**Schizophrenia Bulletin** vol. 44 no. 4 pp. 737–748, 2018
doi:10.1093/schbul/sbx117
Advance Access publication August 31, 2017

# Social Cognition Psychometric Evaluation: Results of the Final Validation Study

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

[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 Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, Miami, FL; [4]Research Service, Miami VA Healthcare System; [5]Department of Psychology, University of North Carolina, Chapel Hill, NC; [6]School of Psychology, Australian Catholic University, Melbourne, VIC

*To whom correspondence should be addressed; tel: (972) 883-4462, fax: 972-883-3491, e-mail: amy.pinkham@utdallas.edu

**Social cognition is increasingly recognized as an important treatment target in schizophrenia; however, the dearth of well-validated measures that are suitable for use in clinical trials remains a significant limitation. The Social Cognition Psychometric Evaluation (SCOPE) study addresses this need by systematically evaluating the psychometric properties of promising measures. In this final phase of SCOPE, eight new or modified tasks were evaluated. Stable outpatients with schizophrenia (*n* = 218) and healthy controls (*n* = 154) completed the battery at baseline and 2–4 weeks later across three sites. Tasks included the Bell Lysaker Emotion Recognition Task (BLERT), Penn Emotion Recognition Task (ER-40), Reading the Mind in the Eyes Task (Eyes), The Awareness of Social Inferences Test (TASIT), Hinting Task, Mini Profile of Nonverbal Sensitivity (MiniPONS), Social Attribution Task—Multiple Choice (SAT-MC), and Intentionality Bias Task (IBT). BLERT and ER-40 modifications included response time and confidence ratings. The Eyes task was modified to include definitions of terms and TASIT to include response time. Hinting was scored with more stringent criteria. MiniPONS, SAT-MC, and IBT were new to this phase. Tasks were evaluated on (1) test-retest reliability, (2) utility as a repeated measure, (3) relationship to functional outcome, (4) practicality and tolerability, (5) sensitivity to group differences, and (6) internal consistency. Hinting, BLERT, and ER-40 showed the strongest psychometric properties and are recommended for use in clinical trials. Eyes, TASIT, and IBT showed somewhat weaker psychometric properties and require further study. MiniPONS and SAT-MC 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/attributional style

## Introduction

The Social Cognition Psychometric Evaluation (SCOPE) study is a five-phase project designed to improve measurement of social cognition in schizophrenia by systematically evaluating the psychometric properties of the most widely used measures. SCOPE implemented a sequential consensus-based, empirical program of research that relied on expert consensus in phases 1–3 and was directed by the study PIs (Harvey, Penn, & Pinkham) in phases 4 and 5. Phases 1 and 2 utilized expert surveys and the RAND Appropriateness Method of consensus development to specify four core domains of social cognition: emotion processing, social perception, theory of mind/mental state attribution, and attributional style/bias. Eight measures representing the best existing tasks of social cognition were also identified.[1]

Phase 3 empirically evaluated these tasks in an initial psychometric study, and a smaller, reconvened RAND panel evaluated each task.[2] Results classified two tasks from the domains of emotion processing and mental state attribution as suitable for immediate use in clinical trials: The Bell Lysaker Emotion Recognition Task (BLERT) and the Hinting task. Three other tasks showing adequate psychometric properties were rated as acceptable but with limitations. The first of these, the Penn Emotion Recognition Task (ER-40), showed relatively poorer prediction of functional outcomes. For the second and third tasks, Reading the Mind in the Eyes (Eyes) and The Awareness of Social Inferences Task (TASIT), concerns were raised about dependence on vocabulary and equivalence of alternate forms, respectively. The final three tasks, addressing the domains of attributional style/bias and social perception, were found to be less psychometrically sound and were not recommended. These included the Ambiguous Intentions Hostility Questionnaire (but see refs[3,4]), Relationships Across Domains task, and Trustworthiness Task.

In phase 4 of SCOPE, the study PIs modified the three promising but somewhat limited tasks and utilized small pilot samples to evaluate the feasibility of the modified tasks.[5] Expansions to the ER-40 focused on increasing predictive utility and included collection of response time for each item (RT) and ratings of confidence in the correctness of each response. Our previous work has indicated that RT for social cognitive decisions can be a strong predictor of functional outcomes[6] and Introspective Accuracy (IA), or the awareness of one's abilities, is a stronger predictor of outcomes than task performance.[7,8] The concept of IA also overlaps with that of social metacognition, which was identified as a potential core domain of social cognition in phases 1 and 2. Social metacognition is broadly defined as evaluating thinking, including both discrete acts such as assessing the correctness of a response and synthetic acts such as integrating thoughts and feelings into complex representations.[9] Thus, confidence ratings (CR) provide a potentially useful method of evaluating the IA component of social metacognition within the context of the other core domains.

Modifications to the Eyes task focused on reducing the dependence of performance on vocabulary and encouraging use of a glossary of terms that is provided with the task. We, therefore, created a version of the task that includes embedded definitions, which can be viewed on the same screen as the stimuli. Modifications to TASIT included the collection of RT and counterbalanced administration of test forms across study visits.

Phase 5, which we report here, selected new measures to represent the domains of attributional style/bias and social perception and recruited large samples of individuals with schizophrenia and healthy controls across three sites in order to conduct a final validation study assessing the reliability and validity of the modified and new measures. The previously recommended measures, BLERT and Hinting, were also included to allow comparisons between tasks and consideration of this group of tasks as a comprehensive battery for assessing social cognition.

To select replacement measures for the domains of attributional style/bias and social perception, the study PIs re-examined task nominations and evaluations collected from our initial expert survey and RAND panel and consulted with experts in the field. For attributional style/bias, there were no viable candidate measures from the RAND Panel or the original pool of nominations. We, therefore, conducted an extensive literature search for tasks developed since our survey and identified the Intentional Bias Task (IBT) as a promising measure.[10] Due to its novelty, psychometric information was limited; however, patients demonstrated greater intentionality bias compared to healthy individuals,[11] supporting sensitivity to group differences.

For social perception, we selected the Mini Profile of Nonverbal Sensitivity (MiniPONS)[12] and the Social Attribution Test-Multiple Choice (SAT-MC).[13] The Half PONS[14] was judged by the RAND panel to be fair/good but ultimately was not selected because of its length. The MiniPONS, thus, appeared to offer a suitable compromise, and it has previously shown sensitivity to group differences.[15] Tasks similar to the SAT-MC were also evaluated by the RAND panel but were rated poorly on utility as a repeated measure and criterion validity because of a lack of available data. The SAT-MC has since been further developed for use in schizophrenia and now also has an alternate form,[16] which may be particularly valuable for clinical trials. Of note, the SAT-MC, and tasks of its kind, has traditionally been considered under the domain of mental state attribution. However, they require the perception of social cues before social or mental state attributions can be made.[17] Thus, we evaluate it here under social perception, but it can also be considered a hybrid task requiring both social perception and mental state attribution. All tasks are described below.

As in our initial psychometric study,[2] and other National Institute of Mental Health measurement initiatives,[18–21] measures were evaluated on metrics that are prioritized for clinical trials. These include (1) test-retest reliability, (2) utility as a repeated measure, (3) relationship to functional outcome, and (4) practicality and tolerability.[22] Sensitivity to change was also identified as a key metric; however, the lack of a treatment component in SCOPE precluded evaluation of this criterion. Given that social cognitive measures are also used widely in nonintervention research, sensitivity to differences between patients and healthy controls was also assessed, and basic psychometric properties are reported for healthy controls. While of limited utility in evaluating measures for use in clinical trials,[23] internal consistency was assessed to ensure that our modifications did not negatively impact the construct validity of the measures and to aid in any future attempts to further develop these measures.

## Method

### *Participants*

Data collection occurred at three sites: The University of Texas at Dallas (UTD), The University of Miami Miller School of Medicine (UM), and The University of North Carolina at Chapel Hill (UNC). Participants were stable outpatients with diagnoses of schizophrenia or schizoaffective disorder ($n$ = 218) and healthy controls ($n$ = 154). UTD patients 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, and UNC patients were recruited from the Schizophrenia Treatment and Evaluation Program (STEP) in Carrboro, NC, and the Clinical Research Unit (CRU) in Raleigh, NC. At all sites, healthy controls were recruited via community advertisements. Inclusion and exclusion criteria were identical to previous phases of SCOPE and are detailed in Supplementary Material.

Groups did not differ on gender, race, ethnicity, age, or parental education. The patient group completed fewer years of education and had lower estimated IQs than the control group. For patients, positive and negative symptom severity was low and stable across visits, but general symptoms decreased slightly at visit 2 ($t(207) = 3.53$, $P = .001$, $d_z = .24$). Demographic and clinical characteristics are provided in table 1.

## Measures

### *Social Cognition Measures New to SCOPE*

*Attributional Style/Bias.* *The Intentional Bias Task.*[10] The IBT assesses the tendency to attribute intentionality to the actions of others. Participants indicated whether 24 brief descriptions of actions (e.g., "He broke the window") occurred "on purpose" or "by accident." In total, 12 trails were presented in a fast condition (2.4 s), and 12 trails were presented in a slow condition (5 s). Intentionality bias was calculated as the percentage of intentional responses across all available trials and ranged from 0 to 1, with higher scores indicating greater intentionality bias.

**Table 1.** Participant Demographic and Clinical Characteristics

| | Patients ($n$ = 218) | | Controls ($n$ = 154) | |
|---|---|---|---|---|
| Characteristic | $n$ | % | $n$ | % |
| Male | 142 | 65 | 97 | 63 |
| Race | | | | |
| Caucasian | 115 | 53 | 80 | 52 |
| African American | 87 | 40 | 62 | 40 |
| Native American | 3 | 1 | 0 | 0 |
| Asian | 6 | 3 | 4 | 3 |
| Other | 7 | 3 | 8 | 5 |
| Ethnicity | | | | |
| Hispanic | 33 | 15 | 26 | 17 |
| Non-Hispanic | 185 | 85 | 128 | 83 |
| Diagnosis | | | | |
| Schizophrenia | 112 | 51 | | |
| Schizoaffective | 106 | 49 | | |
| Medication type[a] | | | | |
| Typical | 25 | 12 | | |
| Atypical | 161 | 74 | | |
| Combination | 16 | 7 | | |
| No antipsychotic | 15 | 7 | | |
| | Mean | SD | Mean | SD |
| Age (years) | 41.72 | 11.64 | 41.95 | 12.42 |
| Education (years)** | 13.04 | 2.49 | 14.19 | 1.91 |
| Maternal education (years) | 13.43 | 3.61 | 13.25 | 2.93 |
| Paternal education (years) | 13.52 | 4.19 | 13.49 | 3.26 |
| WRAT-3** | 94.78 | 14.64 | 101.11 | 11.48 |
| PANSS | | | | |
| Positive total | 15.96 | 5.31 | | |
| Negative total | 14.09 | 5.67 | | |
| General total | 31.63 | 8.09 | | |
| CPZ equivalent | 463.64 | 422.82 | | |

Abbreviations: WRAT, Wide Range Achievement Test; PANSS, Positive and Negative Syndrome Scale; CPZ, Chlorpromazine[a] Medication information was missing for 1 patient.
**$P < .01$.

*Social Perception.* *The Mini Profile of Nonverbal Sensitivity (MiniPONS).*[12] The MiniPONS is a multichannel test of accuracy in decoding interpersonal cues (face, body, and voice tone). Sixty-four two-second auditory or visual segments of a Caucasian female exhibiting facial expressions, voice intonations, and/or gestures were presented. Participants chose which of two behavioral labels best described the situation. Performance was indexed as the total number of correct labels (ranging from 0–64).

*The Social Attribution Task—Multiple Choice version.*[13] Participants viewed a short animation of geometric shapes enacting a social drama. The animation was shown twice, and participants then answered 19 multiple-choice questions about what happened. To reduce memory load, relevant segments of the animation were shown before each question. The two available forms were administered in counterbalanced order across study visits. The number of correct responses (ranging from 0–19) was used as the primary dependent variable.

### *Modified Social Cognition Measures*

The remaining measures were used in phase 3 of SCOPE but were expanded or modified as described above in hopes of addressing the concerns raised from the previous study.

*Emotion Processing.* *Penn Emotion Recognition Test.*[24] The ER-40 includes 40 color photographs of static faces expressing 4 basic emotions (ie, happiness, sadness, anger, or fear) and neutral expressions. Participants chose the correct emotion label for each face. As noted above, expansions included the collection of RT and CR. Participants were, therefore, instructed to respond as quickly as possible without sacrificing accuracy, and after giving their response, to rate their confidence in the accuracy of their response on a scale from 0 (not at all confident) to 100 (extremely confident). Accuracy scores (ranging from 0 to 40), mean RT, and mean CR were the primary dependent variables.

*Bell Lysaker Emotion Recognition Task.*[25] The BLERT measures recognition of seven emotional states: happiness, sadness, fear, disgust, surprise, anger, or no emotion. Participants identified the emotion shown in 21 videos of a male actor providing dynamic facial, vocal-tonal, and upper-body movement cues. This task was also expanded to accommodate RT and CR, and performance was indexed as the total number of correctly identified emotions (ranging from 0 to 21), mean RT, and mean CR.

*Mental State Attribution.* *Reading the Mind in the Eyes Test.*[26] Eyes measures the capacity to understand mental states of others from expressions in the eye region of the face. Participants viewed 36 photos and chose the most accurate descriptor word from four choices for the thought/feeling that was portrayed. As noted above, definitions of the response choices were embedded in the task. The dependent measure was the total number of correct responses, ranging from 0 to 36.

*The Awareness of Social Inferences Test, Part III.*[27] TASIT assesses detection of lies and sarcasm and has two forms that were administered in counterbalanced order. Participants watched short videos of everyday social interactions and answered four standard questions per video probing understanding of the intentions, beliefs, and meanings of the speakers and their exchanges. Outcome variables were total number correct, ranging from 0 to 64, and mean RT.

*Hinting Task.*[28] Hinting examines the ability to infer the true intent of indirect speech. Ten short passages presenting an interaction between two characters were read aloud. Each passage ended with one of the characters dropping a hint, and participants explained what the character truly meant. If the first response provided was inaccurate, a second hint was delivered, allowing participants to earn partial credit. Modifications included refinement of our more stringent scoring criteria (available from AEP upon request). Total scores ranged from 0 to 20.

### *Neurocognitive Measures*

Participants completed a subset of the MARTICS Consensus Cognitive Battery[20] including Trail Making Test-Part A, BACS-Symbol Coding, Category Fluency-Animal Naming, Letter-Number Span, and the Hopkins Verbal Learning Test-Revised. The WRAT-3 Reading subscale provided an estimate of premorbid IQ,[29] and the WASI Vocabulary subtest[30] was used to assess the relation between vocabulary knowledge and Eyes performance.

### *Functional Outcome Measures*

Consistent with phases 3 and 4 of SCOPE, functional capacity was assessed with the UCSD Performance-Based Skills Assessment, Brief (UPSA-B),[31] and social competence was assessed with the Social Skills Performance Assessment (SSPA),[32] which was coded by the same expert rater used previously. Real-world functional outcome was assessed via the 31-item, informant-rated version of the Specific Level of Functioning Scale (SLOF).[33] Informants were high contact clinicians, family members, or close friends identified by the participants.

### *Procedures*

Prior to beginning data collection, all study research assistants at each site participated in a full-day Skype training session lead by Dr. Pinkham that provided detailed instruction regarding task administration, scoring, and data entry. Fidelity to study protocol across sites and across time was assured via monthly conference calls between Dr. Pinkham and all SCOPE research assistants and via periodic data reviews conducted by Dr Pinkham throughout the duration of data collection and entry. Of note, only the Hinting task required subjective decisions on the part of the assessor, and all sites showed adequate inter-rater reliability (ICCs > 0.9 with a gold standard rater).

Participants completed two study visits: baseline and a retest assessment conducted 2–4 weeks after the initial visit (mean interval = 16.69 days). At visit 1, all participants provided informed consent and completed the social cognitive and functional outcome measures. The order of these task blocks was counterbalanced, and within the social cognitive battery, the order of individual tasks was also counterbalanced. For patients, visit 1 also included diagnostic assessment and an evaluation of symptom severity using the Positive and Negative Syndrome Scale.[34] Diagnostic and symptom raters were trained to reliability using established procedures at each site.

At visit 2, symptom severity was reassessed in patients, and all participants completed the neurocognitive assessments and repeated the social cognitive measures in the same order as their first visit. For TASIT and SAT-MC, alternative forms were counterbalanced across visit so that forms A and B were equally likely to be administered at visit 1. Ten patients and six healthy controls did not complete visit 2.

Tolerability and practicality were also assessed for all social cognitive measures. Tolerability was indexed as participant ratings of pleasantness on a scale from 1 (very unpleasant) to 7 (very pleasant). Ratings of 4 indicated neither pleasant nor unpleasant. Practicality was operationalized as total administration time including instructions.

### *Statistical Analyses*

The analytic plan followed that used in our initial psychometric study. Score distributions of the social cognitive measures were first checked for normality by examining skew and kurtosis statistics and visually inspecting histograms. Outliers, defined as ±3 SD from the respective group mean, were evident for several tasks (Supplemental table 1). A total of 34 individuals were identified as outliers on at least one assessment, and the majority was only outliers on a single task. Outlying data points were omitted from all subsequent analyses, and removal of these outliers resulted in normal distributions for all variables. Both TASIT and

SAT-MC were evaluated according to version rather than date of completion, and unless otherwise specified, the reported psychometric properties pertain to form A.

Test–retest reliability was computed using Pearson's *r* correlation coefficients. Cronbach's alpha evaluated internal consistency. Utility as a repeated measure was assessed via 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, correlations were first calculated between visit 1 (or form A) social cognitive and neurocognitive measures and the three outcome measures. Those social cognitive tasks showing a significant correlation with each outcome where entered as a single block into regression models to assess the explanatory power of the tasks as group. Hierarchical regression models were then 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* examined group differences.

### *Process for Final Task Recommendations*

Each of the SCOPE study PIs independently reviewed the psychometric data and, based on the aggregate of these data and a formal classification strategy, classified each task as Acceptable, Acceptable with Reservations, or Not Recommended. Evaluation criteria were identical to those utilized by the RAND Panels in previous phases of SCOPE. For test-retest reliability, values of Pearson's $r \geq .6$ were considered acceptable.[35,36] For utility as a repeated measure, emphasis was placed on the absence of floor/ceiling effects at both visit 1 and visit 2 that may limit the measure's ability to show change. Practice effects (ie, change between visits or forms) were examined but were not viewed as critically given that the inclusion of comparison groups in clinical trials allows for examination of treatment specific effects beyond those related to repeated administrations. In assessing relationship to functional outcome, those measures showing significant correlations with outcomes and/or accounting for variance in outcomes (ie, criterion validity), and particularly those showing incremental validity beyond neurocognition, were viewed more favorably. For practicality and tolerability, administration times under 10 min were considered desirable, as were higher ratings of pleasantness. As in the earlier phases of SCOPE, internal consistency and sensitivity to group differences were given lower priority due to reduced applicability to clinical trials. Classifications were consistent across all PIs for every task except the MiniPONS. A consensus classification was reached for this task following discussion.

## Results

### *Site Effects*

Across sites, patient samples differed in some demographic and clinical factors including race, ethnicity, diagnosis, education, parental education, IQ, and symptoms (all $P < .05$). In general, UNC patients were more likely to be Caucasian and to have higher levels of IQ, education, and general symptoms relative to the other sites. Site differences in patient performance on the social cognitive measures at visit 1 were also evident for several tasks. Differences varied; however, UNC generally showed better scores than UM and UTD (Supplementary table 2).

### *Test–Retest Reliability*

Test–retest reliability was acceptable for the majority of measures. Only IBT and SAT-MC showed inadequate values among patients. For healthy controls, test-retest reliability was generally lower, with ITB, Hinting, SAT-MC, TASIT and TASIT RT all having values below benchmark standards (table 2).

### *Internal Consistency*

For patients, the majority of measures approached an acceptable level of internal consistency ($\alpha = .80$)[37] when indexed via accuracy. Internal consistency was much higher when using CR and RT. The IBT and Hinting task were exceptions, with values of .538 and .681, respectively.

### *Utility as a Repeated Measure*

Among patients, both BLERT and Hinting performance significantly improved from the first to second visit (table 3). IBT, BLERT RT, and ER-40 RT decreased at visit 2. Significant differences were also evident for SAT-MC and TASIT, with patients performing worse on form B of the SAT-MC and showing longer RTs for form B of the TASIT. Effect sizes for all visit/version differences were generally small; however, BLERT, SAT-MC, and RT differences approached medium effects. Evidence for floor and ceiling effects was limited with the exception of form B of the SAT-MC wherein 11% of patients scored at or below chance levels.

Patterns were similar among controls. BLERT, Hinting, IBT, and RTs for BLERT and ER-40 all showed significant practice effects. Performance was also worse on form B of SAT-MC and slower for TASIT form B. Effect sizes ranged from small to medium. Approximately 8% of the sample scored at ceiling on Hinting at visit 2. No other tasks showed notable floor or ceiling effects.

### *Relationship to Functional Outcome*

Correlations between social and neurocognitive tasks and functional outcome measures for patients are presented

**Table 2.** Test-Retest Reliability and Internal Consistency

| | Test-Retest Reliability (Person *r*) | | Internal Consistency (Cronbach's Alpha) | |
|---|---|---|---|---|
| Task | Patients (*n* = 208) | Controls (*n* = 148) | Patients (*n* = 218) | Controls (*n* = 154) |
| BLERT | .809 | .622 | .778 | .570 |
| ER-40 | .710 | .679 | .754 | .555 |
| Eyes | .806 | .716 | .750 | .640 |
| IBT | .587 | .511 | .538 | .503 |
| Hinting | .695 | .509 | .681 | .635 |
| MiniPONS | .721 | .663 | .712 | .656 |
| SAT-MC | .573 | .554 | .786 | .735 |
| TASIT | .636 | .534 | .807 | .825 |
| BLERT CR | .613 | .701 | .962 | .932 |
| BLERT RT | .658 | .660 | .939 | .951 |
| ER-40 CR | .625 | .796 | .973 | .962 |
| ER-40 RT | .662 | .629 | .915 | .914 |
| TASIT RT | .687 | .559 | .920 | .881 |

Note: Due to the time limit on responding, many participants had missed trials on the IBT. Estimates of internal consistency for this task are therefore based on much smaller samples of participants (26 patients and 38 controls) who responded to all items.
Abbreviations: BLERT, Bell Lysaker Emotion Recognition Task; ER-40, Penn Emotion Recognition Test; IBT, Intentionality Bias Task; MiniPONS, Mini Profile of Nonverbal Sensitivity; SAT-MC, Social Attribution Test-Multiple Choice; TASIT, The Awareness of Social Inferences Test; CR, confidence ratings; RT, response time

**Table 3.** Utility as a Repeated Measure

| | $T_1$/Version A | | $T_2$/Version B | | $T_2$-$T_1$ 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* = 208) | | | | | | | | | | |
| BLERT | 13.96 | 3.96 | 14.93 | 3.80 | .97 | 2.40 | 0/3 | 2/5 | 5.82 | <.001 | .40 |
| ER-40 | 31.17 | 4.20 | 31.34 | 4.30 | .17 | 3.24 | 0/0 | 0/0 | .78 | .439 | .05 |
| Eyes | 21.20 | 5.52 | 20.76 | 5.68 | −.44 | 3.49 | 3/0 | 9/0 | −1.81 | .072 | .13 |
| IBT | .44 | .18 | .40 | .18 | −.04 | .16 | — | — | −3.55 | <.001 | .26 |
| Hinting | 13.43 | 3.70 | 13.89 | 4.10 | .47 | 3.07 | 1/2 | 1/4 | 2.20 | .029 | .15 |
| MiniPONS | 42.95 | 6.37 | 43.44 | 6.80 | .49 | 4.94 | 14/0 | 12/0 | 1.42 | .158 | .10 |
| SAT-MC | 11.89 | 4.01 | 10.05 | 4.12 | −1.84 | 3.76 | 9/3 | 24/2 | −7.00 | <.001 | .49 |
| TASIT | 44.56 | 7.43 | 43.73 | 6.80 | −.83 | 6.10 | 9/0 | 10/0 | −1.96 | .052 | .14 |
| BLERT CR | 80.66 | 16.66 | 82.03 | 15.08 | 1.37 | 14.03 | 0/22 | 0/20 | 1.41 | .161 | .10 |
| BLERT RT (s) | 16.02 | 3.74 | 15.03 | 3.83 | -.99 | 3.13 | — | — | −4.55 | <.001 | .32 |
| ER-40 CR | 83.88 | 13.25 | 83.76 | 14.18 | −.11 | 11.90 | 0/25 | 0/19 | −.14 | .891 | .01 |
| ER-40 RT (s) | 3.87 | 1.11 | 3.45 | 1.09 | −.42 | .91 | — | — | −6.55 | <.001 | .46 |
| TASIT RT (s) | 55.79 | 4.52 | 57.65 | 4.57 | 1.86 | 3.60 | — | — | 7.11 | <.001 | .52 |
| | Controls (n = 148) | | | | | | | | | | |
| BLERT | 15.87 | 2.72 | 16.58 | 2.85 | .71 | 2.43 | 0/3 | 0/7 | 3.56 | .001 | .29 |
| ER-40 | 32.86 | 3.21 | 33.20 | 3.50 | .33 | 2.70 | 0/0 | 0/0 | 1.49 | .138 | .12 |
| Eyes | 24.69 | 4.34 | 24.40 | 4.79 | −.29 | 3.46 | 0/0 | 0/0 | −1.02 | .309 | .08 |
| IBT | .40 | .15 | .37 | .16 | −.03 | .15 | — | — | −2.14 | .034 | .18 |
| Hinting | 15.44 | 2.65 | 15.93 | 2.81 | .49 | 2.71 | 0/8 | 0/12 | 2.16 | .033 | .18 |
| MiniPONS | 46.58 | 5.59 | 46.84 | 5.89 | .257 | 4.72 | 3/0 | 2/0 | 0.66 | .509 | .05 |
| SAT-MC | 14.21 | 3.30 | 13.14 | 3.96 | −1.07 | 3.48 | 1/2 | 5/7 | −3.75 | <.001 | .31 |
| TASIT | 50.46 | 6.83 | 49.72 | 7.12 | −.74 | 6.74 | 1/0 | 2/0 | −1.32 | .189 | .11 |
| BLERT CR | 85.20 | 10.55 | 86.65 | 10.68 | 1.45 | 8.21 | 0/5 | 0/9 | 2.15 | .033 | .18 |
| BLERT RT (s) | 15.59 | 3.49 | 13.79 | 3.41 | −1.80 | 2.85 | — | — | −7.69 | <.001 | .63 |
| ER-40 CR | 84.92 | 10.69 | 85.20 | 10.71 | .29 | 6.83 | 0/4 | 0/6 | 0.51 | .610 | .04 |
| ER-40 RT (s) | 3.56 | 1.03 | 3.14 | 0.87 | −.42 | 0.83 | — | — | −6.02 | <.001 | .50 |
| TASIT RT (s) | 53.83 | 3.90 | 55.37 | 3.47 | 1.54 | 3.48 | — | — | 5.24 | <.001 | .44 |

Notes: For SAT-MC and TASIT, scores are for Versions A and B regardless of whether they were administered at time 1 or time 2. Ceiling effects for confidence ratings were defined as an average confidence of 100.

in table 4. With the exception of TASIT RT, all social cognitive indices showed significant correlations with at least one outcome measure. The magnitude of these relations ranged from small to medium (0.18–0.44). Neurocognitive tasks also significantly related to outcomes with comparable magnitudes (0.20–0.42).

Criterion validity was assessed via regression models using those social cognitive variables that were significantly correlated with each outcome as predictors. Social cognitive tasks significantly accounted for 25% of the variance in functional capacity (UPSA-B: adjusted $R^2$ = .246, *F*(9,188) = 8.16, *P* < .001), 31% of the variance in social competence (SSPA: adjusted $R^2$ = .305, *F*(9,191) = 10.74, *P* < .001), and 6% of the variance in real-world functioning (SLOF: adjusted $R^2$ = .059, *F*(5,125) = 2.63, *P* = .03). When restricting the sample to individuals with high-quality informants (ie, professionals with mental health experience, *n* = 53),[38] fewer social cognitive indices were significantly correlated to SLOF scores, but the predictive ability of the social cognitive variables improved to 17% (SLOF-HQ: adjusted $R^2$ = .173, *F*(3,49) = 4.63, *P* = .006). Details are provided in table 5.

Incremental validity of the social cognitive tasks was examined by determining whether they would significantly predict variance above and beyond neurocognitive performance (table 6). Neurocognitive variables alone significantly accounted for 22% of the variance in UPSA-B total scores (adjusted $R^2$ = .220, *F*(5,190) = 11.98, *P* < .001), 14% of the variance in SSPA ratings (adjusted $R^2$ = .140, *F*(5,187) = 7.26, *P* < .001), 4% of the variance in SLOF ratings (adjusted $R^2$ = .044, *F*(2,125) = 3.92, *P* = .02), and 9% of the variance in SLOF-HQ ratings (adjusted $R^2$ = .089, *F*(1,48) = 5.80, *P* = .02). Social cognition, entered after neurocognition as a second block, significantly contributed an additional 11% of variance in UPSA-B scores ($R^2$ change = .106, *P* = .001), 18% of variance in SSPA ratings ($R^2$ change = .176, *P* < .001), and 15% of variance in SLOF-HQ ratings ($R^2$ change = .146, *P* = .043). The incremental increase in variance for SLOF ratings was not significant ($R^2$ change = .057, *P* = .183).

### *Practicality and Tolerability*

Administration time was under 10 min for the majority of tasks. TASIT was a notable exception, taking 18 min on average for both patients and controls. Participants rated all tasks to be pleasant. Both patients and controls rated the MiniPONS lowest (table 7).

### *Group Differences*

Patients were less accurate than controls on all measures (effect sizes ranged from 0.48 to 0.84) and showed greater intentionality bias (*d* = .24; table 8). Patients provided higher BLERT CR (*d* = .32) but did not differ from controls for ER-40 CR (*d* = .08). Patients also did not differ from controls in BLERT RT (*d* = .16) but were slower to respond on ER-40 (*d* = .32) and TASIT (*d* = .47).

### *Final Task Recommendations*

The BLERT, ER-40, and Hinting task were all classified as Acceptable. Eyes, IBT, and TASIT were categorized

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

| | UPSA Total (*n* = 208) | SSPA Average (*n* = 208) | SLOF Community Informant (*n* = 135) | SLOF-HQ Community Informant (*n* = 53) |
|---|---|---|---|---|
| Social cognitive | | | | |
| BLERT | .368*** | .415*** | .208* | .062 |
| ER-40 | .361*** | .410*** | .174* | .088 |
| Eyes | .381*** | .277*** | .154 | .086 |
| IBT | −.189** | −.137 | −.191* | −.004 |
| Hinting | .404*** | .437*** | .192* | .345* |
| MiniPONS | .391*** | .379*** | .169* | .092 |
| SAT-MC | .265*** | .329*** | −.004 | −.028 |
| TASIT | .362*** | .380*** | .106 | −.016 |
| BLERT CR | −.080 | −.030 | .060 | −.412** |
| BLERT RT (sec) | −.029 | −.176* | −.102 | .062 |
| ER-40 CR | −.181** | −.090 | −.030 | −.371** |
| ER-40 RT (sec) | −.110 | −.292*** | .043 | −.167 |
| TASIT RT (sec) | −.018 | −.105 | .089 | −.046 |
| Neurocognitive | | | | |
| TrailsA | −.291*** | −.215** | .022 | −.100 |
| Symbol Coding | .388*** | .290*** | .095 | .255 |
| HVLT-R | .394*** | .337*** | .198* | .328* |
| Letter-Number Span | .423*** | .322*** | .217* | .096 |
| Animal Naming | .236** | .195** | .042 | .026 |

Note: SLOF informant ratings were available for only a subset of the patient sample. SLOF-HQ indicates ratings from high quality informants (ie, professionals with mental health experience).
**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 | .28 | .25 | 8.16 | <.001 | | | | |
| BLERT | | | | | .01 | .05 | .96 | .000 |
| ER-40 | | | | | .12 | 1.34 | .18 | .007 |
| Eyes | | | | | .05 | .56 | .58 | .001 |
| IBT | | | | | −.15 | −2.34 | .020 | .021 |
| Hinting | | | | | .26 | 3.87 | <.001 | .057 |
| MiniPONS | | | | | .14 | 1.66 | .10 | .010 |
| SAT-MC | | | | | .02 | .27 | .79 | .000 |
| TASIT | | | | | .09 | 1.06 | .29 | .004 |
| ER-40 CR | | | | | −.11 | −1.64 | .10 | .010 |
| SSPA average | .34 | .31 | 10.74 | <.001 | | | | |
| BLERT | | | | | .08 | .87 | .39 | .003 |
| ER-40 | | | | | .16 | 1.93 | .06 | .017 |
| Eyes | | | | | −.07 | −.74 | .46 | .002 |
| Hinting | | | | | .26 | 4.04 | <.001 | .057 |
| MiniPONS | | | | | .09 | 1.11 | .27 | .004 |
| SAT-MC | | | | | .03 | .43 | .67 | .001 |
| TASIT | | | | | .11 | 1.38 | .17 | .007 |
| BLERT RT | | | | | .01 | .10 | .92 | .000 |
| ER-40 RT | | | | | −.20 | −2.93 | .004 | .030 |
| SLOF total | .095 | .059 | 2.63 | .03 | | | | |
| BLERT | | | | | .10 | .89 | .38 | .006 |
| ER-40 | | | | | .04 | .40 | .69 | .001 |
| IBT | | | | | −.18 | −2.06 | .04 | .031 |
| Hinting | | | | | .13 | 1.45 | .15 | .015 |
| MiniPONS | | | | | .06 | .58 | .56 | .002 |
| SLOF-HQ total | .221 | .173 | 4.63 | .006 | | | | |
| Hinting | | | | | .23 | 1.63 | .11 | .042 |
| BLERT CR | | | | | −.31 | −1.61 | .11 | .041 |
| ER-40 CR | | | | | −.04 | −.23 | .82 | .001 |

Note: SLOF-HQ indicates ratings from high quality informants (ie, professionals with mental health experience).

as Acceptable with Reservations. The MiniPONS and SAT-MC were classified as Not Recommended. These classifications are discussed below.

## Discussion

In this final phase of the SCOPE study, we examined the psychometric properties of eight new or modified social cognitive measures in order to identify tasks that are well suited for use in clinical trials. The Hinting task, ER-40, and BLERT emerged as the strongest tasks and are recommended for use. Similar to its performance in the initial psychometric study (ie, phase 3), Hinting showed adequate test-retest reliability, small practice effects, and strong relations to functional outcomes including uniquely accounting for variance in outcomes while controlling for other social cognitive tasks. Hinting also showed uniquely significant incremental validity in the prediction of functional capacity and social competence. The task could be administered quickly, was liked by patients, and distinguished patient and control performance. Ceiling effects have previously been reported for Hinting;[39–41] however, none were evident here or in phase 3. This is likely due to the modified, more stringent scoring system. Analyses are currently underway to assess whether the psychometric properties reported here may change when utilizing the original scoring criteria.

The ER-40 showed many of the same strengths as Hinting, and both ER-40 accuracy and RT emerged as uniquely significant predictors of social competence, even when controlling for all other cognitive and social cognitive variables. Confidence ratings for ER-40 were also significantly correlated with real-world functioning reported by high-quality informants. As the ER-40 showed only limited relations to functional outcomes in phase 3, the current findings suggest our modifications have potential to increase the functional utility of this measure. BLERT also demonstrated adequate psychometric properties but showed greater practice effects and reduced criterion/incremental validity in this phase relative to phase 3. Like ER-40, however, BLERT CR was significantly correlated with SLOF-HQ suggesting that more detailed analyses of these modifications would be beneficial[5] and that concurrently assessing performance as well as awareness of that performance (ie, Introspective Accuracy) may be a promising strategy for improving the utility of social cognitive measures. In light of the BLERT's strong showing in phase 3, and current evidence of good test-retest

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

| | UPSA-B ($n$ = 196) | | SSPA ($n$ = 193) | | SLOF ($n$ = 128) | | SLOF-HQ ($n$ = 50) | |
|---|---|---|---|---|---|---|---|---|
| | $b^*$ | $sr^2$ | $b^*$ | $sr^2$ | $b^*$ | $sr^2$ | $b^*$ | $sr^2$ |
| Block 1—Neurocognition | | | | | | | | |
| Trails A | −.06 | .002 | .05 | .002 | — | — | — | — |
| Symbol Coding | .13 | .008 | .07 | .002 | — | — | — | — |
| HVLT-R | .16* | .016* | .09 | .004 | .07 | .003 | .23 | .035 |
| Letter-Number Span | .12 | .008 | .05 | .002 | .09 | .005 | — | — |
| Animal Naming | −.06 | .003 | −.05 | .002 | — | — | — | — |
| Block 2— Social Cognition | | | | | | | | |
| BLERT | −.08 | .002 | .07 | .002 | .03 | .000 | — | — |
| ER-40 | .13 | .009 | .17* | .014* | .06 | .002 | — | — |
| Eyes | .004 | .000 | −.11 | .004 | — | — | — | — |
| IBT | −.12 | .014 | — | | −.19* | .035* | — | — |
| Hinting | .22** | .043** | .25*** | .050*** | .11 | .009 | .18 | .022 |
| MiniPONS | .11 | .006 | .07 | .003 | .05 | .001 | — | — |
| SAT-MC | .03 | .000 | .03 | .000 | — | — | — | — |
| TASIT | .003 | .000 | .08 | .003 | — | — | — | — |
| BLERT CR | — | | — | — | — | — | −.38 | .062 |
| BLERT RT | — | | −.001 | .000 | — | — | — | — |
| ER-40 CR | −.12 | .013 | — | — | — | — | .10 | .004 |
| ER-40 RT | — | | −.19* | .024* | — | — | — | — |
| Overall Model | | | | | | | | |
| *Adjusted $R^2$* | .295*** | | .287*** | | .06* | | .188** | |
| *$R^2$ Change* | .11** | | .18*** | | .06 | | .146* | |

Note: SLOF-HQ indicates ratings from high-quality informants (ie, professionals with mental health experience).
*$P$ < .05, **$P$ < .01, ***$P$ < .001.

**Table 7.** Practicality and Tolerability

| | Practicality (Administration Time in Minutes) | | | | Tolerability (Participant Ratings) | | | |
|---|---|---|---|---|---|---|---|---|
| | Patients ($n$ = 218) | | Controls ($n$ = 154) | | Patients ($n$ = 218) | | Controls ($n$ = 154) | |
| Task | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| BLERT | 9.86 | 1.72 | 9.54 | 1.77 | 5.42 | 1.43 | 5.51 | 1.23 |
| ER-40 | 8.38 | 1.61 | 7.82 | 1.75 | 5.52 | 1.44 | 5.65 | 1.22 |
| Eyes | 6.84 | 3.38 | 5.81 | 2.10 | 5.36 | 1.41 | 5.51 | 1.24 |
| IBT | 5.43 | 1.00 | 5.01 | 0.58 | 5.08 | 1.69 | 5.35 | 1.28 |
| Hinting | 6.85 | 2.05 | 6.76 | 1.44 | 5.35 | 1.54 | 5.75 | 1.06 |
| MiniPONS | 12.17 | 2.37 | 11.08 | 1.76 | 4.65 | 1.79 | 4.76 | 1.58 |
| SAT-MC | 10.26 | 1.75 | 9.58 | 1.22 | 5.22 | 1.58 | 5.55 | 1.29 |
| TASIT | 18.62 | 1.73 | 17.94 | 1.48 | 5.07 | 1.55 | 5.38 | 1.18 |

reliability, limited potential for floor/ceiling effects, sensitivity to group differences, and high practicality and tolerability, we continue to recommend this task for use in clinical trials.

Eyes, TASIT, and IBT generally showed acceptable psychometric properties, but each also had limitations that should be considered carefully. Eyes failed to offer any unique contribution to the prediction of outcomes, and task performance was strongly correlated with WASI Vocabulary scores in patients ($r$ = .63) and controls ($r$ = .47). Previous studies report a correlation of .49 between WASI Vocabulary and Eyes performance in healthy individuals,[42] which suggests our modifications did not successfully reduce this relationship. TASIT also showed only limited relations to functional outcomes, and as in phase 3, had the longest administration time, which may be impractical for some clinical trials. Counterbalancing form administration did, however, appear to reduce the discrepancy between forms noted in phase 3, and thus, we recommend counterbalancing be implemented when using both forms. For IBT, concerns include lower test-retest reliability and increased missing data due to limiting response times. Importantly though, IBT uniquely accounted for variance in both functional

**Table 8.** Group Differences on Social Cognitive Measures

| | Patients ($n$ = 218) | | Controls ($n$ = 154) | | | | |
|---|---|---|---|---|---|---|---|
| Task | Mean | SD | Mean | SD | *t* | *P* | *Cohen's d* |
| BLERT | 13.93 | 4.02 | 15.92 | 2.70 | 5.70 | <.001 | .58 |
| ER-40 | 31.12 | 4.28 | 32.94 | 3.19 | 4.69 | <.001 | .48 |
| Eyes | 21.28 | 5.49 | 24.79 | 4.33 | 6.88 | <.001 | .71 |
| IBT | .44 | .18 | .40 | .15 | −2.09 | .037 | .24 |
| Hinting | 13.36 | 3.71 | 15.38 | 2.68 | 6.05 | <.001 | .62 |
| MiniPONS | 42.88 | 6.47 | 46.69 | 5.52 | 6.10 | <.001 | .63 |
| SAT-MC | 11.91 | 4.00 | 14.24 | 3.28 | 6.09 | <.001 | .64 |
| TASIT | 44.56 | 7.44 | 50.57 | 6.80 | 7.81 | <.001 | .84 |
| BLERT CR | 81.06 | 16.66 | 85.58 | 10.56 | 3.20 | .001 | .32 |
| BLERT RT (s) | 16.04 | 3.69 | 15.45 | 3.56 | −1.54 | .124 | .16 |
| ER-40 CR | 84.08 | 13.48 | 85.05 | 10.72 | .776 | .438 | .08 |
| ER-40 RT (s) | 3.89 | 1.11 | 3.55 | 1.04 | −2.99 | .003 | .32 |
| TASIT RT (s) | 55.91 | 4.56 | 53.91 | 3.98 | −4.24 | <.001 | .47 |

capacity and real-world functional outcome. The IBT, thus, offers promise as a useful measure of attributional style/bias, and more detailed analyses and development (eg, examination of relations to symptoms, consideration of condition effects, etc.) appear warranted.

Two tasks, MiniPONS and SAT-MC, demonstrated psychometric properties that currently preclude them from recommendation for use in clinical trials. Neither task accounted for significant variance in functional outcomes, and both tasks showed the greatest floor effects, particularly form B of SAT-MC. The MiniPONS also had the second longest administration time and was the least liked, and SAT-MC showed poorer test-retest reliability, likely due to nonequivalence of forms. Given the form differences, we did examine correlations between SAT-MC form B and outcomes. These values were not appreciably different from form A.

Overall, results across the phases of SCOPE are notably consistent. Hinting (with the revised scoring criteria) continues to provide a suitable measure addressing the domain of mental state attribution, and the modified BLERT and ER-40 provide good representation for emotion processing. Phase 5 has identified the IBT as a potentially promising measure of attributional bias/style, but as in phase 3, no measures of social perception have been recommended. Given that our initial expert survey identified social perception as an important domain, this underscores the need to continue improving existing measures and to develop new ones. Further clarification of how social perception overlaps with and differs from the other domains of social cognition may be helpful, and it may also be useful to continue testing measures from the social neuroscience literature as Green, Penn, and colleagues have recently done.[19,21] While accuracy in the detection of biological motion did not show strong psychometric properties, it is possible that more nuanced indices such as false alarm rates or detection threshold may perform better. Complementary approaches referencing social psychological frameworks may also be fruitful.

Finally, potential limitations require consideration. The study PIs, rather than a larger panel of experts/stakeholders, made final task recommendations. While the classification strategy was well defined and there was strong agreement, a larger panel may have reached different conclusions. Further, our sample is composed of predominately older, clinically stable individuals in chronic phases of schizophrenia. Psychometric properties may not apply uniformly to other diagnoses or to more symptomatic or early stages of illness.[43] BLERT is recommended for use in clinical trials despite showing significant practice effects. Such effects were not as pronounced in phase 3, but investigators are encouraged to consider this possibility in their data. Likewise, RT indices showed significant practice effects that warrant caution in using RT as a sole index of performance. Due to the psychometric goals of SCOPE, corrections for multiple comparisons were not implemented. Thus, results from the correlation and regression analyses should be interpreted with this limitation in mind. The current study also did not include the MSCEIT, which prevents direct comparisons with this well-established battery of emotional intelligence. Site differences were also evident in social cognitive performance that may be related to demographic differences (eg, higher mean IQ for UNC patients). The inclusion of multiple sites, however, allows for a diverse sample that is more broadly representative of individuals who may be included in clinical trials. As our previous work demonstrates both age and race effects on social cognitive performance,[44] diverse samples should be prioritized.

These limitations notwithstanding, our current psychometric data recommend the Hinting Task, ER-40, and BLERT for use in clinical trials targeting social cognition. These three tasks together assess the domains of emotion processing and mental state attribution, but as no tasks addressing social perception or attributional style are currently recommended, it is premature to consider these tasks a comprehensive battery of social cognition. Results from factor analyses are varied but largely support attributional style as a separable factor.[45,46] Thus, it is particularly important for this domain to be represented. As noted above, the IBT appears promising, and additional work with the Ambiguous Intentions Hostility Questionnaire suggests that it may also have utility.[3,4] However, it is also likely that developing new measures will be required to generate a reliable, valid battery for social cognition. Indeed, combining across all phases, SCOPE has identified more measures that *should not* be used (or used cautiously) rather than those that *should* be used. Social cognition is also still a relatively young field, and it is likely that measurement needs and priorities will evolve as our knowledge grows. We, therefore, strongly encourage an ongoing process within the field geared toward measure development, refinement, and evaluation.

## Supplementary Material

Supplementary data are available at *Schizophrenia Bulletin* online.

## Funding

This work was supported by the National Institute of Mental Health at the National Institutes of Health (R01 MH093432 to P.H.D., D.L.P., and A.E.P.).

## Acknowledgments

We would like to thank all of the individuals who participated in this study and the following individuals for their assistance with data collection and management: Grant Hardaway (UTD), Katie Kemp (UTD), Benjamin Buck (UNC), Neil Hester (UNC), Kelsey Ludwig (UNC), Lana Nye (UNC), Grace Lee Simmons (UNC), Sara Kaplan (UM), and Craig Winter (UM).

## Conflict of Interest Statement

Dr Harvey serves as a consultant/advisory board member for Boeheringer Ingelheim, Lundbeck, Otsuka Digital Heath, Roche, Sanofi, Sunovion, and Takeda. Drs. Penn and Pinkham report no conflicts of interest.

## References

1. 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.
2. Pinkham AE, Penn DL, Green MF, Harvey PD. Social cognition psychometric evaluation: Results of the initial psychometric study. *Schizophr Bull*. 2016;42:494–504.
3. Buck B, Iwanski C, Healey KM, et al. Improving measurement of attributional style in schizophrenia; A psychometric evaluation of the Ambiguous Intentions Hostility Questionnaire (AIHQ). *J Psychiatr Res*. 2017;89:48–54.
4. Buck BE, Pinkham AE, Harvey PD, Penn DL. Revisiting the validity of measures of social cognitive bias in schizophrenia: Additional results from the Social Cognition Psychometric Evaluation (SCOPE) study. *Br J Clin Psychol*. 2016;55:441–454.
5. Cornacchio D, Pinkham AE, Penn DL, Harvey PD. Self-assessment of social cognitive ability in individuals with schizophrenia: Appraising task difficulty and allocation of effort. *Schizophr Res*. 2017;179:85–90.
6. Pinkham AE, Penn DL. Neurocognitive and social cognitive predictors of interpersonal skill in schizophrenia. *Psychiatry Res*. 2006;143:167–178.
7. Gould F, McGuire LS, Durand D, et al. Self-assessment in schizophrenia: Accuracy of evaluation of cognition and everyday functioning. *Neuropsychology.* 2015;29:675.
8. Harvey PD, Pinkham A. Impaired self-assessment in schizophrenia: Why patients misjudge their cognition and functioning: Observations from caregivers and clinicians seem to have the most validity. *Current Psychiatry.* 2015;14:53.
9. Lysaker PH, Dimaggio G. Metacognitive capacities for reflection in schizophrenia: Implications for developing treatments. *Schizophr Bull*. 2014;40:487–491.
10. Rosset E. It's no accident: Our bias for intentional explanations. *Cognition*. 2008;108:771–780.
11. Peyroux E, Strickland B, Tapiero I, Franck N. The intentionality bias in schizophrenia. *Psychiatry Res*. 2014;219:426–430.
12. Bänziger T, Scherer KR, Hall JA, Rosenthal R. Introducing the MiniPONS: A short multichannel version of the Profile of Nonverbal Sensitivity (PONS). *Journal of Nonverbal Behavior.* 2011;35:189–204.
13. Bell MD, Fiszdon JM, Greig TC, Wexler BE. Social attribution test–multiple choice (SAT-MC) in schizophrenia: Comparison with community sample and relationship to neurocognitive, social cognitive and symptom measures. *Schizophr Res*. 2010;122:164–171.
14. Ambady N, Hallahan M, Rosenthal R. On judging and being judged accurately in zero-acquaintance situations. *J Pers Soc Psychol.* 1995;69:518.
15. Walther S, Stegmayer K, Sulzbacher J, et al. Nonverbal social communication and gesture control in schizophrenia. *Schizophr Bull*. 2015;41:338–345.
16. Johannesen JK, Lurie JB, Fiszdon JM, Bell MD. The Social Attribution Task-Multiple Choice (SAT-MC): A Psychometric and equivalence study of an alternate form. *ISRN Psychiatry*. 2013;2013:830825.
17. Klin A. Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: The Social Attribution Task. *J Child Psychol Psychiatry*. 2000;41:831–846.
18. Green MF, Penn DL. Going from social neuroscience to schizophrenia clinical trials. *Schizophr Bull*. 2013;39:1189–1191.
19. 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.
20. 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.
21. 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.

22. 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.
23. Kraemer HC. Toward sound objective evaluation of clinical measures. *Am J Geriatr Psychiatry*. 2013;21:589–595.
24. 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.
25. 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.
26. 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.
27. McDonald S, Flanagan S, Rollins J, Kinch J. TASIT: A new clinical tool for assessing social perception after traumatic brain injury. *J Head Trauma Rehabil*. 2003;18:219–238.
28. 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.
29. 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.
30. Wechsler D. *Wechsler Abbreviated Scale of Intelligence*. Agra, Uttar Pradesh, India: Psychological Corporation; 1999.
31. 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.
32. 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.
33. Schneider LC, Struening EL. SLOF: A behavioral rating scale for assessing the mentally ill. *Soc Work Res Abstr*. 1983;19:9–21.
34. Kay SR, Opler LA, Fiszbein A. *Positive and Negative Syndrome Scale: Manual*. North Tonawanda, NY: Multi-Health Systems, Inc; 1992.
35. Kraemer HC, Kupfer DJ, Clarke DE, Narrow WE, Regier DA. DSM-5: How reliable is reliable enough? *Am J Psychiatry*. 2012;169:13–15.
36. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
37. Nunnally JC. *Psychometric theory*. New York, NY: McGraw-Hill; 1967.
38. 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.
39. Marjoram D, Miller P, McIntosh AM, Cunningham Owens DG, Johnstone EC, Lawrie S. A neuropsychological investigation into 'Theory of Mind' and enhanced risk of schizophrenia. *Psychiatry Res*. 2006;144:29–37.
40. Roberts DL, Penn DL. Social cognition and interaction training (SCIT) for outpatients with schizophrenia: A preliminary study. *Psychiatry Res*. 2009;166:141–147.
41. Versmissen D, Janssen I, Myin-Germeys I, et al. Evidence for a relationship between mentalising deficits and paranoia over the psychosis continuum. *Schizophr Res*. 2008;99:103–110.
42. 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.
43. Ludwig KA, Pinkham AE, Harvey PD, Kelsven S, Penn DL. Social cognition psychometric evaluation (SCOPE) in people with early psychosis: A preliminary study. *Schizophr Res.* 2017.
44. Pinkham AE, Kelsven S, Kouros C, Harvey PD, Penn DL. The effect of age, race, and sex on social cognitive performance in individuals with schizophrenia. *J Nerv Ment Dis*. 2017;205:346–352.
45. Buck BE, Healey KM, Gagen EC, Roberts DL, Penn DL. Social cognition in schizophrenia: Factor structure, clinical and functional correlates. *J Ment Health*. 2016;25:330–337.
46. Mancuso F, Horan WP, Kern RS, Green MF. Social cognition in psychosis: Multidimensional structure, clinical correlates, and relationship with functional outcome. *Schizophr Res*. 2011;125:143–151.