*Psychological Medicine*

**cambridge.org/psm**

# Ecological momentary facial emotion recognition in psychotic disorders

Colin A. Depp[1,2], Snigdha Kamarsu[1], Tess F. Filip[1], Emma M. Parrish[1,3], Philip D. Harvey[4,5], Eric L. Granholm[1,2], Samantha Chalker[2], Raeanne C. Moore[1] and Amy Pinkham[6]

[1]University of California San Diego Department of Psychiatry, San Diego, California, USA; [2]Veterans Affairs San Diego Healthcare System, San Diego, California, USA; [3]San Diego State University/University of California San Diego Joint Doctoral Program in Clinical Psychology, San Diego, California, USA; [4]University of Miami Miller School of Medicine, Miami, FL, USA; [5]Research Service, Bruce W. Carter VA Medical Center, Miami, FL, USA and [6]The University of Texas at Dallas, Dallas, TX, USA

## Original Article

**Cite this article:** Depp CA, Kamarsu S, Filip TF, Parrish EM, Harvey PD, Granholm EL, Chalker S, Moore RC, Pinkham A (2022). Ecological momentary facial emotion recognition in psychotic disorders. *Psychological Medicine* **52**, 2531–2539. https://doi.org/10.1017/S0033291720004419

Received: 16 July 2020
Revised: 14 October 2020
Accepted: 2 December 2020
First published online: 12 January 2021

**Key words:**
Mobile technology; psychometric assessment; psychosis; social cognition; time series analysis

**Author for correspondence:**
Colin Depp, E-mail: cdepp@ucsd.edu

## Abstract

**Background.** Cognitive tasks delivered during ecological momentary assessment (EMA) may elucidate the short-term dynamics and contextual influences on cognition and judgements of performance. This paper provides initial validation of a smartphone task of facial emotion recognition in serious mental illness.

**Methods.** A total of 86 participants with psychotic disorders (non-affective and affective psychosis), aged 19–65, were administered in-lab 'gold standard' affect recognition, neurocognition, and symptom assessments. They subsequently completed 10 days of the mobile facial emotion recognition task, assessing both accuracy and self-assessed performance, along with concurrent EMA of psychotic symptoms and mood. Validation focused on task adherence and predictors of adherence, gold standard convergent validity, and symptom and diagnostic group variation.

**Results.** The mean rate of adherence to the task was 79%; no demographic or clinical variables predicted adherence. Convergent validity was observed with in-lab measures of facial emotion recognition, and no practice effects were observed on the mobile facial emotion recognition task. EMA reports of more severe voices, sadness, and paranoia were associated with worse performance, whereas mood more strongly associated with self-assessed performance.

**Conclusion.** The mobile facial emotion recognition task was tolerated and demonstrated convergent validity with in-lab measures of the same construct. Social cognitive performance, and biased judgements previously shown to predict function, can be evaluated in real-time in naturalistic environments.

## Introduction

A variety of tests across cognitive domains have been evaluated on mobile self-administered platforms (Moore, Swendsen, & Depp, 2017). Potential advantages of mobile cognitive testing include the ability to assess cognitive performance in naturalistic settings and enhance practical access to cognitive testing for research or clinical purposes. When repeated intensively within persons over time, coupled with the reports of experiences in ecological momentary assessment (EMA), it is also possible to evaluate day-to-day and contextual influences on cognition heretofore challenging, if not impossible, to measure. In a small group of studies, ecological momentary cognitive tests have made measurement more precise by reducing error and cognitive performance was linked to variability in activity participation (Allard et al., 2014; Moore et al., 2017). Tasks developed to date have been designed to measure a variety of cognitive domains (e.g. memory, attention, processing speed; Jongstra *et al.*, 2017; Moore *et al.*, 2017, 2020; Schweitzer *et al.*, 2017), but none to our knowledge have focused on social cognition. The purpose of this paper is to detail the initial validation and relationships (e.g. symptoms as measured by EMA) with the performance of a new ecological momentary test of facial emotion recognition.

Social cognition is a growing focus of observational and interventional research in schizophrenia and psychotic disorders (Green, Horan, & Lee, 2015). There is evidence that social cognition abilities in general are separable from non-social cognition, and that social cognition independently predicts community function in schizophrenia and bipolar disorder (Fett, Viechtbauer, Penn, van Os, & Krabbendam, 2011; Hoe, Nakagami, Green, & Brekke, 2012; Lahera et al., 2012; Mehta et al., 2013). A number of longitudinal studies indicate that social cognitive abilities, including facial emotion recognition, appear to be stable over the course of the illness, similar to neurocognition (Comparelli et al., 2013; Green et al., 2012; McCleery et al., 2016). Moreover, social cognitive tasks that have been vetted for psychometric properties,

© The Author(s), 2021. Published by Cambridge University Press

CAMBRIDGE UNIVERSITY PRESS

including facial emotion recognition, have test–retest reliabilities comparable to non-social cognitive tests (Pinkham, Harvey, & Penn, 2018). As such, social cognitive abilities are presumed to be relatively stable trait-like abilities over time and might not be assumed to fluctuate markedly. Nonetheless, it is unclear if this stability is evident in intensively repeated performance outside of the lab setting. In addition, psychotic symptoms may influence social cognitive abilities, and this influence may be greater than that between symptoms and non-social cognition (Fett, Maat, & Investigators, 2013; Pinkham, Harvey, & Penn, 2016; Ventura, Wood, & Hellemann, 2013) including increasing intra-task variation (Hajduk, Harvey, Penn, & Pinkham, 2018). Thus, although social cognitive performance may be generally stable, the influence of symptoms on changes within persons over time is unclear.

In addition to elucidating the influence of symptoms on performance, ecological momentary tests of facial emotion recognition could help specify the influence of social cognition on social behavior and performance within participants over time. However, in addition to no ecological momentary tasks, according to a recent review (Mote & Fulford, 2020), there has been only one study to evaluate the relationship between in-lab social cognitive performance and EMA-derived social behavior. That study indicated a somewhat surprising lack of association between social cognitive performance and EMA measures of social activity (e.g. time spent alone, with others; Janssens et al., 2012). Therefore, assessing social cognition in a manner that is more proximal to social behavior may provide a more sensitive test of this relationship.

An additional dimension that may vary over time in conjunction with affect recognition accuracy is introspective accuracy judgement, or the ability to accurately gauge one's performance (Harvey & Pinkham, 2015). Introspective accuracy is strongly linked to functional outcomes (Gould et al., 2015), and introspective accuracy for facial affect predicts social function above and beyond ability (Gould et al., 2015; Silberstein, Pinkham, Penn, & Harvey, 2018). In particular, overconfidence is particularly pronounced in psychotic disorders (Balzan, Woodward, Delfabbro, & Moritz, 2016; Jones et al., 2020; Moritz et al., 2016). Psychotic and mood symptoms, which vary over time, are associated with overconfidence and underestimation of performance, respectively (Harvey et al., 2019; Harvey, Paschall, & Depp, 2015; Jones et al., 2020; Köther et al., 2012; Moritz et al., 2015). Therefore, simultaneous and intensively repeated evaluation of symptoms, cognitive performance, and introspective accuracy may help to identify if shifts in mood or psychotic symptoms within subjects are associated with changes in introspective accuracy for emotion recognition.

We developed a facial affect recognition measure called Mobile Ecological Test of Emotion Recognition (METER) that is delivered through a web-based smartphone capable program coupled with contemporaneous EMA and real-time accuracy judgements for the task performance. This study aimed to evaluate acceptability, adherence, and convergent validity of the METER task in regard to the following planned analyses: (1) METER adherence and predictors of adherence, (2) patterns of performance and self-assessed performance ratings over time and evidence of practice effects, (3) convergent validity with 'gold standard' facial emotion recognition measures, (4) convergent validity with non-social cognition test performance. We explored associations with psychotic and mood symptoms measured by both in-lab-based testing and with EMA reports as well as patterns of overestimation as identified in prior cross-sectional lab-based research (Jones et al., 2020),

## Methods

### Participants

Data for this study were derived from an ongoing longitudinal study investigating relationships between negative social cognitive biases, psychosis, and suicidal ideation and behavior. Participants were recruited from three sites – the University of California San Diego (UCSD), The University of Texas at Dallas (UTD), and the University of Miami (UM). Recruitment was performed in a stratified fashion based on the presence *v.* absence of active suicidal ideation by the use of the Columbia Suicide Severity Rating Scale (CSSRS; Posner et al., 2011). Participants were included in the study if they (1) were between the ages of 18 and 65; (2) had a current diagnosis of schizophrenia, schizoaffective disorder, bipolar disorder with psychotic features, or major depressive disorder with psychotic features, confirmed by the Structured Clinical Interview for the DSM-V (SCID 5; First, Williams, Karg, & Spitzer, 2015) and Mini International Neuropsychiatric Interview (MINI; Sheehan et al., 1998); (3) had an informant they were regularly in contact with, for safety procedures; (4) were in outpatient, partial hospitalization, or residential care; (5) were proficient in English; and (6) were able to provide informed consent.

Participants were excluded if they (1) had a history of a head trauma with loss of consciousness >15 min; (2) were ever diagnosed with neurological or neurodegenerative disorder; (3) had vision or hearing problems that would interfere with data collection; (4) had an estimated IQ<70, as determined by the Wide Range Achievement Test-4 (WRAT-4; Wilkinson & Robertson, 2006); (5) had a DSM-V diagnosis of a substance use disorder in the past 3 months, excluding cannabis and tobacco, and confirmed by the SCID-V (First et al., 2015). This study was reviewed by each site's Institutional Review Board, and all participants provided written informed consent.

### Procedures

Once deemed eligible, participants completed lab-based assessments examining their social and neurocognitive performance. At the end of this visit, participants were given an option of using their own smartphone (either iPhone or Android) or using a study-provided Samsung Galaxy S8 Android smartphone to complete the EMA surveys and METER tasks. All participants were provided with a 15 min training session at the end of this in-lab visit on operating the study-provided smartphone (if borrowed), and in completing the EMA and METER tasks. During the 10-day remote survey period, research staff conducted weekly or as needed check-ins to maintain adherence and to resolve participant concerns. Once the 10 days were completed, participants returned the smartphone, if borrowed, and were compensated for their completed surveys [participants received $1.67 for each completed survey (30 total)] for a maximum of $50 (in addition to $50 for in-lab testing).

#### In-lab measures of psychopathology

Clinical diagnoses were established through the MINI (Sheehan et al., 1998), the SCID 5 (First et al., 2015), clinical chart reviews, and consensus meetings with the site investigators. Primary current diagnoses were based on both past and present history of diagnoses and symptoms using the methods described above. Psychotic symptom severity was assessed with the Positive and Negative Syndrome Scale subscales for positive and negative

https://doi.org/10.1017/S0033291720004419 Published online by Cambridge University Press

symptoms (PANSS; Kay, Fiszbein, & Opler, 1987). Depressive symptom severity was also measured using the interview-rated Montgomery–Åsberg Depression Rating Scale (MADRS; Montgomery & Åsberg, 1979). Symptoms of mania were assessed using the Young Mania Rating Scale (YMRS; Young, Biggs, Ziegler, & Meyer, 1978). These three symptom assessments measured current (past week) symptom severity.

### Facial emotion recognition measures

Participants completed the Bell Lysaker Emotion Recognition Task (BLERT; Bryson, Bell, & Lysaker, 1997) and the computerized Penn Emotion Recognition Task (ER-40; Kohler et al., 2003). The BLERT displays 21 video segments of one male actor who, through intonation, upper body movement cues, and facial expression, displays one of seven emotions: happiness, sadness, fear, disgust, surprise, anger, or no emotion. Participants were instructed to choose the correctly displayed emotion in this task. A total score was calculated to determine the number of correct emotion choices identified.

The ER-40 measures emotion recognition ability by displaying 40 color photographs expressing one of four emotions: happiness, sadness, anger, fear, or no emotion. Participants were presented one image at a time and asked to select the emotion displayed as quickly and as accurately as possible. Total scores were calculated as a sum of correct responses from 0 to 40.

### Neurocognitive performance measures

Premorbid verbal ability was assessed with the WRAT-4 (Wilkinson & Robertson, 2006). Participants were administered a subset of the MATRICS Consensus Cognitive Battery (MCCB; Nuechterlein et al., 2008) including the Trail Making Test, Part A (TMT-A; Tombaugh, 2004); Brief Assessment of Cognition in Schizophrenia (BACS) Symbol Coding Subtest (Keefe et al., 2004); Category Fluency: Animal Naming (Spreen, 1991), Letter-Number Span (LNS; Gold, Carpenter, Randolph, Goldberg, & Weinberger, 1997), and Hopkins Verbal Learning Test (HVLT; Brandt & Benedict, 2001). In addition to individual subscale scores, age- and education-normed T-scores were calculated and averaged into a Global Composite Score.

### EMA procedures

Participants were sent text notifications to their smartphones (or study provided Android device) to complete the smartphone-based surveys three times daily for 10 days and the METER task once per day. This text notification contained the participant link for the study surveys. Participants selected preferred time slots for the survey notifications, with at least a 2 h increment in between each survey. Participants received the surveys once in the morning, once in the afternoon, and once at night. Upon receiving the link, participants first completed several EMA questions assessing context, mood, and behaviors and then subsequently completed the METER task, if administered, followed by post-task game performance questions. Once the survey was delivered, the link stayed active for 1 h, after which the survey was no longer accessible. Study surveys were linked to the smartphone number, and so were opened only by the device. Participant's data were deidentified and were not stored locally on the devices. Survey data were sent to encrypted, HIPAA compliant, cloud storage in Amazon Web Services (AWS), and responses were recorded even if participants did not complete the entire survey. The AWS system allowed research staff to access participant data in real-time and monitor progress daily. If participants missed more than three surveys in a row, experimenters contacted them to address any technical difficulties or adherence issues.

### Mobile facial emotion task

The mobile facial emotion task (see Fig. 1) was modeled directly after the widely used and validated Penn Emotion Recognition 40 test (Kohler et al., 2003) and was administered concurrently with the EMA surveys once per day. The timing of the task was stratified by time of day (either morning, afternoon, or evening periods). This task was administered immediately following the EMA questions. In METER, participants were presented with a total of 10 faces each session from a pool of 100 unique faces taken from the publicly available University of Pennsylvania Brain Behavior Lab 2D Facial Emotion Stimuli. Those faces were validated by collecting recognition ratings from healthy volunteers, and only those faces identified with accuracy levels exceeding 70% were used (Gur et al., 2002). Each face displayed one of five emotions: happiness, sadness, anger, fear, or no emotion, and two of each category were presented each session. Neither actor identities nor specific stimuli overlapped with those used in the ER-40. Completion time was collected for each emotion choice for every face, aggregated and averaged across the 10-day protocol.

After each session of the METER, participants were asked to rate their self-assessed performance, that is, how many faces they believe they correctly identified from 0 to 10. We then calculated the difference between actual and self-assessed performance and, since this discrepancy score is bimodal with optimal performance in the middle, we categorized each test session into (1) overestimation (estimated > actual), (2) accurate estimation (estimated = actual), and (3) underestimation (estimated < actual).

### EMA mood and symptom items

The EMA survey included items on location, activity, mood, voices and paranoia, and social activities. For this study, we focused on items that corresponded to in-lab symptom measures of psychosis: voices (e.g. 'Since the past alarm, how much have you been bothered by voices?'), and participants' trust in others (e.g. 'Since the last alarm, how much have you had thoughts that you really can't trust other people?'), along with mood state: happiness (e.g. 'Since the past alarm, how much have you felt happy?'), sadness (e.g. 'Since the past alarm, how much have you felt sad or depressed?'). These self-reported items were presented on a seven-point Likert Scale (1 not at all and 7 extremely). Analyses using these variables focused on the epoch in which the METER was administered.

### Statistical analysis

We first evaluated METER adherence, which was calculated as the number of tests that were completed out of the total number possible (i.e. 10). We also evaluated the relative impact of removing low adherent participants on convergent validity. We evaluated the METER's total completion time relationship with the in-lab social cognition measures and METER performance. Parametric or non-parametric correlations (depending on whether variables violated normality assumptions) were used to examine the relationship between adherence and demographics, mental health symptoms, and cognitive variables. Then, mean squared successive difference (MSSD), or the sum of the differences between consecutive observations squared, and then divided by ($N$−1),

https://doi.org/10.1017/S0033291720004419 Published online by Cambridge University Press

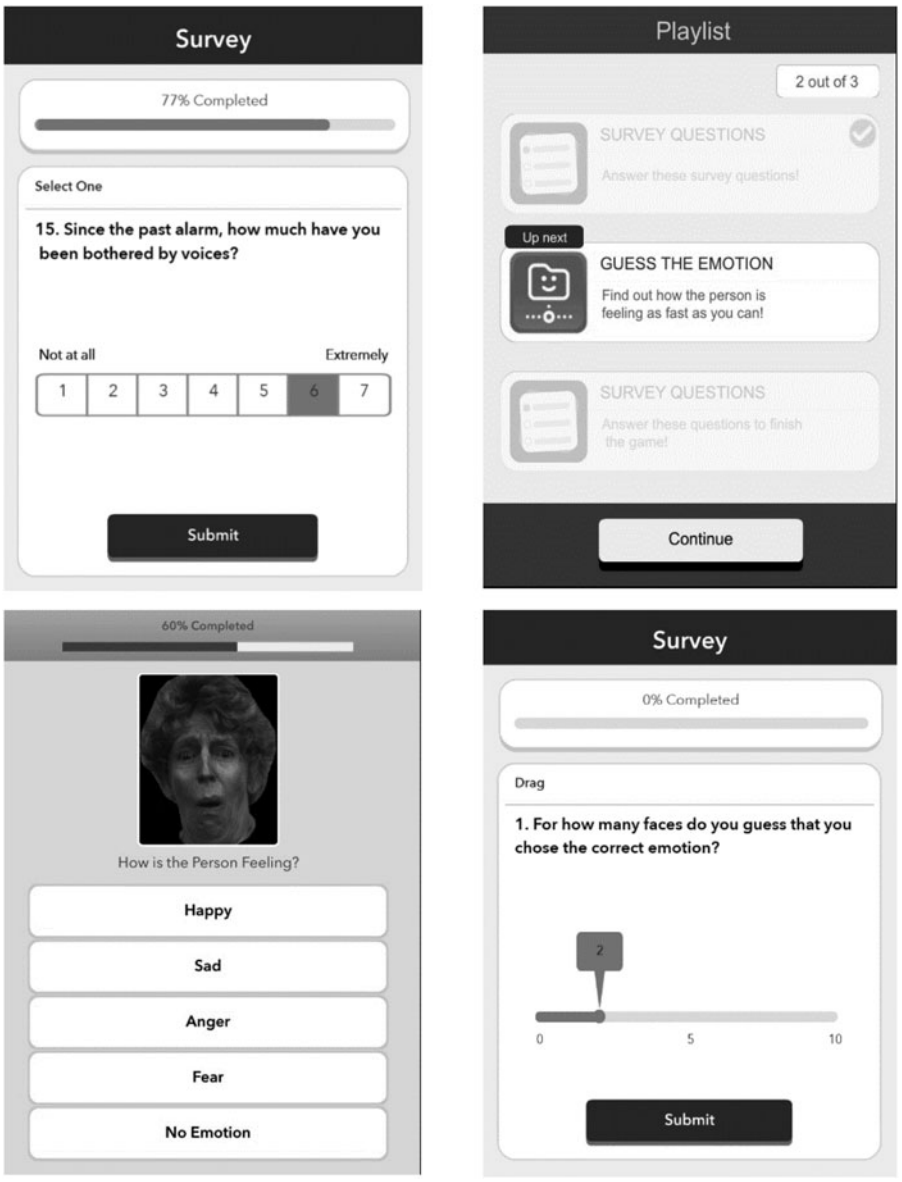



**Fig. 1.** Screenshots of ecological momentary emotion recognition test (METER).

was calculated to understand within-person variability, as was an intraclass correlation coefficient (ICC). We then evaluated performance and self-assessed performance across testing sessions to evaluate practice effects using linear mixed models. We evaluated convergent validity by examining correlations between the person-averaged METER performance and self-assessed performance with in-lab measures of facial emotion recognition (BLERT and ER-40 Total scores), non-social neurocognitive measures (MCCB tasks), and symptoms. These analyses included univariate (correlations) between individual measures and a multivariate regression to determine the contribution of social *v.* non-social cognitive measures to METER performance. With EMA data, we evaluated combined between and within-person associations between EMA measures of happiness, sadness, hearing voices, and trust in others with METER performance and self-assessed performance using linear mixed models (Twisk, 2019). All linear mixed models had a random effect for the intercept (subject). The $\alpha$ value was set at 0.05 and the Bonferroni correction was used for post-hoc pairwise analyses.

## Results

### *Sample characteristics and adherence*

Sample characteristics can be seen in Table 1. As might be expected, diagnostic groups differed on the PANSS Positive Syndrome Scale [$F_{(2,83)} = 0.5$, $p = 0.001$] and Negative Syndrome Scale [$F_{(2,83)} = 6.23$, $p = 0.003$]. Participants with schizoaffective disorder (M = 20.0, S.D. = 5.0) had a higher severity of positive symptoms than those with schizophrenia (M = 18.5, S.D. = 5.24) and the mood disorder group (M = 14.0, S.D. = 5.9). Additionally, the schizophrenia group (M = 14.7, S.D. = 4.2) showed higher severity of negative symptoms than those with schizoaffective disorder (M = 12.5, S.D. = 3.7) and the mood disorder group (M = 11.0, S.D. = 2.8). Groups also differed by depression severity on the MADRS [$F_{(2,82)} = 4.7$, $p = 0.012$], with the group with schizoaffective disorder (M = 18.9, S.D. = 10.9) having more severe depressive symptoms than the group with schizophrenia (M = 11.0, S.D. = 11.6) and the mood disorder group not differing significantly from either psychosis group (M = 19.2, S.D. = 13.8). Overall, the sample had more severe positive

https://doi.org/10.1017/S0033291720004419 Published online by Cambridge University Press

**Table 1.** Sample characteristics ($N = 86$)

| | Mean (S.D.) or % | Range |
|---|---|---|
| Age | 44.1 (11.4) | 19–65 years |
| Gender – % female | 54.7% | |
| Race | | |
| White | 26.7% | |
| Black or African American | 48.9% | |
| Asian | 4.6% | |
| Other | 19.8% | |
| Ethnicity | | |
| % Hispanic | 23.3% | |
| Educational attainment (years) | 12.7 (2.4) | 4–18 years |
| Primary diagnosis (%) | | |
| Schizophrenia | 39.5% | |
| Schizoaffective disorder | 40.7% | |
| Bipolar disorder with psychosis | 17.5% | |
| Major depressive disorder with psychosis | 2.3% | |
| Social cognition variables | | |
| ER-40 Total | 30.2 (5.7) | 6–40 |
| BLERT Total | 14.0 (3.9) | 4–21 |
| Neurocognitive variables | | |
| WRAT-4 Standard Score | 94.4 (10.4) | 76–116 |
| Symbol Digit Coding Total | 44.9 (12.6) | 17–90 |
| HVLT Recall Total Score | 21.2 (5.4) | 9–35 |
| Letter Number Span | 11.6 (3.4) | 4–20 |
| Animal Fluency Total | 20.4 (5.5) | 9–38 |
| Trail Making Test A (in seconds) | 35.4 (15.3) | 14–91 |
| Global Impairment *T* score | 42.2 (7.2) | 23.8–59 |
| Symptoms | | |
| PANSS Positive | 18.2 (5.6) | 7–34 |
| PANSS Negative | 13.1 (4.0) | 7–26 |
| MADRS Total | 15.9 (12.3) | 0–39 |
| YMRS Total | 1.8 (3.7) | 0–16 |

WRAT-4, Wide Range Achievement Test 4; ER-40, Penn Emotion Recognition Task; BLERT, Bell Lysaker Emotion Recognition Task; HVLT, Hopkins Verbal Learning Test; Global Impairment T-Score was calculated by averaging the MCCB age-corrected T-scores; PANSS, Positive and Negative Symptoms Scale; MADRS, Montgomery–Åsberg Depression Scale; YMRS, Young Mania Rating Scale. The ranges are observed from our sample.

symptoms and comparable negative symptoms to prior reports involving social cognition validation (Pinkham et al., 2018) but was otherwise similar in terms of demographic distribution.

### *METER adherence, mean performance, variability, and reaction time*

The mean rate of adherence (number of tests completed/number provided) for the METER task was 79.8% (S.D. = 20.9), ranging from 10% to 100%. Adherence was not correlated with any demographic, cognitive, or symptom variables ($p$'s > 0.05; see online Supplementary Table S1). Adherence was not significantly different across schizophrenia, schizoaffective disorder, and the mood disorder group [$F_{(2,83)} = 0.23$, $p = 0.799$] or by the presence of current suicidal ideation [$F_{(1,84)} = 0.9$, $p = 0.585$].

Mean percent of faces correct on the METER was 75.6% (S.D. = 11.0%), which was very similar to the self-assessed correct number of faces, 76.5% (S.D. = 15.1%). Interestingly, mean actual and self-assessed performance were not correlated ($\rho = 0.151$, $p = 0.159$). In terms of potential practice effects, performance was negatively associated with protocol day, with slight but significant declines in performance over time (estimate = −0.07, S.E. = 0.23, $t = -2.98$, $p = 0.003$), but no significant changes over time were observed in self-assessed performance (estimate = 0.006, S.E. = 0.03, $t = -0.23$, $p = 0.822$). Performance on the METER was negatively correlated with age and positively correlated with education and WRAT-4 Standard Score (Table 2). After removing seven individuals who have <50% adherence on the METER, performance was no longer correlated with WRAT-4 score. There were no correlations between METER participant self-assessed performance and other demographics characteristics ($p$'s > 0.05; Table 2).

In terms of within-person variability, we found that performance on the METER had a higher MSSD (5.21, S.D. = 3.12) than self-assessed performance on the METER (MSSD = 4.08, S.D. = 5.74). Greater variability of performance on the METER was correlated with older age ($\rho = 0.282$, $p = 0.009$). The ICC for performance on the METER was 0.29, whereas the ICC for self-assessed performance on the METER was 0.51. Mean total completion time for the task, aggregated across all 10 lists, was 49.61 s (S.D. = 85.8). The mean total completion time was negatively associated with performance ($\rho = -0.218$, $p = 0.044$). A linear mixed model revealed that there was no effect of day on completion time (estimate = −2.56, S.E. = 2.9, $t = -0.88$, $p = 0.382$).

### *Convergent validity with the METER performance, variability, and reaction time*

Mean performance on the METER was strongly positively associated with the ER-40 total score ($\rho = 0.454$, $p < 0.001$) as well as the BLERT total score ($\rho = 0.592$, $p < 0.001$) (Table 2). METER performance was associated with all non-social MCCB neurocognitive measures with the exception of the HVLT total score. METER correlations with non-social cognition were slightly lower than that of the ER-40 or BLERT. The strength and significance of associations in the subsample of participants with 50% or higher adherence ($N = 79$) was highly similar to that in the entire sample, yet with TMT-A score was no longer significantly associated with METER performance. To evaluate the relative association of METER performance to social and non-social tests, a linear regression predicting METER performance including social cognition and non-social cognition tests was significant overall [$F_{(7,78)} = 11.4$, $p < 0.001$, $R^2 = 0.51$]. BLERT emerged as the only significant predictor ($B = 0.04$, $t = 6.1$, $p < 0.001$) followed by ER40 ($B = 0.01$, $t = 2.0$, $p = 0.053$). Participant self-assessed performance on the METER had no relationship with any of these social or non-social cognitive scores (Table 2).

Greater within-person variability in performance as calculated by MSSDs on METER was associated with worse ER40 performance ($\rho = -0.312$, $p = 0.004$). There were no other correlations between variability of performance and other variables of interest ($p$'s > 0.153), and there were no correlations between variability of

https://doi.org/10.1017/S0033291720004419 Published online by Cambridge University Press

**Table 2.** METER parametric and non-parametric correlations with in-lab variables ($N = 86$)

| | METER actual performance (number of faces correct) | METER self-assessed performance | METER mean total completion time (in seconds) |
|---|---|---|---|
| Age | −0.306 ($p = 0.004$)*[a] | 0.108 ($p = 0.324$) | 0.407 ($p < 0.001$)**a |
| Education (years) | 0.248 ($p = 0.021$)*[a] | −0.098 ($p = 0.370$) | −0.224 ($p = 0.038$)*a |
| WRAT-4 | 0.261 ($p = 0.015$)*[a] | −0.065 ($p = 0.554$) | −0.271 ($p = 0.012$)*a |
| Affect recognition | | | |
| ER-40 Total | 0.454 ($p < 0.001$)**[a] | 0.081 ($p = 0.459$)a | −0.200 ($p = 0.065$)a |
| BLERT Total | 0.592 ($p < 0.001$)**[a] | −0.047 ($p = 0.669$)a | −0.327 ($p = 0.002$)**a |
| Non-social neurocognition | | | |
| Symbol Digit | 0.466 ($p < 0.001$)**[a] | −0.015 ($p = 0.891$) | −0.438 ($p < 0.001$)**a |
| HVLT Total | 0.168 ($p = 0.122$)[a] | −0.107 ($p = 0.326$) | −0.145 ($p = 0.183$)a |
| Letter Number Span | 0.368 ($p < 0.001$)**[a] | −0.068 ($p = 0.534$) | −0.267 ($p = 0.013$)*a |
| Animal Fluency | 0.260 ($p = 0.015$)*[a] | −0.102 ($p = 0.352$) | −0.372 ($p < 0.001$)**a |
| Trail Making Test A | −0.223 ($p = 0.039$)*[a] | 0.021 ($p = 0.850$)a | 0.382 ($p < 0.001$)**a |
| Global Impairment *T* Score | 0.293 ($p = 0.006$)**[a] | −0.001 ($p = 0.994$) | −0.293 ($p = 0.006$)**a |
| Psychosis and mood symptoms | | | |
| PANSS Positive Score | −0.482 ($p < 0.001$)**a | −0.063 ($p = 0.567$) | 0.213 ($p = 0.049$)*a |
| PANSS Negative Score | 0.000 ($p = 0.997$)a | −0.061 ($p = 0.575$)a | 0.070 ($p = 0.524$)a |
| MADRS Total | −0.001 ($p = 0.990$)[a,b] | −0.299 ($p = 0.005$)**[a,b] | −0.090 ($p = 0.414$)[a] |
| YMRS Total | −0.046 ($p = 0.680$)[a,c] | −0.174 ($p = 0.118$)[a,c] | 0.079 ($p = 0.483$)[a] |

WRAT-4, Wide Range Achievement Test 4; ER-40, Penn Emotion Recognition Task; BLERT, Bell Lysaker Emotion Recognition Task; HVLT, Hopkins Verbal Learning Test; Global Impairment T-Score was calculated by averaging the MCCB age-corrected T-scores; PANSS, Positive and Negative Symptoms Scale; MADRS, Montgomery–Åsberg Depression Scale; YMRS, Young Mania Rating Scale.
[a]Non-parametric correlation.
[b]$N = 85$; [c]$N = 82$.
*Significant at $p < 0.05$; **significant at $p < 0.01$.

self-assessed performance and variables of interest ($p$'s > 0.125). Finally, reaction time on the METER Task was associated negatively with BLERT, ER40, and non-social tasks, with longer reaction time associated particularly strongly with worse performance on timed tasks (e.g. Trail Making Test, Symbol Digit).

### Associations with in-lab symptom measures

PANSS positive syndrome score was strongly negatively correlated with METER performance ($\rho = -0.537$, $p < 0.001$) (Table 2). To evaluate whether this effect was confounded with diagnosis, we repeated this analysis with only participants with a diagnosis of schizophrenia or schizoaffective disorder, and found a similar correlation ($\rho = -0.540$, $p < 0.001$). By comparison, the PANSS positive syndrome score was also negatively correlated with the BLERT ($\rho = -0.315$, $p = 0.003$) but not the ER-40 ($\rho = -0.115$, $p = 0.319$). Additionally, the specific association between PANSS positive syndrome scale was significant when adjusting in a partial correlation for PANSS general psychopathology ($r = -0.467$, $p < 0.001$).

There was no significant association between METER performance and the PANSS negative syndrome scale nor the MADRS total score. Self-assessed performance on the METER was negatively correlated with depressive symptoms (MADRS total score; $\rho = -0.30$, $p = 0.005$). The mean total completion time was positively significantly associated with only PANSS positive symptoms ($\rho = 0.213$, $p = 0.049$).

### Associations between METER and time-varying EMA-assessed symptoms

Linear mixed models were used to assess the effects of concurrent EMA-reported psychosis symptoms of hearing voices, and mistrustfulness in others, along with affective ratings of happiness and sadness, on actual and self-assessed performance on the METER. In these models, we simultaneously evaluated the person-averaged effect and momentary differences from average effects. Person-averaged voices (estimate = −0.29, S.E. = 0.07, $t = -4.3$, $p < 0.001$), mistrustfulness (estimate = −0.13, S.E. = 0.06, $t = -2.2$, $p = 0.032$), sadness, (estimate = −0.14, S.E. = 0.06, $t = -2.1$, $p = 0.037$) were associated with reduced accuracy in METER, with no significant effects of momentary changes (there was a trend for momentary increases in voices and reduced performance, estimate = 0.11, S.E. = 0.06, $t = 1.8$, $p = 0.080$). Both person-averaged and momentary psychotic symptoms were not associated with self-assessed performance ($p$'s > 0.05). In contrast, person-averaged self-assessed performance was associated negatively with person-averaged sadness (estimate = 0.33, S.E. = 0.10, $t = -3.5$, $p = 0.001$) and happiness (estimate = 0.33, S.E. = 0.10, $t = 3.42$, $p = 0.001$), with no effect of momentary changes. As seen in Fig. 2, effects that combined both actual and self-assessed performance, overestimation was associated with more severe voices, whereas underestimation was associated with greater sadness and lesser happiness.

https://doi.org/10.1017/S0033291720004419 Published online by Cambridge University Press

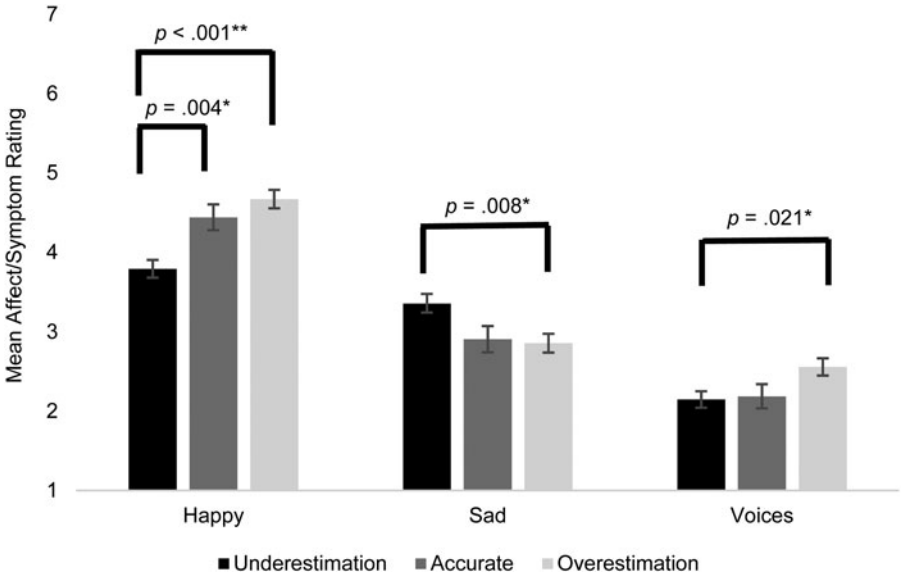



**Fig. 2.** Overestimation and underestimation of performance on the METER by EMA voices severity and mood. *Note.* This figure depicts the concurrent associations between underestimation, overestimation and accurate estimation of the METER tests with self-reported EMA voices severity and mood. As shown, there is a significant association between reported overestimation and reported happiness. The same trend applies to voices severity. However, those that tend to report sadness tend to underestimate their METER performance. Linear mixed models; Overestimation > accurate, *estimate* = 0.23, *S.E.* = 0.20, $p$ = 0.004; Overestimation > underestimation, *estimate* = 0.87, *S.E.* = 0.16, $p < 0.001$; Happy: $F(2,694) = 15.4$, $p < 0.001$; Sad: $F(2,694) = 5.1$, $p = 0.006$; Underestimation > overestimation, *estimate* = 0.49, *S.E.* = 0.16, $p = 0.008$; Voices: $F(2,674) = 4.1$, $p = 0.017$; Overestimation > underestimation, *estimate* = .40, *S.E.* = 0.15, $p = 0.021$ (pairwise contrasts Bonferroni adjusted)

## Discussion

This paper provides preliminary evidence for the validity of a new mobile self-administered ecological momentary emotion recognition test (METER) which enables the evaluation of social cognitive ability and self-assessments of ability in naturalistic settings. The measure was well tolerated, with an average adherence of 79.8%, and all participants contributed data for analyses. Adherence was not correlated with any demographic or symptom variables, indicating broad tolerability. Despite frequent repetition, the measure was not associated with detectable practice effects. In terms of convergent validity, performance on the task was highly associated with gold standard in-lab affect recognition measures as well as other non-social neurocognitive measures. Highlighting the potential utility of intensively repeated tests, concurrently assessed psychosis symptoms (severity of voices and mistrustfulness) and sadness were associated with diminished performance, whereas sadness and positive affect but not psychotic symptoms impacted self-assessed performance. Taken together, these findings extend prior cross-sectional work on the influence of psychotic symptoms on social cognitive ability, and mood symptoms on biased judgements of performance. Thus, the METER could be a useful complement to a variety of applications in social cognition research.

Our findings address many of the dimensions used to evaluate the validity of lab-based social cognitive tasks [see Social Cognition Psychometric Evaluation study (SCOPE); Pinkham et al., 2018; Pinkham, Penn, Green, & Harvey, 2016]. In terms of practicability and tolerability, the METER was associated with a relatively high rate of adherence, which was likely boosted by the practice of micro-payments per survey and check-ins from staff. Baseline symptoms, cognitive, or demographic data did not impact adherence, and so there did not appear to be subgroups who experienced greater challenges with completing the task; in particular, adherence did not vary by level of cognitive impairment, which might be assumed to determine whether individuals can complete self-administered tasks. Adherence also did not appear to markedly impact convergent validity.

Although the task did not display substantial practice effects over the course of 10 days, future research would be needed to evaluate test–retest reliability of the task over separate measurement epochs. Furthermore, the task was associated with gold-standard, in-lab measures of the same construct (BLERT, ER-40) as well as to a lesser extent non-social neurocognition tests. Regression analyses indicated some specificity toward validation to the target of facial emotion recognition. Nonetheless, some psychometric properties remain to be evaluated. In particular, a central aspect of SCOPE was validation against measures of functional outcome, which will be evaluated in future studies examining the METER. In addition, other psychometric properties typical of in-lab measures, such as internal consistency, are challenging to measure with mobile repeated tasks. Each testing epoch contains a relatively small number of stimuli and it is somewhat unclear how internal consistency metrics could be inclusive of repeated administrations over time when the underlying construct under study may also change; we found slight performance declines over time and so task design must take into account performance changes as they correspond to the ordering of stimuli. As such, ecological momentary tasks may also require different kinds of psychometric indices. Further, the lack of control group in this study inhibits the establishment of normative performance.

In addition to establishing the initial validity of the METER, this study showcases some of the potential for EMA to examine how clinical factors might influence social cognition and confidence judgments. Our study is consistent with prior work indicating psychotic symptoms, in particular voices, negatively impact social cognitive ability whereas mood, but not psychotic symptoms, is associated with self-assessment of performance. Disentangling between and within-person effects, these effects were best accounted for between variation rather than day-to-day increases in symptoms. There were surprisingly few associations with negative symptoms, although the sample was likely enriched for positive symptoms given the focus on suicidal ideation. Further, the PANSS may be less optimal for quantifying negative symptoms compared to other instruments such as the Clinical Assessment Interview for Negative Symptoms (CAINS; Kring, Gur, Blanchard, Horan, & Reise, 2013).

Self-assessed performance is an emerging area of research in psychotic disorders because of its link to functional outcome (Gould et al., 2015; Silberstein & Harvey, 2019), and biases in judgements of performance may alter effort, motivation, and

https://doi.org/10.1017/S0033291720004419 Published online by Cambridge University Press

sustainment of goal-directed activities (Cornacchio, Pinkham, Penn, & Harvey, 2017; Gould, Sabbag, Durand, Patterson, & Harvey, 2013; Harvey, Strassnig, & Silberstein, 2019). Extending prior work of in-lab studies (Harvey et al., 2015, 2019; Moritz et al., 2015), our study indicated that overestimation of performance was linked to concurrent severity of voices as measured by EMA, whereas sadness was associated with underestimation of performance (and reduced performance). As with actual performance, these effects were most aligned with between-person variation rather than within-person fluctuation. This study demonstrates that over- and underestimation biases can be studied in real-time. This opens the door for evaluating person and time-varying mechanisms and social-environmental influences on these biases, such as with lagged models that exploit time series, the impact of these biases on everyday social decision making, social avoidance, and behavior. It may also be possible for rehabilitative interventions to attempt to alter biases as they occur, such as with feedback delivered through ecological momentary interventions.

There were several limitations to the study. The sample size was small and so validity should be considered preliminary, and the findings on the strength and direction of associations with in-lab and EMA measures would need to be replicated in a larger sample. The sample was stratified to over-recruit for participants with current suicidal ideation, and the mean level of current depression and psychosis severity were likely higher than that of prior studies of social cognitive tests that recruited more psychiatrically stable outpatient samples. In addition, at this time, we lack data from multiple EMA epochs, and so test–retest reliability across EMA-bursts is unknown. Lastly, the METER is a measure of only one domain of social cognition and future work may evaluate whether other domains (e.g. theory of mind) could be translated to mobile self-administered ecological momentary formats.

In summary, this study provided initial validation of a novel mobile self-administered facial emotion task, with a positive indication of adherence, tolerability, practicability, and lack of observed practice effects, along with convergent validity with gold-standard lab-based measures of the same construct and non-social-related neurocognitive domains. EMA analyses reveal that psychotic symptoms influence facial emotion recognition accuracy but not self-assessed performance, whereas mood had a stronger impact on self-assessed performance. Future work will evaluate test–retest reliability and capitalize on whether and how these observed accuracy deficits and biases influence behavior, including social function and suicidal behavior. Finally, this study provides optimism that other social cognitive tasks could be translated into EMA paradigms.

**Supplementary material.** The supplementary material for this article can be found at https://doi.org/10.1017/S0033291720004419.

**Acknowledgements.** We would like to thank Katelyn Barone, Bianca Tercero, Cassi Springfield, Linlin Fan, Ian Kilpatrick, and Maxine Hernandez for their involvement in data collection and recruitment. We would also like to thank Mayra Cano for her efforts in data collection and in managing the data across the three sites.

**Financial support.** This work was supported by the National Institute of Mental Health (grant number: NIMH R01 MH116902-01A1).

**Conflict of interest.** Dr Raeanne C. Moore is a co-founder of KeyWise AI, Inc. and a consultant for NeuroUX. Dr Philip D. Harvey has received consulting fees or travel reimbursements from Acadia Pharma, Alkermes, Bio Excel, Boehringer Ingelheim, Minerva Pharma, Otsuka Pharma, Regeneron Pharma, Roche Pharma, and Sunovion Pharma during the past year. He receives royalties from the Brief Assessment of Cognition in Schizophrenia. He is the chief scientific officer of i-Function, Inc. He had a research grant from Takeda and from the Stanley Medical Research Foundation. None of these companies provided any information to the authors that is not in the public domain. No other authors have conflicts of interest to report.

**Ethical standards.** The authors assert that all procedures contributing to this project comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

## References

Allard, M., Husky, M., Catheline, G., Pelletier, A., Dilharreguy, B., Amieva, H., … Swendsen, J. (2014). Mobile technologies in the early detection of cognitive decline. *PLoS ONE*, *9*(12), e112197. doi:10.1371/journal.pone.0112197.

Balzan, R. P., Woodward, T. S., Delfabbro, P., & Moritz, S. (2016). Overconfidence across the psychosis continuum: A calibration approach. *Cognitive Neuropsychiatry*, *21*(6), 510–524. doi:10.1080/13546805.2016.1240072.

Brandt, J., & Benedict, R. H. B. (2001). *The Hopkins verbal learning test – revised*. Odessa, FL: Psychological Assessment Resources.

Bryson, G., Bell, M., & Lysaker, P. (1997). Affect recognition in schizophrenia: A function of global impairment or a specific cognitive deficit. *Psychiatry Research*, *71*(2), 105–113. doi:10.1016/s0165-1781(97)00050-4.

Comparelli, A., Corigliano, V., De Carolis, A., Mancinelli, I., Trovini, G., Ottavi, G., … Girardi, P. (2013). Emotion recognition impairment is present early and is stable throughout the course of schizophrenia. *Schizophrenia Research*, *143*(1), 65–69. doi:10.1016/j.schres.2012.11.005.

Cornacchio, D., Pinkham, A. E., Penn, D. L., & Harvey, P. D. (2017). Self-assessment of social cognitive ability in individuals with schizophrenia: Appraising task difficulty and allocation of effort. *Schizophrenia Research*, *179*, 85–90. doi:10.1016/j.schres.2016.09.033.

Fett, A.-K. J., Maat, A., & Investigators, G. (2013). Social cognitive impairments and psychotic symptoms: What is the nature of their association? *Schizophrenia Bulletin*, *39*(1), 77–85. doi:10.1093/schbul/sbr058.

Fett, A.-K. J., Viechtbauer, W., Penn, D. L., van Os, J., & Krabbendam, L. (2011). The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, *35*(3), 573–588. doi:10.1016/j.neubiorev.2010.07.001.

First, M. B., Williams, J. B. W., Karg, R. S., & Spitzer, R. L. (2015). *Structured clinical interview for DSM-5 – research version* (*SCID-5 for DSM-5, research version; SCID-5-RV*). Arlington, VA: American Psychiatric Association.

Gold, J. M., Carpenter, C., Randolph, C., Goldberg, T. E., & Weinberger, D. R. (1997). Auditory working memory and Wisconsin card sorting test performance in schizophrenia. *Archives of General Psychiatry*, *54*, 159–165. doi:10.1001/archpsyc.1997.01830140071013.

Gould, F., McGuire, L. S., Durand, D., Sabbag, S., Larrauri, C., Patterson, T. L., … Harvey, P. D. (2015). Self-assessment in schizophrenia: Accuracy of evaluation of cognition and everyday functioning. *Neuropsychology*, *29*(5), 675–682. doi:10.1037/neu0000175.

Gould, F., Sabbag, S., Durand, D., Patterson, T. L., & Harvey, P. D. (2013). Self-assessment of functional ability in schizophrenia: Milestone achievement and its relationship to accuracy of self-evaluation. *Psychiatry Research*, *207*(1–2), 19–24. doi:10.1016/j.psychres.2013.02.035.

Green, M. F., Bearden, C. E., Cannon, T. D., Fiske, A. P., Hellemann, G. S., Horan, W. P., … Nuechterlein, K. H. (2012). Social cognition in schizophrenia, part 1: Performance across phase of illness. *Schizophrenia Bulletin*, *38*(4), 854–864. doi:10.1093/schbul/sbq171.

Green, M. F., Horan, W. P., & Lee, J. (2015). Social cognition in schizophrenia. *Nature Reviews Neuroscience*, *16*(10), 620–631. doi:10.1038/nrn4005.

Gur, R. C., Sara, R., Hagendoorn, M., Marom, O., Hughett, P., Macy, L., … Gur, R. E. (2002). A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies. *Journal of Neuroscience Methods*, *115*(2), 137–143. doi:10.1016/s0165-0270(02)00006-7.

Hajduk, M., Harvey, P. D., Penn, D. L., & Pinkham, A. E. (2018). Social cognitive impairments in individuals with schizophrenia vary in severity. *Journal of Psychiatric Research*, *104*, 65–71. doi:10.1016/j.jpsychires.2018.06.017.

https://doi.org/10.1017/S0033291720004419 Published online by Cambridge University Press

Harvey, P. D., Deckler, E., Jones, M. T., Jarskog, L. F., Penn, D. L., & Pinkham, A. E. (2019). Autism symptoms, depression, and active social avoidance in schizophrenia: Association with self-reports and informant assessments of everyday functioning. *Journal of Psychiatric Research*, *115*, 36–42. doi:10.1016/j.jpsychires.2019.05.010.

Harvey, P. D., Paschall, G., & Depp, C. (2015). Factors influencing self-assessment of cognition and functioning in bipolar disorder: A preliminary study. *Cognitive Neuropsychiatry*, *20*(4), 361–371. doi:10.1080/13546805.2015.1044510.

Harvey, P. D., & Pinkham, A. E. (2015). Impaired self-assessment in schizophrenia: Why patients misjudge their cognition and functioning. *Current Psychiatry*, *14*(4), 53–59.

Harvey, P. D., Strassnig, M. T., & Silberstein, J. (2019). Prediction of disability in schizophrenia: Symptoms, cognition, and self-assessment. *Journal of Experimental Psychopathology*, *10*(3), 1–20. 2043808719865693..

Hoe, M., Nakagami, E., Green, M., & Brekke, J. (2012). The causal relationships between neurocognition, social cognition and functional outcome over time in schizophrenia: A latent difference score approach. *Psychological Medicine*, *42* (11), 2287–2299. doi:10.1017/S0033291712000578.

Janssens, M., Lataster, T., Simons, C., Oorschot, M., Lardinois, M., Van Os, J., & Myin-Germeys, I. (2012). Emotion recognition in psychosis: No evidence for an association with real world social functioning. *Schizophrenia Research*, *142*(1–3), 116–121. doi:10.1016/j.schres.2012.10.003.

Jones, M. T., Deckler, E., Laurrari, C., Jarskog, L. F., Penn, D. L., Pinkham, A. E., & Harvey, P. D. (2020). Confidence, performance, and accuracy of self-assessment of social cognition: A comparison of schizophrenia patients and healthy controls. *Schizophrenia Research Cognition*, *19*, 002. doi:10.1016/j.scog.2019.01.002.

Jongstra, S., Wijsman, L. W., Cachucho, R., Hoevenaar-Blom, M. P., Mooijaart, S. P., & Richard, E. (2017). Cognitive testing in people at increased risk of dementia using a smartphone app: The iVitality proof-of-principle study. *JMIR mHealth and uHealth*, *5*(5), e68. doi:10.2196/mhealth.6939.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, *13*(2), 261–276. doi:10.1093/schbul/13.2.261.

Keefe, R. S., Goldberg, T. E., Harvey, P. D., Gold, J. M., Poe, M. P., & Coughenour, L. (2004). The brief assessment of cognition in schizophrenia: Reliability, sensitivity, and comparison with a standard neurocognitive battery. *Schizophrenia Research*, *68*(2–3), 283–297. doi:10.1016/j.schres.2003.09.011.

Kohler, C. G., Turner, T. H., Bilker, W. B., Brensinger, C. M., Siegel, S. J., Kanes, S. J., … Gur, R. C. (2003). Facial emotion recognition in schizophrenia: Intensity effects and error pattern. *The American Journal of Psychiatry*, *160*(10), 1768–1774. doi:10.1176/appi.ajp.160.10.1768.

Köther, U., Veckenstedt, R., Vitzthum, F., Roesch-Ely, D., Pfueller, U., Scheu, F., & Moritz, S. (2012). "Don't give me that look" – overconfidence in false mental state perception in schizophrenia. *Psychiatry Research*, *196*(1), 1–8. doi:10.1016/j.psychres.2012.03.004.

Kring, A. M., Gur, R. E., Blanchard, J. J., Horan, W. P., & Reise, S. P. (2013). The clinical assessment interview for negative symptoms (CAINS): Final development and validation. *The American Journal of Psychiatry*, *170*(2), 165–172. doi:10.1176/appi.ajp.2012.12010109.

Lahera, G., Ruiz-Murugarren, S., Iglesias, P., Ruiz-Bennasar, C., Herreria, E., Montes, J. M., & Fernandez-Liria, A. (2012). Social cognition and global functioning in bipolar disorder. *The Journal of Nervous and Mental Disease*, *200*(2), 135–141. doi:10.1097/NMD.0b013e3182438eae.

McCleery, A., Lee, J., Fiske, A. P., Ghermezi, L., Hayata, J. N., Hellemann, G. S., … Green, M. F. (2016). Longitudinal stability of social cognition in schizophrenia: A 5-year follow-up of social perception and emotion processing. *Schizophrenia Research*, *176*(2–3), 467–472. doi:10.1016/j.schres.2016.07.008.

Mehta, U. M., Thirthalli, J., Subbakrishna, D., Gangadhar, B. N., Eack, S. M., & Keshavan, M. S. (2013). Social and neuro-cognition as distinct cognitive factors in schizophrenia: A systematic review. *Schizophrenia Research*, *148*(1–3), 3–11. doi:10.1016/j.schres.2013.05.009.

Montgomery, S. A., & Åsberg, M. (1979). A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry*, *134*(4), 382–389. doi:10.1192/bjp.134.4.382.

Moore, R. C., Campbell, L. M., Delgadillo, J. D., Paolillo, E. W., Sundermann, E. E., Holden, J., … Swendsen, J. (2020). Smartphone-based measurement of executive function in older adults with and without HIV. *Archives of Clinical Neuropsychology*, *35*(4), 347–357. doi:10.1093/arclin/acz084.

Moore, R. C., Swendsen, J., & Depp, C. A. (2017). Applications for self-administered mobile cognitive assessments in clinical research: A systematic review. *International Journal of Methods in Psychiatric Research*, *26*(4), e1562. doi:10.1002/mpr.1562.

Moritz, S., Balzan, R. P., Bohn, F., Veckenstedt, R., Kolbeck, K., Bierbrodt, J., & Dietrichkeit, M. (2016). Subjective versus objective cognition: Evidence for poor metacognitive monitoring in schizophrenia. *Schizophrenia Research*, *178*(1–3), 74–79. doi:10.1016/j.schres.2016.08.021.

Moritz, S., Goritz, A. S., Gallinat, J., Schafschetzy, M., Van Quaquebeke, N., Peters, M. J., & Andreou, C. (2015). Subjective competence breeds overconfidence in errors in psychosis. A hubris account of paranoia. *Journal of Behavior Therapy and Experimental Psychiatry*, *48*, 118–124. doi:10.1016/j.jbtep.2015.02.011.

Mote, J., & Fulford, D. (2020). Ecological momentary assessment of everyday social experiences of people with schizophrenia: A systematic review. *Schizophrenia Research*, *216*, 56–68. doi:10.1016/j.schres.2019.10.021.

Nuechterlein, K. H., Green, M. F., Kern, R. S., Baade, L. E., Barch, D. M., Cohen, J. D., … Marder, S. R. (2008). The MATRICS consensus cognitive battery, part 1: Test selection, reliability, and validity. *The American Journal of Psychiatry*, *165*(2), 203–213. doi:10.1176/appi.ajp.2007.07010042.

Pinkham, A. E., Harvey, P. D., & Penn, D. L. (2016). Paranoid individuals with schizophrenia show greater social cognitive bias and worse social functioning than non-paranoid individuals with schizophrenia. *Schizophrenia Research: Cognition*, *3*, 33–38. doi:10.1016/j.scog.2015.11.002.

Pinkham, A. E., Harvey, P. D., & Penn, D. L. (2018). Social cognition psychometric evaluation: Results of the final validation study. *Schizophrenia Bulletin*, *44*(4), 737–748. doi:10.1093/schbul/sbx117.

Pinkham, A. E., Penn, D. L., Green, M. F., & Harvey, P. D. (2016). Social cognition psychometric evaluation: Results of the initial psychometric study. *Schizophrenia Bulletin*, *42*(2), 494–504. doi:10.1093/schbul/sbv056.

Posner, K., Brown, G. K., Stanley, B., Brent, D. A., Yershova, K. V., Oquendo, M. A., … Shen, S. (2011). The Columbia–Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry*, *168*(12), 1266–1277. doi: 10.1176/appi.ajp.2011.10111704.

Schweitzer, P., Husky, M., Allard, M., Amieva, H., Pérès, K., Foubert-Samier, A., … Swendsen, J. (2017). Feasibility and validity of mobile cognitive testing in the investigation of age-related cognitive decline. *International Journal of Methods in Psychiatric Research*, *26*(3), e1521. doi:10.1002/mpr.1521.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., … Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, *59* (Suppl 20), 22–33, quiz 34–57. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9881538.

Silberstein, J. M., & Harvey, P. D. (2019). Impaired introspective accuracy in schizophrenia: An independent predictor of functional outcomes. *Cognitive Neuropsychiatry*, *24*(1), 28–39.

Silberstein, J. M., Pinkham, A. E., Penn, D. L., & Harvey, P. D. (2018). Self-assessment of social cognitive ability in schizophrenia: Association with social cognitive test performance, informant assessments of social cognitive ability, and everyday outcomes. *Schizophrenia Research*, *199*, 75–82. doi:10.1016/j.schres.2018.04.015.

Spreen, O. (1991). Controlled oral word association (word fluency). In O. Spreen, & E. Strauss (Eds.), *A compendium of neuropsychological tests* (pp. 447–463). Oxford: Oxford University Press.

Tombaugh, T. N. (2004). Trail Making Test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology*, *19*(2), 203–214. doi:10.1016/S0887-6177(03)00039-8.

Twisk, J. W. (2019). *Applied mixed model analysis: A practical guide*. Cambridge, England: Cambridge University Press.

Ventura, J., Wood, R. C., & Hellemann, G. S. (2013). Symptom domains and neurocognitive functioning can help differentiate social cognitive processes in schizophrenia: A meta-analysis. *Schizophrenia Bulletin*, *39*(1), 102–111. doi:10.1093/schbul/sbr067

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide range achievement test 4 professional manual*. Lutz, FL: Psychological Assessment Resources.

Young, R. C., Biggs, J. T., Ziegler, V. E., & Meyer, D. A. (1978). A rating scale for mania: Reliability, validity and sensitivity. *British Journal of Psychiatry*, *133*(5), 429–435. doi:10.1192/bjp.133.5.429.

https://doi.org/10.1017/S0033291720004419 Published online by Cambridge University Press