Psychological Medicine

cambridge.org/psm

Original Article

Cite this article: Badal VD, Depp CA, Harvey PD, Ackerman RA, Moore RC, Pinkham AE (2023). Confidence, accuracy judgments and feedback in schizophrenia and bipolar disorder: a time series network analysis. *Psychological Medicine* **53**, 4200–4209. https://doi.org/10.1017/S0033291722000939

Received: 9 December 2021
Revised: 9 March 2022
Accepted: 22 March 2022
First published online: 28 April 2022

Keywords:
Bipolar disorder; cognition; metacognition; network analysis; schizophrenia; causal time series analysis

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

CAMBRIDGE UNIVERSITY PRESS

Confidence, accuracy judgments and feedback in schizophrenia and bipolar disorder: a time series network analysis

Varsha D. Badal[1,2], Colin A. Depp[1,2,3], Philip D. Harvey[4,5], Robert A. Ackerman[6], Raeanne C. Moore[1,3] and Amy E. Pinkham[6,7]

[1]Department of Psychiatry, University of California San Diego, San Diego, California, USA; [2]Sam and Rose Stein Institute for Research on Aging, University of California San Diego, San Diego, California, USA; [3]VA San Diego Healthcare System, La Jolla, California, USA; [4]Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, Miami, FL, USA; [5]Research Service, Miami VA Healthcare System, Miami, FL, USA; [6]School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA and [7]Department of Psychiatry, University of Texas Southwestern Medical School, Dallas, TX, USA

Abstract

Background. Inaccurate self-assessment of performance is common among people with serious mental illness, and it is associated with poor functional outcomes independent from ability. However, the temporal interdependencies between judgments of performance, confidence in accuracy, and feedback about performance are not well understood.

Methods. We evaluated two tasks: the Wisconsin Card Sorting Test (WCST) and the Penn Emotion recognition task (ER40). These tasks were modified to include item-by-item confidence and accuracy judgments, along with feedback on accuracy. We evaluated these tasks as time series and applied network modeling to understand the temporal relationships between momentary confidence, accuracy judgments, and feedback. The sample constituted participants with schizophrenia (SZ; $N = 144$), bipolar disorder (BD; $N = 140$), and healthy controls (HC; $N = 39$).

Results. Network models for both WCST and ER40 revealed denser and lagged connections between confidence and accuracy judgments in SZ and, to a lesser extent in BD, that were not evidenced in HC. However, associations between feedback regarding accuracy with subsequent accuracy judgments and confidence were weaker in SZ and BD. In each of these comparisons, the BD group was intermediate between HC and SZ. In analyses of the WCST, wherein incorporating feedback is crucial for success, higher confidence predicted worse subsequent performance in SZ but not in HC or BD.

Conclusions. While network models are exploratory, the results suggest some potential mechanisms by which challenges in self-assessment may impede performance, perhaps through hyperfocus on self-generated judgments at the expense of incorporation of feedback.

Introduction

A growing body of literature spanning cognitive science, education, and more recently, psychiatry, has found associations between inaccurate judgments of performance, or poor introspective accuracy (IA), and impairments in everyday functioning (Harvey & Pinkham, 2015). IA can be operationalized in a number of ways but herein we refer to it as the discrepancy between objective, potentially accessible, data and subjective estimation of task performance (Silberstein & Harvey, 2019). The correlation between impaired IA and everyday functioning exceeds the contribution of ability variables (Gould et al., 2015; Silberstein, Pinkham, Penn, & Harvey, 2018). People with psychotic disorders demonstrate poor IA in multiple domains, including cognition, social cognition, functional capacity, and everyday functioning, and across various measurement approaches (Durand et al., 2021; Gould et al., 2015; Harvey & Pinkham, 2015; Tercero et al., 2021).

In some domains, poor IA may be an even more discriminating feature of schizophrenia than task performance, as our recent studies (Badal et al., 2021a; Pinkham, Harvey, & Penn, 2018) indicated greater separation of people with SZ and HC on self-assessment of performance than accuracy alone on a facial emotion recognition task. Further, a generally positive bias in self-assessment (over-confidence) was detected and was correlated with greater impairments in performance on the specific social cognitive (Jones et al., 2019) or neurocognitive (Perez, Tercero, Penn, Pinkham, & Harvey, 2020) tasks. However, from these correlational studies, albeit featuring within-study longitudinal examination of task performance, it is not clear how IA interferes with function, nor or to what extent these challenges in judgment and response biases are specific to schizophrenia *v.* other serious mental illnesses like bipolar disorder.

Evaluation of the dynamic associations between accuracy judgments, confidence, and feedback on accuracy (feedback on correctness) may help to unravel the mechanisms underlying this effect. There are several possible mechanisms through which IA could interfere with task performance and subsequently everyday functioning. Over confidence seems to be associated with lower correlations between self-assessments and ability (Jones et al., 2019) and also to be correlated with diminished ability to adaptively adjust effort (Cornacchio, Pinkham, Penn, & Harvey, 2017). Another explanation is a diminished receptiveness to feedback (Gold, Waltz, Prentice, Morris, & Heerey, 2008; Goldberg, Weinberger, Berman, Pliskin, & Podd, 1987), possibly based on over-reliance on self-generated information compared to externally available information. Similar response biases and self-assessment challenges have been identified in research on the resistance of delusional thinking to counter-evidence (Engh et al., 2010). It has been suggested that people with SZ deploy processing resources more intensely but narrowly, failing to assimilate a wider set of environmental cues which might include correctness feedback (Luck, Hahn, Leonard, & Gold, 2019). This 'hyperfocusing' could lead to discounting externally generated information in favor of internally generated information. A recent study suggests confidence may be utilized as a substitute for information when information is lacking (Ptasczynski, Steinecker, Sterzer, & Guggenmos, 2021). Therefore, several factors may link poor IA to performance problems, but correlational research may make it difficult to understand the processes from which accuracy judgments, confidence, and accuracy arise and how they are expressed on a momentary basis.

To study IA impairments, modifications of existing tasks have been developed that, on an item-by-item basis, ask for: (1) a response to that item, (2) a judgment of whether that response was correct or incorrect, and (3) confidence in that judgment. These questions are followed by feedback about the actual correctness of that item. At the end of the task the participant is also typically asked to generate a global judgment of performance on the task as a whole (Springfield & Pinkham, 2020; Tercero et al., 2021). While most prior research has evaluated inter-relationships among aggregate scores on feedback on accuracy, accuracy judgments, and confidence, performance on the tasks and all these self-assessment variables are ordered in time and evaluating the temporal relationships in the data set may lead to an understanding of the underlying cognitive processes. The application of network models offers promise for untangling some of these processes that may take place. Such networks are often constructed using sets of intensively sampled variables such as from ecological momentary assessment (EMA) data (Badal, Parrish, Holden, Depp, & Granholm, 2021b; Shiffman, Stone, & Hufford, 2008). Going a step further, the EMA data are accompanied by timing information, and the temporal ordering of samples enables the construction of network models that represent contemporaneous as well as time-lagged relationships between variables. Prior work has applied network models to affective experience in schizophrenia (Strauss et al., 2019), but none to our knowledge have evaluated cognitive processes.

Therefore, we evaluated two tasks from a multi-site study examining IA among a sample of people with schizophrenia, bipolar disorder and healthy controls. We applied network modeling techniques to gain insight on the effect of accuracy judgments on actual accuracy over time. The two in-lab tasks we included in the study were modified versions of the Wisconsin Card Sorting Task (WCST) and the Penn Emotion recognition test (ER-40), wherein feedback about the accuracy of each response is provided, after the participant has rendered their judgment regarding accuracy and their confidence in that judgment. We hypothesized that network models exploring temporal links between feedback on accuracy, accuracy judgements and confidence would differ across diagnostic categories; from past correlational analyses, we expected the links between confidence and feedback on accuracy to be weaker in SZ than in BD and HC. We also expected the links between feedback on accuracy to be more tightly linked to confidence in the WCST compared to the ER40 given that centrality of feedback to adequate performance on the WCST. The comparison between the ER40 and WCST provide a contrast of tasks wherein responses to feedback is crucial for successful subsequent performance (i.e. WCST) or is not relevant (i.e. ER40).

Method

Participants

Participants were a part of a larger investigation into IA; they were outpatients recruited from three universities: (1) University of California San Diego (UCSD), including Outpatient Psychiatric Services clinic, a large public mental health clinic, the San Diego VA Medical Center, other local community clinics, (2) The University of Texas at Dallas (UTD), including Metrocare Services, a nonprofit mental health services organization in Dallas County, Texas, and from other local clinics, and (3) The University of Miami (UM), including the Jackson Memorial Hospital-University of Miami Medical Center and the Miami VA Medical Center. The diagnostic groups included schizophrenia, schizoaffective disorder, or bipolar I or II disorder, with or without current or previous psychotic symptoms meeting the criteria defined in Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-V). The study is ongoing with a target size of $n = 450$; this interim analysis includes patients with diagnoses of schizophrenia or schizoaffective disorder (SZ, $n = 144$), bipolar disorder (BD, $n = 140$) and healthy controls (HC, $n = 39$).

Inclusion criteria included (1) a DSM-V diagnosis of Schizophrenia, Schizoaffective Disorder, or Bipolar Disorder (with or without psychotic features) or for HCs, no DSM-V diagnosis, (2) age 18 to 65 years old, (3) English proficiency, (4) outpatient, (5) stable medications for at least 6 weeks, and (6) willingness to provide a high contact informant with no prior psychiatric diagnosis. Exclusion criteria included (1) a history of or current medical or neurological disorders that might affect brain functioning (e.g. stroke, untreated seizure disorder, loss of consciousness greater than 15 min), (2) low estimated verbal IQ (i.e. a standard score less than 70 on the Wide Range Achievement Test 4 Reading test (Wilkinson & Robertson, 2006) (or pervasive developmental disorder according to the DSM-IV criteria, 3) substance use disorder in the past six months (excluding tobacco and cannabis), and (4) visual or hearing impairments that interfere with assessment. Participants were also excluded if they had been hospitalized within the past six weeks. All participants provided written informed consent and the studies were approved by institutional review boards at each of the sites (Harvey et al., 2021).

Measures

Metacognitive WCST (Tercero et al., 2021)

The WCST is a standard neuropsychological test of problem solving and cognitive flexibility, an important component of executive functioning. For this study, a modified version of the WCST was administered (called the Metacognitive WCST) (Tercero et al., 2021). Like the original, this computerized test presents participants with a sequence of 64 cards. The participants are instructed to sort the cards without any predefined criteria for sorting. This version of the test used the standard color-form-number category sequences, changing to the next category after 10 consecutive correct responses. For each item, the response, the participant's judgment on accuracy (Did you get it correct? – Yes/No), along with the participant's confidence in the correctness of their accuracy judgment (on a 5-point scale from 0%–100% confident) was also recorded. Following those three sequential responses from the participant, feedback regarding the actual accuracy of the response was provided to the participant (i.e. correct, incorrect).

Emotion recognition (ER40) (Gur et al., 2002)

The ER-40 is a 40-item computer-based task, with each item involving the identification of an emotion depicted in the photograph of a face. Four basic emotions (i.e. happiness, sadness, anger, or fear) and neutral expressions in equal proportions were presented, one at a time. Participants were required to identify the correct emotion for each face, with a self-assessment (i.e. accuracy judgment and confidence) and feedback procedure that was the same as for the WCST.

Positive and negative syndrome scale (PANSS) (Kay et al., 1987)

The PANSS scale measures the severity of symptoms in people with schizophrenia. It is a 30-item scale comprising 7 points for positive symptoms (such as hallucinations and delusions), 7 points for negative symptoms (such as reduced expression) and 16 for general psychopathology.

Montgomery-Asberg depression rating scale (MADRS) (Montgomery & Åsberg, 1979)

It is a 10 item diagnostic questionnaire used to measure depression severity. The score ranges between 0–60, a score of 35 or above is considered severe.

Analysis

Network analysis

Network analysis was performed using Tigramite (Runge, 2019), a Python implementation of Momentary Conditional Independence [PCMCI; (Runge, Nowack, Kretschmer, Flaxman, & Sejdinovic, 2019)]. The implementation uses high sensitivity and effectively eliminates spurious dependencies. For the WCST, we used a 32-item window ($\tau_{\max} = 32$), and for the ER40 dataset, we used a 20-item window ($\tau_{\max} = 20$), which correspond to half the total number of items and responses in the list, providing a balance of inter- and intra-individual effects when establishing lagged dependencies. The assumptions for ER40 and the WCST tasks, that each item corresponds to fixed time interval are simplifications.

The networks thus constructed were compared across the groups based upon measures of edge density, goodness of fit, and presence of feedback loops. Network density quantifies how interconnected the nodes in the graph are, and the greater the density, the greater the presence of feedback loops and complex behavior of the system. Network density is defined as:

$$D = \frac{\sum_{i,j}^{N} w_{i \to j}}{N(N-1)/2} \quad (1)$$

Where $w_{i \to j}$ is an edge connecting nodes i and j in the network of N nodes.

Goodness of fit (R^2) attempts to measure how similar two network graphs with identical set of nodes are, by computing the square of differences in correlations of corresponding edges, normalizing it, and subtracting it from 1. A measure of 1 implies identical structure, and value closer to 0 would imply little similarity.

$$R^2 = 1 - \sum_{i,j}^{N} (w_{i \to j}^{\mathrm{DX}} - w_{i \to j}^{\mathrm{HC}})^2 \quad (2)$$

Where N is the total number of nodes in the network of participants with diagnosis (DX), i and j are nodes, and $w_{i \to j}$ is a weighted edge between them. Only the edges with p value < 0.05 (the default alpha parameter in Tigramite) and significant p values were represented in graphs.

Feedback loops, wherein a sequence of edges starting and terminating at the same node, were also identified and interpreted. A presence of odd number of negative edges along the path is considered as a negative feedback loop, that is associated with homeostasis. An even number (or zero) of negative edges results in a net positive feedback loop that could imply amplification or attenuation of involved variables.

Network analysis based on small samples are prone to Type 1 errors in correlations. The Benjamini–Hochberg Method (Benjamini & Hochberg, 1995) was used in the PCMCI to effectively control the false discovery rate for all networks, including for the smaller HC ($n = 39$) group. The method has gained high visibility, as it has shown relevance even in very small sample size genome wide studies (Storey & Tibshirani, 2003).

Results

Sample characteristics

Demographic information and clinical characteristics for the groups are presented in Table 1. Of the 323 participants, 44.6% had a diagnosis of SZ, 43.3% had a diagnosis of BD, and 12.1% were HC. Individuals with SZ differed from HC in age, education, race and employment status, while people with BD differed from HCs in gender and employment only. Individuals with SZ were older, had a greater proportion of males, and had fewer years of education compared to participants with BD. Race and employment status were also significantly different between the two groups. Individuals with SZ had more positive and negative symptoms and lesser depressive symptomatology compared to participants with BD. Differences in all other socio-demographics across the groups were insignificant. Generally, mean confidence in performance was highest amongst people with SZ, and the feedback on accuracy was the lowest, although the differences were not significant in between-groups analyses. SZ showed significantly lower IA (correct and estimated sort match) in the WCST task compared to HC.

Lagged network analysis

WCST

The WCST networks included 3 variables: accuracy judgments, confidence, and feedback on accuracy (Fig. 1). The density (number of relationships between variables) of the SZ network was the highest (4.00), followed by BD (3.33) and then HC (2.33) (Fig. 1).

https://doi.org/10.1017/S0033291722000939 Published online by Cambridge University Press

Table 1. Demographic details of participants and scores on ER40 and WCST tasks

	Healthy control HC (*N* = 39)	Bipolar BD (*N* = 140)	Schizophrenia SZ (*N* = 144)	HC *v.* BD		HC *v.* SZ		SZ *v.* BD	
				t Test χ^2	*p* Value	*t* Test χ^2	*p* Value	*t* Test χ^2	*p* Value
Age – mean (s.d.); range	35.2 (12.0); 19–60	38.7 (11.5); 19–64	41.9 (10.8); 20–64	−1.68	0.096	−3.36	**0.001**	−2.41	**0.017**
Gender – female (%)	16 (41.0%)	97 (69.3%)	70 (48.6%)	9.29	**0.002**	0.44	0.509	11.69	**0.001**
Education in years – mean (s.d.)	14.1 (1.8)	14.3 (2.4)	12.6 (2.4)	−0.44	0.661	3.58	**<0.001**	5.84	**<0.001**
Single (%)	19 (48.7%)	68 (48.6%)	86 (59.7%)	0.03	0.869	1.10	0.294	3.12	0.077
Race									
Caucasian (%)	61.5%	60.7%	37.5%	0.01	0.927	6.30	**0.012**	14.39	**<0.001**
African American (%)	23.1%	21.4%	50.0%	0.00	0.999	7.96	**0.005**	23.95	**<0.001**
Other (%)	15.4%	17.9%	12.5%	0.01	0.903	0.04	0.837	1.20	0.274
Ethnicity									
Hispanic (%)	25.6%	30.0%	23.6%	0.13	0.722	0.00	0.959	1.25	0.263
Vocation (some overlap between categories)									
Unemployed	12.8%	49.3%	72.2%	15.26	**<0.001**	42.53	**<0.001**	14.74	**<0.001**
Part time – student	0.0%	2.1%	2.1%						
Full time – student	15.4%	6.4%	1.4%						
Part time – employment	25.6%	23.6%	20.8%						
Full time – employment	51.3%	20.7%	5.6%						
Clinical measures									
WRAT-3 – mean (s.d.)	103.8 (9.2)	102.1 (11.5)	96.0 (11.7)	0.90	0.369	3.88	**<0.001**	4.40	**<0.001**
PANSS positive symptoms – mean (s.d.)	–	11.8 (4.5)	16.7 (4.8)	–	–	–	–	−8.91	**<0.001**
PANSS negative symptoms – mean (s.d.)	–	10.7 (2.6)	13.5 (4.7)	–	–	–	–	−5.18	**<0.001**
MADRS total – mean (s.d.)	–	13.0 (10.7); 0–37	9.6 (9.8); 0–38	–	–	–	–	2.81	**0.005**
ER40 – mean (s.d.); range									
Correct faces (out of 40)	26.4 (13.8); 0–40	26.1 (12.0); 0–40	25.1 (11.7); 0–40	0.16	0.875	0.61	0.540	0.71	0.479
				(Cohen's *d*: 0.03)		(Cohen's *d*: 0.11)		(Cohen's *d*: 0.08)	
Confidence – (100-point scale)	65.3 (35.4); 0–100	66.1 (31.7); 0–100	67.0 (32.1); 0–100	−0.13	0.894	−0.28	0.778	−0.23	0.815
				(Cohen's *d*: −0.02)		(Cohen's *d*: −0.05)		(Cohen's *d*: −0.03)	
Estimated Correct Faces (out of 40)	27.6 (16.9); 0–40	29.1 (15.4); 0–40	28.5 (15.4); 0–40	−0.54	0.588	−0.33	0.739	0.32	0.746
				(Cohen's *d*: −0.10)		(Cohen's *d*: −0.06)		(Cohen's *d*: 0.04)	
Correct and estimated faces match	33.2 (7.3); 0–40	31.5 (7.0); 0–40	30.9 (6.4); 0–40	1.28	0.203	1.91	0.057	0.80	0.425

(*Continued*)

T-he R^2

Table 1. *(Continued.)*

	Healthy control HC (N = 39)	Bipolar BD (N = 140)	Schizophrenia SZ (N = 144)	HC *v.* BD		HC *v.* SZ		SZ *v.* BD	
				t Test χ^2	*p* Value	*t* Test χ^2	*p* Value	*t* Test χ^2	*p* Value
				(Cohen's *d*: 0.23)		(Cohen's *d*: 0.35)		(Cohen's *d*: 0.09)	
WCST – mean (s.d.); range									
Correct sorts (out of 64)	29.6 (18.7); 0–64	30.2 (17.0); 0–64	26.6 (14.5); 0–64	−0.17	0.864	1.08	0.280	1.91	0.057
				(Cohen's *d*: −0.03)		(Cohen's *d*: 0.20)		(Cohen's *d*: 0.23)	
Confidence −100-point scale)	58.3 (34.4); 0–100	61.6 (30.6); 0–100	63.1 (30.0); 0–100	−0.58	0.563	−0.86	0.392	−0.42	0.676
				(Cohen's *d*: −0.10)		(Cohen's *d*: −0.15)		(Cohen's *d*: −0.05)	
Estimated correct sort (out of 64)	39.2 (26.4); 0–64	42.5 (25.2); 0–64	41.4 (24.0); 0–64	−0.70	0.483	−0.48	0.632	0.38	0.703
				(Cohen's *d*: −0.13)		(Cohen's *d*: −0.09)		(Cohen's *d*: 0.05)	
Correct and estimated sort match	45.1 (12.3); 0–64	41.7 (13.1); 0–64	38.3 (12.9); 0–64	1.45	0.149	2.91	**0.004**	2.15	**0.032**
				(Cohen's *d*: 0.26)		(Cohen's *d*: 0.53)		(Cohen's *d*: 0.26)	

WRAT-3, Wide Range Achievement Test 3; PANSS, Positive and Negative Syndrome Scale; MADRS, Montgomery–Åsberg Depression Rating Scale; ER40, Penn Emotion recognition task; WCST, Wisconsin Card Sorting Test.
Values in bold denote statistical significance at the $p < 0.05$ level.

measures suggest that the SZ network ($R^2 = 0.75$) diverged from HC network to a greater extent than the BD network ($R^2 = 0.87$) (Fig. 1). Task accuracy judgments were generally strongly tied to confidence in HC, BD, and SZ. The contemporaneous association between confidence and accuracy judgment was the least for HC (0.23), followed by BD (0.27) and SZ (0.33), suggesting accuracy judgment overlaps concurrently with confidence in the latter groups to a greater degree. The SZ and BD networks included multiple lagged linkages between accuracy judgment and confidence which were not evident in HC, such that previous confidence and accuracy judgements were more highly associated with current values of the same variable.

In contrast, the correlations obtained from network analysis (Table 2) between confidence and feedback on accuracy were strongest in HC (0.11), followed by BD (0.10) and then SZ (0.05), indicating a greater relationship between accuracy and subsequent confidence ratings, suggesting utilization of external feedback. The greatest lagged influence of feedback on accuracy judgment was also displayed by HC (0.11), followed by BD (0.07), and the least in SZ (0.04). These data suggest that the HC display the greatest assimilation of feedback. Notably, the lagged correlations from prior confidence ratings to later feedback on accuracy were negative in SZ (−0.03) indicating that past higher confidence correlates with poorer future accuracy on the WCST task. This link was missing in HC, and positive in BD (0.04).

All three groups show positive feedback loops (not to be confused with feedback on accuracy; the reference here is to a sequence of edges, some with lags, in the graph starting and ending at the same node creating a 'sustained' effect (Borsboom & Cramer, 2013) between feedback on accuracy and accuracy judgments. The lagged link was strongest in HC, followed by BD and SZ, suggesting the incorporation of feedback regarding accuracy is greatest for HC (Fig. 1). Similar feedback patterns exist for feedback on accuracy and confidence. People with SZ are unique in the presence of a negative loop from confidence to feedback on accuracy and back to confidence (Fig. 1*b*) suggesting over confidence is longitudinally associated with poorer performance for the group.

ER40

ER40 networks included the same 3 variables: accuracy judgment, confidence, and feedback on accuracy (Fig. 2). The network densities showed a similar pattern to that in the WCST; SZ network was the highest (2.67), followed by BD (2.00) and HC (1.00) (Fig. 2). The R^2 measures of the SZ network (0.94) and the BD network (0.95) suggest close similarity to HC networks (Fig. 2); networks diverged across diagnoses less than that of the networks derived for the WCST task.

Accuracy judgments were strongly tied to confidence, from least to most in HC (0.27), SZ (0.29) and BD (0.33). In addition, much like the WCST, accuracy judgments and confidence were tightly coupled with each other in SZ with multiple lagged and bidirectional associations. The HC and BD networks lacked these lagged edges.

As in the WCST, confidence correlated with feedback on accuracy most strongly in HC (0.26), followed by BD (0.24) and then SZ (0.20) (Table 2). The lagged correlation from feedback on accuracy to confidence ratings was negative in BD (−0.04) but non-existent in HC and SZ. This suggests a tendency for individuals with BD to not improve in confidence despite receiving positive feedback.

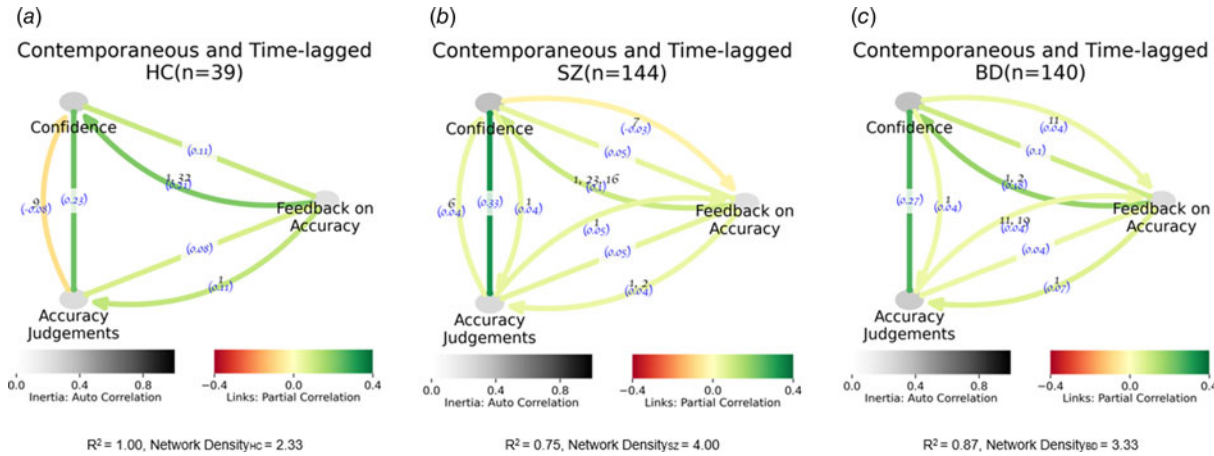

Fig. 1. Network constructed from WCST-IA dataset. Network graphs representing interaction between confidence, accuracy and accuracy judgment for healthy controls (HC), and the two clinical categories [individuals with schizophrenia and schizoaffective disorders (SZ) and bipolar disorder (BD)]. (*a*) Accuracy feedback strongly influences future confidence and accuracy judgment in HC. As accuracy judgment increases, the responses become tentative, this is specific to WCST task due to shifting criteria for sorting. (*b*) Among individuals with SZ, over-confidence is associated with future accuracy negatively. (*c*) BD network displays mostly intermediate characteristics, in between those of HC and SZ. In our figures, positive correlations are represented in green while negative are in red, the depth of color encodes the strength of the association (in blue).

Table 2. Strengths of contemporaneous and time-lagged associations from network analysis for WCST and ER40 tasks (IA dataset)

	Strength of association		
	Schizophrenia	Bipolar	Healthy control
WCST			
Contemporaneous			
Confidence – accuracy judgment	0.33	0.27	0.23
Confidence – accuracy	0.05	0.1	0.11
Accuracy – Accuracy Judgment	0.05	0.04	0.08
Time lagged			
Confidence→accuracy	−0.03	0.04	–
Accuracy→Confidence	0.1	0.18	0.21
Accuracy→accuracy judgment	0.04	0.07	0.11
Accuracy judgment→accuracy	0.05	0.04	–
Accuracy judgment→confidence	0.04	–	−0.08
Confidence→accuracy judgment	0.04	0.04	–
ER-40			
Contemporaneous			
Confidence – accuracy judgment	0.29	0.33	0.27
Confidence – accuracy	0.2	0.24	0.26
Accuracy – accuracy judgment	0.17	0.19	0.18
Time lagged			
Confidence→Accuracy	–	–	–
Accuracy→Confidence	–	−0.04	–
Accuracy→Accuracy judgment	−0.04	–	–
Accuracy judgment→Accuracy	0.06	0.06	–
Accuracy judgment→Confidence	0.05	–	–
Confidence→accuracy judgment	0.06	–	–

Dashes imply no significant correlations exist and hence the network edges do not exist.

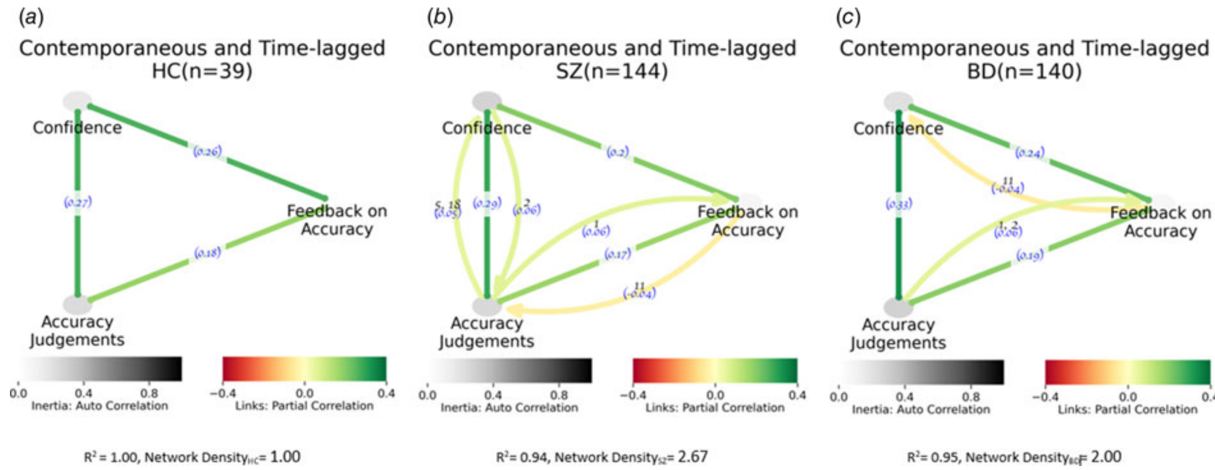

Fig. 2. Network constructed from ER40-IA dataset: (*a*) Confidence has the strongest correlation to accuracy in healthy controls (HC). Accuracy has little learning associated with it. (*b*) Accuracy influenced accuracy judgment negatively among individuals with schizophrenia and schizoaffective disorders (SZ). (*c*) Accuracy influenced confidence negatively among individuals with bipolar disorder (BD).

Unlike the WCST positive feedback loop (here again, we imply a sequence of edges that start and terminate at the same node creating a sustained effect) between feedback on accuracy and accuracy judgments was absent in HC. In ER40, feedback on accuracy was not connected with a lag to confidence or accuracy judgment in HC; consistent with the properties of the task in which there is no learning transferred from one item to other. However, item-by-item feedback predicted future accuracy judgments negatively for SZ (−0.04) (Fig. 2*b*), and predicted future confidence negatively for BD (−0.04) (Fig. 2*c*). This suggests some undue influence of past performance, in a task where items are seemingly unrelated. These lagged effects are absent in HCs (Fig. 2*a*).

Discussion

Several studies (Jones et al., 2019; Silberstein & Harvey, 2019; Sabbag et al., 2012) have identified that there is a much smaller correlation between subjective self-assessments and objective indicators of performance in people with psychotic disorders compared to people without psychotic disorders. Further, across these studies, a greater disconnect between self-assessments and objective data is associated with poorer performance across different measures of neurocognition and social cognition (Jones et al., 2019; Perez et al., 2020) and worse functional outcomes (Gould et al., 2015; Silberstein et al., 2018). However, this is the first study to evaluate the dynamics of momentary self-assessment in relation to task-based performance on the WCST or ER-40 task, using network models adapted for time series data, to further understand how within-person variation in self-assessments (accuracy judgments and confidence in those judgments) influences performance on the later items in the task.

Consistent across the ER-40 and WCST, these findings suggest that (1) among individuals with SZ, confidence was more decoupled from feedback on accuracy compared to these associations in BD and SZ, and accuracy judgments were more associated with confidence ratings in people with BD and SZ, both concurrently and in lagged relationships, and (2) feedback regarding one's performance is more impactful on accuracy judgments and confidence among HC. The degree of deviation from HC along these two dimensions follows a gradient from BD to SZ. In the case of the WCST, wherein incorporation of feedback on accuracy is connected to success on the task, past confidence ratings are correlated negatively with future performance, suggesting for the first time how over-confidence may interfere with cognitive task performance among individuals with SZ.

Understanding that network analyses are inherently exploratory, there are several potential explanations for the relationships observed here across WCST and ER40 that differ across the three groups. It is important to point out the key difference between the two tasks (WCST and ER40). Feedback is key to performance in the WCST, whereas feedback for each item has no bearing on the correct response for the next item in the ER40. In the WCST task, we found that while concurrent confidence and feedback on accuracy are positively correlated in SZ, BD, and HC, past confidence demonstrated a negative lagged effect in SZ but not in HC. That is, greater confidence in self-generated accuracy judgments was associated with lower performance on subsequent trials in SZ. Individuals with SZ diverged from HCs in that among individuals with SZ there was a weaker lagged link between feedback about accuracy and subsequent self-generated accuracy judgments and confidence compared to HCs. In BD and SZ, accuracy judgment was determined by both confidence and feedback on accuracy, whereas in HC, only objective feedback on accuracy was related to accuracy judgment. Potential alternative explanations include differences in attention to the task at hand, and in self-assessments (accuracy judgment and confidence). However, variability in self-assessment ratings was similar across the groups, indicating that people with SZ were indeed altering their ratings on a trial-by-trial level and not simply repeating fixed values. We found that self-assessment ratings (accuracy judgments and confidence) were more interrelated for the SZ and the BD groups than the HC group. The network models suggested that accuracy judgments were more influenced by past confidence in SZ and BD than in HC. In contrast, accuracy judgments in HC showed stronger correlations between feedback on accuracy than on confidence, suggesting accuracy judgments are based more on external cues for the HC group. This, from the perspective of a proposed Bayesian framework (Fleming & Daw, 2017), implies accuracy judgments were more tightly bound to confidence for the two clinical groups at the expense of attention to feedback on accuracy.

Although to our knowledge this is among the first studies to provide evidence for the phenomenon observed between past confidence related to worst subsequent cognitive performance in SZ, related work may have bearing. Computational modeling (Ashinoff, Singletary, Baker, & Horga, 2021) has framed delusions through a Bayesian lens, whereby prior beliefs interfere with subsequent information processing, such as consideration of dis-confirmatory evidence. In our study, it may be that prior confidence is overweighted (i.e. higher self-assessment of performance), which causes diminished ability to focus attention on performance in tasks and updating of internal representations of one's own performance based on feedback. An interpretation of these data is that hyperfocusing on self-generated accuracy judgments or confidence limits other elements of the full spectrum of task performance in this complex, multi-tasking IA task (getting the item correct, understanding when you are correct, generating global judgments on accuracy of performance). Thus, attention to prior confidence or accuracy judgments (Luck et al., 2019) may overstress the already limited working memory capacity required to perform multi-tasking judgments commonly seen in the psychosis spectrum (Harvey, Reichenberg, Romero, Granholm, & Siever, 2006). Similar to belief positive model of delusion (Erdmann & Mathys, 2021; Schmack et al., 2013), patients eventually prioritize internally generated information in direct competition with the actual external contextual information because of an inability to consider all elements of the task situation. The tendency among SZ to report false memories with stronger conviction (or errors in memory monitoring), has also been suggested (Berna, Zou, Danion, & Kwok, 2019), however in a recent study, it was found that SZ relied more upon on recent confidence history in trial-by-trial confidence rating (Zheng et al., 2022). This 'confidence leakage' occurs when previous confidence judgments should have no influence current judgments, yet they do (Shekhar & Rahnev, 2021). These ideas are speculative and require experimental approaches to confirm, but they do point to some possible mechanisms by which inaccurate self-assessment (and perhaps overfocus on self-assessment) may contribute to poor performance on tasks.

Although we have focused much of the discussion on the networks of people with SZ, the networks of participants with BD were intermediate between HC and SZ networks in respect to links between accuracy judgments and confidence and then also to feedback on accuracy. Individuals with SZ displayed the highest network density, followed by BD and HC. The higher densities in SZ and BD are brought about by the increased presence of lagged associations between variables. Density measures suggest BD networks were more like SZ networks than the HC. These findings mirror the intermediate status of the BD cohort (between HC and SZ) in general cognition (Krabbendam, Arts, van Os, & Aleman, 2005). This view is also consistent with the Bipolar and Schizophrenia Network for Intermediate Phenotypes, or BSNIP, findings (Tamminga et al., 2014). It is unclear why differences exist between SZ and BD, these effects might be intrinsic to the aspects of psychopathology such as the severity of psychotic symptoms, and also, of any medication related to the condition. Yet, we do note these findings do however parallel that of general cognitive deficit differences across the groups (Krabbendam et al., 2005) with higher performance on BD than SZ. Over-reliance on prior confidence in particular has been linked to delusional processes more aligned with SZ than BD (Klein & Pinkham, 2020). In our sample, differences in depressive symptoms were also significant between SZ and BD. Greater accuracy and awareness are sometimes associated with mild depression(Alloy & Abramson, 1979; Bortolotti & Antrobus, 2015). In this study, we did not evaluate the effect of symptoms within diagnostic categories. One interesting possibility is the extent to which self-assessment could be targeted for intervention. In people without a diagnosis of serious mental illness, task-based self-assessment accuracy can be enhanced with feedback about judgment accuracy, which evidenced transfer of training to untrained tasks (Carpenter et al., 2019). Although not yet tested in a sample of individuals with SZ, it may be that increasing the accuracy of self-assessment judgments and better aligning these judgments with task demands could have downstream effects on behavior and subsequent performance (Engeler & Gilbert, 2020).

The limitations of our study include that the networks were constructed using itemized ER40 and WCST data with simplifying assumptions of test items being evenly spaced on the time axis, and hence were exploratory. The analysis derives its conclusions only from these two tasks, restricting its scope. The sample constituting HC was considerably smaller ($n = 39$), compared to the samples of individuals with SZ ($n = 144$) and BD ($n = 140$), this was mitigated by the use of Benjamini–Hochberg Method (Benjamini & Hochberg, 1995) to reduce type 1 errors to which small-sample correlation-based studies are most prone. This is evidenced in the fact that expected higher correlations between confidence and feedback were correctly identified by the algorithm despite smaller sample size in HC. It is also important to point out that power analysis of network methods is not straightforward; although the network is constructed using significant correlations, it is unclear how a composite score for the entire network can be calculated. Also, in the conditional independence testing, some edges may be removed. In this study, we did not investigate symptom severity or variation within diagnoses, which would be a worthy topic for independent future study. In comparing BD and SZ groups, scores on depression severity were higher in the BD group, which may have contributed to differences between groups because the presence of mild depression is commonly seen in individuals with more accurate self-assessments (Alloy & Abramson, 1979; Bortolotti & Antrobus, 2015). Future research may benefit from evaluating, within BD, whether differences from healthy comparators or SZ are evident in euthymic states in metacognitive processing. Further, the samples constituted outpatients only, and the findings may not be generalizable to more severely ill patients. Our previous study (Pinkham, Kelsven, Kouros, Harvey, & Penn, 2017) suggests age, race and sex are linked to social cognitive performance in HC, but not in SZ. The difference in racial/ethnic distribution across BD, HC and SZ groups was significant, so it may have impact on the resulting networks.

In summary, this study provides suggestions as to the mechanisms through which inaccurate self-assessment may hinder performance on cognitive and functional tasks. Network analyses revealed patterns in BD and SZ that indicate greater lagged links between confidence and accuracy judgments and weaker associations with feedback on accuracy. On the WCST, where feedback is critical to task performance, greater confidence predicted worse later performance in the SZ group. Experimental approaches to delineate factors that lead to greater attentional allocation to internally generated confidence and accuracy judgments, likely at the expense of external cues, may help to specify novel approaches to improve self-assessment, which may in turn, improve performance on tasks. It should also be noted that problems with IA may be problematic for both patients and healthy individuals, even if it is not directly related to the task performance. For example, over confidence in one's abilities on one task may generalize to poor effort or unsafe behaviors in other domains.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291722000939.

Acknowledgements. We thank the participants who volunteered for this study.

Author contributions. V. D. B: Designed and implemented the timeseries based network analysis, performed data analysis, and interpreted results, edited, prepared the manuscript, and contributed to drafts of the manuscript. C. A, D: Collaborator of the study, oversaw the study, proposed the research questions, performed data analysis, and interpreted results, provided feedback throughout the process, and contributed to drafts of the manuscript. P. D. H: Collaborator of the study, oversaw the study, provided feedback throughout the process, edited, and contributed to drafts of the manuscript. R. A. A: Provided feedback, edited, and contributed to the manuscript. R. C. M: Provided feedback, edited, and contributed to the manuscript. A. E. P: Principal investigator of the study, conceived, designed, and implemented the Inlab and E. M. A study, provided feedback, edited, and contributed to drafts of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Financial support. This work was supported by the National Institute of Mental Health (R01 MH112620 to A. E. P., R01MH116902 to C. A. D.). Salary for V. D. B. was supported by the National Institute of Mental Health (NIMH) T32 Geriatric Mental Health Program MH019934 to Dilip V. Jeste).

Conflict of interest. P. D. H. has received consulting fees or travel reimbursements from Alkermes, Bio Excel, Boehringer Ingelheim, Karuna Pharma, Minerva Pharma, SK Pharma, and Sunovion Pharma during the past year. He receives royalties from the Brief Assessment of Cognition in Schizophrenia (Owned by Verasci, Inc. and contained in the MCCB). He is the chief scientific officer of i-Function, Inc. R. C. M. is a co-founder of KeyWise, Inc. and a consultant for NeuroUX. The terms of these arrangements have been reviewed and approved by UCSD in accordance with its conflict of interest policies. For the remaining authors, no conflicts of interest were declared.

Ethical standards. All participants provided written informed consent and the studies were approved by institutional review boards at each of the sites (Harvey et al., 2021).

References

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General*, *108*, 441.

Ashinoff, B. K., Singletary, N. M., Baker, S. C., & Horga, G. (2021). Rethinking delusions: A selective review of delusion research through a computational lens. *Schizophrenia Research*, S0920-9964(21)00065-7. doi: 10.1016/j.schres.2021.01.023. Epub ahead of print. PMID: 33676820; PMCID: PMC8413395.

Badal, V. D., Depp, C. A., Hitchcock, P. F., Penn, D. L., Harvey, P. D., & Pinkham, A. E. (2021a). Computational methods for integrative evaluation of confidence, accuracy, and reaction time in facial affect recognition in schizophrenia. *Schizophrenia Research: Cognition*, *25*, 100196.

Badal, V. D., Parrish, E. M., Holden, J. L., Depp, C. A., & Granholm, E. (2021b). Dynamic contextual influences on social motivation and behavior in schizophrenia: A case-control network analysis. *NPJ Schizophrenia*, *7*, 62. doi: 10.1038/s41537-021-00189-6. PMID: 34887402; PMCID: PMC8660790.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 289–300.

Berna, F., Zou, F., Danion, J.-M., & Kwok, S. C. (2019). Overconfidence in false autobiographical memories in patients with schizophrenia. *Psychiatry Research*, *279*, 374–375. doi: 10.1016/j.psychres.2018.12.063. Epub 2018 Dec 10. PMID: 30558819.

Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121.

Bortolotti, L., & Antrobus, M. (2015). Costs and benefits of realism and optimism. *Current Opinion in Psychiatry*, *28*, 194.

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, *148*, 51.

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.

Durand, D., Strassnig, M. T., Moore, R. C., Depp, C. A., Ackerman, R. A., Pinkham, A. E., & Harvey, P. D. (2021). Self-reported social functioning and social cognition in schizophrenia and bipolar disorder: Using ecological momentary assessment to identify the origin of bias. *Schizophrenia Research*, *230*, 17–23.

Engeler, N. C., & Gilbert, S. J. (2020). The effect of metacognitive training on confidence and strategic reminder setting. *PLoS One*, *15*, e0240858.

Engh, J. A., Friis, S., Birkenaes, A. B., Jónsdóttir, H., Klungsøyr, O., Ringen, P. A., … Andreassen, O. A. (2010). Delusions are associated with poor cognitive insight in schizophrenia. *Schizophrenia Bulletin*, *36*, 830–835.

Erdmann, T., & Mathys, C. (2021). A generative framework for the study of delusions. *Schizophrenia Research*, S0920-9964(20)30627-7. doi: 10.1016/j.schres.2020.11.048. Epub ahead of print. PMID: 33648810.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*, 91–114.

Gold, J. M., Waltz, J. A., Prentice, K. J., Morris, S. E., & Heerey, E. A. (2008). Reward processing in schizophrenia: A deficit in the representation of value. *Schizophrenia Bulletin*, *34*, 835–847.

Goldberg, T. E., Weinberger, D. R., Berman, K. F., Pliskin, N. H., & Podd, M. H. (1987). Further evidence for dementia of the prefrontal type in schizophrenia? A controlled study of teaching the Wisconsin card sorting test. *Archives of General Psychiatry*, *44*, 1008–1014.

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*, 675.

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*, 137–143.

Harvey, P. D., Miller, M. L., Moore, R. C., Depp, C. A., Parrish, E. M., & Pinkham, A. E. (2021). Capturing clinical symptoms with ecological momentary assessment: Convergence of momentary reports of psychotic and mood symptoms with diagnoses and standard clinical assessments. *Innovations in Clinical Neuroscience*, *18*, 24.

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

Harvey, P. D., Reichenberg, A., Romero, M., Granholm, E., & Siever, L. J. (2006). Dual-task information processing in schizotypal personality disorder: Evidence of impaired processing capacity. *Neuropsychology*, *20*, 453.

Jones, M. T., Deckler, E., Laurrari, C., Jarskog, L. F., Penn, D. L., Pinkham, A. E., … Harvey, P. D. (2019). Confidence, performance, and accuracy of self-assessment of social cognition: A comparison of schizophrenia patients and healthy controls. *Schizophrenia Research: Cognition*, *19*, 002–2. doi: 10.1016/j.scog.2019.01.002. PMID: 31832336; PMCID: PMC6889550.

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

Klein, H., & Pinkham, A. (2020). Metacognition in individuals with psychosis. *Translational Issues in Psychological Science*, *6*, 21.

Krabbendam, L., Arts, B., van Os, J., & Aleman, A. (2005). Cognitive functioning in patients with schizophrenia and bipolar disorder: A quantitative review. *Schizophrenia Research*, *80*, 137–149.

Luck, S. J., Hahn, B., Leonard, C. J., & Gold, J. M. (2019). The hyperfocusing hypothesis: A new account of cognitive dysfunction in Schizophrenia. *Schizophrenia Bulletin*, *45*, 991–1000.

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

Perez, M. M., Tercero, B. A., Penn, D. L., Pinkham, A. E., & Harvey, P. D. (2020). Overconfidence in social cognitive decision making: Correlations with social cognitive and neurocognitive performance in participants with schizophrenia and healthy individuals. *Schizophrenia Research*, *224*, 51–57.

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

Pinkham, A. E., Kelsven, S., Kouros, C., Harvey, P. D., & Penn, D. L. (2017). The effect of age, race, and sex on social cognitive performance in individuals with schizophrenia. *The Journal of Nervous and Mental Disease*, *205*, 346.

Ptasczynski, L. E., Steinecker, I., Sterzer, P., & Guggenmos, M. (2021). The value of confidence: Confidence prediction errors drive value-based learning in the absence of external feedback. *PsyArXiv*. doi:10.31234/osf.io/wmv89.

Runge. (2019). Tigramite package in Python. Access Year: 2020. URL: https://github.com/jakobrunge/tigramite/.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*, eaau4996.

Sabbag, S., Twamley, E. W., Vella, L., Heaton, R. K., Patterson, T. L., & Harvey, P. D. (2012). Predictors of the accuracy of self assessment of everyday functioning in people with schizophrenia. *Schizophrenia Research*, *137*, 190–195.

Schmack, K., de Castro, A. G.-C., Rothkirch, M., Sekutowicz, M., Rössler, H., Haynes, J.-D., … Sterzer, P. (2013). Delusions and the role of beliefs in perceptual inference. *Journal of Neuroscience*, *33*, 13701–13712.

Shekhar, M., & Rahnev, D. (2021). Sources of metacognitive inefficiency. *Trends in Cognitive Sciences*, *25*, 12–23.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1–32.

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

Springfield, C. R., & Pinkham, A. E. (2020). Introspective accuracy for social competence in psychometric schizotypy. *Cognitive Neuropsychiatry*, *25*, 57–70.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*, 9440–9445.

Strauss, G. P., Zamani Esfahlani, F., Visser, K. F., Dickinson, E. K., Gruber, J., & Sayama, H. (2019). Mathematically modeling emotion regulation abnormalities during psychotic experiences in schizophrenia. *Clinical Psychological Science*, *7*, 216–233.

Tamminga, C. A., Pearlson, G., Keshavan, M., Sweeney, J., Clementz, B., & Thaker, G. (2014). Bipolar and schizophrenia network for intermediate phenotypes: Outcomes across the psychosis continuum. *Schizophrenia Bulletin*, *40*, S131–S137.

Tercero, B. A., Perez, M. M., Mohsin, N., Moore, R. C., Depp, C. A., Ackerman, R. A., … Harvey, P. D. (2021). Using a meta-cognitive Wisconsin card sorting test to measure introspective accuracy and biases in schizophrenia and bipolar disorder. *Journal of Psychiatric Research*, 436–442. doi: 10.1016/j.jpsychires.2021.06.016. Epub 2021 Jun 11. PMID: 34147931; PMCID: PMC8319124.

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide range achievement test* (*WRAT4*). Lutz, FL: Psychological Assessment Resources.

Zheng, Y., Wang, L., Gerlofs, D. J., Duan, W., Wang, X., Yin, J., … Wang, J. (2022). Atypical meta-memory evaluation strategy in schizophrenia patients. *Schizophrenia Research: Cognition*, *27*, 100220.