Advance Access publication May 4, 2015

Social Cognition Psychometric Evaluation: Results of the Initial Psychometric Study

Amy E. Pinkham*[,1,2], David L. Penn[3,4], Michael F. Green[5,6], and Philip D. Harvey[7,8]

[1]School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX; [2]Department of Psychiatry, University of Texas Southwestern Medical School, Dallas, TX; [3]Department of Psychology, University of North Carolina, Chapel Hill, NC; [4]Department of Psychology, Australian Catholic University, Melbourne, Victoria, Australia; [5]Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA; [6]Department of Veterans Affairs, Desert Pacific Mental Illness Research, Education, and Clinical Center, Los Angeles, CA; [7]Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, Miami, FL; [8]Research Service, Miami VA Healthcare System, Miami, FL.

*To whom correspondence should be addressed; School of Behavioral and Brain Sciences, The University of Texas at Dallas, 800 West Campbell Road, GR 41, Richardson, TX 75080; tel: 972 883 4462, fax: 972 883 2491, e-mail: amy.pinkham@utdallas.edu

Measurement of social cognition in treatment trials remains problematic due to poor and limited psychometric data for many tasks. As part of the Social Cognition Psychometric Evaluation (SCOPE) study, the psychometric properties of 8 tasks were assessed. One hundred and seventy-nine stable outpatients with schizophrenia and 104 healthy controls completed the battery at baseline and a 2–4-week retest period at 2 sites. Tasks included the Ambiguous Intentions Hostility Questionnaire (AIHQ), Bell Lysaker Emotion Recognition Task (BLERT), Penn Emotion Recognition Task (ER-40), Relationships Across Domains (RAD), Reading the Mind in the Eyes Task (Eyes), The Awareness of Social Inferences Test (TASIT), Hinting Task, and Trustworthiness Task. Tasks were evaluated on: (i) test-retest reliability, (ii) utility as a repeated measure, (iii) relationship to functional outcome, (iv) practicality and tolerability, (v) sensitivity to group differences, and (vi) internal consistency. The BLERT and Hinting task showed the strongest psychometric properties across all evaluation criteria and are recommended for use in clinical trials. The ER-40, Eyes Task, and TASIT showed somewhat weaker psychometric properties and require further study. The AIHQ, RAD, and Trustworthiness Task showed poorer psychometric properties that suggest caution for their use in clinical trials.

Key words: schizophrenia/measurement/reliability/validity/emotion processing/social perception/mental state attribution

Introduction

The importance of social cognition for schizophrenia research is substantial and growing rapidly, largely based on data showing that social cognition predicts functioning[1,2] and that treating social cognitive impairment leads to improvements in real-world social outcomes.[3–5] However, the paucity of well-validated measures of social cognition remains an ongoing challenge for productive research. Most existing measures have poor psychometric properties, or their psychometrics are not known. Inadequate measurement can compromise validity and reproducibility of findings and limits treatment development and evaluation by rendering it difficult to accurately assess treatment response.

The Social Cognition Psychometric Evaluation (SCOPE) Study seeks to address this problem by systematically evaluating the psychometric properties of the most widely used measures of social cognition. Phases 1 and 2 of the project utilized expert surveys and the RAND Appropriateness Method of consensus development to select the best existing measures based on current knowledge of their psychometric properties and their potential for use in clinical trials. Eight measures of social cognition covering 4 domains and 1 "novel" category were identified.[6] In phase 3, large samples of individuals with schizophrenia and healthy controls completed the measures to assess the reliability and validity of each task.

In this article, we report the results of Phase 3 constituting the initial psychometric study of the measures' properties. Consistent with other National Institute of Mental Health measurement initiatives (eg, Measurement and Treatment Research to Improve Cognition in Schizophrenia[7] and Social Cognition and Functioning in Schizophrenia[8–10]), we report data on those characteristics rated most important for evaluation

of measures to be used in clinical trials, including: (i) test-retest reliability, (ii) utility as a repeated measure, (iii) relationship to functional outcome, and (iv) practicality and tolerability.[11] Sensitivity to change was also identified as a key criterion for clinical trials; however, the lack of a treatment component in this study precluded evaluation of this criterion. As the evaluated measures are used extensively in nonintervention research, we also report data on the sensitivity of these measures to differences between patients and healthy controls and internal consistency. Finally, Phase 3 concluded by reconvening a subset of the initial RAND Panelists and the study consultants to review the data and determine which tasks were appropriate for continued investigation. Recommendations of this Panel are presented at the end of the results.

Methods

Participants

The study took place at 2 sites, Southern Methodist University (SMU) and the University of Miami Miller School of Medicine (UM). Patients at the SMU site were recruited from Metrocare Services, a nonprofit mental health services provider organization in Dallas County, TX, and other area clinics. UM patient recruitment occurred at the Miami VA Medical Center and the Jackson Memorial Hospital-University of Miami Medical Center. At both sites, healthy controls were recruited via community advertisements.

To be eligible, patients required a Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) diagnosis of schizophrenia or schizoaffective disorder as confirmed by clinical interview with the Mini International Neuropsychiatric Interview[12] and Structured Clinical Interview for DSM Disorders Psychosis Module.[13] Patients could not have any hospitalizations within the last 2 months and had to be on a stable medication regimen for a minimum of 6 weeks with no dose changes for a minimum of 2 weeks. Healthy controls were screened for history of psychopathology to ensure they did not meet criteria for any major DSM-IV Axis I or II disorders. Exclusion criteria for both groups included: (i) presence or history of pervasive developmental disorder or mental retardation (defined as IQ < 70) by DSM-IV criteria, (ii) presence or history of medical or neurological disorders that may affect brain function (eg, seizures, central nervous system tumors, or loss of consciousness for 15 min or more), (iii) presence of sensory limitation including visual (eg, blindness, glaucoma, vision uncorrectable to 20/40) or hearing impairments that interfere with assessment, (iv) no proficiency in English, (v) presence of substance abuse in the past month, and (vi) presence of substance dependence not in remission for the past 6 months.

Measures

Social Cognition Measures

Attribtuional Style/Bias

The Ambiguous Intentions and Hostility Questionnaire (AIHQ).[14] The AIHQ evaluates hostile social cognitive biases. Participants read 5 hypothetical, negative situations with ambiguous causes (ie, they could be intentional or accidental), imagined the scenario happening to them, and recorded a reason why the scenario occurred. Participants then used Likert scales to rate whether the other person/s performed the action on purpose (1 "definitely no" to 6 "definitely yes"), how angry it made them feel (1 "not angry at all" to 5 "very angry"), and how much they blamed the other person/s (1 "not at all" to 5 "very much"). Finally, the participant wrote down how they would respond to the situation. Responses to open-ended questions were coded by 2 independent raters to compute a hostility bias (HB) index and an aggression bias (AB) index, respectively, ranging from 1 to 5 (Intraclass correlation coefficients, ICC (3,2), ranged from .834 to .967 for the individual items). A Blame Score (BS) was computed by averaging Likert ratings to each question and then summing the 3 averages (range = 3–16).

Emotion Processing

Bell Lysaker Emotion Recognition Task (BLERT).[15] The BLERT measures the ability to correctly identify 7 emotional states: happiness, sadness, fear, disgust, surprise, anger, or no emotion. Participants viewed 21 10-second video clips of a male actor, providing dynamic facial, vocal-tonal, and upper-body movement cues. After viewing each video, participants identified the expressed emotion. Performance was indexed as the total number of correctly identified emotions (ranging from 0 to 21).

Penn Emotion Recognition Text (ER-40).[16] The ER-40 includes 40 color photographs of static faces expressing 4 basic emotions (ie, happiness, sadness, anger, or fear) and neutral expressions. Stimuli are balanced for poser's gender, age, and ethnicity, and for each emotion category, 4 high-intensity and 4 low-intensity expressions are included. Participants viewed 1 image at a time and chose the correct emotion label for each face. Accuracy scores, ranging from 0 to 40, were the primary dependent variable.

Social Perception

Relationships Across Domains (RAD).[17] The RAD measures competence in the perception of 4 relational models: communal sharing, authority ranking, equality matching, and market pricing. The abbreviated version is comprised of 15 vignettes involving different male-female dyads that represent one of the relational models. Participants read each vignette and answered 3 yes/no questions about whether a future behavior was likely to happen given the described relationship. Performance was indexed as the total number of correct responses (ranging from 0 to 45).

Theory of Mind/Mental State Attribution

Reading the Mind in the Eyes Test (Eyes).[18] The Eyes task measures the capacity to discriminate the mental state of others from expressions in the eye region of the face. Participants viewed 36 photos of the eye region of different faces and chose the most accurate descriptor word for the thought/feeling that was portrayed. Four possible options were presented with each photo, and a glossary of mental state terms was provided for reference. The dependent measure was the total number of correct responses, ranging from 0 to 36.

The Awareness of Social Inferences Test, Part III (TASIT).[19] The TASIT assesses detection of lies and sarcasm. Participants watched short videos of everyday social interactions and answered 4 standard questions per video that probed understanding of the intentions, beliefs, and meanings of the speakers and their exchanges. Total number correct indexed performance, and scores ranged from 0 to 64.

Hinting Task.[20] The Hinting Task examines the ability of individuals to infer the true intent of indirect speech. Ten short passages present an interaction between 2 characters, and each passage ends with one of the characters dropping a hint. Passages were read aloud by the experimenter, and participants were asked what the character truly meant. If the first response provided was inaccurate, a second hint was delivered, allowing participants to earn partial credit for that passage. Total scores ranged from 0 to 20.

Novel Category

Trustworthiness Task (Trust)[21]

This task assesses participants' ability to make complex social judgments of trustworthiness. Participants rated 42 faces for trustworthiness on a scale from −3 to 3. Faces were presented in grayscale and represented ethnically diverse males and females. The average rating across all faces served as the primary outcome variable.

Additional Outcome Variables

In addition to the primary outcome variables, tolerability and practicality were assessed for each of the social cognitive measures. Task tolerability referred to the degree to which participants found the task enjoyable and was rated on a scale from 1 (very unpleasant) to 7 (very pleasant). Ratings of 4 indicated neither pleasant nor unpleasant. Practicality was operationalized as administration time.

Neurocognitive Measures

Given our emphasis on social cognition and previous work investigating the domains accounting for the most variance in composite scores of neurocognitive performance,[22] participants completed only a subset of the MARTICS Consensus Cognitive Battery.[7] Assessed domains included speed of processing (Trail Making Test, Part A; BACS: Symbol Coding; and Category Fluency: Animal Naming), working memory (Letter-Number Span), and verbal learning (HVLT-R). The Wide Range Achievement Test-3 Reading subscale provided an estimate of premorbid IQ.[23]

Functional Outcome Measures

UCSD Performance-Based Skills Assessment, Brief (UPSA-B).[24] The UPSA-B is a widely used measure of functional capacity that assesses financial and communication skills required for community living. Total scores could range from 0 to 100.

Social Skills Performance Assessment (SSPA).[25] Social competence was assessed with the SSPA, a role-play measure in which participants were asked to initiate and maintain a conversation in 2 social situations: meeting a new neighbor and negotiating with a landlord to fix a leak. Role-plays were audiotaped and coded by an expert rater blind to diagnosis on the following variables: interest, fluency, clarity, focus, overall abilities, and social appropriateness. The landlord role-play was also coded for negotiation ability and persistence. The mean score across both role-plays was used as the dependent measure and could range from 1 to 5.

Specific Level of Functioning Scale (SLOF).[26] Real-world functional outcome was assessed via the 31-item version of the SLOF, an informant-rated measure of social functioning (interpersonal relationships and social acceptability) and community-living skills (participation in activities and work skills). Informants were identified by the participants and were high contact clinicians, family members, or close friends. Ratings for each item were made on a 1–5 point scale with higher scores indicating better functioning. An average score across the entire measure was used as the dependent variable.

Procedures

All participants completed 2 study visits: baseline and a retest assessment completed 2–4 weeks after the initial visit (mean interval = 17.29 days). At visit 1, all participants provided informed consent and completed the social cognitive, neurocognitive, and functional outcome measures. The order of these task blocks was counterbalanced, and within the social cognitive battery, the order of individual tasks was counterbalanced as well. For patients, visit 1 also included diagnostic assessment and an evaluation of symptom severity using the Positive and Negative Syndrome Scale.[27] Diagnostic and symptom raters were trained to reliability using the established procedures at each site. At visit 2, symptom severity was reassessed in the patients, and all participants repeated the social cognitive measures in the same order as their first visit. For TASIT, an alternative form (TASIT-B) was administered to all subjects at visit 2; however, alternative forms were not available for any other social cognitive task, so these were identical to visit 1. Visit durations were approximately 3.5–4.5 hours for visit 1 and 3 hours for visit 2.

Statistical Analyses

Score distributions of the social cognitive measures were first checked for normality by examining skew and kurtosis statistics and visually inspecting histograms. No measures required transformation; however, 1 control participant was an outlier on the Hinting task (< 3SD from the mean); these data were excluded from further analyses. Test-retest reliability was computed using Pearson's *r* correlation coefficients. Utility as a repeated measure was evaluated by assessing evidence for practice effects (paired-samples *t*-tests with Cohen's d_z) and floor/ceiling effects (number of participants scoring at/below chance levels or scoring 100%).

To examine relationship to functional outcome among patients, 3 steps were followed. First, correlations were calculated between the visit 1 social cognitive and neurocognitive measures and the 3 outcome measures. Second, those social cognitive tasks showing a significant correlation with each outcome where then entered into regression models in a single block to assess the explanatory power of the tasks as group. Third, hierarchical regression models were conducted with neurocognitive variables entered in block 1 and social cognitive variables entered in block 2. Together, these analyses allowed for an examination of criterion validity and incremental validity beyond neurocognitive abilities.

Descriptive statistics assessed practicality and tolerability. Independent samples *t*-tests with Cohen's *d* were used to examine group differences. Finally, internal consistency was evaluated with Cronbach's alpha.

Results

Participants

Across the 2 sites, 179 patients and 104 healthy controls completed visit 1, with 171 and 98 participants, respectively, completing visit 2. Groups did not differ on race, ethnicity, age, parental education, or estimated IQ. Patients completed fewer years of education than controls, and there were more males than females in the patient sample. Patients reported relatively low levels of symptoms at visit 1, and there were slight reductions in positive and general symptoms at visit 2 (positive: $t(170) = 2.05$, $P = .042$, $d_z = .16$; negative: $t(170) = 2.54$, $P = .012$, $d_z = .19$). Demographic and clinical characteristics are provided in table 1.

Site Effects

Site differences in patient performance on the social cognitive measures at visit 1 were examined. SMU patients scored higher than UM patients on both the Eyes task ($t(178) = 4.65$, $P = .032$, $d = .33$) and TASIT ($t(178) = 3.95$, $P = .048$, $d = .31$). No other comparisons were statistically significant.

Table 1. Participant Demographic and Clinical Characteristics

	Patients (n = 179)		Controls (n = 104)	
Characteristic	*N*	(%)	*N*	(%)
Male*	117	65	49	47
Race				
Caucasian	76	42	43	41
African American	94	53	55	53
Native American	1	1	0	0
Asian	4	2	4	4
Other	4	2	2	2
Ethnicity				
Hispanic	37	21	21	20
Non-Hispanic	142	79	83	80
Diagnosis				
Schizophrenia	96	54		
Schizoaffective	83	46		
Medication type[a]				
Typical	26	15		
Atypical	125	70		
Combination	3	2		
	Mean	SD	Mean	SD
Age (years)	42.11	12.32	39.20	13.70
Education (years)*	12.70	2.14	13.43	1.66
Maternal education (years)	12.61	3.22	13.14	2.53
Paternal education (years)	13.04	3.75	13.43	2.49
WRAT-3	93.68	15.88	95.35	13.19
PANSS				
Positive total	16.14	5.79		
Negative total	13.72	5.29		
General total	30.83	7.99		

Note: WRAT, Wide Range Achievement Test; PANSS, Positive and Negative Syndrome Scale.
[a]Nineteen individuals were not taking antipsychotic medications, and medication information was missing for 6 individuals.
*$P < .01$.

Test-Retest Reliability

Adopting a range of Pearson's r values of .6 to .8 as good,[28,29] test-retest reliability was adequate for the majority of measures. Only the 2 bias measures of the AIHQ (ie, hostile and aggressive biases) showed inadequate values among patients. For healthy controls, test-retest reliability was generally lower than in patients, with AIHQ-HB, Hinting task, TASIT and Trust task all having values below benchmark standards (table 2).

Utility as a Repeated Measure

Among patients, all tasks except the Eyes task showed statistically significant differences between the first and second administration (table 3). AIHQ-AB, BLERT, ER-40, Hinting, RAD and Trust all showed improved scores at visit 2, and TASIT, AIHQ-HB, and AIHQ-BS scores significantly decreased between administrations. However, the effect sizes for these differences were all small (range 0.15–0.27). Floor effects were most pronounced RAD, wherein 43% of patients scored at or below chance levels

at visit 1. This percentage decreased to 33% at visit 2. For the remaining tasks, less than 7% of the sample scored at floor or ceiling.

Healthy control performance also showed practice effects for AIHQ, ER-40, and Hinting. TASIT performance again worsened at visit 2. Effect sizes for these differences ranged from small to medium. RAD continued to have a number of participants, 12%, scoring at floor levels. Additionally, approximately 7% of the sample scored at ceiling for the Hinting task at both time points.

Table 2. Test-Retest Reliability and Internal Consistency

	Test-Retest Reliability (Person *r*)		Internal Consistency (Cronbach's Alpha)	
Task	Patients (*n* = 171)	Controls (*n* = 98)	Patients (*n* = 179)	Controls (*n* = 104)
AIHQ				
Hostility bias (HB)	.516	.572	.859	.846
Aggression bias (AB)	.572	.700	.422	.467
Blame Score (BS)	.738	.756	.491	.338
BLERT	.699	.680	.737	.626
ER-40	.753	.753	.808	.645
Eyes	.753	.761	.735	.673
Hinting	.639	.424	.729	.563
RAD	.751	.756	.717	.700
TASIT	.600	.544	.807	.757
Trust	.737	.597	.960	.900

Notes: AIHQ, Ambiguous Intentions Hostility Questionnaire; BLERT, Bell Lysaker Emotion Recognition Task; RAD, Relationships Across Domains; TASIT, The Awareness of Social Inferences Test.

Relationship to Functional Outcome

Correlations between the social and neurocognitive tasks and the functional outcome tasks for patients are presented in table 4. With the exception of the AIHQ and Trust tasks, social cognitive tasks showed significant positive correlations with outcomes. The magnitude of these relations ranged from small to medium (0.20–0.46). Neurocognitive tasks were also significantly related to outcomes with comparable magnitudes (0.17–0.54).

We assessed the explanatory power of social cognitive performance by first entering those tasks showing a significant correlation with outcome as a single block. The social cognitive tasks significantly accounted for 31% of the variance in functional capacity as measured by the UPSA-B (adjusted $R^2 = .308$, $F(6,168) = 13.92$, $P < .001$), 16% of the variance in social competence as measured by the SSPA (adjusted $R^2 = .156$, $F(6,167) = 6.35$, $P < .001$), and 10% of the variance in real-world functioning indexed by the SLOF (adjusted $R^2 = .104$, $F(4,169) = 6.04$,

Table 3. Utility as a Repeated Measure

	T_1		T_2		T_2–T_1 Difference		Number at Floor/Ceiling				
Task	Mean	SD	Mean	SD	Mean	SD	T_1	T_2	*t*	*P* value	Cohen's d_z
Patients (*n* = 171)											
AIHQ-HB	2.38	0.61	2.21	0.64	−0.17	0.62	—	—	−3.57	<.001	0.27
AIHQ-AB	1.88	0.39	1.95	0.44	0.06	0.39	—	—	2.05	.04	0.16
AIHQ-BS	8.76	2.85	8.42	3.06	−0.34	2.15	—	—	−2.06	.04	0.16
BLERT	13.24	3.82	13.91	3.99	0.67	3.04	1/0	0/4	2.87	.005	0.22
ER-40	29.69	5.37	30.42	4.95	0.73	3.65	1/0	0/0	2.62	.01	0.20
Eyes	20.22	5.52	20.66	5.85	0.44	4.00	5/0	4/0	1.43	.15	0.11
Hinting	13.65	3.80	14.25	3.68	0.60	3.18	0/2	0/2	2.46	.02	0.19
RAD	24.79	5.79	25.86	5.70	1.07	4.06	77/0	56/0	3.40	.001	0.26
TASIT	44.55	7.55	42.92	6.36	−1.63	6.31	12/0	9/0	−3.37	.001	0.26
Trust	−0.12	1.13	−0.002	0.91	0.12	0.77	—	—	2.01	.05	0.15
Controls (*n* = 98)											
AIHQ-HB	2.00	0.60	1.78	0.53	−0.22	0.53	—	—	-4.19	<.001	0.42
AIHQ-AB	1.83	0.26	1.82	0.31	−0.01	0.22	—	—	−0.28	.78	0.05
AIHQ-BS	7.08	2.30	6.34	2.41	−0.73	1.65	—	—	−4.41	<.001	0.44
BLERT	15.74	2.89	16.12	2.96	0.38	2.34	0/2	0/1	1.59	.11	0.16
ER-40	32.61	3.53	33.13	3.41	0.52	2.44	0/0	0/0	2.11	.04	0.21
Eyes	23.50	4.71	23.55	5.34	0.05	3.52	0/0	2/0	0.14	.89	0.01
Hinting	16.85	2.01	17.45	1.50	0.59	1.93	0/6	0/7	3.02	.003	0.31
RAD	29.87	5.21	30.45	5.61	0.58	3.80	12/0	9/0	1.52	.13	0.15
TASIT	51.44	5.68	48.21	6.58	−3.22	5.91	0/0	0/0	−5.40	<.001	0.54
Trust	0.18	0.60	0.24	0.58	0.06	0.53	—	—	1.11	.27	0.11

Notes: AIHQ, Ambiguous Intentions Hostility Questionnaire; BLERT, Bell Lysaker Emotion Recognition Task; RAD, Relationships Across Domains; TASIT, The Awareness of Social Inferences Test.

$P < .001$). When restricting the sample to individuals with high-quality informants (ie, professionals with mental health experience, $n = 137$),[30] AIHQ-BS was also significantly correlated to SLOF scores, and the predictive ability of the social cognitive variables improved to 16% variance in SLOF ratings (adjusted $R^2 = .159$, $F(5,131) = 6.13$, $P<.001$). Details are provided in table 5.

Next, we examined the incremental validity of the social cognitive tasks by determining whether they would significantly predict variance above and beyond neurocognitive performance (table 6). For UPSA-B, the neurocognitive variables alone predicted 31% of the variance in scores (adjusted $R^2 = .310$, $F(5,165) = 16.29$, $P < .001$), which significantly increased to 37% when the social cognitive variables were added (adjusted $R^2 = .372$, $F(11,159) = 10.16$, $P < .001$; R^2 change = .082, $P = .002$). Neurocognitive variables also significantly predicted social competence (SSPA) variance (adjusted $R^2 = .146$, $F(4,165) = 8.23$, $P < .001$). The inclusion of the social cognitive variables contributed an additional 7% to the total variance (R^2 change = .068, $P = .032$), constituting a significant increase and bringing the overall model to approximately 19% of variance explained (adjusted $R^2 = .187$, $F(10,159) = 4.87$, $P < .001$). For the SLOF, both models accounted for significant variance (neurocognitive: adjusted $R^2 = .085$, $F(4,166) = 4.93$, $P = .001$ and neurocognitive plus social cognitive: adjusted $R^2 = .112$, $F(8,162) = 3.67$, $P = .001$), but the increase between models was not significant (R^2 change = .047, $P = .065$). When including data only from high-quality informants, neurocognitive performance alone accounted for 5% of the variance (adjusted $R^2 = .050$, $F(4,129) = 2.76$, $P = .03$), and the combined social and neurocognitive variables accounted

Table 4. Correlations between Social Cognitive Tasks and Functional Outcome Measures in Patients

	UPSA Total	SSPA Average	SLOF Total
Social cognitive			
AIHQ-HB	−.071	.063	−.058
AIHQ-AB	.041	.078	−.071
AIHQ-BS	−.005	.094	−.137
BLERT	.317***	.261***	.310***
ER-40	.360***	.240***	.046
Eyes	.425***	.300***	.127
Hinting	.462***	.394***	.197**
RAD	.439***	.243**	.202**
TASIT	.437***	.310**	.304***
Trust	.052	−.030	.043
Neurocognitive			
Trails A	−.270***	−.103	−.237**
Symbol coding	.264***	.301***	.263***
HVLT-R	.421***	.358***	.174*
Letter number span	.544***	.317***	.255**
Animal naming	.174*	.168*	.078

Notes: AIHQ, Ambiguous Intentions Hostility Questionnaire; BLERT, Bell Lysaker Emotion Recognition Task; RAD, Relationships Across Domains; SLOF, Specific Level of Functioning Scale; TASIT, The Awareness of Social Inferences Test.
*$P < .05$, **$P < .01$, ***$P < .001$.

Table 5. Regression Models Demonstrating the Overall Contribution of the Social Cognitive Tasks to Outcomes

	R^2	Adjusted R^2	F	P	b*	t	P	sr^2
UPSA total	.332	.308	13.92	<.001				
BLERT					−.08	−.88	.382	.003
ER-40					.11	1.28	.20	.007
Eyes					.09	.99	.32	.004
Hinting					.29	3.91	<.001	.06
RAD					.19	2.29	.02	.02
TASIT					.13	1.47	.15	.008
SSPA average	.186	.156	6.35	<.001				
BLERT					.04	.39	.70	.000
ER-40					.05	.48	.63	.001
Eyes					.08	.77	.44	.002
BLERT					.21	2.35	.02	.03
Hinting					.04	.52	.60	.001
RAD					−.02	−.26	.79	.000
TASIT					.19	1.94	.05	.02
SLOF-HQ	.190	.159	6.13	<.001				
AIHQ-BS					−.12	−1.52	.13	.01
BLERT					.38	3.83	<.001	.09
Hinting					−.01	−.08	.94	.000
RAD					−.05	−.47	.64	.001
TASIT					.08	.72	.47	.003

Notes: SLOF-HQ indicates ratings from high quality informants (ie, professionals with mental health experience). AIHQ, Ambiguous Intentions Hostility Questionnaire; BLERT, Bell Lysaker Emotion Recognition Task; RAD, Relationships Across Domains; TASIT, The Awareness of Social Inferences Test. b* indicates standardized coefficients.

for 13% variance (adjusted R^2 = .133, $F(9,124)$ = 3.27, P = .001). The social cognitive tasks accounted for an additional 11% of variance above and beyond neurocognition alone (R^2 change = .113, P = .006).

Practicality and Tolerability

Ratings of practicality and tolerability were implemented approximately midway through the study duration, and thus 95 patients and 59 controls provided data. As seen in table 7, administration time was under 8 minutes for the majority of tasks. The RAD and TASIT each took more than twice as long as the other tasks. As expected, completion time for patients was longer than controls. Participants also rated all tasks to be pleasant with comparable ratings between tasks. The RAD and TASIT received the lowest ratings from both patients and controls.

Group Differences

Patient and control performance significantly differed on all measures and indices except for AIHQ-AB (table 8). As expected, on AIHQ-HB and AIHQ-BS, patients scored higher than controls, and on the remaining measures,

Table 6. Final Regression Models Accounting for Additional Variance in Outcome beyond Neurocognitive Performance

	UPSA-B		SSPA		SLOF		SLOF-HQ	
	*b**	sr^2	*b**	sr^2	*b**	sr^2	*b**	sr^2
Block 1—Neurocognition								
Trails A	−.157*	.016*	—	—	−.109	.008	−.060	.002
Symbol coding	−.106	.006	.136,	.013	.069	.003	.007	.000
HVLT-R	.087	.004	.191*	.021*	−.028	.000	.005	.000
Letter number span	.310**	.046**	.071	.003	.087	.004	.072	.003
Animal Naming	−.050	.002	−.008	.000	—	—	—	—
Block 2—Social cognition								
AIHQ-BS	—	—	—	—	—	—	−.094	.008
BLERT	−.116	.007	.005	.000	.154	.015	.353**	.076**
ER-40	.094	.005	−.022	.000	—	—	—	—
Eyes	.043	.000	.052	.001	—	—	—	—
Hinting	.242**	.041**	.258**	.047**	.038	.001	−.020	.000
RAD	.082	.003	−.095	.004	−.060	.002	−.071	.003
TASIT	.090	.004	.059	.002	.158	.013	.065	.002
Overall model								
Adjusted R^2	.372***		.187***		.112**		.133**	
R^2 change	.082**		.068*		.047		.113**	

Notes: SLOF-HQ indicates ratings from high quality informants (ie, professionals with mental health experience). AIHQ, Ambiguous Intentions Hostility Questionnaire; BLERT, Bell Lysaker Emotion Recognition Task; RAD, Relationships Across Domains; TASIT, The Awareness of Social Inferences Test. *b** indicates standardized coefficients.
*P < .05, **P < .01, ***P < .001.

Table 7. Practicality and Tolerability

	Practicality (Administration Time in Minutes)				Tolerability (Participant Ratings)			
	Patients (n = 95)		Controls (n = 59)		Patients (n = 95)		Controls (n = 59)	
Task	Mean	SD	Mean	SD	Mean	SD	Mean	SD
AIHQ	6.35	2.01	5.82	1.61	5.54	1.30	5.73	1.20
BLERT	7.09	1.50	6.94	0.99	5.14	1.72	5.54	1.58
ER-40	3.21	1.02	2.73	0.73	5.55	1.40	5.59	1.41
Eyes	6.56	3.56	5.45	2.58	5.43	1.59	5.31	1.33
Hinting	6.13	1.89	5.33	1.46	5.38	1.44	5.60	1.50
RAD	15.84	4.45	13.82	3.15	4.74	1.78	4.70	1.53
TASIT	17.92	3.93	17.46	2.12	5.04	1.59	4.83	1.67
Trust	4.46	2.78	3.48	1.29	5.28	1.66	5.19	1.76

Notes: AIHQ, Ambiguous Intentions Hostility Questionnaire; BLERT, Bell Lysaker Emotion Recognition Task; RAD, Relationships Across Domains; TASIT, The Awareness of Social Inferences Test.

Table 8. Group Differences on Social Cognitive Measures

	Patients (n = 179)		Controls (n = 104)				
Task	Mean	SD	Mean	SD	t	P	Cohen's d
AIHQ-HB	2.38	0.60	1.99	0.60	5.29	<.001	.65
AIHQ-AB	1.89	0.38	1.83	0.26	1.46	.147	.18
AIHQ-BS	8.74	2.81	7.02	2.31	5.29	<.001	.67
BLERT	13.17	3.88	15.75	2.88	−6.38	<.001	.76
ER-40	29.55	5.40	32.80	3.54	−6.10	<.001	.71
Eyes	20.15	5.46	23.55	4.62	−5.58	<.001	.67
Hinting	13.59	3.87	16.82	2.05	−9.14	<.001	1.04
RAD	24.76	5.76	29.82	5.16	−7.37	<.001	.93
TASIT	44.43	7.64	51.48	5.62	−8.89	<.001	1.05
Trust	−0.09	1.14	0.16	0.62	−2.33	.02	.27

Notes: AIHQ, Ambiguous Intentions Hostility Questionnaire; BLERT, Bell Lysaker Emotion Recognition Task; RAD, Relationships Across Domains; TASIT, The Awareness of Social Inferences Test.

control participants outperformed patients. Effect sizes varied but were generally large (range: 0.65–1.05) with the exception of the trust task which showed only a small effect (d = 0.27).

Internal Consistency

Kraemer has pointed out that Cronbach's alpha should not be used as a reliability coefficient and is therefore of limited utility in the evaluation of measures for use in clinical trials.[31] Cronbach's alpha can however be informative during the development of a measure, and we have therefore chosen to include this information in the event that it may aid in the further development of these measures.

A Cronbach's alpha of .80 is considered appropriate for research tools.[32] For patients, almost all measures either approached or exceeded this value (table 2). The one notable exception was the AIHQ for which both the AB and BS indices were much lower, 0.42 and 0.49, respectively. These scales included the fewest number of items, which could explain the lower values.

Recommendations from Follow-up Panel

A subset of the original RAND Panel (participants listed in Appendix) and the study consultants reviewed the psychometric data and classified each task as: Acceptable As Is, Acceptable with Modifications, or Not Recommended for Further Consideration. Consensus for these classifications was achieved via a conference call. No panelist reported financial conflicts of interest; however, Drs Green and Lysaker each reported involvement in the development of one of the measures under consideration. In addition, one of the Principal Investigators, who did not participate in the Rand Panel (DLP), also developed one of the measures under consideration (AIHQ).

BLERT, ER-40, and Hinting were all classified as Acceptable As Is. The panel did however express some concern about overlap between the BLERT and ER-40 (r = .59 at both time points; supplementary table 1), commenting that the BLERT is currently more suitable because it predicted real-world functional outcome whereas the ER-40 did not. The Eyes task was also rated as acceptable, but a concern included the potential dependence of performance on vocabulary and the somewhat limited relation with outcomes. The TASIT was also rated as acceptable, but the Panel noted a need to clarify whether the differences between performance at visits 1 and 2 were due to interference from previous administration or non-equivalence between the 2 task forms.

AIHQ, RAD, and Trust were classified as Not Recommended for Further Consideration. For AIHQ, the primary concern was the limited relation with functional outcomes. Concerns about the RAD included length, patient tolerability, and difficulty as a high proportion of patients performed at chance levels. Lack of a unique contribution to outcomes was also considered. For Trust, concerns included a reduced ability to distinguish patients from controls and the lack of a relation with functional outcomes.

Discussion

In the third phase of the Social Cognition Psychometric Evaluation study, we conducted a systematic investigation of the reliability and validity of 8 social cognitive measures and presented our findings to a panel of experts to obtain consensus on which tasks could currently be recommended for use in clinical trials. Five social cognitive tasks (BLERT, ER-40, Eyes, Hinting, and TASIT) displayed acceptable psychometric characteristics, while 3 tasks (AIHQ, RAD, and Trust) showed weaker characteristics, suggesting that they may be of more limited use in clinical trials. The BLERT and Hinting task were the strongest of the 5 acceptable tasks. Both showed adequate test-retest reliability, small practice effects, and limited potential for floor/ceiling effects. They also distinguished patient performance from controls, were well tolerated by patients, and could be administered relatively quickly. Notably these 2 tasks showed the strongest relation to functional outcomes and uniquely predicted variance in outcomes while controlling for all other social and neurocognitive variables. The Hinting task emerged as a uniquely significant predictor of functional capacity and social competence, and BLERT was the only uniquely significant predictor of real-world outcomes as rated by high-quality informants. The psychometric properties of these tasks therefore recommended them for use in clinical trials seeking to improve those aspects of social cognition which have strong links to functioning; improvement on these tasks may be considered an intermediate target for treatments focusing on functional outcomes.

The ER-40, Eyes, and TASIT also generally showed adequate psychometric characteristics, but these tasks had some limitations. While each task was correlated with some outcomes and was included in the block of social cognitive variables that provided incremental validity for the prediction of outcomes beyond neurocognition, none emerged as uniquely significant. This raises questions of redundancy between tasks, particularly when tasks are drawn from the same social cognitive domain. For example, it is currently unclear whether the ER-40 offers a unique contribution beyond the BLERT. Additionally, previous work indicates that performance on Eyes requires a heavy demand on vocabulary,[33] which may interfere with accurate measurement of social cognitive ability. For TASIT, concerns exist about the equivalence of test forms, and the longer administration time may be prohibitive for some clinical trials.

Three tasks, AIHQ, RAD, and Trust, showed less consistent properties that warrant caution for their use in clinical trials. The AIHQ bias scores (but not the blame score) showed low test-retest reliability, and the RAD showed considerable floor effects along with lower patient tolerability and one of the longer administration times. The Trust task only weakly discriminated patient and control performance, and all 3 tasks showed limited correlations with outcomes. For these reasons, the reconvened expert panel recommended that these tasks be omitted from further study under the SCOPE project.

The ramifications of these omissions require consideration. First, removal of the AIHQ and the Trust task eliminates the only bias measures in the battery. Attention has recently been drawn to the distinction between social cognitive capacities (eg, the ability to generate emotional state representations) and social cognitive biases (eg, the tendency to interpret any negative emotion as angry) with the idea that both are likely involved in social dysfunction and should be targets of remediation.[34] The weaker psychometric properties of these tests do not negate the potential importance of studying biases. In fact, the functional significance of attributional biases may be most salient on outcomes that assess aggressive behavior, rather than general social/instrumental function,[35] and distrust biases are associated with important symptom dimensions (eg, paranoia).[36,37] Thus, while these tasks cannot currently be recommended for use in clinical trials (if general functional outcomes are the target), we encourage continued development in this area. Second, removal of the RAD currently leaves the domain of social perception unrepresented. The omission of these tasks raises questions about how best to assess the full range of social cognitive processes in a clinical trial.

The future phases of SCOPE will attempt to address this question as well as those limitations identified by the panel for the Eyes task and TASIT. The goals of phase 4 will be 2-fold. First, given that experts in our previous survey considered the domains of attribtuional style and social perception to be important for the study of social cognition in schizophrenia,[6] we will make one more attempt to find suitable assessments for these domains. We will re-examine the data gathered from our previous expert survey and consult directly with experts in the field to identify other measures that may be suitable replacements. Second, phase 4 will also collect new pilot data on a modified protocol. For Eyes, we have created a version of the task that provides definitions on the same screen as the stimuli, which may reduce dependence on vocabulary. We will also administer TASIT forms in a counterbalanced order to determine the equivalence of forms. Additionally, we will collect response time (RT) data on BLERT, ER-40, Eyes, and TASIT. These tasks are well suited for the collection of RT data, and this information may aid in the prediction of functional outcomes. The final phase of SCOPE, Phase 5, will recruit another large sample of patients and controls to investigate the psychometric properties of the new attributional style and social perception measures. Modifications to the remaining tasks that appear successful based on the pilot data from phase 4 will also be validated with this larger sample.

Finally, potential limitations of this study should be considered. The tasks evaluated here were identified based upon expert surveys and a consensus process, and it is currently unclear if these tasks are truly the best measures of social cognition. Future efforts may benefit from utilizing brain-based or social psychological frameworks to identify tasks for evaluation. Additionally, within the patient group, all tasks except the Eyes task showed small but significant practice effects. It is possible that our short test-retest interval contributed to these improvements; however, the presence of these effects highlights the need for equivalent alternative forms of these measures. Likewise, although the BLERT and Hinting task were judged by the panel to be ready for use, the test-retest reliabilities of these tasks fell at the lower end of the "good" range. Users of these tasks should therefore interpret their data with these lower values in mind. Thought should also be given to how the present results compare to previous reports using these measures. The SCOPE patient sample may differ in key ways from other samples, eg, in terms of educational or IQ levels, and this might have impacted some of the psychometric scores. However, the large, diverse sample used here is representative of individuals targeted for clinical trials.[38] Thus, current psychometric data indicate that the BLERT and Hinting task are appropriate for use in clinical trials seeking to improve social cognition in individuals with schizophrenia.

Supplementary Material

Supplementary material is available at http://schizophreniabulletin.oxfordjournals.org.

Funding

National Institute of Mental Health at National Institutes of Health (R01 MH093432 to P.H.D., D.L.P., and A.E.P.).

Acknowledgments

We would like to thank all the individuals who participated in this study and the following individuals for their assistance with data collection and management: Skylar Kelsven (SMU), Isis Nelson-Graham (SMU), Kelsey Ludwig (UNC), Gabriela Vargas (UM), and Belinda Robertson (UM). Dr Pinkham has received consulting fees from Otsuka America Pharmaceutical, Inc. Dr Harvey serves as a consultant/advisory board member for Boeheringer Ingelheim, Forest Labs, Forum Pharma Genentech, Otsuka America, Roche, Sanofi Sunovion, and Takeda. Dr Green has received consultant fees from AbbVie, DSP, Forum, Mnemosyne (scientific board), and Roche. He also reports receiving past research support from Amgen. Dr Green is an officer in the non-profit MATRICS Assessment, Inc., but receives no financial compensation. Dr Penn reports no conflicts of interest.

Appendix

RAND Panel Members
William Horan, PhD (University of California, Los Angeles)
Paul Lysaker, PhD (Roudebush VA Medical Center, Indianapolis, IN and Indiana University School of Medicine)
Helena Kramer, PhD (Stanford University)
Keith Payne, PhD (University of North Carolina at Chapel Hill)
Study Consultants
Michael Green, PhD (University of California, Los Angeles)
Jean Addington, PhD (University of Calgary)

References

1. Couture SM, Penn DL, Roberts DL. The functional significance of social cognition in schizophrenia: a review. *Schizophr Bull*. 2006;32:S44–SS63.
2. Fett AK, Viechtbauer W, Dominguez MD, Penn DL, van Os J, Krabbendam L. The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: A meta-analysis. *Neurosci Biobehav Rev*. 2011;35:573–588.
3. Lindenmayer JP, McGurk SR, Khan A, et al. Improving social cognition in schizophrenia: a pilot intervention combining computerized social cognition training with cognitive remediation. *Schizophr Bull*. 2013;39:507–517.
4. Roberts DL, Combs DR, Willoughby M, et al. A randomized, controlled trial of Social Cognition and Interaction Training (SCIT) for outpatients with schizophrenia spectrum disorders. *Br J Clin Psychol*. 2014;53:281–298.
5. Kurtz MM, Richardson CL. Social cognitive training for schizophrenia: a meta-analytic investigation of controlled research. *Schizophr Bull*. 2012;38:1092–1104.
6. Pinkham AE, Penn DL, Green MF, Buck B, Healey K, Harvey PD. The social cognition psychometric evaluation study: results of the expert survey and RAND panel. *Schizophr Bull*. 2014;40:813–823.
7. Nuechterlein KH, Green MF, Kern RS, et al. The MATRICS Consensus Cognitive Battery, part 1: test selection, reliability, and validity. *Am J Psychiatry*. 2008;165:203–213.
8. Green MF, Penn DL. Going from social neuroscience to schizophrenia clinical trials. *Schizophr Bull*. 2013;39:1189–1191.
9. Kern RS, Penn DL, Lee J, et al. Adapting social neuroscience measures for schizophrenia clinical trials, Part 2: trolling the depths of psychometric properties. *Schizophr Bull*. 2013;39:1201–1210.
10. Olbert CM, Penn DL, Kern RS, et al. Adapting social neuroscience measures for schizophrenia clinical trials, part 3: fathoming external validity. *Schizophr Bull*. 2013;39:1211–1218.
11. Green MF, Nuechterlein KH, Gold JM, et al. Approaching a consensus cognitive battery for clinical trials in schizophrenia: the NIMH-MATRICS conference to select cognitive domains and test criteria. *Biol Psychiatry*. 2004;56:301–307.
12. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry*. 1998;59 (suppl 20):22–33.
13. First MB, Spitzer RL, Gibbon M, Williams JBW. *Structured Clinical Interview for DSM-IV-TR Axis I disorders, research version, patient edition with psychotic screen (SCID-I/P W/ PSY SCREEN)*. New York, NY: Biometrics Research, New York State Psychiatric Institute; 2002.
14. Combs DR, Penn DL, Wicher M, Waldheter E. The Ambiguous Intentions Hostility Questionnaire (AIHQ): a new measure for evaluating hostile social-cognitive biases in paranoia. *Cogn Neuropsychiatry*. 2007;12:128–143.
15. Bryson G, Bell M, Lysaker P. Affect recognition in schizophrenia: a function of global impairment or a specific cognitive deficit. *Psychiatry Res*. 1997;71:105–113.
16. Kohler CG, Turner TH, Bilker WB, et al. Facial emotion recognition in schizophrenia: intensity effects and error pattern. *Am J Psychiatry*. 2003;160:1768–1774.
17. Sergi MJ, Fiske AP, Horan WP et al. Development of a measure of relationship perception in schizophrenia. *Psychiatry Res*. 2009;166:54–62.
18. Baron-Cohen S, Wheelwright S, Hill J, Raste Y, Plumb I. The 'Reading the mind in the eyes' Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J Child Psychol Psychiatry*. 2001;42:241–251.
19. McDonald S, Flanagan S, Rollins J, Kinch J. TASIT: A new clinical tool for assessing social perception after traumatic brain injury. *J Head Trauma Rehab*. 2003;18:219–238.
20. Corcoran R, Mercer G, Frith CD. Schizophrenia, symptomatology and social inference: investigating "theory of mind" in people with schizophrenia. *Schizophr Res*. 1995;17:5–13.
21. Adolphs R, Tranel D, Damasio AR. The human amygdala in social judgment. *Nature*. 1998;393:470–474.
22. Keefe RS, Bilder RM, Harvey PD, et al. Baseline neurocognitive deficits in the CATIE schizophrenia trial. *Neuropsychopharmacology*. 2006;31:2033–2046.

23. Weickert TW, Goldberg TE, Gold JM, Bigelow LB, Egan MF, Weinberger DR. Cognitive impairments in patients with schizophrenia displaying preserved and compromised intellect. *Arch Gen Psychiatry*. 2000;57:907–913.
24. Mausbach BT, Harvey PD, Goldman SR, Jeste DV, Patterson TL. Development of a brief scale of everyday functioning in persons with serious mental illness. *Schizophr Bull*. 2007;33:1364–1372.
25. Patterson TL, Moscona S, McKibbin CL, Davidson K, Jeste DV. Social skills performance assessment among older patients with schizophrenia. *Schizophr Res*. 2001;48:351–360.
26. Schneider LC, Struening EL. SLOF: a behavioral rating scale for assessing the mentally ill. *Social Work Research Abstracts* 1983;19:9–21.
27. Kay SR, Opler LA, Fiszbein A. *Positive and Negative Syndrome Scale: Manual*. North Tonawanda, NY: Multi-Health Systems, Inc.; 1992.
28. Kraemer HC, Kupfer DJ, Clarke DE, Narrow WE, Regier DA. DSM-5: how reliable is reliable enough? *Am J Psychiatry*. 2012;169:13–15.
29. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
30. Sabbag S, Twamley EM, Vella L, Heaton RK, Patterson TL, Harvey PD. Assessing everyday functioning in schizophrenia: not all informants seem equally informative. *Schizophr Res*. 2011;131:250–255.
31. Kraemer HC. Toward sound objective evaluation of clinical measures. *Am J Geriatr Psychiatry*. 2013;21:589–595.
32. Nunnally JC. *Psychometric Theory*. New York, NY: McGraw-Hill; 1967.
33. Peterson E, Miller SF. The eyes test as a measure of individual differences: how much of the variance reflects verbal IQ? *Front Psychol*. 2012;3:220.
34. Roberts DL, Pinkham AE. The future of social cognition in schizophrenia: Implications for the normative literature. In: Roberts DL, Penn DL, eds. *Social Cognition in Schizophrenia*. New York, NY: Oxford University Press; 2013:401–414.
35. Harris ST, Oakley C, Picchioni MM. A systematic review of the association between attributional bias/interpersonal style, and violence in schizophrenia/psychosis. *Aggress Violent Behav*. 2014;19:235–241.
36. Couture SM, Penn DL, Losh M, Adolphs R, Hurley R, Piven J. Comparison of social cognitive functioning in schizophrenia and high functioning autism: more convergence than divergence. *Psychol Med*. 2010;40:569–579.
37. Pinkham AE, Hopfinger JB, Pelphrey KA, Piven J, Penn DL. Neural bases for impaired social cognition in schizophrenia and autism spectrum disorders. *Schizophr Res*. 2008;99:164–175.
38. Buchanan RW, Keefe RS, Umbricht D, Green MF, Laughren T, Marder SR. The FDA-NIMH-MATRICS guidelines for clinical trial design of cognitive-enhancing drugs: what do we know 5 years later? *Schizophr Bull*. 2011;37:1209–1217.