., & SPARK Consortium. (2020). Psychiatric and medical profiles of autistic adults in the SPARK cohort. *Journal of Autism and Developmental Disorders*, *50*, 3679–3698. https://doi.org/10.1007/s10803-020-04414-6.

Gourlay, C., Collin, P., Caron, P. O., D'Auteuil, C., & Scherzer, P. B. (2020). Psychometric assessment of social cognitive tasks. Applied Neuropsychology: Adult. Advance online publication. https://doi.org/10.1080/23279095.2020.1807348

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*, 243–259. https://doi.org/10.2307/1416950.

Hurley, R. S. E., Losh, M., Parlier, M., Reznick, J. S., & Piven, J. (2007). The broad autism phenotype questionnaire. *Journal of Autism and Developmental Disorders*, *37*, 1679–1690. https://doi.org/10.1007/s10803-006-0299-3.

Ingersoll, B., Hopwood, C. J., Wainer, A., & Donnellan, M. B. (2011). A comparison of three self-report measures of the broader autism phenotype in a non-clinical sample. *Journal of Autism and Developmental Disorders*, *41*, 1646–1657. https://doi.org/10.1007/s10803-011-1192-2.

Isaksson, J., Westeinde, A. V., Cauvet, E., Kuja-Halkola, R., Lundin, K., Neufeld, J., Willfors, C., & Bolte, S. (2019). Social cognition in autism and other neurodevelopmental disorders: a co-twin control study. *Journal of Autism and Developmental Disorders*, *49*, 2838–2848. https://doi.org/10.1007/s10803-019-04001-4.

Johannesen, J. K., Fiszdon, J. M., Weinstein, A., & Ciosek, D. (2018). The Social Attribution Task – multiple choice (SAT-MC): psychometric comparison with social cognitive measures for schizophrenia research. *Psychiatry Research*, *262*, 154–161. https://doi.org/10.1016/j.psychres.2018.02.011.

Karakolidis, A., O'Leary, M., & Scully, D. (2021). Animated videos in assessment: comparing validity evidence from and test-takers' reactions to an animated and a text-based situational judgment test. *International Journal of Testing*, *21*, 57–79. https://doi.org/10.1080/15305058.2021.1916505.

Kittel, A. F. D., Olderbak, S., & Wilhelm, O. (2022). Sty in the mind's eye: a meta-analytic investigation of the nomological network and internal consistency of the "Reading the mind in the Eyes" test. *Assessment*, *29*, 872–895. https://doi.org/10.1177/1073191121996469.

Klin, A. (2000). Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: the social attribution task. *Journal of Child Psychiatry and Allied Disciplines*, *41*, 831–846. https://doi.org/10.1111/1469-7610.00671.

Kramer, J. H., Mungas, D., Possin, K. L., Rankin, K. P., Boxer, A. L., Rosen, H. J., Bostrom, A., Sinha, L., Berhel, A., & Widmeyer, M. (2014). NIH EXAMINER: conceptualization and development of an executive function battery. *Journal of the International Neuropsychological Society*, *20*, 11–19. https://doi.org/10.1017/S1355617713001094.

Lee, H. S., Corbera, S., Poltorak, A., Park, K., Assaf, M., Bell, M. D., Wexler, B. E., Cho, Y. I., Jung, S., Brocke, S., & Choi, K. H. (2018). Measuring theory of mind in schizophrenia research: cross-cultural validation. *Schizophrenia Research*, *201*, 187–195. https://doi.org/10.1016/j.schres.2018.06.022.

Lievens, F., & Chan, D. (2010). Practical intelligence, emotional intelligence, and social intelligence. In J. L. Farr, & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 342–364). New York, NY: Routledge.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: a versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433–442. https://doi.org/10.3758/s13428-016-0727-z.

Liu, M., Rea-Sandin, G., Foerster, J., Fritsche, L., Brieger, K., Clark, C., Li, K., Pandit, A., Zajac, G., Abecasis, G. R., & Vrieze, S. (2020). Validating online measures of cognitive ability in genes for good, a genetic study of health and behavior. *Assessment*, *27*, 136–148. https://doi.org/10.1177/1073191117744048.

Livingston, L. A., Carr, B., & Shah, P. (2019). Recent advances and new directions in measuring theory of mind in autistic adults. *Journal of Autism and Developmental Disorders*, *49*, 1738–1744. https://doi.org/10.1007/s10803-018-3823-3.

Livingston, L. A., & Happé, F. (2017). Conceptualising compensation in neurodevelopmental disorders: reflections from autism spectrum disorder. *Neuroscience and Biobehavioral Reviews*, *80*, 729–742. https://doi.org/10.1016/j.neubiorev.2017.06.005.

Livingstone, L. A., Shah, P., White, S. J., & Happé, F. (2021). Further developing the Frith-Happé animations: a quicker, more objective, and web-based test of theory of mind for autistic and neurotypical adults. *Autism Research*, *14*, 1905–1912. https://doi.org/10.1002/aur.2575.

Ludwig, N. N., Hecht, E. E., King, T. Z., Revill, K. P., Moore, M., Fink, S. E., & Robins, D. L. (2020). A novel social attribution paradigm: the dynamic interacting shape clips (DISC). *Brain and Cognition*, *138*, 105507. https://doi.org/10.1016/j.bandc.2019.105507.

Luyten, P., Campbell, C., Allison, E., & Fonagy, P. (2020). The mentalizing approach to psychopathology: state of the art and future directions. *Annual Review of Clinical Psychology*, *16*, 297–325. https://doi.org/10.1146/annurev-clinpsy-071919-015355.

MacCann, C., Joseph, D. L., Newman, D. A., & Roberts, R. D. (2014). Emotional intelligence is a second-stratum factor of intelligence: evidence from hierarchical and bifactor models. *Emotion*, *14*, 358–374. https://doi.org/10.1037/a0034755.

Martinez, G., Mosconi, E., Daban-Huard, C., Parellada, M., Fananas, L., Gaillard, R., Fatjo-Vilas, M., Krebs, M. O., & Amado, I. (2019). "A circle and a triangle dancing together": Alteration of social cognition in schizophrenia compared to autism spectrum disorders. *Schizophrenia Research*, *210*, 94–100. https://doi.org/10.1016/j.schres.2019.05.043.

Mayer, J. D., & Salovey, P. (1993). The intelligence of emotional intelligence. *Intelligence*, *17*, 433–442. https://doi.org/10.1016/0160-2896(93)90010-3.

McDonald, S. (2012). New frontiers in neuropsychological assessment: assessing social perception using a standardised instrument, the awareness of Social Inference Test. *Australian Psychologist*, *47*, 39–48. https://doi.org/10.1111/j.1742-9544.2011.00054.x.

Morrison, K. E., Pinkham, A. E., Kelsven, S., Ludwig, K., Penn, D. L., & Sasson, N. J. (2019). Psychometric evaluation of social cognitive measures for adults with autism. *Autism Research*, *12*, 766–778. https://doi.org/10.1002/aur.2084.

Murano, D., Lipnevich, A. A., Walton, K. E., Burrus, J., Way, J. D., & Anguiano-Carrasco, C. (2020). Measuring social and emotional skills in elementary students: Development of self-report Likert, situational judgment test, and forced choice items. Personality and Individual Differences. Advance online publication. https://doi.org/10.1016/j.paid.2020.110012

Neubauer, A. C., & Hofer, G. (2022). Retest-) reliable and valid despite low alphas? An example from a typical performance situational judgment test of emotional management. *Personality and Individual Differences*, *189*, 111511. https://doi.org/10.1016/j.paid.2022.111511.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Oakley, B. F. M., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: a cautionary note on the reading

the mind in the eyes test. *Journal of Abnormal Psychology*, *125*, 818–823. https://doi.org/10.1037/abn0000182.

Olderbak, S., & Wilhelm, O. (2020). Overarching principles for the organization of socioemotional constructs. *Current Directions in Psychological Science*, *29*, 63–70. https://doi.org/10.1177/0963721419884317.

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: toward a brief form for research and applied settings. *Frontiers in Psychology*, *6*, 1503–1520. https://doi.org/10.3389/fpsyg.2015.01503.

Peterson, E., & Miller, S. F. (2012). The eyes test as a measure of individual differences: how much of the variance reflects verbal IQ? *Frontiers in Psychology*, *3*, 220. https://doi.org/10.3389/fpsyg.2012.00220.

Pinkham, A. E., Harvey, P. D., & Penn, D. L. (2018). Social cognition psychometric evaluation: results of the final validation study. *Schizophrenia Bulletin*, *44*, 737–748. https://doi.org/10.1093/schbul/sbx117.

Pinkham, A. E., Kelsven, S., Kouros, C., Harvey, P. D., & Penn, D. L. (2017). The effect of age, race, and sex on social cognitive performance in individuals with schizophrenia. *Journal of Nervous Mental Disorders*, *205*, 346–352. https://doi.org/10.1097/NMD.0000000000000654.

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, *15*, 384–396. https://doi.org/10.1177/1745691619896607.

Ratajska, A., Brown, M. I., & Chabris, C. F. (2020). Attributing social meaning to animated shapes: a new experimental study of apparent behavior. *American Journal of Psychology*, *133*, 295–312. https://doi.org/10.5406/amerjpsyc.133.3.0295. https://www.jstor.org/stable/.

Revelle, W., Dworak, E. M., & Condon, D. (2020). Cognitive ability in everyday life: the utility of open-source measures. *Current Directions in Psychological Science*, *29*, 358–363. https://doi.org/10.1177/0963721420922178.

Rossel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48, 1–36. URL http://www.jstatsoft.org/v48/i02/

Salter, G., Seigal, A., Claxton, M., Lawrence, K., & Skuse, D. (2008). Can autistic children read the mind of an animated triangle? Autism, 12,349–371. https://doi.org/10.1177/1362361308091654

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, *19*, 65–72. https://doi.org/10.1016/j.tics.2014.11.007.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350–353. https://doi.org/10.1037/1040-3590.8.4.350.

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., & Kanske, P. (2021). Toward a hierarchical model of social cognition: a neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, *147*, 293–327. https://doi.org/10.1037/bul0000303.

Schwartzman, B. C., Wood, J. J., & Kapp, S. K. (2016). Can the five factor model of personality account for the variability of autism symptom expression? Multivariate approaches to behavioral phenotyping in adult autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *46*, 253–272. https://doi.org/10.1007/s10803-015-2571-x.

Sliwinski, M. J., Mogle, J. A., Hyun, J., Munoz, E., Smyth, J. M., & Lipton, R. B. (2018). Reliability and validity of ambulatory cognitive assessments. *Assessment*, *25*, 14–30. https://doi.org/10.1177/1073191116643164.

Soto, C. J., Napolitano, C. M., & Roberts, B. W. (2020). Taking skills seriously: Toward an integrative model and agenda for social, emotional, and behavioral skills. Current Directions in Psychological Science. Advance online publication. https://doi.org/10.1177/0963721420978613

The SPARK Consortium. (2018). SPARK: a US cohort of 50,000 families to accelerate Autism research. *Neuron*, *97*, 488–493. https://doi.org/10.1016/j.neuron.2018.01.015.

Tippins, N. T. (2015). Technology and assessment in selection. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 551–582. https://doi.org/10.1146/annurev-orgpsych-031413-091317.

Türközer, H. B., & Öngür, D. A. (2020). A projection for psychiatry in the post-COVID-19 era: potential trends, challenges, and directions. *Molecular Psychiatry*, *25*, 2214–2219. https://doi.org/10.1038/s41380-020-0841-2.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–69. https://doi.org/10.1177/109442810031002.

Vandewouw, M. M., Safar, K., Mossad, S. I., Lu, J., Lerch, J. P., Anagnostou, E., & Taylor, M. J. (2021). Do shapes have feelings? Social attribution in children with autism spectrum disorder and attention-deficit/hyperactivity disorder. *Translational Psychiatry*, *11*, 493. https://doi.org/10.1038/s41398-021-01625-y.

Velikonja, T., Fett, A. K., & Velthorst, E. (2019). Patterns of nonsocial and social cognitive functioning in adults with autism spectrum disorder. *JAMA Psychiatry*, *76*, 135–151. https://doi.org/10.1001/jamapsychiatry.2018.3645.

Warrier, V., Grasby, K. L., Uzefovsky, F., Toro, R., Smith, P., Chakrabarti, B., et al. (2018). Genome-wide meta-analysis of cognitive empathy: heritability, and correlates with sex, neuropsychiatric conditions and cognition. *Molecular Psychiatry*, *23*, 1402–1409. https://doi.org/10.1038/mp.2017.122.

Webster, E. S., Paton, L. W., Crampton, P. E. S., & Tiffin, P. A. (2020). Situational judgment test validity for selection: a systematic review and meta-analysis. *Medical Education*, *54*, 888–902. https://doi.org/10.1111/medu.14201.

White, S. J., Coniston, D., Rogers, R., & Frith, U. (2011). Developing the Frith-Happé animations: a quick and objective test of theory of mind for adults with autism. *Autism Research*, *4*, 149–154. https://doi.org/10.1002/aur.174.

Wilson, A. C. (2021). Do animated triangles reveal a marked difficulty among autistic people with reading minds? *Autism*, *25*, 1175–1186. https://doi.org/10.1177/1362361321989152.

Wright, A. J. (2020). Equivalence of remote, digital administration and traditional, in-person administration of the Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V). *Psychological Assessment*, *32*, 809–817. https://doi.org/10.1037/pas0000939.