Schizophrenia Research 190 (2017) 136–143

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

Social cognition psychometric evaluation (SCOPE) in people with early psychosis: A preliminary study

Kelsey A. Ludwig [a,*], Amy E. Pinkham [b,c], Philip D. Harvey [d,e], Skylar Kelsven [f], David L. Penn [a,g]

[a] Department of Psychology and Neuroscience, University of North Carolina, Chapel Hill, NC, United States
[b] School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, United States
[c] Department of Psychiatry, University of Texas Southwestern Medical School, Dallas, TX, United States
[d] Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, Miami, FL, United States
[e] Research Service, Miami VA Healthcare System, United States
[f] Department of Psychology, San Diego State University, The University of California at San Diego, San Diego, CA, United States
[g] School of Psychology, Australian Catholic University, Melbourne, VIC, Australia

ARTICLE INFO

Article history:
Received 10 November 2016
Received in revised form 26 February 2017
Accepted 1 March 2017
Available online 14 March 2017

Keywords:
First episode psychosis
Schizophrenia
Social cognition
Measurement
Psychometric evaluation

ABSTRACT

Social cognition is an important outcome in schizophrenia research. Unfortunately, there has been a lack of consensus regarding which measures of social cognition best capture this domain of functioning. The Social Cognition Psychometric Evaluation (SCOPE) study was developed to address the need for a battery of measures that have sound psychometric properties and can be implemented in clinical trials for individuals with chronic schizophrenia. The current study expands upon the SCOPE study by examining the psychometric properties of the eight candidate measures administered to individuals early in the course of psychosis. Thirty-eight stable outpatients with first episode psychosis and thirty-nine healthy controls completed the battery at baseline and one-month follow-up assessments. The SCOPE battery was evaluated on a collection of psychometric properties, including: (1) Reliability – including test-retest and internal consistency, (2) Between group differences, (3) Utility as a repeated measure, (4) Relationship to social and occupational functioning, (5) Incremental validity – variance in functioning beyond neurocognition, and (6) Feasibility – including practicality of administration and tolerability. Social cognition accounted for substantially more variance in functional outcome than neurocognition. Only one measure, the Hinting task, displayed adequate psychometric properties to be recommended for use in clinical research with first episode psychosis. The remaining candidate measures would require modifications before implementation or cannot be recommended for use in clinical research with first episode psychosis.

1. Introduction

Social cognition, defined as the mental processes underlying people's capacity to perceive, process and comprehend social information, is related to quality of life, daily living skills and occupational functioning in schizophrenia (Frith, 2008; Green et al., 2012, Kunda, 1999; Mancuso et al., 2011). Social cognition accounts for additional variance in functioning than various cognitive factors (Brüne et al., 2007), and mediates the relationship between neurocognition and functioning in psychosis (Fett et al., 2011; Schmidt et al., 2011). Based on its relation to functional outcome, social cognition in schizophrenia has garnered considerable research interest over the past few decades and is increasingly considered a viable target for treatment (Couture et al., 2006; Fett et al., 2011; Green and Leitman, 2008; Penn et al., 1997).

Despite burgeoning interest in studying social cognition, studies investigating this construct vary greatly in the tasks employed, many of which may lack a strong empirical foundation and involve unknown or questionable psychometric properties (Couture and Penn, 2012; Fett et al., 2011; Savla et al., 2013; Thompson et al., 2011). The absence of a validated battery of social cognitive measures is problematic as inadequate and inconsistent measurement can jeopardize the validity, reproducibility, and comparability of findings, and may lead to effective treatments being discarded or ineffective treatments pursued (Drost, 2011).

To address this need, an ongoing NIMH project called the Social Cognition Psychometric Evaluation (SCOPE) study was initiated (Pinkham et al., 2014; Pinkham et al., 2015). SCOPE is a multiphase project that involves identifying the currently accepted domains of social cognition, selecting the best available measures to assess these domains, and

* Corresponding author at: Department of Psychology and Neuroscience, The University of North Carolina at Chapel Hill, 235 E. Cameron Avenue, 266 Davie Hall, Chapel Hill, NC 27599, United States.
E-mail address: Kelsey_Ludwig@med.unc.edu (K.A. Ludwig).

http://dx.doi.org/10.1016/j.schres.2017.03.001
0920-9964/

administering tasks to a large sample of stable outpatients with schizophrenia and demographically-matched controls.

Findings from the initial validation study suggested the Bell-Lysaker Emotion Recognition Task (*BLERT*; Bell et al., 1997), Penn Emotion Recognition Task (*ER-40*; Kohler et al., 2003), Reading the Mind in the Eyes Test (*Eyes*; Baron-Cohen et al., 2001), The Awareness of Social Inferences Test (*TASIT*; McDonald et al., 2003), and Hinting Task (*Hinting*; Corcoran et al., 1995), displayed acceptable reliability and validity for implementation in clinical research. Remaining measures, including: Ambiguous Intentions Hostility Questionnaire (*AIHQ*; Combs et al., 2007), Relationships Across Domains (*RAD*; Sergi et al., 2009), and Trustworthiness Task (*Trust*; Adolphs et al., 1998), demonstrated weaker characteristics and were deemed inadequate for use in clinical trials targeting social cognition (Pinkham et al., 2015), although subsequent findings support continued development and use of the AIHQ Blame Score (Buck et al., 2017, in press).

Importantly, SCOPE included a predominantly middle-aged, chronic sample typical of many treatment studies. There is some debate as to whether first episode psychosis (FEP) and chronic schizophrenia patients should exhibit the same types and degree of social cognitive impairment (Savla et al., 2013; Thompson et al., 2011; Ventura et al., 2015). Some research suggests attenuated or unremarkable deficits earlier in the course of illness (An et al., 2010; Bora and Pantelis, 2013; Romero-Ferreiro et al., 2016; Sprong et al., 2007), though findings are mixed (Barkl et al., 2014; Green et al., 2012; Horan et al., 2012; Zaytseva et al., 2013). FEP samples may also be more heterogeneous than many chronic schizophrenia samples (Birchwood et al., 1998), and differences in social cognition across phase of illness may stem from variations in clinical stability (Bora and Pantelis, 2013; Green et al., 2012) and age-related changes in neurocognitive abilities (Hartshorne and Germine, 2015). Consequently, the results of SCOPE may not accurately represent younger individuals with FEP.

The purpose of the current study was to extend Pinkham et al.'s (2015) psychometric investigation of the SCOPE battery with a younger FEP sample. Paralleling SCOPE, we report on: (1) Reliability: test-retest, internal consistency, (2) Between-group differences, (3) Utility as a repeated measure, (4) Relationship to social/occupational functioning, (5) Incremental validity: variance in functioning beyond neurocognition, and (6) Feasibility: practicality of administration, tolerability.

2. Method

2.1. Participants

The study took place at the University of North Carolina at Chapel Hill. FEP patients were primarily recruited from the Outreach and Support Intervention Services (OASIS) clinic in Carrboro, NC. Patients required a diagnosis of schizophrenia, schizoaffective disorder, schizophreniform disorder, or psychosis NOS, confirmed by the Structured Clinical Interview for DSM-IV Axis-I Disorders, Patient Edition (*SCID-P*; First et al., 2002). OASIS clinicians and/or a trained research assistant at UNC-CH conducted all diagnostic interviews.

Participants were excluded if diagnosed with psychosis for greater than five years, or had been hospitalized within the last two-months. Deterioration is most common before illness onset and during the first few years of psychosis (Birchwood et al., 1998). Furthermore, evidence indicates a subsequent illness "plateau," during which a level of relative stability is established 2–5 years after illness onset (Srihari et al., 2012). Thus, a cut-off of five years for illness duration was used. Participants were required to be on a stable medication regimen for a minimum of the two-month hospitalization-free period, though they were not excluded if psychiatrically stable while not receiving antipsychotics.

Control participants were recruited through community flyers and online advertisements. Controls were selected for similarities in age/gender to outpatient participants. Controls were precluded from participation for meeting criteria for any Axis I/II disorder according to the DSM-IV, or if they had a first-degree family member with a history of psychosis.

All participants were considered ineligible based on: 1) presence/history of mental retardation, 2) presence/history of brain injury and/or neurological disorder (e.g., seizures, multiple sclerosis), 3) sensory limitation that would interfere with assessment (e.g., blindness/deafness), and/or 4) evidence of non-nicotine substance dependence in the past six-months, with substance use not being exclusionary. Evidence of substance dependence was collected from patients' healthcare providers, via chart review, and/or through substance use disorder modules from the SCID (DSM-IV; First et al., 2002).

2.2. Measures

2.2.1. Social cognition

We administered identical versions of eight candidate measures of social cognition from SCOPE, including: one attributional style measure– *Ambiguous Intentions Hostility Questionnaire*, abbreviated (*AIHQ*; Combs et al., 2007); two emotion processing tasks– *Bell-Lysaker Emotion Recognition Task* (*BLERT*; Bell et al., 1997) and *Penn Emotion Recognition Test* (*ER-40*; Kohler et al., 2003); three theory of mind measures– *Reading the Mind in the Eyes Test* (*Eyes*; Baron-Cohen et al., 2001), *The Awareness of Social Inferences Test, Part-III* (*TASIT*; McDonald et al., 2003), and *Hinting Task* (*Hinting*; Corcoran et al., 1995); one measure of social perception– *Relationships Across Domains*, abbreviated (*RAD*; Sergi et al., 2009); and one novel task that does not fit neatly under the four aforementioned domains– *Trustworthiness Task* (*Trust*; Adolphs et al., 1998).

2.2.2. Neurocognition

Neurocognition was measured using a subset of *The MATRICS Consensus Cognitive Battery* (*MCCB*): Trail-Making Test-Part A, BACS-Symbol Coding, Category Fluency-Animal Naming, Letter-Number Span, and Hopkins Verbal Learning Test-Revised (Nuechterlein et al., 2008). Consistent with SCOPE, subtests were selected according to correlations with composite scores of neurocognitive performance (Keefe et al., 2006; Pinkham et al., 2015). Composite scores were calculated using the standardized mean of corrected *t*-scores for each subtest (See Appendix A). *The Wide Range Achievement Test* (*WRAT-3*) reading-subscale provided an estimate of IQ (Weickert et al., 2000).

2.2.3. Social and occupational functioning

Social skills were assessed with *The Social Skills Performance Assessment* (*SSPA*; Patterson et al., 2001). Participant and experimenter acted out two social situations: meeting a new neighbor and persuading a landlord to fix a bathroom leak. Scenes were audio-recorded and rated by a blind-to-diagnosis, expert coder involved in all previous ratings for SCOPE.

Community and daily living skills were assessed using *The UCSD Performance-Based Skills Assessment-Brief* (*UPSA-B*; Mausbach et al., 2007), a performance-based measure of functional capacity, and *The Specific Level of Functioning Scale, Self-Report* (*SLOF*; Schneider and Struening, 1983).

2.2.4. Symptomatology

Symptom severity was measured using *The Positive and Negative Syndrome Scale* (*PANSS*; Kay et al., 1987).

2.2.5. Feasibility

Practicality was operationalized as administration time. To assess tolerability, participants rated candidate measures on a Likert-scale from 1-*very unpleasant* to 7-*very pleasant*.

2.3. Procedure

Participants completed two assessments: baseline and a retest assessment scheduled to occur approximately 4 weeks later. With the project approved by the UNC-CH Institutional Review Board, participants provided signed informed consent and completed social cognitive, neurocognitive, and functional outcome measures at baseline. Task block order and the order of individual tasks within the social cognitive battery were counterbalanced. A rater trained using the same procedures employed in SCOPE conducted diagnostic and symptomatic interviews. Symptoms were reassessed in patients at retest. With the exception of TASIT, for which an alternative form was available, identical social cognitive tasks were administered at retest, in a different counterbalanced order. In accord with the original SCOPE protocol, Version-A was administered at baseline, Version-B at retest.

2.4. Statistical analyses

Data analyses were performed using the Statistical Package for the Social Sciences (SPSS, version 23). Statistical significance was defined as $p < 0.05$. We followed the psychometric validation process employed by the initial validation study (Pinkham et al., 2015). See Appendix A for details.

2.5. Decision-making process regarding SCOPE battery for FEP

To determine the suitability of the SCOPE battery for FEP, we emphasized test-retest reliability, relationship to functional outcome, and ability to distinguish patient and control performance. **Acceptable at Present** signifies the measure displayed acceptable reliability and validity in the current study, and would not require modifications before use. **Acceptable with Concerns** indicates specific attributes of the task were concerning and warrant further investigation before implementation. **Currently Unacceptable** signifies a task displayed weak psychometrics overall and was not recommended.

3. Results

3.1. Participants

Thirty-eight patients and 39 controls completed the baseline assessment. Thirty-five patients and 36 controls returned to complete visit two. Average time between administrations was comparable for both groups ($M_{FEP} = 33.08$ days, $SD = 5.65$; $M_{HC} = 31.61$ days, $SD = 4.81$; $t(70) = -1.190$, $P = 0.238$). Groups did not differ in regard to gender, race, ethnicity, age, or estimated IQ (see Table 1). Patients completed significantly fewer years of education than controls, whereas patients' parents completed significantly more years of education than the control sample. Patients reported relatively low levels of symptoms at baseline.

There were significant reductions in positive symptoms from baseline to retest ($t(37) = 3.137$, $P = 0.003$, $dz = 0.55$, $r = 0.379$). Negative and general symptoms also decreased, though reductions were non-significant (negative: $t(37) = 1.494$, $P = 0.144$, $dz = 0.20$, $r = 0.338$; general: $t(37) = 1.959$, $P = 0.058$, $dz = 0.26$, $r = 0.376$)(Table 1).

Table 1
Participant demographic and clinical characteristics.*

Characteristic	Patients ($n = 38$)		Controls ($n = 39$)	
	n	(%)	n	(%)
Male	33	86.7	32	82.1
Race				
Caucasian	29	76.3	29	74.4
African American	5	13.2	5	12.8
Asian	2	5.3	2	5.1
Other	2	5.3	3	7.7
Ethnicity				
Hispanic	2	5.3	6	15.4
Non-Hispanic	36	94.7	33	84.6
Diagnosis				
Schizophrenia	25	65.8		
Schizoaffective	6	15.8		
Psychosis NOS	7	18.4		
Medication type				
Typical	1	2.6		
Atypical	32	84.2		
Combination	2	5.3		
Unmedicated	3	7.9		
	Mean	SD	Mean	SD
Age (years)	23.45	3.01	23.77	3.39
Education (years)*	14.03	1.52	15.44	1.80
Maternal education (years)*	16.21	2.27	14.85	1.99
Paternal education (years)*	17.33	2.33	15.53	2.81
WRAT-3	105.87	9.35	107.82	8.91
UPSA-B	70.55	11.63	80.53	9.59
SSPA-Avg.	4.10	0.39	4.68	0.21
$SLOF_{SR}$-Avg.	4.25	0.46	4.61	0.24
PANSS (Visit 1)				
Positive total	17.53	4.91		
Negative total	16.58	3.96		
General total	36.00	5.95		
Overall total	70.11	10.37		
PANSS (Visit 2)				
Positive total	14.63	5.28		
Negative total	15.21	5.58		
General total	32.92	10.20		
Overall total	62.76	18.69		

Note: Patients reported relatively low levels of symptoms at visit 1, and there were moderate reductions in positive symptoms at visit 2 ($t(37) = 3.137$, $P = 0.003$, $d_z = 0.55$).
* $p < 0.05$.

3.2. Psychometric properties

3.2.1. Test-retest reliability

Hinting, RAD, and AIHQ (BS) demonstrated acceptable levels of test-retest reliability (Pearson's r values ≥0.6) for patients. BLERT, ER-40, Eyes, RAD, and two AIHQ subscales (AB/BS) showed adequate values among controls (Table 2).

3.2.2. Internal consistency

For patients, few candidate measures approached/exceeded acceptable Cronbach's alpha values ($\alpha \geq 0.8$). Exceptions included Trust (0.943), TASIT (0.795) and AIHQ-BS (0.857). Internal consistency was generally lower for controls. Excluding Trust (0.821), values for all tasks administered to controls fell below target standards (Table 2).

Table 2
Test-retest reliability and internal consistency.

Task	Test-retest reliability (Pearson r)		Internal consistency (Cronbach's Alpha)	
	Patients ($n = 34$)	Controls ($n = 36$)	Patients ($n = 38$)	Controls ($n = 39$)
AIHQ				
Hostility Bias (HB)	0.529	0.394	0.497	0.387
Aggression Bias (AB)	0.238	0.664	0.259	0.242
Blame Score (BS)	0.737	0.680	0.857	0.742
BLERT	0.455	0.665	0.740	0.411
ER-40	0.496	0.705	0.599	0.538
Eyes	0.534	0.708	0.488	0.630
Hinting	0.735	0.204	0.685	0.493
RAD	0.753	0.735	0.683	0.558
TASIT	0.314	0.338	0.795	0.691
Trust	0.218	0.537	0.943	0.821

Note: With the outlier included in the analyses, test-retest reliability for the BLERT was 0.490 ($n = 35$).

3.2.3. Group differences

Significant group differences were observed only for ER40, Hinting, and TASIT (Table 3). The magnitude of between-group differences spanned the moderate-to-large range.

3.2.4. Utility as a repeated measure

For patients, three of eight measures demonstrated statistically significant differences between assessments (Table 4). Patient performance on ER-40 and Hinting improved, whereas TASIT performance worsened from baseline to retest. Effect sizes were moderate (Cohen's d_z range: 0.414–0.642). Compared to the initial psychometric evaluation (Pinkham et al., 2015), floor/ceiling effects were less evident for this sample. A maximum of two patients (<6%) received perfect or chance-level scores on any measure.

Alternatively, controls performed significantly better on BLERT, ER-40, RAD, and AIHQ-BS at retest. Similar to patients, control performance across versions of the TASIT worsened significantly from baseline to retest. Practice effects varied, with effect sizes in the small-to-medium range (Cohen's d_z range: 0.212–0.732). Only one control scored at/below chance levels on any task (Eyes) during either visit.

With the exception of the second administration of BLERT and first administration of Hinting, <8% of controls scored at ceiling for any candidate measure. Five (12.8%) received perfect scores on BLERT (visit 2), whereas four (10.3%) scored at ceiling on Hinting (visit 1).

3.2.5. Relationship to functional outcome

Correlations between social cognitive and neurocognitive tasks, and functional outcome measures for FEP are provided in Table 5. With the exception of BLERT, Trust, and AIHQ-AB, most measures demonstrated significant relationships with one or more outcome measures. Significant associations were in the expected direction, and of medium magnitude (range: $r = 0.344$–0.473).

3.2.6. Incremental validity

Social cognition explained additional variance in functional outcome above and beyond neurocognition. Simple linear regression analyses indicated neurocognition, when entered as a single predictor variable, accounted for 22% of variance in UPSA-B total scores (adjusted $R^2 = 0.218$, $F(1,36) = 11.33$, $P < 0.01$) and 12% of variance in SSPA ratings (adjusted $R^2 = 0.123$, $F(1,35) = 6.039$, $P < 0.05$), but was not a significant predictor of SLOF self-report values (adjusted $R^2 = -0.012$, $F(1,36) = 0.550$, $P > 0.05$)(Table 6). Sequential regression analyses revealed social cognition, entered after neurocognition as a second block, accounted for an additional 20% of variance in community living skills (UPSA-B; R^2 change = 0.199, $P < 0.05$), 19% of variance in social skills (SSPA; R^2 change = 0.193, $P < 0.05$), and 21% of variance in real-world functioning (SLOF-SR; R^2 change = 0.214, $P < 0.01$)(Table 6).

Table 3
Group differences on social cognitive measures.*

	Patients ($n = 38$)		Controls ($n = 39$)				
Task	Mean	SD	Mean	SD	*t*	*p*	*Cohen's d*
AIHQ-HB*	2.02	0.590	1.89	0.493	1.081	0.283	−0.239
AIHQ-AB	1.74	0.225	1.77	0.198	−0.672	0.504	0.142
AIHQ-BS	7.09	2.15	7.09	1.471	−0.007	0.995	0.000
BLERT	17.14	2.12	17.59	2.04	−0.953	0.344	0.216
ER-40**	32.05	3.54	33.67	3.00	−2.159	0.034	0.494
Eyes	25.16	3.67	26.54	4.08	−1.558	0.123	0.356
Hinting***	15.82	2.82	17.72	1.78	−3.554	<0.001	0.806
RAD	32.34	4.92	33.87	3.95	−1.506	0.136	0.343
TASIT***	51.89	6.24	57.31	3.95	−4.560	<0.001	1.038
Trust	0.17	0.83	0.39	0.39	−1.502	0.137	0.339

Note: BLERT for patients (mean = 16.74, SD = 2.04) when the outlier is included in the analyses.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

3.2.7. Practicality and tolerability

Excluding BLERT administration time ($t(75) = 5.78$, $P = 0.019$, $d = 0.499$) and TASIT enjoyability ratings ($t(74) = 5.06$, $P = 0.027$, $d = -0.379$), practicality and tolerability did not differ significantly between patients and controls (Supplementary Table 1). Most measures required <8 min to complete. Participants rated all tasks as relatively pleasant (range: M = 4.29–5.62).

3.3. Recommendations

Regarding suitability for FEP, Hinting was the sole measure to be considered **Acceptable at Present**. RAD was categorized as **Acceptable with Concerns**. Remaining candidate measures (AIHQ, BLERT, ER40, Eyes, TASIT, Trust) were regarded as **Currently Unacceptable**. A comparison between our recommendations and the outcome of the initial psychometric evaluation has been provided in Supplementary Table 2. A discussion of convergence and divergence between studies is provided below.

4. Discussion

The current study evaluated the psychometric properties of the SCOPE battery for FEP. Our findings suggest one measure, the Hinting task, was considered **Acceptable at Present**, or appropriate for use with FEP patients. In addition to displaying adequate test-retest reliability and effectively distinguishing between patients/controls, Hinting also exhibited significant relationships with both performance-based measures of functioning.

The RAD was classified as Acceptable with Concerns and may be cautiously considered for use with FEP. This measure demonstrated adequate test-retest reliability, a significant relationship to functioning, minimal practice effects, and limited floor/ceiling effects. However, this task was one of the longest to administer and was rated the least enjoyable by patients and controls. RAD's failure to distinguish patients from controls also tempers enthusiasm for this measure. Modification efforts to develop a shortened version may prove beneficial.

The remaining candidate measures were deemed Currently Unacceptable for use with FEP and warrant careful consideration if employed in future clinical trials. Though relatively quick and easy to administer, BLERT and Trust displayed the weakest psychometric properties overall, including poor test-retest reliability, failure to differentiate individuals with/without psychosis, and limited relation to functioning. For the ER40 and TASIT, the primary concern was inadequate test-retest reliability. Based on moderate and significant practice effects observed for the TASIT, there was also concern about possible interference or non-equivalence between versions.

Though AIHQ and Eyes demonstrated significant associations with real-world functioning, predominantly low test-retest reliability estimates and inability to distinguish patients from controls precluded these tasks from recommendation. Notably, however, one subscale of the AIHQ, the BS, was strong on all metrics except group differences. Prior research indicates this subscale of the AIHQ demonstrates adequate psychometric properties, including acceptable internal consistency and test-retest reliability estimates; distinguishes patients from controls; displays significant relationships to functional outcome, and exhibits associations with relevant clinical variables in chronic samples (e.g., hostility and suspiciousness)(Buck et al., 2017, in press). The AIHQ-BS may therefore benefit from further examination and use; however it will be important to determine if it can be used independently of the rest of the measure.

Note that measures were required to demonstrate adequate test-retest reliability among patients, distinguish patients from controls, and exhibit significant relationships to functioning in order to be

Table 4
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 ($n = 34$)											
AIHQ-HB	1.99	0.577	1.86	0.561	−0.125	0.552	–	–	−1.359	0.183	0.226
AIHQ-AB	1.73	0.226	1.77	0.268	−0.047	0.307	–	–	−0.924	0.362	0.153
AIHQ-BS	7.10	2.19	7.07	2.29	−0.028	−1.626	–	–	−0.102	0.919	0.017
BLERT	17.23	2.14	17.91	1.74	0.686	2.055	1/0	0/1	1.974	0.057	0.351
ER-40	32.22	3.55	33.97	2.65	1.750	3.210	0/0	0/0	3.271	0.002	0.545
Eyes	25.25	3.74	25.89	3.46	0.639	3.482	2/0	2/0	1.101	0.278	0.184
Hinting	15.83	2.87	17.08	2.13	1.250	1.948	2/1	0/2	3.851	<0.001	0.642
RAD	32.64	4.88	32.19	5.52	−0.444	3.707	1/0	2/0	0.719	0.477	0.120
TASIT[a]	52.03	6.36	49.00	6.12	−3.028	7.307	1/0	0/1	−2.486	0.018	0.414
Trust	0.172	0.834	0.198	0.531	0.026	0.885	–	–	0.179	0.859	0.029
Controls ($n = 36$)											
AIHQ-HB	1.86	0.460	1.69	0.539	−0.169	0.55	–	–	−1.836	0.075	0.307
AIHQ-AB	1.77	0.195	1.74	0.222	−0.036	0.17	–	–	−1.255	0.218	0.212
AIHQ-BS	7.13	1.482	6.59	1.819	−0.544	1.36	–	–	−2.41	0.021	0.400
BLERT	17.61	2.032	18.67	1.805	1.056	1.58	0/2	0/5	3.997	<0.001	0.668
ER-40	33.67	3.089	34.58	2.781	0.917	2.27	0/0	0/0	2.420	0.021	0.404
Eyes	26.78	4.134	27.78	3.958	1.000	3.098	1/0	1/0	1.936	0.061	0.323
Hinting	17.92	1.538	18.00	1.639	0.083	2.005	0/4	0/3	0.249	0.805	0.041
RAD	34.08	4.003	36.19	3.984	2.111	2.906	0/0	0/0	4.359	<0.001	0.726
TASIT[a]	57.75	3.667	53.47	5.955	−4.278	5.843	0/0	0/2	−4.392	<0.001	0.732
Trust	0.394	0.392	0.327	0.413	−0.067	0.388	–	–	−1.091	0.282	0.173

Note: These are the BLERT values when the outlier is included in the analyses.

16.81	3.30	17.83	1.78	1.028	2.883	1/0	0/1	2.139	0.040	0.357

[a] Alternate forms were used for the TASIT

Table 5
Correlations between social cognitive and neurocognitive tasks and functional outcome measures in patients.

	UPSA total	SSPA average	$SLOF_{SR}$ average
Social cognitive			
AIHQ-HB	−0.096	−0.162	−0.360*
AIHQ-AB	0.136	0.069	−0.253
AIHQ-BS	0.158	0.053	−0.372*
BLERT	0.265	0.159	0.138
ER-40	0.337*	0.435**	−0.101
Eyes	0.326*	0.234	0.407*
Hinting	0.372*	0.473**	0.189
RAD	0.456**	0.344*	0.020
TASIT	0.475**	0.179	0.205
Trust	−0.037	−0.252	0.161
Neurocognitive			
MCCB Composite	0.489**	0.384*	0.123

Note$_a$: There was an error with scene two for the SSPA role-play for one SCZ participant. This particular individual's data – the average for scene 1 only – were included in the above analyses.
Note$_b$: All participants completed the self-report (SR) version of the SLOF. Informants were identified for each SCZ participant, though only 25 individuals successfully completed the informant version of the measure. Neither performance on the social cognition measures nor scores on the social functioning measures were significantly related to the informant version of the SLOF.
* $p < 0.05$.
** $p < 0.01$.

Table 6
Regression models summarizing independent and combined contributions of neurocognition and social cognition to outcomes.

Neurocognition only

MCCB composite	R^2	Adjusted R^2	F	p
UPSA total	0.239	0.218	11.334	0.002
SSPA average	0.147	0.123	6.039	0.019
$SLOF_{SR}$ average	0.015	−0.012	0.550	0.463

Social cognition only

SC tasks	R^2	Adjusted R^2	F	p	b^*	t	p	sr^2
UPSA total	0.392	0.297	4.12	0.005				
ER-40					0.198	1.241	0.224	0.029
Eyes					0.143	0.963	0.343	0.018
Hinting					0.007	0.041	0.968	0.000
RAD					0.233	1.486	0.147	0.042
TASIT					0.357	2.099	0.044	0.084
SSPA average	0.323	0.262	5.26	0.004				
ER-40					0.254	1.588	0.122	0.052
Hinting					0.325	2.054	0.048	0.086
RAD					0.172	1.124	0.269	0.026
$SLOF_{SR}$ average	0.298	0.236	4.82	0.007				
AIHQ-HB					−0.169	−0.991	0.329	0.020
AIHQ-BS					−0.250	−1.478	0.149	0.045
Eyes					0.356	2.442	0.020	0.123

Neurocognition and social cognition

	UPSA-B		SSPA		$SLOF_{SR}$	
	b^*	sr^2	b^*	sr^2	b^*	sr^2
Block 1 - neurocognition						
MCCB composite	0.373**	0.120**	0.274	0.069	−0.023	0.000
Block 2 – social cognition						
AIHQ-HB	–	–	–	–	−0.168	0.020
AIHQ-AB	–	–	–	–	–	–
AIHQ-BS	–	–	–	–	−0.251	0.045
ER-40	0.216	0.035	0.257	0.053	–	–
Eyes	0.043	0.001	–	–	0.363*	0.115*
Hinting	−0.030	0.001	0.294	0.070	–	–
RAD	0.169	0.022	0.106	0.009	–	–
TASIT	0.371*	0.091*	–	–	–	–
Overall model						
Adjusted R^2	0.417*		0.316*		0.214**	
R^2 change	0.199*		0.193*		0.214**	

* $p < 0.05$.
** $p < 0.01$.

recommended for use in clinical trials targeting social cognition in FEP. Given the small sample size, we suggest careful consideration of these recommendations and thoughtful interpretation of the present findings. In particular, our recommendations may not be as applicable to other research goals (e.g., cross-sectional designs).

Consistent with Pinkham et al., our data demonstrate that Hinting is a psychometrically valid theory of mind measure that should be considered appropriate for implementation in psychosis research regardless of stage of illness. Importantly, both Pinkham et al. and the present study utilized a more stringent scoring manual. We emphasize the reported psychometric properties as limited to this revised scoring system (available from AEP upon request). Analyses are underway to determine whether the psychometric properties of the task may change if the original scoring criteria are utilized.

Also consistent with the original SCOPE study, our findings substantiate the claim that social cognition accounts for more variance in functional outcome than various cognitive factors. When measures of social cognition were included in the analyses, the explanatory power of neurocognition dropped significantly. These findings corroborate previous research suggesting social cognition mediates the relationship between neurocognition and functioning in psychosis (Fett et al., 2011; Schmidt et al., 2011). Together, findings from this study provide strong support for the importance of social cognition in FEP.

In contrast, our findings diverge from the initial psychometric evaluation in a number of ways. Although the BLERT displayed some of the strongest properties in SCOPE, it was one of the weakest measures when administered to FEP outpatients. Whereas only two AIHQ subscales (HB/AB) showed inadequate test-retest reliability among patients in SCOPE, Hinting, RAD and one subscale of the AIHQ (BS) were the only measures to reach acceptable levels when administered to FEP. Reliability estimates were generally lower for controls than patients in SCOPE, while the opposite was observed in our sample. In addition, excluding one AIHQ subscale (AB), all social cognitive tasks adequately differentiated between clinical and normative groups in Pinkham et al. Alternatively, significant group differences were observed for fewer than half the battery when administered to a younger sample.

Certain procedural incongruences and sample differences between our study and the original SCOPE study may have contributed to lower test-retest reliability estimates, differential sensitivity to group differences, and limited relationship to functional outcome. Effect sizes indicating meaningful changes in performance between visits suggested clinically relevant practice effects for half the battery when administered to a younger sample (Table 4). Memory and practice effects have been shown to adversely affect test-retest reliability (Abner et al., 2012; Broglio et al., 2007; Greig et al., 2004). In fact, post-hoc independent samples *t*-tests indicated educational attainment and general intelligence for our patient sample were significantly higher than chronic patients in SCOPE (Equal variances assumed, Education: $t = -3.55$, $P < 0.001$; IQ: $t = -4.56$, $P < 0.001$).

Remarkably, FEP performance was more comparable to controls than patients in SCOPE (Supplementary Table 3). Higher levels of general intelligence in the FEP sample may explain the absence of floor effects for the RAD observed in the initial psychometric evaluation. It is also plausible that learning and memory influenced performance at retest and weakened reliability estimates. Implementing dual-baseline assessments, establishing a "learning plateau," and/or employing truly equivalent alternate forms may bolster test-retest reliability of these measures (Beglinger et al., 2003; Beglinger et al., 2005).

Psychosis onset typically occurs during late adolescence and early adulthood, a period of developmental transition and social/lifestyle changes that may contribute to less stable social cognition early in the course of illness (Horan et al., 2012). To assess the possibility that changes in symptom severity between visits may have impacted social

cognitive performance, we recalculated test-retest correlations controlling for symptom fluctuations. Values were unchanged, thus indicating it is unlikely symptom variability accounted for lower test-retest reliability estimates.

Specific differences between our clinical sample and that of SCOPE may also explain why measures did not reliably differentiate patients and controls, and clarify the limited value of most tasks as independent predictors of functional outcome. Post-hoc analyses revealed our FEP sample outperformed chronic patients on all tasks (Supplementary Table 3). It is plausible that social cognitive deficits are less prominent early in the course of illness and/or the outcome measures used to assess social functioning and daily living skills are inappropriate for younger patients.

Finally, certain limitations must be considered. First, the inclusion of a relatively small sample, especially compared to the original SCOPE study, is a noteworthy limitation of the present study. Additionally, data were collected from a relatively homogenous sample of predominantly white, well-educated males from one of the fastest-growing metropolitan areas in the United States. FEP patients were also recruited from a coordinated specialty care clinic focused on early intervention and recovery, and may qualitatively differ from clinical samples recruited from more traditional community mental health centers. Thus, interpretations of the present findings should be regarded with caution.

In summary, the present study indicates social cognitive assessment needs to be approached differently for individuals early in the course of illness, and investigators should use caution when employing tasks that have been used primarily with chronic samples. This underscores the need for the development of new measures for use with FEP, as well as a better understanding of how social cognition and functioning may differ across stage of illness. In addition to improving the validity, reproducibility, and comparability of research findings, we may use this information to tailor treatment and develop targeted interventions for FEP.

Role of the funding source

This work was supported by the National Institute of Mental Health (R01 MH093432 awarded to Drs. Harvey, Penn and Pinkham).

Contributors

PH, DP and AP (Principal Investigators) designed the study and wrote the protocol for this project. KL wrote the first draft of the manuscript, conducted all statistical analyses, and certified the accuracy of the results. All authors provided edits and revisions to the manuscript, and are in agreement with the final version.

Conflicts of interest

Dr. Harvey serves as a consultant/advisory board member for Acadia Pharma, Boehringer Ingelheim, Lundbeck, Otsuka Digital Health, Roche, Sanofi, Sunovion, and Takeda. The remaining authors declare no conflicts of interest pertinent to this study.

Acknowledgments

We would like to thank all individuals who participated in this investigation, including all research subjects and research assistants at the participating sites for their significant effort and contribution to the SCOPE study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.schres.2017.03.001.

References

Abner, E.L., Dennis, B.C., Mathews, M.J., Mendiondo, M.S., Caban-Holt, A., Kryscio, R.J., Schmitt, F.A., Crowley, J.J., 2012. Practice effects in a longitudinal, multi-center Alzheimer's disease prevention clinical trial. Trials 13 (1), 217.

Adolphs, R., Tranel, D., Damasio, A.R., 1998. The human amygdala in social judgment. Nature 393 (6684), 470–474.

An, S.K., Kang, J.I., Park, J.Y., Kim, K.R., Lee, S.Y., Lee, E., 2010. Attribution bias in ultra-high risk for psychosis and first-episode schizophrenia. Schizophr. Res. 118 (1), 54–61.

Barkl, S.J., Lah, S., Harris, A.W., Williams, L.M., 2014. Facial emotion identification in early-onset and first-episode psychosis: a systematic review with meta-analysis. Schizophr. Res. 159 (1), 62–69.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., Plumb, I., 2001. 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 42 (2), 241–251.

Beglinger, L.J., Ahmed, S., Derby, M.A., Siemers, E., Fastenau, P.S., Crawford-Miller, J., Shekhar, A., Kareken, D.A., 2003. Neuropsychological practice effects and change detection in people with schizophrenia. Schizophr. Res. 62 (1), 191–194.

Beglinger, L.J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D.A., Crawford, J., Fasteneau, P.S., Siemers, E.R., 2005. Practice effects and the use of alternate forms in serial neuropsychological testing. Arch. Clin. Neuropsychol. 20 (4), 517–529.

Bell, M., Bryson, G., Lysaker, P., 1997. Positive and negative affect recognition in schizophrenia: a comparison with substance abuse and normal control subjects. Psychiatry Res. 73 (1–2), 73–82.

Birchwood, M., Todd, P., Jackson, C., 1998. Early intervention in psychosis: the critical-period hypothesis. Int. Clin. Psychopharmacol. 172 (33), 53–59.

Bora, E., Pantelis, C., 2013. Theory of mind impairments in first-episode psychosis, individuals at ultra-high risk for psychosis and in first-degree relatives of schizophrenia: systematic review and meta-analysis. Schizophr. Res. 144 (1–3), 31–36.

Broglio, S.P., Ferrara, M.S., Macciocchi, S.N., Baumgartner, T.A., Elliott, R., 2007. Test-retest reliability of computerized concussion assessment programs. J. Athl. Train. 42 (4), 509–514.

Brüne, M., Abdel-Hamid, M., Lehmkamper, C., Sonntag, C., 2007. Mental state attribution, neurocognitive functioning, and psychopathology: what predicts poor social competence in schizophrenia best. Schizophr. Res. 92 (1), 151–159.

Buck, B., Iwanski, C., Healey, K.M., Green, M.F., Horan, W.P., Kern, R.S., Lee, J., Marder, S.R., Reise, S.P., Penn, D.L., 2017. Improving measurement of attributional style in schizophrenia; a psychometric evaluation of the Ambiguous Intentions Hostility Questionnaire (AIHQ). J. Psychiatr. Res. (In press).

Combs, D.R., Penn, D.L., Wicher, M., Waldheter, E., 2007. The ambiguous intentions hostility questionnaire (AIHQ): a new measure for evaluating hostile social-cognitive biases in paranoia. Cogn. Neurodyn. 12 (2), 128–143.

Corcoran, R., Mercer, G., Frith, C.D., 1995. Schizophrenia, symptomatology and social inference: investigating "theory of mind" in people with schizophrenia. Schizophr. Res. 17 (1), 5–13.

Couture, S.M., Penn, D.L., 2012. Introduction to social cognition in schizophrenia. In: Roberts, D.L., Penn, D.L. (Eds.), Social Cognition in Schizophrenia: From Evidence to Treatment. Oxford University Press, New York, pp. 1–16.

Couture, S.M., Penn, D.L., Roberts, D.L., 2006. The functional significance of social cognition in schizophrenia: a review. Schizophr. Bull. 32 (Suppl. 1), S44–S63.

Drost, E.A., 2011. Validity and reliability in social science research. Educ. Res. Perspect. 38 (1), 105.

Fett, A.J., Viechtbauer, W., Dominguez, M., Penn, D.L., van Os, J., Krabbendam, L., 2011. The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: a meta-analysis. Neurosci. Biobehav. Rev. 35 (3), 573–588.

First, M.B., Spitzer, R.L., Gibbon, M., Williams, J.B.W., 2002. Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition (SCID-I/P). John Wiley & Sons, New York.

Frith, S., 2008. Social cognition. Philos. Trans. R. Soc. B. Biol. Sci. 363 (1499), 2033–2039.

Green, M.F., Leitman, D.I., 2008. Social cognition in schizophrenia. Schizophr. Bull. 34 (4), 670–672.

Green, M.F., Bearden, C.E., Cannon, T.D., Fiske, A.P., Hellemann, G.S., Horan, W.P., Kee, K., Kern, R.S., Lee, J., Sergi, M.J., Subotnik, K.L., Sugar, C.A., Ventura, J., Yee, C.M., Nuechterlein, K.H., 2012. Social cognition in schizophrenia, part 1: performance across phase of illness. Schizophr. Bull. 38 (4), 854–864.

Greig, T.C., Nicholls, S.S., Wexler, B.E., Bell, M.D., 2004. Test–retest stability of neuropsychological testing and individual differences in variability in schizophrenia outpatients. Psychiatry Res. 129 (3), 241–247.

Hartshorne, J.K., Germine, L.T., 2015. When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. Psychol. Sci. 26 (4), 433–443.

Horan, W.P., Green, M.F., DeGroot, M., Fiske, A., Hellemann, G., Kee, K., Kern, R.S., Lee, J., Sergi, M.J., Subotnik, K.L., Sugar, C.A., 2012. Social cognition in schizophrenia, part 2: 12-month stability and prediction of functional outcome in first-episode patients. Schizophr. Bull. 38 (4), 865–872.

Kay, S.R., Flszbein, A., Opfer, L.A., 1987. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. Schizophr. Bull. 13 (2), 261–276.

Keefe, R.E., Bilder, R.M., Harvey, P.D., Davis, S.M., Palmer, B.W., Gold, J.M., Meltzer, H.Y., Green, M.F., Miller, D.D., Canive, J.M., Adler, L.W., Manschreck, T.C., Swartz, M., Rosenheck, R., Perkins, D.O., Walker, T.M., Stroup, T.S., McEvoy, J.P., Lieberman, J.A., 2006. Baseline neurocognitive deficits in the CATIE schizophrenia trial. Neuropsychopharmacology 31 (9), 2033–2046.

Kohler, C.G., Turner, T.H., Bilker, W.B., Brensinger, C.M., Siegel, S.J., Kanes, S.J., Gur, R.E., Gur, R.C., 2003. Facial emotion recognition in schizophrenia: intensity effects and error pattern. Am. J. Psychiatry 160 (10), 1768–1774.

Kunda, Z., 1999. Social Cognition: Making Sense of People. first ed. MIT Press, Cambridge.

Mancuso, F., Horan, W.P., Kern, R.S., Green, M.F., 2011. Social cognition in psychosis: multidimensional structure, clinical correlates, and relationship with functional outcome. Schizophr. Res. 125 (2), 143–151.

Mausbach, B.T., Harvey, P.D., Goldman, S.R., Jeste, D.V., Patterson, T.L., 2007. Development of a brief scale of everyday functioning in persons with serious mental illness. Schizophr. Bull. 33 (6), 1364–1372.
McDonald, S., Flanagan, S., Rollins, J., Kinch, J., 2003. TASIT: a new clinical tool for assessing social perception after traumatic brain injury. J. Head Trauma Rehabil. 18 (3), 219–238.
Nuechterlein, K.H., Green, M.F., Kern, R.S., Baade, L.E., Barch, D.M., Cohen, J.D., Essock, S., Fenton, W.S., Frese III, F.J., Gold, J.M., Goldberg, T., 2008. The MATRICS Consensus Cognitive Battery, part 1: test selection, reliability, and validity. Am. J. Psychiatry 165 (2), 203–213.
Patterson, T.L., Moscona, S., McKibbin, C.L., Davidson, K., Jeste, D.V., 2001. Social skills performance assessment among older patients with schizophrenia. Schizophr. Res. 48 (2), 351–360.
Penn, D.L., Corrigan, P.W., Bentall, R.P., Racenstein, J.M., Newman, L., 1997. Social cognition in schizophrenia. Psychol. Bull. 121 (1), 14–132.
Pinkham, A.E., Penn, D.L., Green, M.F., Buck, B., Healey, K., Harvey, P.D., 2014. The social cognition psychometric evaluation study: results of the expert survey and RAND panel. Schizophr. Bull. 40 (4), 813–823.
Pinkham, A.E., Penn, D.L., Green, M.F., Harvey, P.D., 2015. Social cognition psychometric evaluation: results of the initial psychometric study. Schizophr. Bull. 42 (2), 494–504.
Romero-Ferreiro, M.V., Aguado, L., Rodriguez-Torresano, J., Palomo, T., Rodriguez-Jimenez, R., Pedreira-Massa, J.L., 2016. Facial affect recognition in early and late-stage schizophrenia patients. Schizophr. Res. 172 (1), 177–183.
Savla, G., Vella, L., Armstrong, C., Penn, D.L., Twamley, E., 2013. Deficits in domains of social cognition in schizophrenia: a meta-analysis of the empirical evidence. Schizophr. Bull. 39 (5), 979–992.
Schmidt, S.J., Mueller, D.R., Roder, V., 2011. Social cognition as a mediator variable between neurocognition and functional outcome in schizophrenia: empirical review and new results by structural equation modeling. Schizophr. Bull. 37 (Suppl. 2), S41–S54.
Schneider, L.C., Struening, E.L., 1983. SLOF: a behavioral rating scale for assessing the mentally ill. J. Soc. Work Res. Abstr. 19 (3), 9–21.
Sergi, M.J., Fiske, A.P., Horan, W.P., Kern, R.S., Kee, K.S., Subotnik, K.L., Nuechterlein, K.H., Green, M.F., 2009. Development of a measure of relationship perception in schizophrenia. Psychiatry Res. 166 (1), 54–62.
Sprong, M., Schothorst, P., Vos, E., Hox, J., van Engeland, H., 2007. Theory of mind in schizophrenia: meta-analysis. Br. J. Psychiatry 19 (1), 5–13.
Srihari, V.H., Shah, J., Keshavan, M.S., 2012. Is early intervention for psychosis feasible and effective? Psychiatr. Clin. North Am. 35 (3), 613–631.
Thompson, A.D., Bartholomeusz, C., Yung, A.R., 2011. Social cognition deficits and the ultra-high risk for psychosis population: a review of the literature. Early Interv. Psychiatry 5 (3), 192–202.
Ventura, J., Ered, A., Gretchen-Doorly, D., Subotnik, K.L., Horan, W.P., Hellemann, G.S., Nuechterlein, K.H., 2015. Theory of mind in the early course of schizophrenia: stability, symptom and neurocognitive correlates, and relationship with functioning. Psychol. Med. 45 (10), 2031–2043.
Weickert, T.W., Goldberg, T.E., Gold, J.M., Bigelow, L.B., Egan, M.F., Weinberger, D.R., 2000. Cognitive impairments in patients with schizophrenia displaying preserved and compromised intellect. Arch. Gen. Psychiatry 57 (9), 907–913.
Zaytseva, Y., Burova, V., Garakh, Z., Gurovich, I.Y., 2013. Attributional style in first episode of schizophrenia and schizophrenia spectrum disorders with and without paranoid ideation. Psychiatr. Danub. 25 (2), 329–331.