Schizophrenia Bulletin vol. 39 no. 6 pp. 1201–1210, 2013
doi:10.1093/schbul/sbt127
Advance Access publication September 26, 2013

Adapting Social Neuroscience Measures for Schizophrenia Clinical Trials, Part 2: Trolling the Depths of Psychometric Properties

Robert S. Kern[*,1,2], **David L. Penn**[3], **Junghee Lee**[1], **William P. Horan**[1,2], **Steven P. Reise**[1], **Kevin N. Ochsner**[4], **Stephen R. Marder**[1,2], **and Michael F. Green**[1,2]

[1]Department of Psychiatry and Biobehavioral Sciences, UCLA Semel Institute for Neuroscience & Human Behavior, David Geffen School of Medicine, Los Angeles, CA; [2]Department of Veterans Affairs VISN 22 Mental Illness Research, Education, and Clinical Center, Los Angeles, CA; [3]Department of Psychology, University of North Carolina-Chapel Hill, Chapel Hill, NC; [4]Department of Psychology, Columbia University, New York, NY

*To whom correspondence should be addressed; VA Greater Los Angeles Healthcare Center (MIRECC 210 A), Building 210, Room 116, 11301 Wilshire Boulevard, Los Angeles, CA 90073, US; tel: 310-478-3711, ext. 49229, fax: 310-268-4056, e-mail: rkern@ucla.edu

The psychometric properties of 4 paradigms adapted from the social neuroscience literature were evaluated to determine their suitability for use in clinical trials of schizophrenia. This 2-site study (University of California, Los Angeles and University of North Carolina) included 173 clinically stable schizophrenia outpatients and 88 healthy controls. The social cognition battery was administered twice to the schizophrenia group (baseline, 4-week retest) and once to the control group. The 4 paradigms included 2 that assess perception of nonverbal social and action cues (basic biological motion and emotion in biological motion) and 2 that involve higher level inferences about self and others' mental states (self-referential memory and empathic accuracy). Each paradigm was evaluated on (1) patient vs healthy control group differences, (2) test-retest reliability, (3) utility as a repeated measure, and (4) tolerability. Of the 4 paradigms, empathic accuracy demonstrated the strongest characteristics, including large between-group differences, adequate test-retest reliability (.72), negligible practice effects, and good tolerability ratings. The other paradigms showed weaker psychometric characteristics in their current forms. These findings highlight challenges in adapting social neuroscience paradigms for use in clinical trials.

Key words: social neuroscience/schizophrenia/psychometrics

Introduction

Studies of social cognitive processes in schizophrenia have yielded important new findings concerning their relationship with community functioning,[1–6] formation of psychotic symptoms,[7–10] and aberrant brain functioning.[11–13] For these reasons, social cognitive impairments are increasingly regarded as promising targets for pharmacological and behavioral interventions.[14] However, a prominent obstacle for treatment development in this area is the absence of standardized measures of specific subprocesses with established reliability and validity that are suitable for clinical trials.

Social cognition is often assessed in schizophrenia using measures that were developed several decades ago. Examples include identifying an emotion depicted in a still photograph or reading a vignette depicting a social interaction.[15–18] Not surprisingly, many of the tests were borrowed from developmental psychology, including studies with autistic children.[19,20] Unfortunately, these tests fail to capture the dynamic nature and complexities involved in human social interaction such as the changes in facial expression, voice tone, or gestures that are central to communication and convey meaning apart from the content of speech.[21] Additionally, tests that rely heavily on written materials introduce potential confounds associated with reading and comprehension ability. Hence, there is a clear need to develop and refine new social cognitive assessment measures that are appropriate for adults with schizophrenia.

During the Measurement and Treatment Research for Improving Cognition in Schizophrenia (MATRICS) consensus process,[22] experts agreed that tests considered for use as endpoints in clinical trials research should be evaluated on the following 5 characteristics: (1) discriminant validity (ie, differences between patients and healthy controls), (2) test-retest reliability, (3) utility as a repeated measure, (4) tolerability and practicality, and (5) relationship to community functioning. A sixth criterion, sensitivity to change, was also considered during the MATRICS meetings, but it was acknowledged that lack of data did not permit this criterion to be adequately assessed. Tests that fail to discriminate between patients and controls are unlikely targets for intervention because such tests either indicate a relatively preserved area of functioning or insensitivity to group differences.[23] A test with poor test-retest reliability yields reduced statistical power in clinical trials and may undermine the ability to detect significant treatment effects.[24] Likewise, tests with high practice effects that yield scores close to ceiling (ie, highest possible) would be undesirable for the same reason. Regarding tolerability, tests that patients do not like to take or are impractical to administer and score may lead to early dropout or missing data. Finally, because the ultimate goal of new treatments for social cognition is to improve the quality of life and functioning of individuals, it is hoped that new tests in this area would show a relationship to functionally meaningful outcomes. For paradigms drawn from the neuroscience literature, this may present a more vexing issue.

For the Social Cognition and Functioning in Schizophrenia (SCAF) project, we selected measures from the social neuroscience literature that potentially meet the above criteria. This approach seemed like a good starting point for finding tasks, given that they had been used in neuroscientific investigations and had identified neural substrates. Such knowledge is important for guiding the development of new psychopharmacological treatments for social cognitive impairments.[25] A potential obstacle to adapting tasks from the social neuroscience literature, however, is that activation tasks that work perfectly well in the scanner with college students may fail to satisfy the criteria noted above for use in clinical trials, even with careful efforts at adaptation.

The primary aim of this part of the SCAF project was to evaluate psychometric properties of 4 such paradigms to inform possible use in clinical trials that assess treatment-related changes in social cognition in schizophrenia. Two of the paradigms assessed perception of nonverbal social and action cues, and 2 assessed inferences about others' mental states. For each measure, we examined (1) group differences in performance between patients vs healthy controls, (2) test-retest reliability, (3) utility as a repeated measure (eg, practice effects, ceiling or floor effects), and (4) tolerability (taking the test from the patients' perspective) and administration time. Relationship to functional outcome will be addressed in the succeeding article (Olbert et al, this issue).[26]

Methods

Participants

Schizophrenia participants were recruited from 2 sites: (1) University of California, Los Angeles (UCLA)—outpatient treatment facilities in the Los Angeles area and mental health clinics at the VA Greater Los Angeles Healthcare System and (2) University of North Carolina (UNC)-Chapel Hill Schizophrenia Treatment and Evaluation Program and community mental health clinics in the Chapel Hill area (total across both sites, n = 173). Healthy controls were recruited through ads placed on the internet (total n = 88).

Selection criteria for schizophrenia participants included (1) Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) diagnosis of schizophrenia based on clinical interview, (2) age 18–60 years, (3) able to understand spoken English sufficiently to comprehend testing procedures, (4) no clinically significant neurological disease as determined by medical history (eg, epilepsy), (5) no history of serious head injury (ie, loss of consciousness longer than 1 h, no neuropsychological sequelae, no cognitive rehabilitation treatment after head injury), (6) no evidence of substance or alcohol dependence in the past 6 months and no evidence of substance or alcohol abuse in past month, (7) no sedatives or benzodiazepines within 12 h of testing, (8) no history of mental retardation or developmental disability based on chart review, and (9) clinically stable (ie, no inpatient hospitalizations for 3 months prior to enrollment, no changes in antipsychotic medication type in the 4 weeks prior to enrollment). Antipsychotic medication type or dose was not controlled in the study but was left to the discretion of the patients' treating psychiatrist. All patients were administered the Structured Clinical Interview for DSM-IV (SCID-I/P)[27] by trained diagnosticians according to the training quality assurance procedures used at the respective site. Final diagnosis was determined by the site principal investigator following review of interview data and collateral information (eg, medical records, informants). Scale for the Assessment of Negative Symptoms (SANS) and Brief Psychiatric Rating Scale (BPRS) interviewers were trained according to established procedures that included a library of videotaped interviews developed by the Treatment Unit of the Department of Veterans Affairs VISN 22 Mental Illness Research, Education, and Clinical Center. Raters were trained to a minimum κ of 0.80.

Selection criteria for healthy controls included (1) no psychiatric history involving schizophrenia spectrum disorder (including avoidant, paranoid, schizotypal, or schizoid personality disorders) according to the SCID-II and no psychotic or recurrent major mood axis I disorder according to the SCID-I, (2) no family history of a psychotic disorder among first-degree relatives based on participant report, and (3) no history of substance or alcohol dependence and no current substance use. Criteria concerning age, ability to understand English, neurological disease, head trauma, and sedative or benzodiazepine use were the same as listed for patients above. After providing a complete description of the study to prospective study participants, written informed consent was obtained prior to participation.

Procedures

Schizophrenia participants were administered the battery of social neuroscience paradigms twice (baseline, 4-week retest); healthy controls were administered the battery once. Severity of symptoms was assessed at both testing occasions for patients. The social neuroscience paradigms were grouped to form 2 roughly equivalent sets. Administration of the 2 sets was counterbalanced across subjects to minimize possible confounding effects of fatigue or interference from previously administered paradigms on task performance. A fifth paradigm assessing situational context effects on facial affect perception was dropped based on an interim analysis due to the absence of patient vs healthy control group differences. The results on this paradigm appear in a separate article.[23]

Basic Human Biological Motion. Basic human biological motion was measured using clips of point-light walkers[28] administered in 2 blocks of trials (figure 1). Difficulty level was manipulated by adjusting the percentage of dots moving randomly vs coherently. For each trial, clips were presented for 1 s, and participants were asked to decide whether the clip resembled human movement or not by pressing a corresponding button. In the first block, stimuli were either 100% coherent movement or 100% random. Clips depicting movement type were presented in random order with 10 trials of each movement type. In the second block, the stimulus set consisted of 3 levels of difficulty: 0% coherent, 70% coherent, and 85% coherent. These clips were also presented in random order with 40 trials of each movement type. Adaptation of this task for use in clinical trials was accomplished by making parameter adjustments that yielded stimuli with 15% and 30% random motion, which added to difficulty level and allowed measurement of signal-to-noise sensitivity. The primary dependent measure was an index of sensitivity (*d′*) per level of difficulty (100% coherent, 85% coherent, and 70% coherent).

Emotion in Biological Motion. The ability to perceive emotion in biological motion was assessed using the point-light walker stimuli developed by Heberlein et al.[29] We adapted this task for clinical trials use by selecting a subset of stimuli that captured a range of commonly displayed emotions. Thirty point-light walker clips of 5–10 s in length were presented on a computer screen. Participants were asked which of 5 emotional states (fear, anger, happiness, sadness, or neutral) best described the movement of the walker. The 5 choices for emotional state were presented on the computer screen immediately after presentation of the clip. The primary dependent measure was accuracy measured as percent correct.

Self-Referential Memory. The current paradigm used the methods of Kelley et al[30] and Macrae et al.[31] There were 2 task phases (encoding and delayed recognition). During the encoding phase, participants completed 3 types of trials in which they judged (1) whether a trait word described themselves ("me" or "not me"; self-referential condition), (2) whether the word is a generally desirable trait ("desirable" or "not desirable"; other condition), or (3) whether it is upper case ("uppercase" or "lowercase"; physical condition). Four versions of the task were developed. For each version, 3 lists of words (total of 78) were randomly assigned to the encoding phase (set A) and 3 lists were used as new words for the recognition phase (set B). Word length, number of syllables, valence, and frequency ratings were equated across word sets. Half of the trait adjectives were presented in uppercase and half in lowercase; half were positive and half were negative in valence.

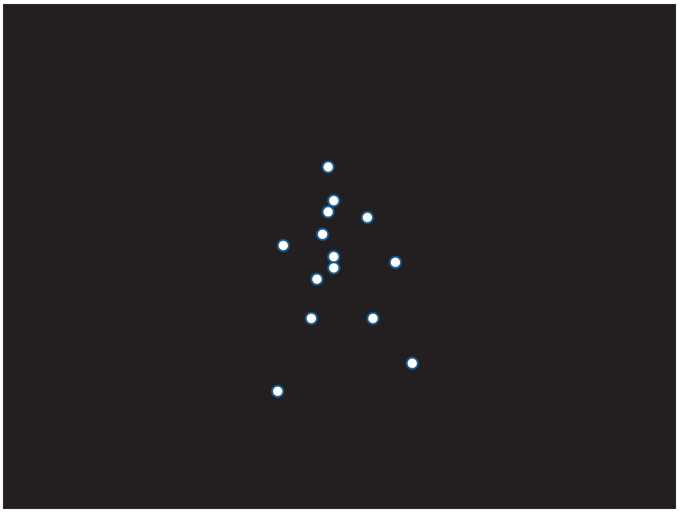

Fig. 1. Example of a point-light walker stimuli.

Each encoding trial consisted of 2 events: (1) first, a central fixation point appeared for 2.5 s, which was followed by (2) presentation of a trait adjective appearing below the fixation point and an instructional cue word above it for 2.5 s. The recognition phase took place 20 min later and consisted of a yes/no recognition test (3-s trait presentation, 3-s inter-stimulus interval) for the 78 old words and an equal number of unseen new trait words. The dependent measure was an index of sensitivity (*d′*) for recognition of words from the self-referential and other conditions.

Empathic Accuracy. The study used 2 versions of an empathic accuracy task with roughly half the subjects tested on each version. In the first part of the data collection, we used a variant of a classic empathic accuracy paradigm of Levenson and Ruef[32] that was developed by Zaki and colleagues.[33] This was the initial version of the task that has been used previously in studies of schizophrenia.[34,35] Although this version performed well in previous studies with schizophrenia, it was limited in a couple of ways. First, it was primarily geared for studies with young adults (eg, college students), and it had limited diversity in terms of age, race, and ethnicity. Thus, it was not well suited for older chronic patients with schizophrenia who would be typical participants in clinical treatment trials.

In addition, the initial version of the task was not created with a type of permission that would enable us to distribute the task broadly to interested investigators. Hence, we developed a new version at UCLA using a new set of social targets who gave permissions for broader use, including use in clinical trials research. This new version also allowed us to capture broader age, racial, and ethnic diversity. It was administered to 82 patients and 59 healthy controls.

In this new UCLA version, participants watched 13 video clips (7 positive and 6 negative), each lasting for 2.0–2.5 min. Each clip showed the head and shoulders of 1 of 6 individuals (targets) while he/she discussed a positive or negative autobiographical event. For each clip, participants were instructed to press 1 of 2 response keys on a computer keyboard to adjust how positive or negative they believed the individual was feeling throughout the duration of the clip based on a 9-point scale (ranging from 1 = extremely negative to 9 = extremely positive). The participant could adjust their ratings as frequently as they felt necessary during the clip to adjust for changes in emotion. The primary dependent measure was the mean correlation across clips between the participant's ratings of the targets' emotions and the targets' ratings of their own emotions calculated in 2-s time epochs throughout the clip. Four clips yielded extreme variability in correlation coefficients and were subsequently dropped. For the analyses, we included a short 6-clip version and a longer 9-clip version. Both versions included positive and negative valence clips.

Tolerability and Administration Time

Tolerability refers to the participant's view of a test (ie, how much they liked or did not like taking the test) and can be influenced by the length of the test, degree of difficulty, or monotony. Patients were asked immediately after they took each paradigm to point to a number on a 7-point Likert scale to indicate how unpleasant or pleasant they found it to be (1 = extremely unpleasant; 7 = extremely pleasant). We also measured administration time for each paradigm to gauge feasibility for use in clinical trials.

Symptom Assessments

Raters were trained to reliability following procedures used by the respective sites. The expanded BPRS[36] was used to assess the presence and severity of psychiatric symptoms. The dependent measures were factor scores for positive and negative symptoms.[37] The SANS[38] was used as an additional measure of negative symptoms. The SANS covers 5 areas of negative symptoms based on interview and reports of the past month: affective flattening, alogia, avolition-apathy, anhedonia-asociality, and attention. The dependent measures were global scores for each subscale, except attention.

Statistical Analyses

Initially, the social neuroscience paradigms were examined for normality of score distribution by examining skewness indices and histograms. None of the measures required transformation. For measures yielding *d′* indices, we considered *d′* below −0.5 to be invalid (ie, below chance). These scores were dropped from the analyses. For basic human biological motion, there were a total of 8 patient outliers over the 2 testing occasions (6 at T_1; 2 at T_2), and there were no healthy control outliers; for self-referential memory, there were a total of 5 patient outliers over the 2 testing occasions (2 at T_1; 3 at T_2), and there was 1 healthy control outlier. Based on examination of score distributions for the empathic accuracy task, we considered clips yielding individual patient scores of $r < -.3$ to be invalid. These clips were dropped from the final 9-clip and shorter 6-clip versions used in the analyses. Group differences between patients and healthy controls were examined by using independent *t* tests; effect sizes were calculated using Cohen's *d*. Correlational analyses were used to examine test-retest reliability in the schizophrenia sample with the Pearson *r* correlation coefficient used as the index of measurement. Practice effects were examined by using paired-samples *t* tests; within-group effect sizes were calculated by dividing the mean difference score by its SD. Measurement of tolerability and administration time was descriptive.

Results

Participants

Across the 2 sites, 173 schizophrenia participants were assessed at baseline and 161 at the 4-week retest (93.1% retention rate). Table 1 provides the demographic characteristics for patients and healthy controls. The 2 groups did not differ in age, parental education, sex, or ethnicity. There was a nonsignificant trend level difference in race with the patient group being comprised of a greater percentage of black/African Americans relative to controls. As expected, patients had significantly lower education and total Wide Range Achievement Test score than controls. At baseline assessment, 76.9% were taking a second-generation antipsychotic, 10.4% a first-generation antipsychotic, 6.4% were taking both, and 1.7% were taking other psychoactive medications only; current medication type was unknown for 4.6%. Symptom levels were low in this clinically stable sample of outpatients and did not differ over the 2 assessments (baseline and 4-week retest). At the initial assessment, the mean BPRS positive symptom factor score was 2.1 (SD = 0.9), and the mean BPRS negative symptom factor score was 1.8 (SD = 0.8). The scores were similar at the 4-week retest (mean positive score = 2.0 [SD = 0.9]; mean negative score = 1.9 [SD = 0.9]). Information on the comparison of the initial version of the empathic accuracy task with

Table 1. Demographic Characteristics

	Patients	Controls
	n = 173	n = 88
Age	42.8 (12.6)	42.6 (10.1)
Education*	12.8 (1.8)	14.7 (1.9)
Parental education	13.3 (3.1)	13.4 (2.7)
Age of onset (y)	21.7 (7.6)	
BPRS positive	2.1 (0.9)	
BPRS negative	1.8 (0.8)	
SANS affective flattening	1.8 (1.3)	
SANS alogia	0.9 (1.2)	
SANS avolition-apathy	2.7 (1.1)	
SANS anhedonia-asociality	2.3 (1.2)	
WRAT*	46.0 (5.9)	50.6 (5.2)
Sex (% men)	71.7 (n = 124)	64.8 (n = 57)
Ethnicity (% Hispanic)	11.6 (n = 20)	11.4 (n = 10)
Race (%)**		
White	51.4 (n = 89)	63.6 (n = 56)
Black/African American	42.2 (n = 73)	29.5 (n = 26)
Asian or Pacific Islander	1.7 (n = 3)	4.5 (n = 4)
More than 1 race	4.6 (n = 8)	2.3 (n = 2)

Note: BPRS, Brief Psychiatric Rating Scale; SANS, Scale for the Assessment of Negative Symptoms; WRAT, Wide Range Achievement Test. *$P < .05$, **$P < .10$.

the UCLA version appears in the online supplementary data and supplementary table S1.

Site Effects

Site differences were examined in patient performance on each of the social neuroscience paradigms. There were significant site differences on the 85% coherent movement condition of the basic biological motion task and a nonsignificant trend level difference ($P = .09$) on the self condition of the self-referential memory test with higher scores at the UCLA site compared to UNC. No other comparisons were statistically significant.

Patient vs Healthy Control Group Differences

Patients showed statistically significant differences from healthy controls on each measure except the “other” condition of the self-referential memory task (table 2). The largest between-group difference was seen on empathic accuracy with both the 6- and 9-clip versions of the task yielding large effect sizes (Cohen’s $d = 0.79$). In contrast, the self-referential memory task yielded the smallest between-group differences with small and small-medium effect size differences on the “other” and “self” conditions, respectively.

Test-Retest Reliability

Test-retest reliability data are summarized in table 3. Generally a Pearson $r = .70$ or higher is considered to be acceptable/desirable level for clinical trials. Only the 9-clip version of the empathic accuracy task met acceptable test-retest reliability standards (Pearson $r = .72$) with the 6-clip version slightly below ($r = .67$). These levels compare favorably with those observed on a more standard measure of social cognition, the Mayer-Salovey-Caruso Emotional Intelligence Test-Managing Emotions branch, included in the MATRICS Consensus Cognitive Battery (MCCB; intraclass correlation coefficient = 0.73).[22] The basic biological motion task had poor values on this criterion with Pearson rs ranging from .35 to .45 across the 3 conditions. The emotion in biological motion and self-referential memory tasks yielded higher measures of test-retest reliability than basic biological motion, but the strength of the correlation coefficients still fell short of acceptable standards (emotion in biological motion: $r = .52$; self-referential memory: rs = .59 and .58 for “self” and “other” conditions, respectively).

Utility as a Repeated Measure

Tests are considered useful for repeated assessments in clinical trials if they do not have problematic practice effects; ie, if practice effects do exist, they do not raise scores to levels approaching ceiling. The strongest measure in this regard was the empathic accuracy task, which showed negligible practice effects from baseline to the 4-week retest (effect size = 0.0), and there were no scores at floor or ceiling. The “other” condition of the self-referential memory task also behaved well in this regard. The “self” condition of this task yielded a nonsignificant

Table 2. Mean Group Differences Between Schizophrenia Participants and Healthy Controls on Social Neuroscience Paradigms

	Patient		Control				
Social Neuroscience Paradigm	Mean	SD	Mean	SD	*t*	*P* Value	Effect Size (Cohen's *d*)
Basic biological motion (*d*′)							
100% Coherent movement	1.72	0.92	2.02	0.86	−2.51	.01	0.34
85% Coherent movement	1.66	0.88	2.32	0.80	−5.81	.001	0.78
70% Coherent movement	1.07	0.65	1.53	0.66	−5.30	.001	0.70
Emotion in biological motion (% accuracy)	0.69	0.12	0.77	0.11	−4.97	.001	0.69
Self-referential memory (*d*′)							
Self	1.30	0.77	1.55	0.67	−2.60	.01	0.35
Other	1.10	0.72	1.25	0.61	−1.58	.12	0.22
Empathic accuracy (*r*)							
6 Clips	0.58	0.17	0.69	0.10	−4.38	.001	0.79
Positive	0.60	0.21	0.71	0.14	−3.36	.001	0.62
Negative	0.54	0.24	0.66	0.15	−3.30	.001	0.60
9 Clips	0.59	0.17	0.70	0.10	−4.32	.001	0.79
Positive	0.63	0.20	0.74	0.13	−3.60	.001	0.65
Negative	0.54	0.22	0.66	0.14	−3.43	.001	0.65

Table 3. Test-Retest Reliability

Social Neuroscience Paradigm	Test Score Used	Pearson *r*
Basic biological motion	*d*′	
100% Coherent movement		.35
85% Coherent movement		.45
70% Coherent movement		.45
Emotion in biological motion	% Accuracy	.52
Self-referential memory	*d*′	
Self		.59
Other		.58
Empathic accuracy	Pearson *r*	
6 Clips		.67
Positive		.50
Negative		.52
9 Clips		.72
Positive		.64
Negative		.74

trend level difference from baseline to retest, but the within-group effect size was small (0.15). The emotion in biological motion task yielded a significant difference between assessment points, but the within-group effect size was small for this paradigm as well (0.17). In contrast, each condition of the basic biological motion task yielded significant within-group differences between testing occasions, and the effect sizes ranged from 0.50 to 0.61. Also, the 100% coherent movement condition yielded 11 cases at ceiling at baseline (6.6%), which rose to 23 cases at the 4-week retest (14.9%) (table 4).

Tolerability and Administration Time

As presented in table 5, patients considered each measure tolerable with little difference noted among tasks. Mean schizophrenia participant ratings ranged from 5.0 to 5.4 across paradigms (scale range: 1 = extremely unpleasant to 7 = extremely pleasant). Administration time for the majority of measures also appeared acceptable for clinical trials with mean administration times ranging from 7.5 to 11.7 min. We could not directly measure administration time for the 6- and 9-clip versions of empathic accuracy because administration time was only measured for the overinclusive 13-clip original version. We provide the cumulative presentation times of the clips, which were 14.8 and 21.3 min for the 6- and 9-clip versions, respectively. Hence, full administration time including instructions to patients would be longer, and this paradigm was the longest to administer.

Discussion

Social neuroscience is a rich scientific field from which new tests can be selected for use as endpoints in clinical trials. There are clear advantages to selecting paradigms from this field in that the cognitive subprocesses and neural substrates associated with task performance are already established. However, social neuroscience paradigms have been largely restricted to use in small sample neuroimaging studies (typically with college undergraduates), and their suitability for use in clinical trials requiring repeated assessments over time has been largely unexplored.

The current evaluation of social neuroscience paradigms coincides with recent efforts conducted by the Cognitive Neuroscience Test Reliability and Clinical applications for Schizophrenia consortium on a broader set of cognitive neuroscience constructs[25,39] as well as the Social Cognition Psychometric Evaluation study on more standard tests of social cognition.[40] Results from each of these efforts underscore the difficulties in finding psychometrically

Table 4. Utility as a Repeated Measure

	T_1		T_2		$T_2{-}T_1$ Difference		Number of Scores at Floor/ Ceiling				
Social Neuroscience Paradigm	Mean	SD	Mean	SD	Mean	SD	T_1	T_2	t	P Value	Effect Size
Basic biological motion (d')											
100% Coherent movement	1.72	0.91	2.24	0.86	0.52	1.01	0/11	0/23	6.35	.001	0.51
85% Coherent movement	1.65	0.87	2.19	0.80	0.54	0.88	0/0	0/0	7.53	.001	0.61
70% Coherent movement	1.08	0.65	1.41	0.65	0.34	0.68	0/0	0/0	6.04	.001	0.50
Emotion in biological motion (% accuracy)	0.68	0.12	0.71	0.12	0.02	0.12	0/0	0/0	2.37	.02	0.17
Self-referential memory (d')											
Self	1.33	0.77	1.21	0.81	−0.11	0.72	0/0	0/0	−1.92	.06	0.15
Other	1.12	0.73	1.06	0.70	−0.05	0.66	0/0	0/0	−0.99	.32	0.08
Empathic accuracy (r)											
6 Clips	0.57	0.17	0.57	0.20	0.00	0.16	0/0	0/0	−0.08	.94	0.00
Positive	0.60	0.21	0.60	0.23	0.00	0.22	0/0	0/0	0.13	.90	0.00
Negative	0.54	0.22	0.56	0.21	0.01	0.21	0/0	0/0	0.59	.56	0.05
9 Clips	0.58	0.18	0.58	0.18	0.00	0.13	0/0	0/0	−0.24	.81	0.00
Positive	0.63	0.20	0.62	0.24	−0.01	0.19	0/0	0/0	−0.63	.53	0.05
Negative	0.56	0.20	0.57	0.19	0.01	0.14	0/0	0/0	0.46	.65	0.07

Table 5. Tolerability and Administration Time

	Tolerability[a]			
	Participants' Ratings		Administration Time (min)	
Social Neuroscience Paradigm	Mean	SD	Mean	SD
Basic biological motion	5.2	1.4	7.5	2.2
Emotion in biological motion	5.4	1.4	8.0	1.4
Self-referential memory				
Encoding	5.3	1.3	11.7	2.5
Recognition	5.0	1.5	9.0	2.9
Empathic accuracy[b]	5.4	1.4	–	–

[a] Rated on a 7-point Likert scale with 1 = extremely unpleasant and 7 = extremely pleasant.
[b] Total presentation time for 6-clip version was 14.8 min and for 9-clip version was 21.3 min.

strong tests in this area for clinical trials use. Compared to the well-established MCCB, the set of social neuroscience paradigms evaluated in the current study generally showed weaker sensitivity to patient-control group differences and weaker test-retest reliability.[22,41] It should be noted, however, that the tests that comprise the MCCB were selected with consideration of their psychometric properties from over 90 nominated tests.[22] These social neuroscience paradigms were selected because they were ones representative of social cognitive domains whose neural substrates could be reliably identified and because they were relevant to social functioning in schizophrenia. Their psychometric properties were largely unknown. Going forward, it is possible that the psychometric characteristics of the current paradigms can be enhanced and made comparable to tests within the MCCB with further consideration of each measure's methods or item characteristics and modifying the test accordingly.

The results from this study showed that only 1 paradigm, empathic accuracy, currently showed sufficiently strong psychometric characteristics to warrant its recommendation as a candidate for use as a clinical trials endpoint. The empathic accuracy paradigm highly discriminated performance between patients and healthy controls, showed adequate test-retest reliability, had virtually no practice effects, and was well tolerated by patients. The lone downside of the paradigm was its length.

We also administered a computerized version of a facial affect identification test[42] in the current study, one typical of those commonly used in schizophrenia research, as an additional benchmark for psychometric comparison. It discriminated between patients and healthy controls (Cohen's d = 0.59), showed adequate test-retest reliability (Pearson r = .74), and had statistically significant, but small, practice effects (effect size = 0.20). The psychometric characteristics of the empathic accuracy task compared favorably with this social cognition measure. From a clinical trials perspective, one might wonder why consider empathic accuracy if facial affect identification is similar in psychometric characteristics plus takes less time to administer. However, it is noteworthy that results from neuroimaging studies indicate that these 2 tests measure distinct social cognitive domains with their own established social cognitive subprocesses.[43–46] For treatment to advance in this area, it is likely that multiple social cognition domains will need to be targeted for intervention.

The other paradigms examined in this study all had limitations for use in clinical trials, at least without further adaptation. Self-referential memory and emotion in biological motion showed a mixed pattern of strengths and weaknesses. The conditions of the self-referential memory paradigm that involved judgments about people had relatively low test-retest reliability, did not strongly discriminate between patients and healthy controls, but yielded small practice effects and had good tolerability. Emotion in biological motion had weaker test-retest reliability than self-referential memory, but better discriminated patients from healthy controls and yielded small practice effects and was well tolerated. The poorest performing paradigm was basic biological motion. It had low test-retest reliability and large practice effects, and so would not be recommended for clinical trials at this time. One potential influence on test-retest reliability and utility as a repeated measure is the paradigm's novelty. In tasks like basic biological motion, participants' performance on the more difficult conditions with 15% and 30% random motion commonly improves over the first few trials because they gain familiarity with the highly novel test stimuli and processing demands of the task. Psychometric limitations affected by task novelty have been noted previously in other social cognition paradigms (eg, a social animation task[47]). Such paradigms are not optimal candidates for use in clinical trials or other investigations that involve repeat assessments over time without further manipulation of methodological procedures to decrease the confound of novelty effects on performance (eg, adding practice trials).

Indeed, these are perilous waters, and careful consideration should be given to addressing the psychometric adaptation challenges indicated by these results. The relatively poor reliability of several of the measures raises concerns about their use in clinical trials and could reflect several factors. For example, it could indicate that some social neural subprocesses can be measured more reliably than others. Another possibility is that a number of paradigms drawn from the social neuroscience field are simply psychometrically unstable at this point in their development. The psychometric characteristics of social neuroscience paradigms have thus far gone largely unexamined, which is not surprising given time and cost considerations associated with psychometric studies of imaging paradigms. While the absence of strong test-retest reliability limits confidence in the use of several of these paradigms as clinical trials endpoints, the extent to which this is a major concern for other types of research is less clear. One perspective is that within-subject reliability is essential to validity for any use of a task in research, including activation tasks in neuroimaging. An alternative view is that activation tasks could have neural construct validity (eg, they activate the same neural circuits across labs), but poor test-retest reliability due to other factors such as practice effects, task novelty, or state-related effects. Such a task could be suitable for investigating activation in a cross-sectional assessment but unsuitable for repeated assessments in clinical trials. This is a topic worthy of further evaluation.

On the positive side, this paradigmatic shift in selecting new tests for clinical trials has considerable potential given the measures' proximal ties to neural substrates—if the psychometric obstacles can be overcome. The challenge is to refine these paradigms, so that they pass rigorous psychometric evaluation. Such an endeavor appears to be more difficult than initially thought. The following article in this issue examines the external validity of these paradigms.

Supplementary Material

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

Funding

National Institute of Mental Health (MH087618, MH43292, MH065707 to M.F.G.); Amgen (to M.F.G.); Amgen, Psychogenics, and Sunovion (to S.R.M.).

Acknowledgments

The authors thank Jejoong Kim, PhD, for the stimuli used for the basic biological motion paradigm; Dr Heberlein for the stimuli used for the emotion in biological motion paradigm; and Kellie M. Smith, MA, for assistance in preparing the manuscript. Requests for the paradigms used in this study should be directed to the first author. Dr Kern is an officer for MATRICS Assessment, Inc. and receives financial compensation for his role in that nonprofit organization. Dr Green reports having been a consultant to Abbott Laboratories (AbbVie), Biogen, and Roche; he is a member of the scientific board for Mnemosyne. Dr Marder reports having been a consultant to Abbott, Roche, Genentech, Otsuka, Bristol Meyers Squibb, Pfizer, Lundbeck, and Boehringer Ingelheim. The rest of the authors report no financial interests or potential conflicts of interest.

References

1. Penn DL. Cognitive rehabilitation of social deficits in schizophrenia: a direction of promise or following a primrose path? *Psychosoc Rehabil J*. 1991;15:27–41.
2. Sergi MJ, Rassovsky Y, Nuechterlein KH, Green MF. Social perception as a mediator of the influence of early visual processing on functional status in schizophrenia. *Am J Psychiatry*. 2006;163:448–454.
3. Brekke J, Kay DD, Lee KS, Green MF. Biosocial pathways to functional outcome in schizophrenia. *Schizophr Res*. 2005;80:213–225.
4. Vauth R, Rüsch N, Wirtz M, Corrigan PW. Does social cognition influence the relation between neurocognitive deficits and vocational functioning in schizophrenia? *Psychiatry Res*. 2004;128:155–165.
5. Addington J, Saeedi H, Addington D. Facial affect recognition: a mediator between cognitive and social functioning in psychosis? *Schizophr Res*. 2006;85:142–150.
6. Couture SM, Penn DL, Roberts DL. The functional significance of social cognition in schizophrenia: a review. *Schizophr Bull*. 2006;32(suppl 1):S44–S63.
7. Garety PA, Freeman D. Cognitive approaches to delusions: a critical review of theories and evidence. *Br J Clin Psychol*. 1999;38(pt 2):113–154.
8. Garety PA, Freeman D, Jolley S, et al Reasoning, emotions, and delusional conviction in psychosis. *J Abnorm Psychol*. 2005;114:373–384.
9. Bentall RP, Swarbrick R. The best laid schemas of paranoid patients: autonomy, sociotropy and need for closure. *Psychol Psychother*. 2003;76:163–171.
10. Randall F, Corcoran R, Day JC, Bentall RP. Attention, theory of mind, and causal attributions in people with persecutory delusions: a preliminary investigation. *Cogn Neuropsychiatry*. 2003;8:287–294.
11. Pinkham AE, Penn DL, Perkins DO, Lieberman J. Implications for the neural basis of social cognition for the study of schizophrenia. *Am J Psychiatry*. 2003;160:815–824.
12. Phillips ML, Drevets WC, Rauch SL, Lane R. Neurobiology of emotion perception II: implications for major psychiatric disorders. *Biol Psychiatry*. 2003;54:515–528.
13. Gur RE, McGrath C, Chan RM, et al. An fMRI study of facial emotion processing in patients with schizophrenia. *Am J Psychiatry*. 2002;159:1992–1999.
14. Green MF, Olivier B, Crawley JN, Penn DL, Silverstein S. Social cognition in schizophrenia: recommendations from the measurement and treatment research to improve cognition in schizophrenia new approaches conference. *Schizophr Bull*. 2005;31:882–887.
15. Happé FG. Communicative competence and theory of mind in autism: a test of relevance theory. *Cognition*. 1993;48:101–119.
16. Langdon R, Coltheart M. Mentalizing, schizotypy, and schizophrenia. *Cognition* 2004;71:43–71.
17. Corcoran R, Cahill C, Frith CD. The appreciation of visual jokes in people with schizophrenia: a study of 'mentalizing' ability. *Schizophr Res*. 1997;24:319–327.
18. Mo S, Su Y, Chan RC, Liu J. Comprehension of metaphor and irony in schizophrenia during remission: the role of theory of mind and IQ. *Psychiatry Res*. 2008;157:21–29.
19. Baron-Cohen S, Leslie AM, Frith U. Does the autistic child have a "theory of mind"? *Cognition*. 1985;21:37–46.
20. Doody GA, Götz M, Johnstone EC, Frith CD, Owens DG. Theory of mind and psychoses. *Psychol Med*. 1998;28:397–405.
21. Zaki J, Ochsner K. The need for a cognitive neuroscience of naturalistic social cognition. *Ann N Y Acad Sci*. 2009;1167:16–30.
22. 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.
23. Lee J, Kern RS, Harvey PO, et al. An intact social cognitive process in schizophrenia: situational context effects on perception of facial affect. *Schizophr Bull*. 2013;39:640–647.
24. 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.
25. Carter CS, Barch DM. Cognitive neuroscience-based approaches to measuring and improving treatment effects on cognition in schizophrenia: the CNTRICS initiative. *Schizophr Bull*. 2007;33:1131–1137.
26. Olbert CM, Penn DL, Kern RS, et al. Adapting social neuroscience measures for schizophrenia clinical trials, Part 3: fathoming external validity. *Schizophr Bull*. This issue.
27. First MB, Spitzer RL, Gibbon M, Williams JBW. *Structured Clinical Interview for DSM-IV Axis I Disorders: Patient Edition*. New York, NY: Biometrics Research Department, New York State Psychiatric Institute; 1997.
28. Puce A, Perrett D. Electrophysiology and brain imaging of biological motion. *Philos Trans R Soc Lond B Biol Sci*. 2003;358:435–445.
29. Heberlein AS, Adolphs R, Tranel D, Damasio H. Cortical regions for judgments of emotions and personality traits from point-light walkers. *J Cogn Neurosci*. 2004;16:1143–1158.
30. Kelley WM, Macrae CN, Wyland CL, Caglar S, Inati S, Heatherton TF. Finding the self? An event-related fMRI study. *J Cogn Neurosci*. 2002;14:785–794.
31. Macrae CN, Moran JM, Heatherton TF, Banfield JF, Kelley WM. Medial prefrontal activity predicts memory for self. *Cereb Cortex*. 2004;14:647–654.
32. Levenson RW, Ruef AM. Empathy: a physiological substrate. *J Pers Soc Psychol*. 1992;63:234–246.
33. Zaki J, Bolger N, Ochsner K. It takes two: the interpersonal nature of empathic accuracy. *Psychol Sci*. 2008;19:399–404.
34. Harvey PO, Zaki J, Lee J, Ochsner K, Green MF. Neural substrates of empathic accuracy in people with schizophrenia. *Schizophr Bull*. 2013;39:617–628.
35. Lee J, Zaki J, Harvey PO, Ochsner K, Green MF. Schizophrenia patients are impaired in empathic accuracy. *Psychol Med*. 2011;41:2297–2304.
36. Lukoff D, Nuechterlein KH, Ventura J. Appendix A: manual for the expanded Brief Psychiatric Rating Scale (BPRS). *Schizophr Bull*. 1986;12:594–602.
37. Kopelowicz A, Ventura J, Liberman RP, Mintz J. Consistency of Brief Psychiatric Rating Scale factor structure across a broad spectrum of schizophrenia patients. *Psychopathology*. 2008;41:77–84.
38. Andreasen NC. *The Scale for the Assessment of Negative Symptoms (SANS)*. Iowa City, IA: The University of Iowa; 1984.
39. Gold JM, Barch DM, Carter CS, et al. Clinical, functional, and intertask correlations of measures developed by the Cognitive Neuroscience Test Reliability and Clinical Applications for Schizophrenia Consortium. *Schizophr Bull*. 2012;38:144–152.

40. Pinkham AE, Penn DL, Green MF, et al. The Social Cognition Psychometric Evaluation (SCOPE) study: results of the expert survey and RAND Panel. *Schizophr Bull*. In press.
41. Kern RS, Gold JM, Dickinson D, et al. The MCCB impairment profile for schizophrenia outpatients: results from the MATRICS psychometric and standardization study. *Schizophr Res*. 2011;126:124–131.
42. Kohler CG, Walker JB, Martin EA, Healey KM, Moberg PJ. Facial emotion perception in schizophrenia: a meta-analytic review. *Schizophr Bull*. 2010;36:1009–1019.
43. Anderson AK, Christoff K, Panitz D, De Rosa E, Gabrieli JD. Neural correlates of the automatic processing of threat facial signals. *J Neurosci*. 2003;23:5627–5633.
44. Adolphs R, Tranel D, Damasio AR. The human amygdala in social judgment. *Nature*. 1998;393:470–474.
45. Völlm BA, Taylor AN, Richardson P, et al. Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task. *Neuroimage*. 2006;29:90–98.
46. Zaki J, Weber J, Bolger N, Ochsner K. The neural bases of empathic accuracy. *Proc Natl Acad Sci USA*. 2009;106:11382–11387.
47. 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.