

Psychometric assessment of social cognitive tasks

Catherine Gourlay , Pascal Collin , Pier-Olivier Caron , Camille D'Auteuil & Peter B. Scherzer

To cite this article: Catherine Gourlay , Pascal Collin , Pier-Olivier Caron , Camille D'Auteuil & Peter B. Scherzer (2020): Psychometric assessment of social cognitive tasks, Applied Neuropsychology: Adult, DOI: [10.1080/23279095.2020.1807348](https://doi.org/10.1080/23279095.2020.1807348)

To link to this article: <https://doi.org/10.1080/23279095.2020.1807348>



Published online: 25 Aug 2020.



Submit your article to this journal [↗](#)






View related articles [↗](#)



View Crossmark data [↗](#)

Psychometric assessment of social cognitive tasks

Catherine Gourlay^a , Pascal Collin^a, Pier-Olivier Caron^b , Camille D'Auteuil^a, and Peter B. Scherzer^a 

^aDepartment of Psychology, Université du Québec à Montréal, Montréal, Canada; ^bUniversité TÉLUQ, Montréal, Canada

ABSTRACT

Although there has been a marked increase in interest in social cognition (SC) in recent years, psychometric data relating to many tasks used to measure its components remain limited in healthy populations with only five articles published to date. It is accordingly premature to speak of a consensus concerning the specific components, or best tests of the components, and possible cultural differences. The present study sought to partially fill that gap, examining the psychometric properties of a battery of SC tasks in a sample of 100 healthy adults aged 18–85 years old. Initially, nine tasks assessing four SC components were selected: emotion recognition, theory of mind, attributional bias, and social judgment. Construct validity and criterion-related validity were assessed using factor and correlational analyses. Performance across age and sex groups was also investigated. Reliability was assessed through internal consistency, interrater and intercoder agreement. Results indicated satisfactory properties for the Ambiguous Intentions Hostility Questionnaire-blame score, the Social Judgment Task, the Facial Emotions Recognition Test, and a modified version of the Strange Stories Task. Statistically significant differences were found between the groups with regard to age and sex after accounting for demographic and cognitive factors. However, the correlations of these measures with relationship quality were mostly very low, raising questions about their concomitant validity. Other tasks showed sub-optimal properties, suggesting that some frequently used tests require further validation or modifications to ensure the quality of research findings. Based on the results, recommended measures for future studies and limitations are discussed.

KEYWORDS

Emotion recognition; psychometric; social cognition; social judgment; theory of mind

Social cognition (SC) focuses on intraindividual cognitive processes to explain social interactions and interpersonal functioning. In recent years, there has been increasing interest in identifying key domains, or components involved in social information processing. Clinical research in autism spectrum disorders (ASD) and schizophrenia has made significant contributions and resulted in important advances in improving our understanding of various aspects of SC. In this context, the National Institute of Mental Health (NIHM; Green et al., 2008) identified five distinct domains that are thought to capture the complex phenomenology of SC deficits in schizophrenia: theory of mind, social perception, emotional processing, social knowledge, and attributional bias.

Social perception refers to the identification of social cues in the environment, such as faces, voices, and gestures (Green et al., 2015). Emotion processing can be parsed as the ability to perceive and recognize emotions from immediate and observable cues, such as facial expressions and prosody (Adolphs, 2002). Theory of mind (ToM) has been described as the ability to reason and attribute mental and emotional states to oneself and others by integrating multiple sources of information. ToM includes first-order

(inferring someone's mental state) and second-order inferences (a belief that someone knows what someone else believes or thinks) (Baron-Cohen, 2001; Sabbagh, 2004). Social knowledge refers to the knowledge about social norms, rules and relations that guide appropriate social behavior (Beer & Ochsner, 2006), which can then be used to judge the appropriateness of behavior in a specific context based on social conventions and standards. Attributional style is typically regarded in schizophrenia research as the individual's tendency to make inferences about the causes of positive and negative events as being the results of internal (personal), external (another person), or situational circumstances. Two attributional biases are of particular interest (see Achim et al., 2016; Kinderman & Bentall, 1997; Langdon et al., 2006, 2013): externalizing bias (i.e., the tendency to attribute negative causes to the actions of others and positive causes to one's own actions), and personalizing bias (i.e., attributing adverse events to oneself).

In line with the NIHM developments, the Social Cognition Psychometric Evaluation (SCOPE) study (Pinkham et al., 2014) identified similar components (emotion processing, social perception, theory of mind, attributional style/bias) and assessed the quality of SC tasks based on various criteria to

reach a consensus on measurement. Their subsequent psychometric evaluations (Buck et al., 2016, 2017; Ludwig et al., 2017; Pinkham et al., 2016) yielded mixed results, as several measures that were examined showed inconsistent results for both clinical and control groups. This led the authors to suggest further investigation of their candidate measures as well as of new SC tasks. These results support the need to (1) continue the examination process for some promising measures, such as the Ambiguous Intentions and Hostility Questionnaire (Combs et al., 2007); (2) reconsider the initial measures selected for evaluation by the RAND panelists of the SCOPE project (e.g., Strange Stories, Happé, 1994; Faux Pas test, Baron-Cohen et al., 1999; Internal Personal and Situational Attributions Questionnaire, Kinderman & Bentall, 1997); (3) further examine the psychometric properties of other instruments found in the literature that could be useful to assess the aspects of SC that have previously been identified (e.g., Picture Sequencing task, Langdon et al., 2014; Task of Attribution of Intention to Others, Brunet et al., 2003; Interpersonal Reactivity Index, Davis, 1983).

Furthermore, investigation of the SC components and how they relate to social behavior and functional outcomes remains limited in healthy populations. Their performance across SC components is typically assessed through comparison with clinical samples with the primary goal of measuring group differences. Previous studies have consistently shown that healthy subjects obtain significantly higher scores on various SC measures when compared with patients diagnosed with disorders such as attention deficit disorder (ADHD), ASD and schizophrenia (Bora & Pantelis, 2016; Chung et al., 2014; Eack et al., 2013; Healey et al., 2016; Savla et al., 2013), anxiety (Plana et al., 2014), depression and bipolar disorder (Weightman et al., 2014; Samamé et al., 2012), and neurodegenerative disorders (Elamin et al., 2012). Beyond the pathology-focused approach, there is a need to build a more inclusive and normative approach of SC. However, there is an important lack of information concerning the psychometric properties of many of these tests, especially but not limited to the general population, which makes it difficult to understand and accurately interpret the results of many studies. Most measures were designed for use in clinical settings (e.g., autism) to identify and quantify SC impairments in many disorders. The large discrepancy in the distribution of scores between patients and healthy subjects frequently led to measurement limitations (e.g., ceiling, floor effects) in non-clinical samples, which may reduce data quality and hamper accurate evaluation of their SC abilities. Floor and ceiling effects negatively affect test sensitivity and statistical power, and consequently, tend to increase the rate of false negatives and positives (Cramer & Howitt, 2005). As well, these effects indicate scores that have minimal variation among individuals and little or no range for improvement or decline in performance in borderline high-functioning clinical or non-clinical individuals. Furthermore, many studies do not include an assessment of functional outcome, a measure of concomitant validity, thus limiting our understanding of the impact of SC abilities in everyday life situations. Additional SC measures or further

validation of existing measures in healthy populations is clearly needed.

Prior research provided evidence for associations between age and facets of SC, with older subjects performing appreciably worse than younger individuals on a variety of emotion identification (e.g., Gonçalves et al., 2018) and ToM tasks (e.g., Moran, 2013). It is not clear, however, whether there are sex differences in SC: while some results indicate no between-sex variations (Di Tella et al., 2020), other studies suggest differences in emotion recognition (Kirkland et al., 2013; Montagne et al., 2005) and ToM tasks involving faux pas understanding (Ahmed & Stephen Miller, 2011), cartoons (Russell et al., 2007), and video-based scenarios (Wacker et al., 2017). Investigating age and sex differences to further establish discriminant validity may thus be of importance in a validation process. There is also empirical support for an overlapping relationship between SC and neurocognition (Adolphs, 2001), which makes it important to control for cognitive abilities while investigating SC performance in different groups.

Objectives

The present study was undertaken to examine the psychometric properties of a battery of SC tests in a sample of healthy francophone adults aged 18–85 years old. These properties consist of the reliability and validity of nine measures related to the domains of SC identified by the NIHM and SCOPE study. A second objective was to examine age and sex-related variations across different components of SC.

Methods

Participants

One hundred subjects aged 18–85 years old were recruited to take part in the study. The majority of the subjects' education had to be in French. In actuality, the first language, home language and language of daily use of all participants was French. Exclusion criteria included a history of neurological disorders such as head injury or epilepsy, or psychiatric disorders, as determined by participants' answers to an intake questionnaire administered by phone. Subjects were recruited through electronic and printed advertisements posted in and around the city of Montreal, and through word of mouth. Adults aged 50 years old and older were screened using the Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005). None of the participants retained for the study scored below the cut-off score of 26. Table 1 shows the sociodemographic characteristics of the participants. Participation was voluntary and transportation expenses were reimbursed. The study was approved by the Human Sciences scientific and ethics committee of the *Université du Québec à Montréal*.

Table 1. Sociodemographic characteristics.

Variables	Participants <i>M (SD) or Nb. (%)</i>
Age (years)	<i>47.7 (18.18)</i>
Range	18–85
Sex	
Male	50 (50.00)
Female	50 (50.00)
Ethnicity	
Caucasian	98 (98.00)
Caribbean/African American	1 (1.00)
Asian	1 (1.00)
Vocational status	
Employed	51 (51.00)
Student	16 (16.00)
Retired	33 (33.00)
Education (years)	<i>15.34 (2.92)</i>
Education in French (years)	<i>14.90 (3.10)</i>
Highest education level	
High school	20 (20.00)
General and vocational college	30 (30.00)
University	50 (50.00)

Note: $N = 100$. Means and standard errors are italicized.

Procedure

Forward and back translation

A translation-back translation procedure was performed on the SC measures originally in English: the Ambiguous Intentions Hostility Questionnaire (Combs et al., 2007), Interpersonal Reactivity Index (Davis, 1983), Social Judgment Task (Langdon et al., 2014), Internal Personal and Situational Attributions Questionnaire (Kinderman & Bentall, 1997), and Picture Sequencing task (Langdon et al., 2014). The Faux Pas test (Baron-Cohen et al., 1999) and Strange Stories task (Happé, 1994; White et al., 2009) were previously used in another study (Scherzer et al., 2012, 2015) and were found to be acceptable to all subjects. The tests were translated into French by two authors (CG, PC) and back-translated into English by the bilingual coauthor whose first language is English (PS). All discrepancies between researchers were discussed in order to find agreement on a common version of every translated test. A consensus was reached on the first translated version. The measures were then pilot-tested on 10 French-speaking adults to ensure that potential cross-cultural differences were addressed. If necessary, adjustments were made to strengthen semantic and conceptual equivalence. The final version of each test was used for data collection. Since there were only slight changes in wording between the pilot and final versions, it was decided to include the pilot-tested individuals in the main study. Two tasks initially developed for use in French were administered with no changes: Task of Attribution of Intention to Others (Brunet et al., 2003) and Facial Emotions Recognition Task (Gaudelus et al., 2014).

Data collection

Participants were tested individually for two 1.5-h sessions on the SC tasks, as well as a battery of neurocognitive tests. The sessions were conducted using a counterbalanced measures design, within and between sessions to limit order and fatigue effects. Only one enrollee had incomplete data values, which were coded as missing values. Initial data

obtained by preliminary analysis of 34 subjects ($M_{\text{age}} = 39.18 \pm 16.72$, 50% men) then 51 subjects ($M_{\text{age}} = 42.27 \pm 17.48$, 57% men) led to the decision to reject three measures (see the “Results” section). One-third of the sample was asked to be re-tested 3-weeks later in order to obtain a measure of test-retest reliability, but all declined. Given that participation was voluntary, test fatigue combined with the fact that there was no monetary incentive to participate in another 3 h of testing may explain this decision. Average time between session 1 and 2 was 14.96 ± 17.65 days.

Measures

The selected tasks consisted of nine tests of different components of SC identified by the NIMH (Green et al., 2008) and SCOPE study (Pinkham et al., 2014). In addition to the tests that were evaluated by the RAND panelists of the SCOPE project, several tests were added for their relevance to SC (i.e., the degree to which a test may assess one of the key domains that have been previously identified). Moreover, two novel tasks were selected on the basis of their clinical and ecological features: Facial Emotions Recognition Test (TREF; Gaudelus et al., 2014) and Social Judgment Task (SJT; Langdon et al., 2014). In addition to accuracy of emotion recognition, the TREF includes a detection threshold for each emotion as a measure of intragroup and intergroup sensitivity to basic emotions. This measure was found to have adequate discriminant validity (clinical versus controls) in the original report by the authors.

The SJT is a measure of social judgment, one of the SC components identified by the SCOPE panel (Pinkham et al., 2014). This judgment has been associated with a broad range of deficits across various neurodevelopmental conditions such as intellectual disability (American Psychiatric Association, 2013), ADHD (Demopoulos et al., 2013; Humphreys et al., 2016; Whalen et al., 1990) and ASD (Loveland et al., 2001). Criterion-related (concurrent) validity was assessed through a validated questionnaire used to measure the quality of interpersonal relationships (see Table 2 for the battery). Detailed information about the tests and complete administration procedures can be found in the original publications. Two tests (Strange Stories and Faux Pas test, see “Adapted Measures” section) initially developed for clinical trials in autism research were adapted to more reliably assess the range of performance in healthy adult subjects. The adaptations were made to reflect the increase in complexity in ToM development, beyond childhood (Wellman, 2018).

Several other tasks inventoried in the literature were developed in non-clinical settings (e.g., Geneva Emotion Recognition Test, Schlegel et al., 2014; Situational Test of Emotion Understanding and Situational Test of Emotion Management, MacCann & Roberts, 2008; Mayer-Salovey-Caruso Test of Emotional Intelligence, Mayer et al., 2003). However, the tests that were evaluated in the present study were selected, in part based on the existing repertoire of frequently used tasks in SC, in order to identify those that demonstrate suboptimal qualities and should be modified and validated for future utilization.

Finally, neurocognitive measures were added in order to distinguish between, and control for the contribution of these abilities and SC abilities in group comparisons. Three subtests from the Wechsler Adult Intelligence Scale-4th edition (WAIS-IV; Wechsler, 2008) were used as control for fluid abilities (Matrix Reasoning, abstract thinking and problem-solving), and verbal abilities (Vocabulary, crystallized knowledge; Similarities, verbal reasoning).

Adapted measures

Strange Stories-Revised (SS-R). The Strange Stories test was initially developed by Happé (1994) to discriminate between control subjects and autistic children, adolescents and young adults in terms of ToM (i.e., inferences about others' mental states, detection of sarcasm/irony, lies or bluffs). Although Happé was successful in distinguishing between the clinical and control groups in her study, many studies found a ceiling effect (i.e., 90%+) in control groups of healthy children, adolescents or adults (see for example Kaland et al., 2005; Ouellet et al., 2010; Rogers et al., 2007). To reduce the likelihood of a ceiling effect, we adapted the questions used by White et al. (2009) and some details within the stories (e.g., several character's names to fit the cultural context) to fit the current sample (adults aged 18–85 years old).

The SS-R consists of eight ToM stories interspersed by seven physical comprehension stories. All stories were presented to the participants in a written format and remained in front of them while reading and responding to minimize the demands placed on memory and attention. A control question was used at the end of each ToM story to identify any comprehension deficit. Each ToM story required 1st and 2nd order inferences that were rated by two judges (CG, CD). In the version used by White et al. (2009), two questions were asked after each ToM story: "Is it true, what the character says?" and "Why did the character do this?." In order to prevent measurement limitation with healthy subjects, we increased the level of difficulty in terms of mental state attribution in including 1st order items (e.g., What does X think?) and more complex 2nd order opened-questions (e.g., What does X think of the woman's reaction?) after each ToM story. Half of the ToM vignettes contained a question related to affective ToM (e.g., How does X feel?). Explicit 2nd order inferences were rated 2 points, 1st order beliefs and partial responses without clear evidence of a 2nd order process were rated 1 point. Non-relevant answers and the absence of mental state inferences were rated 0 points. A ToM composite score was calculated by summing the scores obtained in all ToM stories (maximum of 59 points, excluding the control questions). A Comprehension score was calculated as the sum of the scores obtained on the physical comprehension stories (maximum of 11 points).

Faux Pas Recognition Test-Revised (Faux Pas-R). The Faux Pas Recognition Test was created by Baron-Cohen et al. (1999) to assess ToM through social faux pas (i.e., recognizing that someone made a social mistake) in children with autism. The test was slightly modified for use with brain-injured adults (Stone et al., 1998) and patients with

dementia (Gregory et al., 2002). In both studies involving adults, however, most or all control subjects ToM score was 100%, revealing a limitation in test variability. In the current study, the test was adapted to measure 1st and 2nd order inferences of intentions and emotions. The questions were modified, but the stories remained the same as those previously used in the adult version except for minor changes in characters' names for cultural adaptation.

In the version published by Stone et al. (1998), each vignette was followed by eight questions that required subjects to Q1- and Q2- detect a faux pas (Did anyone say something they shouldn't have said or something awkward?; Who said something they shouldn't have said or something awkward?); Q3- understand inappropriate behaviors (Why shouldn't he/she have said it or why was it awkward?); Q4- infer a character's intentions (Why do you think he/she said it?); Q5- infer a character's belief (Did X know/realize that Y?); Q6- infer a character's feelings (How do you think X felt?); Q7- and Q8- understand the story (comprehension questions). The Faux Pas-Revised was composed of a set of 10 stories used by Baron-Cohen et al. (1999) and Stone et al. (1998), all of which were followed by questions inquiring about one's beliefs about the characters' mental states following the faux pas. For each story, all the questions had the same format. Specifically, after each story, participants were asked how they think the characters felt after the faux pas (questions #1, #4—inference of emotions: How does X feel? How does Y feel?) and why the main protagonist acted this way (question #2—inference of intentions: According to X, why did Y say that?). Participants were also asked why the two characters felt this way after they realize the protagonist's gaffe (questions #3, #5—inference of emotions: Why does X feel this way? Why does Y feel this way?). At the end of each story, a comprehension question (question #6) was included, as in the original version. The stories were placed in front of the participants according to the instructions provided by Stone et al. (1998) so that subjects could refer to them while answering the questions. The stories were presented randomly to control for a possible order effect. Two points were awarded for each 2nd order response and one point was given for every 1st order response. The absence of an inference was scored zero. The scores were summed to provide a total ToM score (maximum of 80 points), excluding the comprehension questions. The ratio of correct answers relative to the correct comprehension answers had to be equal to or greater than 70%, for the ToM composite score to be considered valid. See Table 2 for a description of each test and questionnaire. A summary of their psychometric characteristics found in the available literature is presented in Table 3. Only studies involving more than 100 healthy adults were included in the summary to limit the probability of sampling bias. Most of the reported studies used convenience samples of college students aged 20–25 years old.

Data analysis

Statistical analyses were performed using IBM SPSS 26. Data were examined to detect potential measurement limitations and percentages of extreme scores were calculated to identify

Table 2. Social cognition tasks and outcome measure.

Test	Description	Component evaluated	Range of scores
Social cognition			
Facial Emotions Recognition Task (TREF) (Gaudelus et al., 2014)	Identification of six emotions (happiness, anger, sadness, fear, disgust, contempt) associated with facial expression of emotions, at nine intensity levels ranging from 20–100% in a set of 54 photos. In addition to accuracy scores, a detection threshold was derived, operationalized as the lowest level of intensity that participants could detect an emotion. The test is computer administered.	Emotion processing	<ul style="list-style-type: none"> Accuracy score ranging from 0 to 9 for each emotion and 0–54 for total accuracy performance Detection threshold ranging from 20% to 100% for each emotion
Task of Attribution of Intention to Others (TAIO) (Brunet et al., 2003)	Attribution of intentions to comic strips characters along with two control conditions: <i>Physical causality</i> and <i>Causality with characters</i> .	ToM	<ul style="list-style-type: none"> Total score ranging from 0 to 14 for each condition
Picture Sequencing Task (PST) (Langdon et al., 2014)	Attribution of intentions to story characters acting on the basis of false beliefs in a series of comic-strip-like pictures presented in a random order, to be rearranged in chronological order. There are four types of stories: <i>Attribution of Intentions</i> , <i>Mechanical</i> , <i>Social Script</i> , <i>Capture</i> .	ToM	<ul style="list-style-type: none"> Mean score ranging from 0 to 6 for each type of sequence
Interpersonal Reactivity Index (IRI) (Davis, 1983)	Self-report measure containing four seven-item subscales, each assessing an aspect of empathy: fantasy, perspective taking, empathic concern, personal distress. Items are rated on a five-point Likert-type scale ranging from “does not describe me well” to “describes me very well.”	Empathy	<ul style="list-style-type: none"> Total score ranging from 0 to 28 for each subscale
Internal Personal and Situational Attributions Questionnaire (IPSAQ) (Kinderman & Bentall, 1997)	Sixteen positive and 16 negative situations are categorized by respondents as being something due to themselves (internal attribution, 1pt), to others (external-personal, 2pts), or circumstances (external-situational, 3pts). A mean score was computed for positive and negative scales. Two attributional biases were coded according to authors' guidelines: externalizing bias (EB; positive scores refer to a tendency to attribute negative events to external causes) and personalizing bias (PB; scores higher than .5 refer to greater attributions of negative events to personal than to situational factors).	Attribution style/bias	<ul style="list-style-type: none"> Mean scores ranging from 1 (internal) to 3 (situational) for positive, negative scales Externalizing score ranging from -16 to +16 and Personalizing score ranging from 0 to 1
Ambiguous Intentions Hostility Questionnaire (AIHQ) (Combs et al., 2007)	Participants read clips of 15 short negative social situations that vary in terms of intentionality (intentional, accidental, or ambiguous intention). The cause of each situation and the way subjects would respond are coded by two raters to obtain hostility and aggression mean scores. The extent to which respondents think the other person acted on purpose, how angry it would make them feel, and how much they would blame the other person is rated on a Likert-type scale and averaged in a blame index.	Attribution style/bias	<ul style="list-style-type: none"> Hostility and Aggression biases: mean scores ranging from 1 to 5 Blame score: mean score ranging from 3 to 16
Social Judgment Task (SJT) (Langdon et al., 2014)	Make a judgment on whether the behaviors described in five short stories are normal, unusual, or shocking. Behaviors are labeled as: 1-socially appropriate; 2-violation of social norms; 3-inappropriate but understandable if the characters' thoughts are taken into account.	Social judgment	<ul style="list-style-type: none"> Mean score of correct judgments Percentage of judgments (normal, unusual or shocking) in each category of behaviors
Strange Stories-Revised	see “Adapted Measures” section for description	ToM	<ul style="list-style-type: none"> Comprehension score ToM total score
Faux Pas-Revised	see “Adapted Measures” section for description	ToM	<ul style="list-style-type: none"> ToM total score
Outcome Measure			
Interpersonal Relationship Quality Scale (IRQS) (Senécal et al., 1992)	Brief questionnaire containing 20 items that assess an individual's quality of interpersonal relationships through five subscales: family, love partner, friends, other students/colleagues, people in general.	Interpersonal relations	<ul style="list-style-type: none"> Total score ranging from 0 to 16 for each subscale Mean total score for all domains of relationships

Note. In the translated version of the Social Judgment Task, the term “a pair of underwear” (Story 2) was replaced by “toothbrush” for cultural adaptation considering its high degree of inappropriateness in the Quebec-French culture.

possible floor or ceiling effects. Floor and ceiling effects were defined as the proportion of participants with scores higher than 90% (ceiling) or less than 10% (floor) on a given test. A *p*-value of .05 or less was considered statistically significant throughout the analyses, and reliability was assessed using Cronbach's alpha. For coefficient interpretation, we used the most frequently cited acceptable range of Cronbach's alpha of .70 or above (Nunnally, 1978). Intraclass correlation coefficients (ICCs; -1 to +1) and 95% confidence intervals (CI) were used in intercoder reliability analyses (i.e., the degree of agreement between coders/judges). ICCs were interpreted in terms of absolute

agreement using a two-way random effect model. As suggested by Portney and Watkins (2000), ICCs were interpreted as follows: >0.75 = good; from 0.50 to 0.75 = moderate, and < 0.50 = poor. Cohen's κ coefficient was used to report the degree of agreement between participants on categorical scales (interrater reliability). Cohen's κ coefficients were categorized as follows: values ≤ 0 indicating no agreement, .01–.20 as slight, .21–.40 as fair, .41–.60 as moderate, .61–.80 as substantial, and .81–1.00 as almost perfect (Cohen, 1960). Construct validity of two tasks was tested through data reduction techniques, and correlational analyses (Pearson's *r* values, which varies between -1

Table 3. Summary of the psychometric characteristics of the SC tasks in past studies comprising samples of 100+ non-clinical adults.

Test	Study	Participants	Reliability	Validity	Normality of distribution
Facial Emotions Recognition Task of Attribution of Intention to Others Picture Sequencing Task	Gaudelus et al. (2014)*	Sample of healthy adult subjects < 100	N/A	N/A	N/A
	Brunet et al. (2003)*	Sample of healthy adult subjects < 100	N/A	N/A	N/A
	Langdon et al. (2014)*	Sample of healthy adult subjects < 100	N/A	N/A	N/A
	Schneider et al. (2020)	Netherlands/Belgium. Study 2: N = 101, age = 40.8, sex = 28.7% men; Study 3: N = 349, age = 38.6, sex = 43.8%	Unreported	Discriminant validity: Patients performed worse than controls on the false-belief scores when controlled for age and sex, but the effect was no longer significant when controlling for IQ and Control scores Convergent validity: PST not associated with social functioning	Authors reported ceiling effects: 31.68% had maximum scores on ToM; 51.49% had maximum scores on control conditions
Interpersonal Reactivity Index	Davis (1980)	USA. (1980, study 3). N = 1161 college students, 49.9% men	Internal consistency: Men: FS, $r = .78$; EC, $\alpha = .72$; PD, $\alpha = .78$; PT, $\alpha = .75$; Women: FS, $\alpha = .75$; EC, $\alpha = .70$; PD, $\alpha = .78$; PT, $\alpha = .78$ Test-retest: Men: FS, $r = .79$; EC, $r = .72$; PD, $r = .68$; PT, $r = .61$; Women: FS, $r = .81$; EC, $r = .70$; PD, $r = .76$; PT, $r = .62$	Construct validity: Four factors confirmed but model fit indices were not reported	Unreported
	Davis (1983)*	USA. (1983). N = 1344 college students, sex = 50.4% men	Unreported	Discriminant validity: Sex differences in all subscales, with woman displaying higher scores on each scale	Unreported
Task of Attribution of Intention to Others Picture Sequencing Task	Pulos et al. (2004)	USA. N = 409 college students, age = 22.1, sex = 30.1% men	Internal consistency: FS, $\alpha = .82$; EC, $\alpha = .80$; PD, $\alpha = .75$; PT, $\alpha = .79$	Convergent/divergent validity: All subscales were related to other measures (interpersonal functioning, self-esteem, emotionality, sensitivity to others, empathy) in men and women	Unreported
	De Corte et al. (2007)	Dutch. N = 651, age of men = 27.4, age of women = 27.4, sex = 46.0% men	Internal consistency for total sample: FS, $\alpha = .83$; EC, $\alpha = .73$; PD, $\alpha = .77$; PT, $\alpha = .73$	Construct validity: Four first-order factors corresponding to the four scales, and two second-order factors corresponding to empathy, emotional control Construct validity: Using CFA, original four-factor model showed reasonable fit to data. Improvement was needed; Modified four-factor model showed improved fit Convergent/divergent validity: All scales related to other instruments (emotional quotient, personality traits, self-esteem, machiavellism)	Unreported
	Huang et al. (2012)	China. N = 930 teachers (samples 1-3), age = [29.3-32.1], sex = 29.9% men	Internal consistency for samples 1-3: FS, $\alpha = [.76-.85]$; EC, $\alpha = [.70-.83]$; PD, $\alpha = [.71-.79]$; PT, $\alpha = [.67-.74]$	Discriminant validity: Sex differences in all subscales, with women scoring higher than men on all scales Construct validity: CFA indicated acceptable fit in four-factor model Convergent/divergent validity: All scales related to other constructs (social anxiety, shyness, transgression, self-esteem, agreeableness) Discriminant validity: Sex differences in FS, EC, PD; For PT, EC, men teachers and of general population scored higher than prisoners, and male teachers scored higher than men of general population	Unreported
Task of Attribution of Intention to Others Picture Sequencing Task	Fernández et al. (2011)	Chile. N = 435 college students, age = 20.1, sex = 46.2% men	Internal consistency for total sample: FS, $\alpha = .76$; EC, $\alpha = .73$; PD, $\alpha = .70$; PT, $\alpha = .73$ Test-retest: Men: FS, $r = .82$; EC, $r = .89$; PD, $r = .81$; PT, $r = .67$; Women: FS, $r = .76$; EC, $r = .81$; PD, $r = .78$; PT, $r = .67$	Construct validity: Four-factor model and second-order model indices suggested good model fit, except for CFI (<.90) in both cases Predictive validity: EC, PT, PD associated with other instruments assessing self-esteem, trait anxiety, aggression, social avoidance and distress, and emotionality Discriminant validity: Sex differences favoring woman in FS, EC, PD	Unreported

Chiang et al. (2014)	Taiwan. <i>N</i> = 516 college students, age = 24.5, <i>ed</i> = 14.6, sex = 43.7% men	Internal consistency: FS, α = .75; EC, α = .71; PD, α = .75; PT, α = .73 Test-retest: FS, r = .72; EC, r = .80; PD, r = .76; PT, r = .80	Construct validity: EFA indicated a four-factor model, improvement was needed; After deleting items, modified four-factor model showed greater variance explained Convergent/divergent validity: PT, EC scales related to self-esteem and aggression Discriminant validity: Differences by sex in EC, PD, PT; Difference in EC between control and patients after controlling for education and IQ Construct validity: CFA of a one-factor model was poor; Two-factor model showed poor fit to data; Original four-factor model showed acceptable indices, except for CFI (<.90) Convergent validity: EC, PT, PD related to emotional quotient Discriminant validity: Sex differences in FS, EC; Differences by age groups in FS, PD; Interaction between age, sex significant for PD only	Unreported	Authors mentioned that skewness, kurtosis for total score was close to zero, suggesting normality of distribution
Gilet et al. (2013)	Switzerland. <i>N</i> = 322, age = 49.5, sex = 41.0% men	Internal consistency: FS, α = .81; EC, α = .70; PD, α = .78; PT, α = .71 Test-retest: FS, ICC = .86; EC, ICC = .77; PD, ICC = .85; PT, ICC = .71	Construct validity: CFA of a two-factor model (affective, cognitive empathy) showed poor fit to data; Hierarchical (PT, FS, EC) and original four-factor models showed improved TLI/CFI values Discriminant validity: Sex differences in all subscales	Unreported	Unreported
Chrysikou and Thompson (2016)	USA. <i>N</i> = 417, age = 33.2, sex = 41.0% men	Unreported	Construct validity: CFA of a two-factor model (affective, cognitive empathy) showed poor fit to data; Hierarchical (PT, FS, EC) and original four-factor models showed improved TLI/CFI values Discriminant validity: Sex differences in all subscales	Unreported	Unreported
Lucas-Molina et al. (2017)	Spain. Sample 1: <i>N</i> = 2499 college students, age = 21.1, sex = 28.8% men; Sample 2. <i>N</i> = 1438 adults, age = 40.0, sex = 42.2% men	Internal consistency for students: FS, α = .79; EC, α = .72; PD, α = .72; PT, α = .74	Construct validity: Acceptable fit of the CFA four-factor model for student sample except for CFI/TLI values (<.90); ESEM four-factor model showed improved fit to data, especially in students; Evidence of equivalence of four-factor structure between sex in sample of students Discriminant validity: Sex differences favoring women in FS, EC, PD	Unreported	Authors mentioned that data were normally distributed
Kinderman and Bentall (1997)*	Sample of healthy adult subjects < 100	N/A	Discriminant validity: Sex differences favoring women in FS, EC, PD	N/A	N/A
Larøi and Bédard (2001)	Belgium. <i>N</i> = 243 undergraduate students, age = 24.8, sex = 28.0% men	Internal consistency: PI, α = .69; PP, α = .60; PS, α = .57; NI, α = .77; NP, α = .62; NS, α = .73; EB, .71; PB, .72	Concurrent validity: PI, NI, EB related to other attributional style variables	Unreported	Unreported
Gao et al. (2018)	China. <i>N</i> = 200, age = 20.7, sex = 29.0% men	Internal consistency: PI, α = .71; PP, α = .68; PS, α = .69; NI, α = .69; NP, α = .67; NS, α = .74 Interrater: PI, r = .57; PP, r = .58; PS, r = .61; NI, r = .80; NP, r = .72; NS, r = .56; EB, r = .80; PB, r = .54; all <i>p</i> -values < .001	Discriminant validity: Difference in EB score between more/less depressed nonclinical groups; Difference between less depressed nonclinical sample and delusional patients on NS score	Unreported	Unreported
Combs et al. (2007)*	USA. <i>N</i> = 322 undergraduate students, age = 19.6, <i>ed</i> = 13.7, sex = 51.6% men	Internal consistency: BS-intentional, α = .85; BS-ambiguous, α = .86; BS-accidental, α = .84 Interrater: HB, ICC = [.91-.99]; AB, ICC = [.93-.99]	Incremental validity: Blame scores predicted incremental variance in paranoia over demographics, attribution style, and psychosis proneness; Blame, hostility scores in ambiguous scenarios were significant individual predictors of paranoia Convergent/divergent validity: 5/9 scales related to other constructs (paranoia, hostility, personalizing bias, perceptual aberration, magical ideation) Discriminant validity: Sex differences in HB for ambiguous, intentional scenarios, and in AB for intentional scenarios	Unreported	Mean scores, ranges of scores provided, but information referring to normality of distribution not provided

(continued)

Table 3. Continued.

Test	Study	Participants	Reliability	Validity	Normality of distribution
	Jeon et al. (2013)	South Korea. $N = 263$, age = 21.1, ed = 13.4, sex = 50.6% men	Internal consistency: BS, $\alpha = .61$ –.68] across scenarios	Convergent validity: HB in ambiguous situations related to ToM, anger; BS in ambiguous situations related to anger, trait anxiety Discriminant validity: Sex differences in blame scores in ambiguous situations Convergent, incremental validity: Information not reported for controls	Authors mentioned that skewness, kurtosis of all scores were in acceptable range (<1.0)
	Pinkham et al. (2016)	USA. $N = 104$, age = 39.2, ed = 13.4, sex = 47.0% men	Internal consistency: HB, $\alpha = .85$; BS, $\alpha = .34$; AB, $\alpha = .47$ Test-retest: HB, $r = .57$; BS, $r = .76$; AB, $r = .70$	Discriminant validity: Group differences in HB, BS, with patients scoring higher than controls	Authors mentioned that score distributions were checked for normality (skewness, kurtosis, visual inspection) and no measures required transformation
	Zajenkowska et al. (2018)	$N = 161$, age = 36.7, sex = 35.4% men	Internal consistency in three-factor model: HB, $\alpha = .33$; BS, $\alpha = .83$; AB, $\alpha = .50$; Internal consistency in six-factor model: $\alpha = [.50$ –.87] Interrater: HB and AB items, ICC = [0.83–0.97] Unreported	Construct validity: CFA of the three-factor model (hostility, blame, aggression) showed poor fit to data; EFA suggested a six-factor-solution based on types of scenarios Discriminant validity: Group differences in levels of anger, blame, aggression in some scenarios Construct validity: Satisfactory fit of the CFA five-factor model (only blame-related items were included); Some support for five-factor structure across cultures Discriminant validity/Cultural variations: Differences in patterns of hostile attributions across cultures based on type of social relationship involved in the scenarios N/A	Authors indicated that data were screened for normality of distribution when performing EFA Unreported
Social Judgment Task	Zajenkowska et al. (2020)	USA/Poland/Japan. N Poland = 203, age = 25.8, sex = 45.3% men; N USA = 230, age = 18.7, sex = 25.2% men; N Japan = 274, age = 19.8, sex = 80.3% men Sample of healthy adult subjects < 100	N/A	N/A	N/A
Strange Stories	Happé (1994)*	Sample of healthy adult subjects < 100	N/A	N/A	N/A
	Ahmed and Stephen Miller (2011)—short version	$N = 123$, age = 19.0, sex = 42.7% men	Interrater: Total score, $\rho = [.87$ –.89]	Convergent validity: FP scores unrelated to other ToM tasks	Authors indicated a violation of normality assumption
Faux Pas	Baron-Cohen et al. (1999)*	Sample composed of children only	N/A	N/A	N/A
	Ferguson and Austin (2010)	$N = 162$, age = 34.1, sex = 29.0% men	Internal consistency: FP score, $\alpha = .95$	Convergent validity: FP scores related to other domains or constructs (complex emotion recognition, emotional intelligence, agreeableness) Convergent validity: Total scores unrelated to other ToM tasks	Unreported
	Ahmed and Stephen Miller (2011)	$N = 123$, age = 19.0, sex = 42.7% men	Interrater: Total score, $\rho = [.89$ –.96]	Convergent validity: Total scores unrelated to other ToM tasks	Authors indicated a violation of normality assumption
	Faisca et al. (2016)	Portugal. $N = 200$, age = 33.0, sex = 37.5% men	Internal consistency in one-factor structure: FP detection, $\alpha = .82$; FP score, $\alpha = .83$; FP rejection, $\alpha = .57$	Construct validity: EFA and parallel analysis indicated a one-factor model for both detection and FP scores Discriminant validity: Sex differences in detection and FP questions score Convergent validity: FP scores related to another ToM task	Distributions of FP and FP total scores were asymmetric, with a ceiling effect in FP detection and rejection scores
	Lever and Geurts (2016)—short version	Netherlands. $N = 118$, age = 47.7, sex = 70.3% men	Interrater: Concordance rate of 97.5%	Discriminant validity: Group differences in FP scores with controls scoring higher than patients, but differences no longer observed in older adults; Age-related differences in FP total scores	Authors mentioned a violation of normality assumption for almost all dependant variables

Phillips et al. (2015)	N total = 116, N young adults = 40, age = 25.2; N middle-aged = 40, age = 53.4; N older adults = 36, age = 73.9	Unreported	Discriminant validity: No age-related variations in FP scores Convergent validity: FP scores related to the ability to understand sarcasm	Unreported
Negrão et al. (2016)—short version	Brazil. N = 152, age = 22.0, sex = 48.0% men	Internal consistency: $\alpha = .94$	Discriminant validity: Group differences in FP scores, with controls scoring higher than patients	Authors indicated that normality approximations were rejected by statistical test
Zhang et al. (2018)—short version	China. N total = 171; N young adults = 87, age = 25.6, sex = 43.0% men; N older adults = 84, age = 65.5, sex = 40.0% men	Unreported	Discriminant validity: No single effect of age on FP scores in complete sample; Age-related differences in groups not receiving enhanced motivation Convergent validity: FP scores related to another ToM task	Authors reported that younger adults in the control condition exhibited ceiling effects

Note: *: authors' test; N: total number of participants; age: mean age of participants, rounded; ed: education in years, rounded; sex: sex of participants, % rounded; CFA: confirmatory factor analysis; EFA: exploratory factor analysis; CFI: comparative fit index; TLI: Tucker–Lewis index; PT: perspective taking; EC: empathic concern; PD: personal distress; FS: fantasy; PI: positive-internal scale; PP: positive-situational scale; NI: negative-internal scale; NP: negative-personal scale; NS: negative-situational scale; EB: externalizing bias; PB: personalizing bias; HB: hostility bias; BS: blame score; AB: aggression bias; FP: Faux Pas; ToM: theory of mind.

and +1) were used to examine the concurrent validity. Confirmatory factor analyses (CFA) were carried out in Mplus 8 (Muthén & Muthén, 1998–2017) and model fit indices were interpreted following the criteria recommended by Hu and Bentler (1999) and Caron (2018): chi-squared index (χ^2), root mean squared error approximation (RMSEA) < .05, comparative fit index (CFI) and Tucker-Lewis Index (TLI) > .95, and standardized root mean squared residual (SRMR) < .08. In addition, exploratory factor analyses (EFA) were used to identify the underlying structure of the SS-R and IRI. Finally, independent *t*-tests and analysis of covariance (ANCOVA) were used for comparisons of means between age groups (18–49 years old; 50–85 years old) with and without control variables (biological sex, education and neurocognition). Comparisons of means between sex groups (men, women) were performed using education and neurocognition as covariates.

Results

Distributions and rejected measures

During the initial phase of the study, preliminary analyses revealed sub-optimal characteristics in three tasks (Picture Sequencing task, Task of Attribution of Intention to Others, Faux Pas-R). For reasons indicated in Table 4, these tasks were not included in the final battery, and no further data were collected using these tests after this stage. After completion of data collection, data were reexamined. Most variables were normally distributed: SS-R ToM and Comprehension scores, SJT-correct judgments, all subscales included in the IRI (PT, PD, EC, FS), all scores in the AIHQ (Hostility, Blame, Aggression) and all subscales in the IPSAQ (positive, negative). Detection thresholds and five accuracy scores in the TREF (disgust, contempt, anger, sadness, fear) along with the TREF total accuracy score were normally distributed, while happiness tended to cluster further from the mean with higher data values. Some variables included in the SJT were broadly concentrated in one area (see Figure 1), denoting a convergence in judgment. Considering the nature of the test (clustering is thought to represent a natural tendency in judgments rather than an artifact considering the shared standards on which comparisons and judgments are based; Mussweiler, 2003) and how data were analyzed, situations where cases were grouped did not affect the interpretation of results. Proportions of individuals scoring at floor/ceiling on the measures retained for full validation are presented in Table 5. Intercorrelations among SC variables are shown in Table 6.

Psychometric properties of the final set of measures examined in the battery

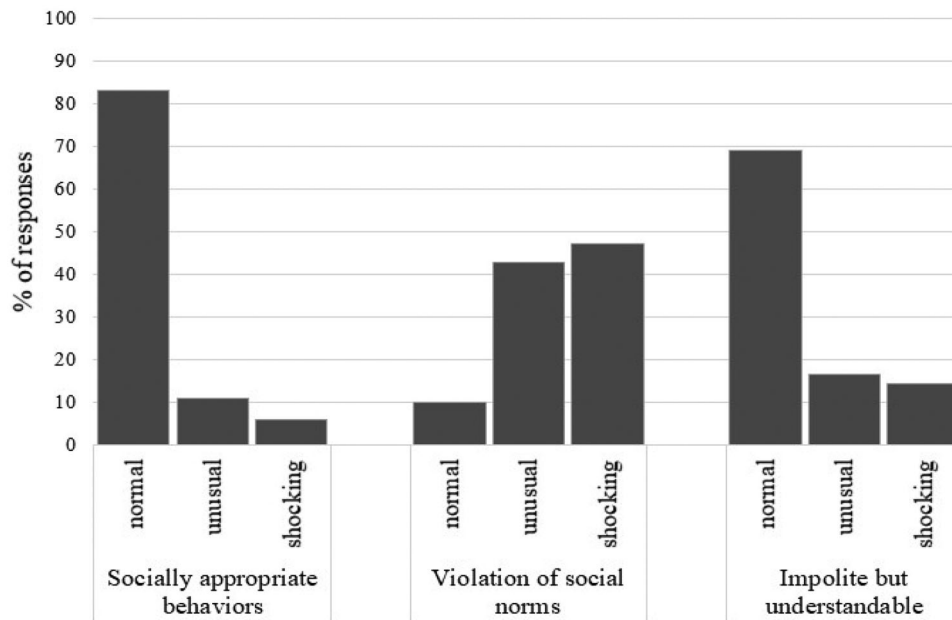
Reliability

Internal consistency. Internal consistency was calculated for five questionnaires. Cronbach's coefficient for the AIHQ-Blame score showed strong reliability ($\alpha = .90$), but the Aggression ($\alpha = .61$) and Hostility subscales ($\alpha = .39$) showed low values. Internal consistency for the negative

Table 4. Summary of rejected measures.

Task	Reason of rejection	Data
Attribution of Intention to Others Task ^a	Large number of extreme scores in all three conditions suggesting ceiling effects	84.2% of subjects obtained a ToM score $\geq 80\%$ and 52.9% had a ToM score $\geq 90\%$ ($M = 11.82$, $SD = 2.61$); 96.0% of subjects obtained a Physical Causality score $\geq 90\%$ ($M = 13.65$, $SD = 1.71$); 94.2% obtained a Causality with Characters score $\geq 90\%$ ($M = 13.61$; $SD = 1.80$).
Picture Sequencing Task ^b	Large number of extreme scores in three conditions suggesting ceiling effects	70.0% of subjects obtained a ToM score $\geq 80\%$; 52.0% of participants obtained a ToM score $\geq 90\%$ ($M = 4.94$, $SD = 1.25$); 82.0% had a Mechanical score $\geq 90\%$ ($M = 5.66$, $SD = .80$); 86.0% obtained a Social Script score $\geq 90\%$ ($M = 5.84$, $SD = 0.40$); 26.0% obtained a Capture score $\geq 80\%$; 6.0% had a score of $\geq 90\%$ ($M = 3.88$; $SD = 1.13$).
Faux Pas-R ^c	Large number of high ToM scores suggesting a ceiling effect	100% of subjects obtained a Comprehension score $>70\%$ ($M = 9.97$; $SD = .17$) thus all ToM scores were valid; 76.1% of subjects obtained a ToM score $\geq 80\%$; 51.6% had a ToM score $\geq 90\%$ ($M = 68.18$; $SD = 9.68$).

Note. ^a $n = 51$ subjects; ^b $n = 50$ subjects; ^c $n = 34$ subjects.

**Figure 1.** Percentage of judgments across categories of behaviors in the Social Judgment Task.**Table 5.** Summary of retained measures for complete validation process.

Task	Range of observed scores	% of participants scoring at ceiling	% of participants scoring at floor
TREF accuracy			
Happiness	3–9	30.00	0
Disgust	2–8	0	0
Sadness	2–9	15.00	0
Fear	3–9	16.00	0
Contempt	0–8	0	4.00
Anger	0–9	4.00	1.00
Total score	16–48	0	0
SS-R			
ToM (raw score)	36–56	4.04	0
Comprehension	7–11	5.05	0
IRI			
Perspective taking	9–28	9.09	0
Empathic concern	10–28	12.12	0
Personal distress	1–25	0	2.02
Fantasy	4–24	0	0
AIHQ			
Hostility bias	1.2–2.67	6.00	0
Aggression bias	1.27–2.93	4.00	0
Blame score	4.8–11.0	0	0
IPSAQ			
Positive events	1.13–2.13	0	0
Negative events	1.25–2.75	0	0
Personalizing bias	0–1	5.00	8.00
Externalizing bias	–13–4	0	1.00
SJT			
Correct judgments	.58–1.00	9.00	0

Note: TREF: Facial Emotions Recognition Task; SS-R: Strange Stories-Revised; IRI: Interpersonal Reactivity Index; AIHQ: Ambiguous Intentions Hostility Questionnaire; IPSAQ: Internal Personal and Situational Attributions Questionnaire; SJT: Social Judgment Task.

Table 6. Intercorrelations among SC variables.

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1 TREF-Happiness	—																				
2 TREF-Disgust	.14	—																			
3 TREF-Sadness	.25*	.29**	—																		
4 TREF-Fear	-.14	.24*	.35**	—																	
5 TREF-Contempt	.04	.15	.31**	.25*	—																
6 TREF-Anger	.34**	.53**	.68**	.62**	.22*	—															
7 TREF-Total	.12	.30**	.50**	.31**	.29**	.60**	—														
8 SS-R ToM	.03	.28**	.30**	.25*	.27**	.21*	.55**	—													
9 IRI-EC	-.07	.09	.07	.03	.21*	.20	.18	.11	.43**	—											
10 IRI-PD	.08	-.07	.12	.03	-.04	-.04	.02	.03	-.07	.02	—										
11 IRI-FS	.08	.05	.15	.21*	.20*	.13	.25*	.23*	.18	.25*	.22*	—									
12 AIHQ-HB	.06	-.10	-.22*	-.10	-.15	-.27**	-.25*	-.13	-.28**	-.10	.26**	.10	—								
13 AIHQ-AB	-.07	-.02	-.04	.02	.00	-.03	-.04	-.03	-.29**	-.01	.20*	.20*	.25*	—							
14 AIHQ-BS	.22	-.05	-.05	.03	.09	-.03	.06	.14	-.14	-.14	.05	.34	.41	.41	—						
15 IPSAQ-Positive	-.09	.01	.15	.23*	.00	-.01	.08	-.05	.14	-.05	.12	-.03	-.11	.03	-.08	—					
16 IPSAQ-Negative	.12	.05	-.03	.00	-.06	.00	.00	-.12	.16	.16	.09	.09	-.22*	-.10	-.21*	.31**	—				
17 IPSAQ-PB	.08	-.03	.04	.10	.07	.00	.08	.07	.10	.08	-.05	.07	.14	-.09	.13	-.02	-.36**	—			
18 IPSAQ-EB	-.17	.00	.15	.14	.01	.00	.07	.08	.00	-.01	-.02	-.05	.06	.15	.12	.42**	-.62**	.04	—		
19 SJT-Correct judgments	-.15	.22*	.15	-.01	.26**	.33**	.27**	.17	.01	.07	-.04	.03	-.24*	-.02	-.02	-.04	-.02	.02	.03	—	

Note: * $p < .05$; ** $p < .01$; TREF: Facial Emotions Recognition Task (Accuracy); SS-R: Strange Stories-Revised; IRI: Interpersonal Reactivity Index; PT: Perspective taking; EC: Empathic concern; PD: Personal distress; FS: Fantasy; AIHQ: Ambiguous Intentions Hostility Questionnaire; HB: Hostility bias; AB: Aggression bias; BS: Blame score; IPSAQ: Internal Personal and Situational Attributions Questionnaire; PB: Personalizing bias; EB: Externalizing bias; SJT: Social Judgment Task.

scale of the IPSAQ was acceptable ($\alpha = .71$), but positive events showed a lower value ($\alpha = .41$). Two subscales of the IRI showed moderate reliability (IRI-PD: $\alpha = .75$; IRI-PT: $\alpha = .69$), while the remaining subscales tended to show reduced values of alpha (IRI-EC: $\alpha = .60$; IRI-FS: $\alpha = .47$) therefore reflecting possible conceptual heterogeneity or greater measurement error. Cronbach's coefficient for the SJT correct judgments showed low reliability for behaviors labeled as Socially appropriate ($\alpha = .27$), Violation of social norms ($\alpha = .36$), and Impolite but understandable ($\alpha = .47$). Similarly, internal consistency was low for TREF accuracy scores: anger ($\alpha = .62$), contempt ($\alpha = .56$), fear ($\alpha = .52$), sadness ($\alpha = .51$), happiness ($\alpha = .47$), and disgust ($\alpha = .24$), except for the Total score, which showed an acceptable value ($\alpha = .74$).

The α for the SS-R was .63. Although this score is below the agreed threshold of .70, some researchers (see for example Hair et al., 2010) suggested a lower limit of acceptability of .60 for exploratory research. Since the SS-R was modified, adapted and used with an initial sample of 100 subjects, the coefficient of .63 at this stage, can be considered to be acceptable.

Interrater reliability. TREF reliability was assessed using Cohen's κ calculations. The results showed high percentages of agreement between participants in recognizing facial emotions: 95.31% for happiness ($\kappa = .94$), 87.04% for fear ($\kappa = .84$), 84.07% for sadness ($\kappa = .81$), 68.15% for anger ($\kappa = .62$), and 66.91% for disgust ($\kappa = .60$). Agreement between participants was reduced for contempt (47.65%) with $\kappa = .37$. The overall κ agreement for all the emotions together was substantial (74.90% agreement, $\kappa = .61$).

Cohen's κ was also used to assess the reliability of each type of behavior in the SJT, based on expected versus observed values. Percentages of responses in each category of behaviors are shown in Figure 1. There was a substantial agreement between subjects in judging Socially appropriate behaviors (83.4% agreement, $\kappa = .75$), which were predominantly rated as "normal." Fair agreement was seen in behaviors labeled Violations of social norms (44.3% agreement, $\kappa = .21$); observers judged them almost equally being "unusual" and "shocking." Finally, while it was expected that participants would rate Impolite but understandable behaviors as "unusual," instead they judged them as being "normal." The resulting κ was consequently worse than expected (22.0% agreement, $\kappa = -.25$), most likely describing a tendency to normalize ambiguity rather than a complete absence of agreement. The degree of agreement was substantial for all judgments taken together (45.4% agreement, $\kappa = .68$).

Intercoder reliability. Intercoder reliability was assessed for two instruments using intraclass correlations. For the AIHQ hostility and aggression scores, a high degree of agreement was found between raters. The average measure ICC for the Hostility subscale was .959 (95% CI: .935–.973, $F(99, 99) = 25.989, p < .001$), while the average measure ICC for the Aggression subscale was .964 (95% CI: .947–.976, $F(99, 99) = 27.757, p < .001$). The average ICC measure of the SS-R

ToM score was .929 with a 95% CI ranging from .892 to .953, $F(98, 98) = 14.656$, $p < .001$, while both raters reached an ICC = .909 with a 95% CI ranging from .862 to .940, $F(98, 98) = 11.621$, $p < .001$, for the Comprehension score.

Construct validity and factor structure

Strange Stories-R. An EFA using principal components analysis was conducted on the ToM score. The scree test indicated a one-factor model that explains 31.47% of the total variance. Table 7 shows that all loadings were above .400. Parallel analysis (Horn, 1965) and the Next Eigenvalue Sufficiency Tests (NEST; Achim, 2017) were used to test the dimensionality of the data set and both confirmed a single component. That component, labeled ToM, is likely to reflect the capacity to attribute mental states to others' thoughts, emotions, and intentions. The ToM raw score was not significantly related to the Comprehension score ($r = .05$; $p = .602$).

Interpersonal Reactivity Index. To assess the fit between the dataset and the four-factor model established in the original IRI (Davis, 1983), a CFA was conducted. The results indicated that the initial four-factor model was a poor fit to the data, $\chi^2(339) = 471.54$, $p < .001$, RMSEA = .066, $p = .043$, CFI = .756, TLI = .728, SRMR = .097. This model misfit may be explained by an insufficient subject-to-item ratio, which in the present study was lower than the recommended 10:1 (Nunnally, 1978). Given that the CFA results failed to confirm a four-factor model, an EFA with varimax rotation was conducted to examine alternative structures. Parallel analysis and EFA results indicated a five-factor model explaining 47.26% of the total variance. This model remained unsatisfactory conceptually and statistically.

Differential performance on SC tasks

Comparison analyses using independent t -tests between two age groups (18–49 years old, $n = 52$; 50–85 years old, $n = 48$) were performed. Such a categorization differs from mainstream research using extreme age groups in comparison analyses, typically young adults in their 20's and seniors aged 65 years old and over. Including middle-aged adults in the younger group aimed at broadening the scope of findings, as suggested by Hess (2006), to better reflect age-related changes occurring before 65 years of age in SC processes (Charlton et al., 2009; Pardini & Nichelli, 2009).

Results indicated age differences in several SC abilities. Specifically, older adults demonstrated significantly lower TREF total accuracy score, $t(98) = 4.64$, $p < .001$, $d = .93$. Detection thresholds were significantly higher (i.e., more difficulty discriminating at low intensity) for negative emotions in older adults.

In terms of inferences, older adults demonstrated significantly lower total SS-R ToM scores than younger adults, $t(97) = 4.04$, $p < .001$, $d = .81$, along with an increased AIHQ-Hostility index, $t(98) = -2.00$, $p = .049$, $d = .40$, and an elevated score on the IPSAQ negative scale, $t(98) = -2.45$, $p = .016$, $d = .49$. Conversely, younger participants

Table 7. Exploratory factor analysis of the SS-R.

Item	Loadings	Item description	Type of mentalistic story
1	.535	The ping-pong paddle	Lie
3	.443	The prisoner	Double bluff
5	.591	Brian's favorite meal	Pretend/Lie
7	.687	The kittens	Persuade
9	.696	Aunt Jane's hat	White lie/Sarcasm
11	.467	The Christmas gift	White lie
13	.447	Mrs. Peabody walking home	Misunderstanding
15	.558	The burglar	Misunderstanding

demonstrated a significantly higher externalization bias (IPSAQ), $t(98) = 3.00$, $p = .003$, $d = .60$.

Older participants were less accurate in social judgments, $t(98) = 2.46$, $p = .016$, $d = .49$, and they rated significantly more behaviors labeled impolite but understandable as being "normal," $t(89.29) = 3.71$, $p < .001$, $d = .75$.

An analysis of covariance (ANCOVA) was used to explore the differences among age groups using sex, years of education and neurocognitive variables as covariables. Results in Table 8 indicated that most variables remained significant after accounting for demographic and cognitive factors. Effect sizes ranged in magnitude from small to large (partial eta squared; $\eta^2 = .02-.20$). Sex differences were also investigated (Table 8) using the following covariates: years of education and cognitive variables. Significant sex-related variations were found in two emotions in the TREF (fear, contempt), and in the SS-R ToM score. Differences were also detected in the empathic concern subscale of the IRI and a subscale in the SJT. Effect sizes (η^2) were moderate (.05–.09).

Concurrent validity and association with the quality of social relationships

The SC tasks were examined in relation to the quality of different types of relationships included in the IRQS (romantic relationship, family, friendships, other students/colleagues, people in general). Parametric tests were used in correlational analysis for all variables. Although visual inspection suggested that the TREF-Happiness subscale was slightly right-skewed, descriptive statistics were not problematic (i.e., acceptable skewness and kurtosis). As for the SJT subscales that were not normally distributed, they were dichotomized to perform the analysis.

None of the total scores were found to correlate significantly, but there were numerous associations between SC subscales and the quality of different aspects of relationships. Significant correlations ranged in magnitude from small to medium ($r = .12-.37$). Although not significant, the association between ToM ability and quality of friendships showed a trend (SS-R; $r = .20$, $p = .052$). Conversely, the QAIPS did not show an association or statistical trend with outcomes as depicted in Table 9.

Discussion

Recent studies focused on understanding the content of SC processes and their association with social functioning (Silberstein & Harvey, 2019). However, the limited amount

Table 8. ANCOVA for the SC performances of two age and sex groups using education and neurocognition as covariates.

Task	18–49 y. <i>M (SD)</i>	50–85 y. <i>M (SD)</i>	<i>p</i>	η^2	Men <i>M (SD)</i>	Women <i>M (SD)</i>	<i>p</i>	η^2
TREF								
<i>Accuracy</i>								
Happiness	7.72 (1.37)	7.69 (1.15)	.645	<.01	7.60 (1.28)	7.84 (1.24)	.319	.01
Disgust	5.50 (1.20)	5.29 (1.22)	.350	.01	5.36 (1.43)	5.48 (0.95)	.517	.01
Sadness	7.38 (1.32)	6.17 (1.62)	<.001	.16	6.82 (1.61)	6.80 (1.57)	.563	<.01
Fear	7.44 (1.40)	6.62 (1.63)	.002	.10	6.80 (1.76)	7.30 (1.28)	.042	.05
Contempt	4.42 (1.89)	3.19 (1.99)	.003	.10	3.62 (2.06)	4.10 (1.99)	.041	.05
Anger	6.02 (1.65)	4.98 (2.12)	.067	.04	5.78 (1.72)	5.26 (2.13)	.395	.01
Total score	38.48 (4.25)	33.94 (5.82)	<.001	.20	35.98 (5.91)	36.78 (5.22)	.068	.04
<i>Detection threshold</i>								
Happiness	26.40 (9.85)	26.04 (8.69)	.920	<.01	26.80 (10.58)	25.40 (7.62)	.488	.01
Disgust	45.40 (13.88)	45.42 (13.36)	.882	<.01	46.60 (14.79)	44.00 (11.95)	.341	.01
Sadness	29.40 (10.58)	28.33 (9.07)	.599	<.01	30.60 (11.32)	27.20 (7.57)	.117	.03
Fear	27.20 (9.04)	31.88 (11.97)	.138	.02	29.00 (11.29)	30.00 (10.10)	.954	<.01
Contempt	38.80 (20.47)	50.63 (24.36)	.022	.06	45.80 (25.16)	42.80 (20.80)	.173	.02
Anger	37.60 (13.79)	44.38 (19.01)	.140	.02	38.20 (13.51)	43.60 (19.03)	.146	.02
IRI								
Perspective taking	20.20 (4.61)	20.04 (4.24)	.818	<.01	20.14 (7.18)	20.16 (4.11)	.862	<.01
Empathic concern	20.64 (4.06)	21.06 (4.04)	.747	<.01	19.71 (4.10)	22.02 (3.65)	.003	.09
Personal distress	10.52 (5.54)	10.79 (4.77)	.967	<.01	9.94 (4.92)	11.26 (5.33)	.279	.01
Fantasy	14.54 (4.88)	13.90 (4.27)	.211	.02	13.59 (4.12)	14.92 (4.93)	.097	.03
SS-R								
ToM score	47.52 (3.95)	44.38 (4.05)	<.001	.15	45.55 (4.48)	46.56 (4.17)	.026	.05
AIHQ								
Hostility bias	1.81 (0.24)	1.93 (0.29)	.120	.03	1.87 (0.27)	1.87 (0.27)	.707	<.01
Blame score	8.00 (1.34)	7.66 (1.21)	.218	.02	7.93 (1.23)	7.75 (1.34)	.843	<.01
Aggression bias	2.01 (0.31)	1.93 (0.34)	.084	.03	1.96 (0.36)	1.99 (0.32)	.477	0.1
IPSAQ								
Positive events	1.63 (0.25)	1.60 (0.26)	.982	<.01	1.62 (0.25)	1.61 (0.26)	.971	<.01
Negative events	1.99 (0.34)	2.16 (0.35)	.021	.06	2.01 (0.36)	2.14 (0.33)	.121	.03
Personalizing bias	0.35 (0.26)	0.40 (0.30)	.474	<.01	0.38 (0.28)	0.38 (0.28)	.880	<.01
Externalizing bias	−3.54 (3.51)	−5.54 (3.20)	.013	.07	−3.78 (3.75)	−5.22 (3.02)	.116	.03
SJT								
Correct judgments	0.83 (0.10)	0.78 (0.10)	.060	.04	0.80 (0.09)	0.80 (0.11)	.503	.01
% “normal” in SA	84.40 (16.31)	80.83 (15.41)	.210	.02	81.60 (16.58)	84.40 (15.27)	.183	.02
% “unusual” in SA	12.00 (16.66)	10.42 (10.91)	.792	<.01	13.60 (16.38)	8.40 (10.76)	.052	.04
% “shocking” in SA	3.60 (8.75)	8.75 (14.24)	.048	.04	4.80 (10.34)	7.20 (13.25)	.607	<.01
% “normal” for VSN	12.00 (14.24)	7.74 (11.00)	.087	.03	12.20 (14.82)	8.29 (11.21)	.253	.01
% “unusual” for VSN	45.43 (22.11)	39.88 (22.24)	.082	.03	39.14 (19.32)	46.57 (24.46)	.048	.04
% “shocking” for VSN	42.57 (22.44)	52.68 (23.97)	0.10	.07	48.85 (20.42)	45.43 (26.40)	.230	.02
% “normal” in IU	76.29 (17.70)	61.30 (22.05)	.005	.09	71.71 (19.90)	66.28 (21.94)	.600	<.01
% “unusual” in IU	14.00 (15.41)	19.35 (16.00)	.264	.01	15.43 (15.77)	18.00 (15.76)	.665	<.01
% “shocking” in IU	9.71 (12.73)	19.34 (17.31)	.011	.07	12.86 (14.50)	15.71 (16.89)	.822	<.01

Note: TREF: Facial Emotions Recognition Task; SS-R: Strange Stories-Revised; AIHQ: Ambiguous Intentions Hostility Questionnaire; IPSAQ: Internal Personal and Situational Attributions Questionnaire; SJT: Social Judgment Task; SA: Socially appropriate behaviors; VSN: Violation of social norms; IU: Impolite but understandable if the characters' thoughts are taken into account. Bold values denote statistical significance.

of empirically validated measures of SC limits our ability to understand and make decisions based on test scores. This study is an attempt to contribute to our knowledge of the psychometric qualities of a battery of SC tests in a sample of 100 healthy adults, and to propose a more reliable and valid battery of such tests. The results add to those of previous studies (Ludwig et al., 2017; Pinkham et al., 2014, 2016), as well as serve to underline the need for more cross-cultural studies of SC (cf Lim et al., 2020). Two tests were adapted for the study to better reflect the distribution of these abilities in a nonclinical sample of young and older adults. The tasks were examined through the assessment of internal consistency, interrater and intercoder reliability, factor structure, concurrent validity, and group comparisons.

The study yielded evidence of unsatisfactory properties for some tests, notably the Task of Attribution of Intention to Others (Brunet et al., 2003), Picture Sequencing task (Langdon et al., 2014), and Faux Pas-R. We did not find sufficient empirical evidence of satisfactory psychometric

properties of the AIHQ hostility and aggression bias despite their association with the quality of relationships and high intercoder reliability. Although not wholly unsatisfactory, additional psychometric evidence is needed before using these scales in clinical and research settings. In addition, the IPSAQ-positive scale and the two attributional biases, personalizing and externalizing, did not show satisfactory results, reliability is low for the positive scale, and no significant association was found between these three scales and any outcomes.

Other tests yielded satisfactory properties. Internal consistency is high for the AIHQ-blame score, making it a suitable candidate for further research on attributional bias in different populations. Consistency is also satisfactory for the IPSAQ negative scale, and two scales in the IRI, personal distress and perspective-taking. Using CFA, we could not, however, confirm the original four-factor structure of the IRI, neither could we identify strong and consistent factor contributors to the variance using EFA. Further

Table 9. Correlations between sociocognitive variables and the *Interpersonal Relationship Quality Scale (IRQS)*.

IRQS domains	Romantic relationship	Family	Friendships	Colleagues	People in general	All domains
Sample size	56	100	99	63	100	100
Task						
TREF						
Accuracy						
Happiness	-.29*	-.15	-.05	-.15	-.25*	-.22*
Disgust	.21	.15	.20	.01	.23*	.22*
Sadness	.03	.05	.05	-.08	-.05	.01
Fear	-.20	-.07	-.02	-.14	-.22*	-.20*
Contempt	.24	.06	.21*	.03	.10	.14
Anger	.23	.03	.24*	.27*	.04	.18
Total score	.09	.03	.20*	.01	-.03	.06
Detection threshold						
Happiness	.21	.06	.03	.07	.14	.11
Disgust	-.16	-.29**	-.15	-.13	-.27**	-.24*
Sadness	.03	-.12	.05	.07	-.08	-.06
Fear	.08	-.00	.07	.18	.16	.13
Contempt	-.35**	-.10	-.28**	.02	-.05	-.17
Anger	-.21	-.03	-.07	-.16	-.02	-.08
SS-R						
Factor score	.07	.17	.20	-.05	.00	.15
IRI						
Perspective taking	.18	.13	.07	-.06	.22*	.16
Empathic concern	-.04	.06	-.01	.18	.06	.07
Personal distress	-.35**	-.06	-.01	-.15	-.06	-.13
Fantasy	-.22	-.03	-.04	.00	.09	-.11
AIHQ						
Hostility bias	-.20	-.13	-.05	-.01	-.11	-.10
Aggression bias	-.32*	-.02	-.07	-.08	-.25*	-.13
Blame score	-.09	-.07	.12	.05	-.21*	-.05
IPSAQ						
Positive events	-.16	.03	-.14	-.13	.01	-.09
Negative events	-.11	.07	-.11	-.03	.00	-.04
Personalizing bias	.11	-.07	.03	.08	.01	.02
Externalizing bias	-.01	-.02	.00	-.02	.00	-.01
SJT						
Correct judgments	.32*	.12	.14	-.07	-.02	.09
% of "normal" in SA	.160	-.03	.03	-.04	-.12	-.07
% of "unusual" in SA	-.12	-.00	-.03	.19	.14	.06
% of "shocking" in SA	-.06	.04	-.00	-.23	.01	.02
% of "normal" for VSN	.04	.09	.04	.04	.14	.12
% of "unusual" for VSN	.07	-.03	-.03	-.07	-.13	-.09
% of "shocking" for VSN	-.09	-.02	.00	.04	.05	.02
% of "normal" in IU	.37**	.25*	.18	-.02	.12	.22*
% of "unusual" in IU	-.27*	-.26**	-.08	.05	-.18	-.22*
% of "shocking" in IU	-.25	-.07	-.17	-.02	.02	-.08

Note: *= $p < .05$; **= $p < .01$; TREF: Facial Emotions Recognition Task; SS-R: Strange Stories-Revised; IRI: Interpersonal Reactivity Index; AIHQ: Ambiguous Intentions Hostility Questionnaire; IPSAQ: Internal Personal and Situational Attributions Questionnaire; SJT: Social Judgment Task; SA: Socially appropriate behaviors; VSN: Violation of social norms; IU: Impolite but understandable if the characters' thoughts are taken into account. Bold values denote statistical significance.

investigation with larger samples is recommended. Reliability and factorial analysis performed on the SS-R show positive results with exploratory research methods. This represents a promising approach for assessing ToM across adulthood.

Socially appropriate behaviors in the SJT show a high interrater agreement. Judgments made of impolite behaviors and violations of social norms, in turn, show patterns of judgments that reflect a shift toward acceptance and normalization of nonconformity, and to a certain degree, of social transgression. This may be because of a possible sampling bias or cultural differences. Social judgment is fundamentally influenced by social norms and rules, which in turn vary as a function of cultural conditions and shared practices in a given context (Kitayama & Uskul, 2011). However, it is unclear whether these responses resulted from a greater tolerance toward ambiguous behaviors and social transgressions in the more urban population of Quebec, or a

response bias toward positive judgments. Regardless, cultural as well as urban versus rural factors merit further study. Internal consistency for all categories of behaviors in the SJT showed low alpha values, probably because each subscale comprised a small number of items. Cross-cultural studies are still limited in SC research among adult populations. Nevertheless, recent evidence suggests cultural variation in complex emotion recognition (Adams et al., 2010), emotion recognition and face processing (Rule et al., 2013), attribution of blame (Combs et al., 2007), false-belief understanding in ToM (Aival-Naveh et al., 2019), and cross-cultural mental state attributions (Perez-Zapata et al., 2016). Results obtained in the present study suggest lines of further investigation in order to specify how the components of SC vary within and between cultures.

Although internal consistency was low for nearly all TREF scores, which is consistent with Schlegel et al.'s (2017) meta-analytic data regarding emotion recognition accuracy,

interrater agreement was substantial for almost all emotions in the TREF. The results are particularly relevant as other than discriminant validity, there was no psychometric data available to date. According to the results, only contempt showed an unacceptable recognition accuracy (>50%) in all ages. Such findings are consistent with results of previous studies using other tests (Matsumoto, 2005; Tracy & Robins, 2008), which revealed relatively low recognition rates of contempt possibly due to its complex social nature, close to anger and disgust, but more related to social transgressions and exclusion (Fischer & Giner-Sorolla, 2016). Emotion processing is important for successful interactions, and impaired recognition of facial affect has been consistently demonstrated in many conditions with differential patterns of deficits such as in schizophrenia (Bediou et al., 2007; Kohler et al., 2010; Morris et al., 2009), major depressive disorder (Dalili et al., 2015), and dementia (Bora et al., 2016; Kumfor & Piguet, 2013). Given its clinical relevance, a more normally distributed, valid and reliable measure such as the TREF could be an effective means for its assessment in clinical and research settings.

Additionally, age-related differences were found in some SC tasks (SS-R, TREF, IPSAQ, SJT) after adjustment for potential confounders. The observed age variations are consistent with those obtained in previous studies reporting developmental changes in SC abilities (Henry et al., 2013; Ruffman et al., 2008). In addition to the few sex differences that were found (TREF, SS-R, IRI, and SJT) among the variables, these findings constitute evidence of “known-groups validity.” Given that the cross-sectional design used in the present study makes it difficult to distinguish between age and cohort effects, further studies using longitudinal data could clarify the nature of the variations.

Lastly, in most of the measures that were examined, associations were found between SC subscores and different types of social relationships. This is in line with prior research in schizophrenia that found interactions between SC processes, social competence, and the ability to maintain satisfying interpersonal relations (Couture et al., 2006; Fett et al., 2011; Mancuso et al., 2011; Poole et al., 2000). Of particular interest in the present study are the emotion recognition accuracy and social judgment components, which seem to interact with multiple levels of relationships. This is an important area for future research, particularly given their significance in everyday life. The lack of association between total scores and domains of relationships, however, highlights the need to further improve the predictive validity of the instruments.

Limitations

The principal limitations of this study are the absence of a measure of test-retest reliability, and sample size. Given the importance of test-retest reliability in assessing psychometric properties, it would be necessary in future studies to test the stability of the different SC measures over time. Second, in the evaluation of concurrent validity, the samples were

relatively small for two types of relationships, romantic and work, thereby limiting statistical power.

Conclusion

This study reviewed the psychometric properties of several SC tools with healthy adults. Based on the psychometric data obtained from this sample across a wide span of ages, reliability and validity were limited for some instruments (TAIO, PST, IRI, Faux Pas-R, AIHQ-Hostility/Aggression, IPSAQ). On the other hand, the TREF, SS-R, AIHQ-Blame, and SJT showed satisfactory properties. Although none of them performed well across all indexes, the multiplicity of quantitative evidence across analysis (structural dimensionality, criterion relevance, performance comparisons, reliability) adds confidence to support their validity (Messick, 2005; Sartori & Pasini, 2007). In sum, this work contributed at this early stage, to the development of a psychometrically acceptable battery of tests measuring the principal components of SC, for research in healthy and diverse populations. Not all tests targeted in this study show adequate psychometric standard (e.g., IPSAQ, Faux Pas) although some of them are still widely used in clinical research, indicating that care should be taken when selecting a battery of social cognitive instruments. The results could help researchers and clinicians select an appropriate test battery and encourage the development of more valid versions of existing tests. Given that SC is a multifaceted construct, this work, similar to the SCOPE study, highlights the importance of assessing a more representative sample of the components of SC in order to get a more complete appreciation of an individual's social cognitive abilities.

Acknowledgments

The authors would like to express their sincere gratitude to Dr. F. Happé for the opportunity to use and adapt the Strange Stories Task, as well as Dr. D. L. Penn, Dr. P. Kinderman, Dr. R. Langdon, Dr. M. Davis, Mr. B. Gaudelus, Dr. E. Brunet-Gouet, Dr. R. Vallerand, and Dr. S. Baron-Cohen (Autism Research Centre) for the access to the tests. Also, the authors would like to thank Francis Germain (Université du Québec à Montréal) for his contribution to data collection.

Disclosure statement

Authors declare that they have no conflicts of interest concerning this article.

ORCID

Catherine Gourlay  <http://orcid.org/0000-0002-8688-5182>
 Pier-Olivier Caron  <http://orcid.org/0000-0001-6346-5583>
 Peter B. Scherzer  <http://orcid.org/0000-0003-4962-4407>

References

- Achim, A. (2017). Testing the number of required dimensions in exploratory factor analysis. *The Quantitative Methods for Psychology*, 13(1), 64–74. <https://doi.org/10.20982/tqmp.13.1.p064>

- Achim, A. M., Sutliff, S., Samson, C., Montreuil, T. C., & Lecomte, T. (2016). Attribution bias and social anxiety in schizophrenia. *Schizophrenia Research. Cognition*, 4(1), 1–3. <https://doi.org/10.1016/j.scog.2016.01.001>
- Adams, R. B., Rule, N. O., Franklin, R. G., Wang, E., Stevenson, M. T., Yoshikawa, S., Nomura, M., Sato, W., Kveraga, K., & Ambady, N. (2010). Cross-cultural reading the mind in the eyes: An fMRI investigation. *Journal of Cognitive Neuroscience*, 22(1), 97–108. <https://doi.org/10.1162/jocn.2009.21187>
- Adolphs, R. (2001). The neurobiology of social cognition. *Current Opinion in Neurobiology*, 11(2), 231–239. [https://doi.org/10.1016/S0959-4388\(00\)00202-6](https://doi.org/10.1016/S0959-4388(00)00202-6)
- Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion in Neurobiology*, 12(2), 169–177. [https://doi.org/10.1016/S0959-4388\(02\)00301-X](https://doi.org/10.1016/S0959-4388(02)00301-X)
- Ahmed, F. S., & Stephen Miller, L. (2011). Executive function mechanisms of theory of mind. *Journal of Autism and Developmental Disorders*, 41(5), 667–678. <https://doi.org/10.1007/s10803-010-1087-7>
- Aival-Naveh, E., Rothschild-Yakar, L., & Kurman, J. (2019). Keeping culture in mind: A systematic review and initial conceptualization of mentalizing from a cross-cultural perspective. *Clinical Psychology: Science and Practice*, 26(4), e12300. <https://doi.org/10.1111/cpsp.12300>
- American Psychiatric Association. (2013). *Intellectual disabilities. In Diagnostic and statistical manual of mental disorders* (5th ed.). Author.
- Baron-Cohen, S. (2001). Theory of mind in normal development and autism. *Prisme*, 34, 174–183.
- Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & 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(5), 407–418. <https://doi.org/10.1023/A:1023035012436>
- Bediou, B., Asri, F., Brunelin, J., Krolak-Salmon, P., D'Amato, T., Saoud, M., & Tazi, I. (2007). Emotion recognition and genetic vulnerability to schizophrenia. *The British Journal of Psychiatry: The Journal of Mental Science*, 191(2), 126–130. <https://doi.org/10.1192/bjp.bp.106.028829>
- Beer, J. S., & Ochsner, K. N. (2006). Social cognition: A multi level analysis. *Brain Research*, 1079(1), 98–105. <https://doi.org/10.1016/j.brainres.2006.01.002>
- Bora, E., & Pantelis, C. (2016). Meta-analysis of social cognition in attention-deficit/hyperactivity disorder (ADHD): Comparison with healthy controls and autistic spectrum disorder. *Psychological Medicine*, 46(4), 699–716. <https://doi.org/10.1017/s0033291715002573>
- Bora, E., Velakoulis, D., & Walterfang, M. (2016). Meta-analysis of facial emotion recognition in behavioral variant frontotemporal dementia: Comparison with Alzheimer disease and healthy controls. *Journal of Geriatric Psychiatry and Neurology*, 29(4), 205–211. <https://doi.org/10.1177/0891988716640375>
- Brunet, E., Sarfati, Y., & Hardy-Baylé, M.-C. (2003). Reasoning about physical causality and other's intentions in schizophrenia. *Cognitive Neuropsychiatry*, 8(2), 129–139. <https://doi.org/10.1080/13546800244000256>
- Buck, B. E., Pinkham, A. E., Harvey, P. D., & Penn, D. L. (2016). Revisiting the validity of measures of social cognitive bias in schizophrenia: Additional results from the Social Cognition Psychometric Evaluation (SCOPE) study. *The British Journal of Clinical Psychology*, 55(4), 441–454. <https://doi.org/10.1111/bjc.12113>
- Buck, B., Iwanski, C., Healey, K. M., Green, M. F., Horan, W. P., Kern, R. S., Lee, J., Marder, S. R., Reise, S. P., & Penn, D. L. (2017). Improving measurement of attributional style in schizophrenia: A psychometric evaluation of the Ambiguous Intentions Hostility Questionnaire (AIHQ). *Journal of Psychiatric Research*, 89, 48–54. <https://doi.org/10.1016/j.jpsychires.2017.01.004>
- Caron, P.-O. (2018). *La modélisation par équations structurelles avec Mplus*. Presses de l'Université du Québec.
- Charlton, R. A., Barrick, T. R., Markus, H. S., & Morris, R. G. (2009). Theory of mind associations with other cognitive functions and brain imaging in normal aging. *Psychology and Aging*, 24(2), 338–348. <https://doi.org/10.1037/a0015225>
- Chiang, S.-K., Hua, M.-S., Tam, W.-C. C., Chao, J.-K., & Shiah, Y.-J. (2014). Developing an alternative Chinese version of the Interpersonal Reactivity Index for normal population and patients with schizophrenia in Taiwan. *Brain Impairment*, 15(2), 120–131. <https://doi.org/10.1017/BrImp.2014.15>
- Chryssikou, E. G., & Thompson, W. J. (2016). Assessing Cognitive and Affective Empathy Through the Interpersonal Reactivity Index: An Argument Against a Two-Factor Model. *Assessment*, 23(6), 769–777. <https://doi.org/10.1177/1073191115599055>
- Chung, Y. S., Barch, D., & Strube, M. (2014). A meta-analysis of mentalizing impairments in adults with schizophrenia and autism spectrum disorder. *Schizophrenia Bulletin*, 40(3), 602–616. <https://doi.org/10.1093/schbul/sbt048>
- Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Combs, D. R., Penn, D. L., Wicher, M., & Waldheter, E. (2007). The Ambiguous Intentions Hostility Questionnaire (AIHQ): A new measure for evaluating hostile social-cognitive biases in paranoia. *Cognitive Neuropsychiatry*, 12(2), 128–143. <https://doi.org/10.1080/13546800600787854>
- Couture, S. M., Penn, D. L., & Roberts, D. L. (2006). The functional significance of social cognition in schizophrenia: A review. *Schizophrenia Bulletin*, 32(Supplement 1), S44–S63. <https://doi.org/10.1093/schbul/sbl029>
- Cramer, D., & Howitt, D. L. (2005). *The SAGE dictionary of statistics: A practical resource for students in the social sciences* (3rd ed.). SAGE.
- Dalili, M. N., Penton-Voak, I. S., Harmer, C. J., & Munafò, M. R. (2015). Meta-analysis of emotion recognition deficits in major depressive disorder. *Psychological Medicine*, 45(6), 1135–1144. <https://doi.org/10.1017/S0033291714002591>
- Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, 10, 85.
- Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1), 113–126. <https://doi.org/10.1037/0022-3514.44.1.113>
- De Corte, K., Buysse, A., Verhofstadt, L. L., Roeyers, H., Ponnet, K., & Davis, M. H. (2007). Measuring empathic tendencies: Reliability and validity of the Dutch version of the Interpersonal Reactivity Index. *Psychologica Belgica*, 47(4), 235–260. <https://doi.org/10.5334/pb-47-4-235>
- Demopoulos, C., Hopkins, J., & Davis, A. (2013). A comparison of social cognitive profiles in children with autism spectrum disorders and attention-deficit/hyperactivity disorder: A matter of quantitative but not qualitative difference? *Journal of Autism and Developmental Disorders*, 43(5), 1157–1170. <https://doi.org/10.1007/s10803-012-1657-y>
- Di Tella, M., Miti, F., Ardito, R. B., & Adenzato, M. (2020). Social cognition and sex: Are men and women really different? *Personality and Individual Differences*, 162, 110045. <https://doi.org/10.1016/j.paid.2020.110045>
- Eack, S. M., Baborik, A. L., McKnight, S. A. F., Hogarty, S. S., Greenwald, D. P., Newhill, C. E., Phillips, M. L., Keshavan, M. S., & Minshew, N. J. (2013). Commonalities in social and non-social cognitive impairments in adults with autism spectrum disorder and schizophrenia. *Schizophrenia Research*, 148(1–3), 24–28. <https://doi.org/10.1016/j.schres.2013.05.013>
- Elamin, M., Pender, N., Hardiman, O., & Abrahams, S. (2012). Social cognition in neurodegenerative disorders: A systematic review. *Journal of Neurology, Neurosurgery, and Psychiatry*, 83(11), 1071–1079. <https://doi.org/10.1136/jnnp-2012-302817>
- Faísca, L., Afonseca, S., Brüne, M., Gonçalves, G., Gomes, A., & Martins, A. T. (2016). Portuguese Adaptation of a Faux Pas Test and a Theory of Mind Picture Stories Task. *Psychopathology*, 49(3), 143–152. <https://doi.org/10.1159/000444689>

- Ferguson, F. J., & Austin, E. J. (2010). Associations of trait and ability emotional intelligence with performance on Theory of Mind tasks in an adult sample. *Personality and Individual Differences*, 49(5), 414–418. <https://doi.org/10.1016/j.paid.2010.04.009>
- Fernández, A. M., Dufey, M., & Kramp, U. (2011). Testing the psychometric properties of the Interpersonal Reactivity Index (IRI) in Chile. *European Journal of Psychological Assessment*, 27(3), 179–185. <https://doi.org/10.1027/1015-5759/a000065>
- Fett, A.-K. J., Viechtbauer, W., Dominguez, M.-G., Penn, D. L., van Os, J., & Krabbendam, L. (2011). The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: A meta-analysis. *Neuroscience and Biobehavioral Reviews*, 35(3), 573–588. <https://doi.org/10.1016/j.neubiorev.2010.07.001>
- Fischer, A., & Giner-Sorolla, R. (2016). Contempt: Derogating others while keeping calm. *Emotion Review*, 8(4), 346–357. <https://doi.org/10.1177/1754073915610439>
- Gao, B., Wang, Y., Zhu, Y., Tian, Q., Chen, Z., Cohen, Z., Landa, Y., & Mueser, K. T. (2018). A psychometric investigation of the Chinese version of the Internal, Personal and Situational Attributions Questionnaire (C-IPSAQ). *Translational Psychiatry*, 8(1), 256. <https://doi.org/10.1038/s41398-018-0314-4>
- Gaudelus, B., Virgile, J., Peyroux, E., Leleu, A., Baudouin, J. Y., Franck, N. (2014). Mesure du déficit de reconnaissance des émotions faciales dans la schizophrénie: étude préliminaire du Test de Reconnaissance des Émotions Faciales (TREF). *L'Encéphale*, 41, 251–259. <https://doi.org/10.1016/j.enceph.2014.08.013>
- Gilet, A.-L., Mella, N., Studer, J., Grün, D., & Labouvie-Vief, G. (2013). Assessing dispositional empathy in adults: A French validation of the Interpersonal Reactivity Index (IRI). *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, 45(1), 42–48. <https://doi.org/10.1037/a0030425>
- Gonçalves, A. R., Fernandes, C., Pasion, R., Ferreira-Santos, F., Barbosa, F., & Marques-Teixeira, J. (2018). Effects of age on the identification of emotions in facial expressions: A meta-analysis. *PeerJ*, 6, e5278. <https://doi.org/10.7717/peerj.5278>
- Green, M. F., Horan, W. P., & Lee, J. (2015). Social cognition in schizophrenia. *Nature Reviews Neuroscience*, 16(10), 620–631. <https://doi.org/10.1038/nrn4005>
- Green, M. F., Penn, D. L., Bentall, R., Carpenter, W. T., Gaebel, W., Gur, R. C., Kring, A. M., Park, S., Silverstein, S. M., & Heinsen, R. (2008). Social cognition in schizophrenia: An NIMH workshop on definitions, assessment, and research opportunities. *Schizophrenia Bulletin*, 34(6), 1211–1220. <https://doi.org/10.1093/schbul/sbm145>
- Gregory, C., Lough, S., Stone, V., Erzincliglu, S., Martin, L., Baron-Cohen, S., & Hodges, J. R. (2002). Theory of mind in patients with frontal variant frontotemporal dementia and Alzheimer's disease: Theoretical and practical implications. *Brain: A Journal of Neurology*, 125(Pt 4), 752–764. <https://doi.org/10.1093/brain/awf079>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Prentice Hall.
- Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24(2), 129–154. <https://doi.org/10.1007/bf02172093>
- Healey, K. M., Bartholomeusz, C. F., & Penn, D. L. (2016). Deficits in social cognition in first episode psychosis: A review of the literature. *Clinical Psychology Review*, 50, 108–137. <https://doi.org/10.1016/j.cpr.2016.10.001>
- Henry, J. D., Phillips, L. H., Ruffman, T., & Bailey, P. E. (2013). A meta-analytic review of age differences in theory of mind. *Psychology and Aging*, 28(3), 826–839. <https://doi.org/10.1037/a0030677>
- Hess, T. M. (2006). Adaptive aspects of social cognitive functioning in adulthood: Age-related goal and knowledge influences. *Social Cognition*, 24(3), 279–309. <https://doi.org/10.1521/soco.2006.24.3.279>
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185. <https://doi.org/10.1007/bf02289447>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Huang, X., Li, W., Sun, B., Chen, H., & Davis, M. H. (2012). The Validation of the Interpersonal Reactivity Index for Chinese teachers from primary and middle schools. *Journal of Psychoeducational Assessment*, 30(2), 194–204. <https://doi.org/10.1177/0734282911410588>
- Humphreys, K. L., Galán, C. A., Tottenham, N., & Lee, S. S. (2016). Impaired social decision-making mediates the association between ADHD and social problems. *Journal of Abnormal Child Psychology*, 44(5), 1023–1032. <https://doi.org/10.1007/s10802-015-0095-7>
- Jeon, I. H., Kim, K. R., Kim, H. H., Park, J. Y., Lee, M., Jo, H. H., Koo, S. J., Jeong, Y. J., Song, Y. Y., Kang, J. I., Lee, S. Y., Lee, E., & An, S. K. (2013). Attributional style in healthy persons: Its Association with "theory of mind" skills. *Psychiatry Investigation*, 10(1), 34–40. <https://doi.org/10.4306/pi.2013.10.1.34>
- Kaland, N., Møller-Nielsen, A., Smith, L., Mortensen, E. L., Callesen, K., & Gottlieb, D. (2005). The strange stories test-A replication study of children and adolescents with Asperger syndrome. *European Child & Adolescent Psychiatry*, 14(2), 73–82. <https://doi.org/10.1007/s00787-005-0434-2>
- Kinderman, P., & Bentall, R. P. (1997). Causal attributions in paranoia and depression: Internal, personal, and situational attributions for negative events. *Journal of Abnormal Psychology*, 106(2), 341–345. <https://doi.org/10.1037/0021-843X.106.2.341>
- Kirkland, R. A., Peterson, E., Baker, C. A., Miller, S., & Pulos, S. (2013). Meta-analysis reveals adult female superiority in "Reading the Mind in the Eyes" Test. *North American Journal of Psychology*, 15(1), 121–146.
- Kitayama, S., & Uskul, A. K. (2011). Culture, mind, and the brain: Current evidence and future directions. *Annual Review of Psychology*, 62(1), 419–449. <https://doi.org/10.1146/annurev-psych-120709-145357>
- Kohler, C. G., Walker, J. B., Martin, E. A., Healey, K. M., & Moberg, P. J. (2010). Facial emotion perception in schizophrenia: A meta-analytic review. *Schizophrenia Bulletin*, 36(5), 1009–1019. <https://doi.org/10.1093/schbul/sbn192>
- Kumfor, F., & Pignatelli, O. (2013). Emotion recognition in the dementias: Brain correlates and patient implications. *Neurodegenerative Disease Management*, 3(3), 277–288. <https://doi.org/10.2217/nmt.13.16>
- Langdon, R., Connors, M. H., & Connaughton, E. (2014). Social cognition and social judgment in schizophrenia. *Schizophrenia Research. Cognition*, 1(4), 171–174. <https://doi.org/10.1016/j.scog.2014.10.001>
- Langdon, R., Corner, T., McLaren, J., Ward, P. B., & Coltheart, M. (2006). Externalizing and personalizing biases in persecutory delusions: The relationship with poor insight and theory-of-mind. *Behaviour Research and Therapy*, 44(5), 699–713. <https://doi.org/10.1016/j.brat.2005.03.012>
- Langdon, R., Still, M., Connors, M. H., Ward, P. B., & Catts, S. V. (2013). Attributional biases, paranoia, and depression in early psychosis. *The British Journal of Clinical Psychology*, 52(4), 408–423. <https://doi.org/10.1111/bjc.12026>
- Larøi, F., & Brédard, S. (2001). Presentation of a French version of the Internal, Personal and Situational Attributions Questionnaire. *European Review of Applied Psychology*, 52, 133–141.
- Lever, A. G., & Geurts, H. M. (2016). Age-related differences in cognition across the adult lifespan in autism spectrum disorder. *Autism Research: Official Journal of the International Society for Autism Research*, 9(6), 666–676. <https://doi.org/10.1002/aur.1545>
- Lim, K., Lee, S.-A., Pinkham, A. E., Lam, M., & Lee, J. (2020). Evaluation of social cognitive measures in an Asian schizophrenia sample. *Schizophrenia Research. Cognition*, 20, 100169. <https://doi.org/10.1016/j.scog.2019.100169>
- Loveland, K. A., Pearson, D. A., Tunali-Kotoski, B., Ortegón, J., & Gibbs, M. C. (2001). Judgments of social appropriateness by children and adolescents with autism. *Journal of Autism and Developmental Disorders*, 31(4), 367–376. <https://doi.org/10.1023/A:1010608518060>

- Lucas-Molina, B., Pérez-Albéniz, A., Ortuño-Sierra, J., & Fonseca-Pedrero, E. (2017). Dimensional structure and measurement invariance of the Interpersonal Reactivity Index (IRI) across gender. *Psicothema*, 29(4), 590–595. <https://doi.org/10.7334/psicothema2017.19>
- Ludwig, K. A., Pinkham, A. E., Harvey, P. D., Kelsven, S., & Penn, D. L. (2017). Social cognition psychometric evaluation (SCOPE) in people with early psychosis: A preliminary study. *Schizophrenia Research*, 190, 136–143. <https://doi.org/10.1016/j.schres.2017.03.001>
- MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion (Washington, D.C.)*, 8(4), 540–551. <https://doi.org/10.1037/a0012746>
- Mancuso, F., Horan, W. P., Kern, R. S., & Green, M. F. (2011). Social cognition in psychosis: Multidimensional structure, clinical correlates, and relationship with functional outcome. *Schizophrenia Research*, 125(2–3), 143–151. <https://doi.org/10.1016/j.schres.2010.11.007>
- Matsumoto, D. (2005). Scalar ratings of contempt expressions. *Journal of Nonverbal Behavior*, 29(2), 91–104. <https://doi.org/10.1007/s10919-005-2742-0>
- Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion (Washington, D.C.)*, 3(1), 97–105. <https://doi.org/10.1037/1528-3542.3.1.97>
- Messick, S. (2005). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5–8. <https://doi.org/10.1111/j.1745-3992.1995.tb00881.x>
- Montagne, B., Kessels, R. P. C., Frigerio, E., de Haan, E. H. F., & Perrett, D. I. (2005). Sex differences in the perception of affective facial expressions: Do men really lack emotional sensitivity? *Cognitive Processing*, 6(2), 136–141. <https://doi.org/10.1007/s10339-005-0050-6>
- Moran, J. M. (2013). Lifespan development: The effects of typical aging on theory of mind. *Behavioural Brain Research*, 237, 32–40. <https://doi.org/10.1016/j.bbr.2012.09.020>
- Morris, R. W., Weickert, C. S., & Loughland, C. M. (2009). Emotional face processing in schizophrenia. *Current Opinion in Psychiatry*, 22(2), 140–146. <https://doi.org/10.1097/ycp.0b013e328324f895>
- Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review*, 110(3), 472–489. <https://doi.org/10.1037/0033-295x.110.3.472>
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.
- Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4), 695–699. <https://doi.org/10.1111/j.1532-5415.2005.53221.x>
- Negrão, J., Akiba, H. T., Lederman, V. R. G., & Dias, Á. M. (2016). Faux Pas Test in schizophrenic patients. *Jornal Brasileiro de Psiquiatria*, 65(1), 17–21. <https://doi.org/10.1590/0047-2085000000098>
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.
- Ouellet, J., Scherzer, P. B., Rouleau, I., Métras, P., Bertrand-Gauvin, C., Djerroud, N., Boisseau, E., & Duquette, P. (2010). Assessment of social cognition in patients with multiple sclerosis. *Journal of the International Neuropsychological Society: JINS*, 16(2), 287–296. <https://doi.org/10.1017/s1355617709991329>
- Pardini, M., & Nichelli, P. F. (2009). Age-related decline in mentalizing skills across adult life span. *Experimental Aging Research*, 35(1), 98–106. <https://doi.org/10.1080/03610730802545259>
- Perez-Zapata, D., Slaughter, V., & Henry, J. D. (2016). Cultural effects on mindreading. *Cognition*, 146, 410–414. <https://doi.org/10.1016/j.cognition.2015.10.018>
- Phillips, L. H., Allen, R., Bull, R., Hering, A., Kliegel, M., & Channon, S. (2015). Older adults have difficulty in decoding sarcasm. *Developmental Psychology*, 51(12), 1840–1852. <https://doi.org/10.1037/dev0000063>
- Pinkham, A. E., Penn, D. L., Green, M. F., Buck, B., Healey, K., & Harvey, P. D. (2014). The social cognition psychometric evaluation study: Results of the expert survey and RAND panel. *Schizophrenia Bulletin*, 40(4), 813–823. <https://doi.org/10.1093/schbul/sbt081>
- Pinkham, A. E., Penn, D. L., Green, M. F., & Harvey, P. D. (2016). Social cognition psychometric evaluation: Results of the initial psychometric study. *Schizophrenia Bulletin*, 42(2), 494–504. <https://doi.org/10.1093/schbul/sbv056>
- Plana, I., Lavoie, M.-A., Battaglia, M., & Achim, A. M. (2014). A meta-analysis and scoping review of social cognition performance in social phobia, posttraumatic stress disorder and other anxiety disorders. *Journal of Anxiety Disorders*, 28(2), 169–177. <https://doi.org/10.1016/j.janxdis.2013.09.005>
- Poole, J. H., Tobias, F. C., & Vinogradov, S. (2000). The functional relevance of affect recognition errors in schizophrenia. *Journal of the International Neuropsychological Society: JINS*, 6(6), 649–658. <https://doi.org/10.1017/s135561770066602x>
- Portney, L. G., & Watkins, M. P. (2000). *Foundations of clinical research: Applications to practice* (2nd ed.). Prentice Hall Health.
- Pulos, S., Elison, J., & Lennon, R. (2004). The hierarchical structure of the Interpersonal Reactivity Index. *Social Behavior and Personality: An International Journal*, 32(4), 355–359. <https://doi.org/10.2224/sbp.2004.32.4.355>
- Rogers, K., Dziobek, I., Hassenstab, J., Wolf, O. T., & Convit, A. (2007). Who cares? Revisiting empathy in Asperger syndrome. *Journal of Autism and Developmental Disorders*, 37(4), 709–715. <https://doi.org/10.1007/s10803-006-0197-8>
- Ruffman, T., Henry, J. D., Livingstone, V., & Phillips, L. H. (2008). A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neuroscience and Biobehavioral Reviews*, 32(4), 863–881. <https://doi.org/10.1016/j.neubiorev.2008.01.001>
- Rule, N. O., Freeman, J. B., & Ambady, N. (2013). Culture in social neuroscience: A review. *Social Neuroscience*, 8(1), 3–10. <https://doi.org/10.1080/17470919.2012.695293>
- Russell, T. A., Tchanturia, K., Rahman, Q., & Schmidt, U. (2007). Sex differences in theory of mind: A male advantage on Happé's "cartoon" task. *Cognition & Emotion*, 21(7), 1554–1564. <https://doi.org/10.1080/02699930601117096>
- Sabbagh, M. A. (2004). Understanding orbitofrontal contributions to theory-of-mind reasoning: Implications for autism. *Brain and Cognition*, 55(1), 209–219. <https://doi.org/10.1016/j.bandc.2003.04.002>
- Samamé, C., Martino, D. J., & Strejilevich, S. A. (2012). Social cognition in euthymic bipolar disorder: Systematic review and meta-analytic approach. *Acta Psychiatrica Scandinavica*, 125(4), 266–280. <https://doi.org/10.1111/j.1600-0447.2011.01808.x>
- Sartori, R., & Pasini, M. (2007). Quality and quantity in test validity: How can we be sure that psychological tests measure what they have to? *Quality & Quantity*, 41(3), 359–374. <https://doi.org/10.1007/s11135-006-9006-x>
- Savla, G. N., Vella, L., Armstrong, C. C., Penn, D. L., & Twamley, E. W. (2013). Deficits in Domains of social cognition in schizophrenia: A meta-analysis of the empirical evidence. *Schizophrenia Bulletin*, 39(5), 979–992. <https://doi.org/10.1093/schbul/sbs080>
- Scherzer, P., Achim, A., Léveillé, E., Boisseau, E., & Stip, E. (2015). Evidence from paranoid schizophrenia for more than one component of theory of mind. *Frontiers in Psychology*, 6, 1643. <https://doi.org/10.3389/fpsyg.2015.01643>
- Scherzer, P., Léveillé, E., Achim, A., Boisseau, E., & Stip, E. (2012). A study of theory of mind in paranoid schizophrenia: A theory or many theories? *Frontiers in Psychology*, 3, 432. <https://doi.org/10.3389/fpsyg.2012.00432>
- Schlegel, K., Boone, R. T., & Hall, J. A. (2017). Individual differences in interpersonal accuracy: A multi-level meta-analysis to assess whether judging other people is one skill or many. *Journal of Nonverbal Behavior*, 41(2), 103–137. <https://doi.org/10.1007/s10919-017-0249-0>
- Schlegel, K., Grandjean, D., & Scherer, K. R. (2014). Introducing the Geneva Emotion Recognition Test: An example of Rasch-based test development. *Psychological Assessment*, 26(2), 666–672. <https://doi.org/10.1037/a0035246>

- Schneider, M., Myin, E., & Myin-Germeys, I. (2020). Is theory of mind a prerequisite for social interactions? A study in psychotic disorder. *Psychological Medicine*, 50(5), 754–760. <https://doi.org/10.1017/S0033291719000540>
- Senécal, C. B., Vallerand, R. J., & Vallières, E. F. (1992). Construction et validation de l'Échelle de la Qualité des Relations Interpersonnelles (EQRI). *Revue Européenne de Psychologie Appliquée*, 42(4), 315–322.
- Silberstein, J., & Harvey, P. D. (2019). Cognition, social cognition, and self-assessment in schizophrenia: Prediction of different elements of everyday functional outcomes. *CNS Spectrums*, 24(1), 88–93. <https://doi.org/10.1017/S1092852918001414>
- Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience*, 10(5), 640–656. <https://doi.org/10.1162/089892998562942>
- Tracy, J. L., & Robins, R. W. (2008). The automaticity of emotion recognition. *Emotion*, 8(1), 81–95. <https://doi.org/10.1037/1528-3542.8.1.81>
- Wacker, R., Bölte, S., & Dziobek, I. (2017). Women know better what other women think and feel: Gender effects on mindreading across the adult life span. *Frontiers in Psychology*, 8, 1324. <https://doi.org/10.3389/fpsyg.2017.01324>
- Wechsler, D. (2008). *Wechsler Adult Intelligence Scale-4th ed.* Pearson.
- Weightman, M. J., Air, T. M., & Baune, B. T. (2014). A review of the role of social cognition in major depressive disorder. *Frontiers in Psychiatry*, 5, 179. <https://doi.org/10.3389/fpsyg.2014.00179>
- Wellman, H. M. (2018). Theory of mind across the lifespan? *Zeitschrift Für Psychologie*, 226(2), 136–138. <https://doi.org/10.1027/2151-2604/a000330>
- Whalen, C. K., Henker, B., & Granger, D. A. (1990). Social judgment processes in hyperactive boys: Effects of methylphenidate and comparisons with normal peers. *Journal of Abnormal Child Psychology*, 18(3), 297–316. <https://doi.org/10.1007/bf00916567>
- White, S., Hill, E., Happé, F. G., & Frith, U. (2009). Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child Development*, 80(4), 1097–1117. <https://doi.org/10.1111/j.1467-8624.2009.01319.x>
- Zajenkowska, A., Bower Russa, M., Rogoza, R., Park, J., Jasielska, D., & Skrzypek, M. (2020). Cultural influences on social information processing: Hostile attributions in the United States, Poland, and Japan. *Journal of Personality Assessment*, 1–9. <https://doi.org/10.1080/00223891.2020.1774380>
- Zajenkowska, A., Prusik, M., & Szulawski, M. (2018). What does the Ambiguous Intentions Hostility Questionnaire really measure? The importance of context in evaluating hostility bias. *Journal of Personality Assessment*, 102(2), 205–213. <https://doi.org/10.1080/00223891.2018.1525389>
- Zhang, X., Lecce, S., Ceccato, I., Cavallini, E., Zhang, L., & Chen, T. (2018). Plasticity in older adults' theory of mind performance: The impact of motivation. *Aging & Mental Health*, 22(12), 1592–1598. <https://doi.org/10.1080/13607863.2017.1376313>