OXFORD

Social Cognitive and Affective Neuroscience, 2021, 608–620

doi: 10.1093/scan/nsab028
Advance Access Publication Date: 4 March 2021
Original Manuscript

The representation of mental state information in schizophrenia and first-degree relatives: a multivariate pattern analysis of fMRI data

David Dodell-Feder,[1,2] Laura M. Tully,[3] Emily Dudek,[1] and Christine I. Hooker[4]

[1]Department of Psychology, University of Rochester, Rochester, NY 14627, USA, [2]Department of Neuroscience, University of Rochester Medical Center, Rochester, NY 14642, USA, [3]Department of Psychiatry and Behavioral Sciences, UC Davis, Davis, CA 95817, USA, and [4]Department of Psychiatry and Behavioral Sciences, Rush University Medical Center, Chicago, IL 60612, USA

Correspondence should be addressed to David Dodell-Feder, Department of Psychology, University of Rochester, 453 Meliora Hall, Rochester, NY 14627, USA. E-mail: d.dodell-feder@rochester.edu.

Abstract

Individuals with a schizophrenia-spectrum disorder (SSD) and those at familial high risk (FHR) for SSDs experience social difficulties that are related to neural abnormalities in the network of brain regions recruited during theory of mind (ToM). Prior work with these groups has focused almost exclusively on characterizing the involvement of these regions in ToM. Here, we examine the representational content of these regions using multivariate pattern analysis. We analyzed two previously collected datasets of SSD, FHR and control participants who, while undergoing functional magnetic resonance imaging, completed the false-belief task in which they read stories describing beliefs or physical representations (e.g. photographs). Univariate and multivariate analyses were performed in regions of interest to evaluate group differences in task-based activation and representational content, respectively. Compared to non-SSDs, SSDs showed reduced decoding accuracy for the category of mental states in the right temporo-parietal junction—which was related to false-belief accuracy—and the dorsal medial prefrontal cortex (DMPFC) and reduced involvement of DMPFC for mental state understanding. FHR showed no differences in decoding accuracy or involvement compared to non-FHR. Given prior studies of disrupted neural involvement in FHR and the lack of decoding differences observed here, the onset of illness may involve processes that corrupt how mental state information is represented.

Key words: theory of mind; fMRI; multivariate pattern analysis; schizophrenia; familial risk

Introduction

Our ability to form meaningful social relationships and stay socially connected to others carries profound consequences for our health and well-being (House *et al.*, 1988; Holt-Lunstad *et al.*, 2015, 2017; Yang *et al.*, 2016; Snyder-Mackler *et al.*, 2020). Successfully navigating the social world and forming such connections hinge upon our ability to attribute and reason about the mental states (i.e. beliefs, desires and intentions) of others—a process

Received: 16 June 2020; **Revised:** 18 December 2020; **Accepted:** 3 March 2021

known as theory of mind (ToM). The importance of ToM is well illustrated in cases where ToM is impaired. One such case is schizophrenia-spectrum disorders (SSDs), which are associated with marked and persistent impairments in behavioral measures of ToM (Bora *et al.*, 2009; Ventura *et al.*, 2015). In support of the notion that ToM facilitates successful social interactions, the extent of these behavioral impairments are cross-sectionally and longitudinally associated with the extent of social functioning impairments (Couture *et al.*, 2006; Fett *et al.*, 2011; Schmidt *et al.*, 2011; Horan *et al.*, 2012), which are also marked and persistent for those with an SSD (Velthorst *et al.*, 2017). Increasing research has demonstrated that ToM impairments are not merely the result of factors secondary to the illness (e.g. socioeconomic consequences and medication); individuals at familial high risk (FHR) SSDs—i.e. those with a first-degree relative with the illness—also demonstrate ToM impairments (Bora and Pantelis, 2013; Lavoie *et al.*, 2013), as well as accompanying deficits in social functioning (Tarbox and Pogue-Geile, 2011). Given that FHR is far more likely to develop an SSD than non-FHR individuals (Gottesman, 1991; Rasic *et al.*, 2014), these findings suggest that ToM impairments are present prior to illness onset and may even contribute to illness onset (Tarbox and Pogue-Geile, 2008; Kim *et al.*, 2011), a notion reflected in prominent etiological theories of SSDs (van der Gaag, 2006; Hoffman, 2007; Selten *et al.*, 2017).

Toward better understanding the nature of ToM impairment in SSDs and FHR and identifying associated neurobiological markers of SSD-related risk and conversion, increasing work has evaluated the functional properties of the neural network subserving ToM in these groups. This network—often called the 'ToM network'—most commonly includes the right and left temporo-parietal junctions (RTPJ and LTPJ), right superior temporal sulcus (RSTS), medial prefrontal cortex (MPFC) and precuneus (PC) (Mar, 2011; Schurz *et al.*, 2014; Molenberghs *et al.*, 2016). Specifically, these brain regions show preferential activation for mental state *vs* non-mental state information across a variety of tasks (e.g. requiring explicit, conscious mental state reasoning and implicit, spontaneous mental state attribution), presented through a variety of modalities (e.g. reading vignettes and watching videos). In SSDs, these regions respond abnormally to mental state information (Kronbichler *et al.*, 2017; Jáni and Kašpárek, 2018). Two recent meta-analyses found that compared to non-SSDs, SSDs showed reduced ToM-related neural activity in the MPFC, PC and aspects of the temporal cortex as well as increased ToM-related activity in TPJ (Kronbichler *et al.*, 2017; Jáni and Kašpárek, 2018), although several studies have also found reduced TPJ activity (Walter *et al.*, 2009; Lee *et al.*, 2011; e.g. Das *et al.*, 2012; Dodell-Feder, Tully, *et al.*, 2014; Lee *et al.*, 2016). Increasing work has also shown ToM-related neural abnormalities in FHR (Marjoram *et al.*, 2006; de Achával *et al.*, 2012; Villarreal *et al.*, 2014; Dodell-Feder, DeLisi, *et al.*, 2014a; Mohnke *et al.*, 2016; Herold *et al.*, 2018). A recent qualitative review found altered ToM-related neural activity in FHR groups characterized by both hypo- and hyper-activation in these same regions of the ToM network (Kozhuharova *et al.*, 2020).

When taken together, these data provide strong support for the view that the ToM network is functionally altered in the schizophrenia spectrum from latent liability to manifest illness. However, mixed findings regarding the major locus (e.g. TPJ *vs* MPFC) and nature (e.g. hyper- versus hypo-activation) of the abnormality in both SSDs and FHR make it difficult to draw strong conclusions regarding how the network changes from latent liability to manifest illness, and what becomes altered, in an information-processing sense, in the schizophrenia spectrum. Moreover, the existing literature largely addresses a single idea—namely, that regions of the ToM network show aberrant levels of involvement in mental state attribution, that is, specific regions show more or less activation during ToM in SSDs and FHR compared to healthy control participants. An alternative, uninvestigated possibility is that beyond aberrant levels of involvement, the representational content of these regions is disturbed, that is, the information about mental states contained or processed in these regions is somehow corrupted. Further, changes from latent liability to manifest illness may be best characterized by relative changes in activation and/or representational content.

This distinction between involvement and information is one that has borne important insights into neural function in SSDs (Yoon *et al.*, 2008), other disorders characterized by social impairment, such as autism spectrum disorder (Gilbert *et al.*, 2009; Coutanche *et al.*, 2011; Koster-Hale *et al.*, 2013; Richardson *et al.*, 2020), and the ToM network more generally (Skerry and Saxe, 2015; Tamir *et al.*, 2016; Koster-Hale *et al.*, 2017). In line with providing complementary yet distinct information about neural function, a key distinction between studies of neural involvement and information is the statistical frameworks they are based on (Hebart and Baker, 2018). While activation-based studies of neural involvement typically rely on univariate analysis to test for differences between conditions in a single voxel or single region (in which activation magnitudes are averaged across voxels), studies of representational content are multivariate in nature and evaluate the pattern of neural activity in response to different stimuli across voxels within a given region (Haynes and Rees, 2006; Kriegeskorte and Bandettini, 2007; Mur *et al.*, 2009; Hebart and Baker, 2018). These multivoxel activity patterns are subjected to classifiers (e.g. linear support vector machine) to determine whether experimental conditions are discriminable; that is, whether there's sufficient information contained in the activity patterns that allows for accurate decoding of condition. By jointly analyzing multiple voxels, this approach, termed multivoxel or multivariate pattern analysis (MVPA), affords better sensitivity at detecting condition or group differences than standard univariate analysis (Haynes and Rees, 2006; Norman *et al.*, 2006; Hebart and Baker, 2018) and has been shown to exhibit regional sensitivity to experimental conditions that go undetected with standard activation-based univariate analysis (Kriegeskorte *et al.*, 2006; Raizada *et al.*, 2010). Despite prior work demonstrating the utility of using MVPA to characterize neural representations of visual objects in SSDs (Yoon *et al.*, 2008), and, separately, the representation of social information in ToM-related brain regions (Skerry and Saxe, 2015; Tamir *et al.*, 2016; Koster-Hale *et al.*, 2017), to our knowledge, there has, yet, to be a study using MVPA towards characterizing the ToM network in SSDs and FHR.

Thus, here, we evaluate whether and how the representational content of mental state information is disturbed in SSDs and FHR towards better characterizing ToM-related functional abnormalities in the schizophrenia spectrum, and possible changes in the ToM network from latent liability to manifest illness. Towards that goal, we re-analyzed data from two prior task-based functional magnetic resonance imaging (fMRI) studies of the ToM network in SSDs and FHR (Dodell-Feder, DeLisi, *et al.* 2014a; Dodell-Feder, Tully, *et al.*, 2014). Both participant groups performed the false belief (FB) task (Saxe and Kanwisher, 2003; Dodell-Feder *et al.*, 2011), which is one of the most widely used tasks in neuroimaging studies of ToM (Schurz *et al.*, 2014;

Molenberghs *et al.*, 2016) known to robustly recruit the ToM network and has been used in prior neuroimaging studies of SSDs and other clinical populations (Dufour *et al.*, 2013; Dodell-Feder, Tully, *et al.*, 2014). In the false-belief task, participants read and answer true/false questions about two stories types: (i) those describing outdated (i.e. false) beliefs, and (ii) those describing outdated physical depictions of the world (i.e. as might occur in an outdated photograph or map). Both story types require the concurrent representation of a representation (i.e. a belief or photograph/map/painting) and reality, and so, in theory, are similar in non-ToM-related task demands (e.g. working memory). The two story types are also similar in linguistic features such as number of words, Flesch readability, causal content (i.e. the extent to which a story conveys causal information as indexed by causal verbs, which is related to story coherence and comprehensibility) and lexical concreteness (i.e. mean concreteness of the content words), among other linguistic features, and are also similar in conceptual features including the extent to which the story provoke thinking about physical objects and causal interactions between those objects, and the ease of mentally visualizing the story events. On the other hand, false-belief stories provoke greater thinking about mental states (e.g. beliefs, desires, emotions) and social information (e.g. social status and social roles) (see Dodell-Feder *et al.* (2011) for a more detailed description of the stories). These features make the task well-suited to addressing questions related to mental state understanding. We perform both univariate and multivariate region of interest (ROI) analyses in a priori regions and exploratory whole-brain analysis towards evaluating activation-based and information-based alterations in the schizophrenia spectrum. Further, we explore brain-behavior associations, evaluating the relation between univariate activity, multivariate pattern information, FB task performance and symptoms.

Methods

Participants

The current study involved re-analyzing two previously acquired datasets. As these studies were designed and conducted to address a separate set of hypotheses, the analyses described herein should be considered exploratory and were not pre-registered. The schizophrenia dataset included 38 participants between the ages of 18 and 58 years; 20 individuals with schizophrenia (*n* = 16, 80%) or schizoaffective disorder (*n* = 4, 20%; hereafter, SSD) and 18 non-schizophrenia control participants (non-SSD) with no current or past Axis I disorder or first-degree relative with a psychotic disorder (Table 1). All participants were administered the Structured Clinical Interview for DSM-IV Disorders (First *et al.*, 2002) to assess psychiatric illness, the Weschler Abbreviated Scale of Intelligence to assess IQ (two-subtest form, Wechsler, 2011), as well as several other measures not analyzed for the purposes of the current study. SSD and non-SSD participants did not differ in demographic characteristics or IQ. SSD participants were also administered the Positive and Negative Syndrome Scale (PANSS) to assess current symptom severity (Kay *et al.*, 1987). For a more detailed description of these participants, please see Dodell-Feder, Tully, *et al.* (2014).

The FHR dataset included 20 individuals with two or more relatives with a psychotic-spectrum disorder (at least one of which was a first-degree relative to schizophrenia or schizoaffective disorder) and 19 controls (non-FHR) with no family history of psychotic disorder, psychiatric hospitalization or suicide. All participants were between the ages of 20 and 35 years (Table 1). Personal and family history of psychiatric illness was assessed with the Diagnostic Interview for Genetic Studies (Nurnberger, 1994) and Family Interview for Genetic Studies (Maxwell, n.d.), respectively. All participants were additionally assessed with the Structured Interview for Prodromal Syndromes (Miller *et al.*, 2003) to assess psychotic symptoms. Exclusion criteria for all participants included current or past history of psychotic disorder or treatment with antipsychotic or mood-stabilizing medications. Given that familial risk status is associated with increased prevalence of psychiatric illness (Erlenmeyer-Kimling, 1997; Chang *et al.*, 2002; Faridi *et al.*, 2009; Dean *et al.*, 2010), participants were not excluded for current or past history of psychiatric illness in order to increase external validity. However, only a minority of participants met lifetime criteria for a non-SSD psychiatric illness (*n* = 9 FHR and *n* = 2 non-FHR). FHR and non-FHR participants did not differ in demographic characteristics or IQ. For a more detailed description of these participants, please see Dodell-Feder, DeLisi, *et al.* (2014a).

Table 1. Participant characteristics

	SSD dataset			FHR dataset		
	SSD	Non-SSD	Group difference	FHR	Non-FHR	Group difference
n	20	18		20	19	
Age, years	38.8 (9.7)	32.4 (12.1)	$t(36) = 1.78$, $P = 0.084$	27.2 (3.9)	26.1 (3.9)	$t(37) = 0.91$, $P = 0.367$
Sex, male/female (*n*)	12/8	12/6	χ^2 $(1, n = 38) = 0.18$, $P = 0.671$	14/6	15/4	χ^2 $(1, n = 39) = 0.07$, $P = 0.785$
Education, years	15.0 (2.3)	14.2 (2.6)	$t(36) = 1.00$, $P = 0.326$	16.0 (1.5)	16.3 (0.7)	$t(28) = 0.71$, $P = 0.486$
IQ	108.7 (13.4)	107.4 (10.7)	$t(36) = 0.03$, $P = 0.763$	115.6 (10.7)	118.3 (11.4)	$t(35) = 0.76$, $P = 0.454$
PANSS						
Positive	15.6 (5.7)					
Negative	11.8 (4.1)					
Disorganized	7.6 (4.0)					
SIPS						
Positive				2.8 (2.5)	0.1 (0.3)	$t(19) = 4.65$, $P < 0.001$
Negative				1.9 (2.0)	0.1 (0.2)	$t(19) = 4.03$, $P < 0.001$
Disorganized				2.1 (1.6)	0.4 (0.6)	$t(24) = 4.44$, $P < 0.001$
General				1.7 (1.7)	0.2 (0.6)	$t(22) = 3.45$, $P = 0.002$

Values represent *M* (SD) unless otherwise noted. SIPS = Structured Interview for Prodromal Syndromes.

Additional exclusion criteria for all studies included being a non-native English speaker, IQ < 70, neurological or major medical illness, history of head trauma and MRI contraindicator. Both studies were approved by the Harvard University Committee on the Use of Humans Subjects.

FMRI experiment: false-belief task

All participants performed an optimized version of the false-belief task (Dodell-Feder *et al.*, 2011, derived from Saxe and Kanwisher, 2003) while undergoing fMRI. In this task, participants read two types of short stories: (i) FB stories described a character's false (i.e. outdated) belief (e.g. 'The morning of the high school dance, Barbara placed her high heel shoes under the dress and then went shopping. That afternoon, her sister borrowed the shoes and later put them under Barbara's bed.'), and (ii) FPR stories described outdated physical states in the world as depicted in photographs, maps and paintings (e.g. 'Old maps of the islands near Titan are displayed in the Maritime Museum. Erosion has since taken its toll, leaving only the three largest islands.'). Following each story, participants are presented with a true/false question (e.g. FB: 'Barbara gets ready assuming her shoes are under the dress'; FPR: 'Near Titan today, there are many islands'). The full stimulus set and presentation code is available online (http://saxelab.mit.edu/use-our-efficient-false-belief-localizer).

Participants saw a total of 10 stories per condition divided into two functional runs (five stories per condition per run). The order of stories was pseudo-randomized in two orders, which were seen in approximately equal amounts between participant groups. Stimuli were presented visually in white text on a black background in the following sequence: fixation on a central cross for 12 s, story for 11 s and true/false question for 6 s (each run ended with an additional 12 s of fixation). MATLAB and the Psychophysics Toolbox (Brainard, 1997; Kleiner *et al.*, 2007) were used to present the task and collect behavioral responses.

MRI data acquisition

All MRI data were acquired with a 3 T Siemens TimTrio scanner at Harvard University. A 32-channel head coil was used to collect the SSD dataset, and a 12-channel coil was used to collected the FHR dataset. Anatomical images were acquired with a T1-weighted multi-echo MPRAGE sequence in 176 sagittal slices (voxel size = 1 mm^3). Functional data were acquired with a T2*-weighted echo-planar imaging sequence with parallel imaging (acceleration factor = 2, 47 slices, voxel size = 3 mm^3, TR = 2560 ms, TE = 30 ms and flip angle = 85°) for the SSD dataset, and a T2*-weighted echo-planar imaging sequence (40 slices, voxel size = 3 mm^3, TR = 2560 ms, TE = 30 ms and flip angle = 85°) for the FHR dataset. In both sequences, the first several volumes consisted of dummy scans that were discarded prior to analysis to allow for steady-state magnetization.

MRI data analysis

Preprocessing. Both datasets were re-preprocessed in SPM12 (http://www.fil.ion.ucl.ac.uk/spm) using the same preprocessing steps and parameters. Functional images were re-aligned to the first image of the first run, co-registered to the anatomical scan, normalized to the MNI template and smoothed using an 8 mm FWHM Gaussian kernel. Prior work has shown that spatial smoothing does not decrease the sensitivity of MVPA (Op de Beeck, 2010). We used the Artifact Detection Tools (ARTs; https://www.nitrc.org/projects/artifact_detect/, Whitfield-Gabrieli *et al.*, 2011) to identify signal artifacts (timepoints with signal that exceeded 3 SD of the global signal) and motion artifacts (timepoints that exceeded the prior timepoint in composite motion by 1 mm), which were included as nuisance regressors in the univariate analyses (see below).

Regions of interest. ROIs were defined from an independent dataset reported in Dufour *et al.* (2013) of 462 neurotypical participants who completed the FB task (available at http://saxelab.mit.edu/use-our-theory-mind-group-maps). Specifically, ROIs were defined as 6 mm spheres around peak coordinates from a whole-brain random-effects analysis of FB > FPR (voxel-level threshold $t > 3$, $k > 10$): dorsal medial prefrontal cortex (DMPFC; MNI coordinate center *x*, *y*, z: 2, 54, 22), LTPJ; −48, −56, 22, PC; 2, −56, 36, RSTS; 58, −10, −14 and RTPJ; 54, −52, 22; see Figure 1A. These regions have been demonstrated by meta-analysis to be most consistently recruited by the FB task (Schurz *et al.*, 2014). Restricting our analyses to these five regions specifically allowed us to test our hypotheses in areas defined a priori as being selective for mental state information and reduced the number of tests we performed, limiting the possibility of Type I error.

Univariate analysis. FB task data were first analyzed at the individual-subject level in the whole brain using a general linear model (GLM), which included a term for condition convolved with the standard hemodynamic response function, and nuisance regressors for the movement parameters and movement and signal outlier timepoints identified by ART. Data were high-pass filtered at 128 s. Individual subject contrasts were generated for each condition versus baseline and FB > FPR.

Data were submitted to ROI and whole-brain analysis. Findings from the SSD dataset were reported in Dodell-Feder, Tully, *et al.* (2014). We note that a different set of ROIs were used in that study. To make these prior findings more comparable with the multivariate findings reported in the current study, we re-ran the univariate ROI analysis with the same ROIs used in the current study and report these findings in the supplementary materials (no differences were observed between the ROI analysis performed in the original study and the current study). For the FHR dataset, we conducted ROI analysis using the Dufour *et al.* (2013) ROIs, and a performed a second-level random-effects whole-brain analysis comparing FHR to non-FHR with a two-sample *t*-test. These data are reported in the supplementary materials. For both datasets, ROI analysis was conducted by extracting the beta values for FB > baseline and FPR > baseline contrasts and submitting these values to repeated-measures analysis of variance (ANOVAs) that included terms for group, condition and their interaction. Follow-up tests to evaluate between group differences in condition were conducted with Welch's *t*-tests (Delacre *et al.*, 2017). These tests and follow-up tests on extracted univariate and multivariate (see below) ROI values were performed in R Statistical Software (R Core Team, 2018).

Multivariate analysis. MVPA was conducted in MATLAB using The Decoding Toolbox (Hebart *et al.*, 2015). Our primary aim

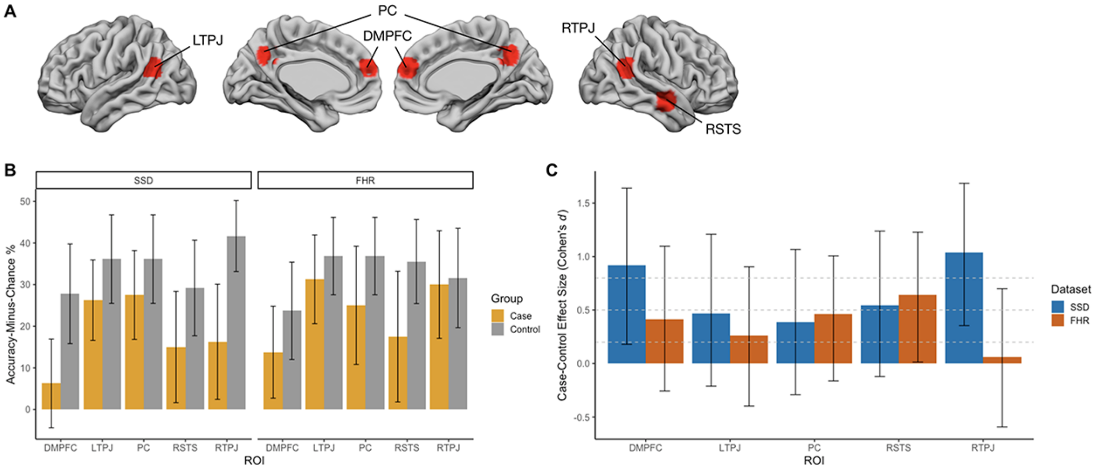

Fig. 1. Multivariate ROI analysis and accompanying effect sizes. (A) Depiction of ROIs. (B) Accuracy-minus-chance percentage (chance = 50%) for the SSD dataset (left panel) and FHR dataset (right panel). Case group (SSD, FHR) depicted in orange and control group (non-SSD, non-FHR) in gray. Error bars depict 95% confidence intervals. (C) Cohen's *d* effect sizes for case minus control with SSD–non-SSD in blue and FHR–non-FHR in red. Error bars depict BCa 95% confidence intervals derived from 10 000 bootstrap samples. Horizontal dashed gray lines represent effect size benchmarks corresponding to small, medium and large effects.

was to characterize group differences in classification accuracy in regions of the brain selective for mental state information. Towards that goal, for each participant, we submitted the beta images for FB and FPR generated from the first-level GLMs described above to a leave-one-out cross-validation scheme using a linear support vector machine as a classifier. This generated a single accuracy-minus-chance (chance = 50%) value per ROI per participant. These values were compared against zero within each group using a one-sample *t*-test, and then between groups using a two-sample Welch's *t*-test. We report false-discovery rate (FDR) adjusted *P*-values (i.e. *q*-values) adjusting for five ROI tests conducted within groups and between groups. Effect sizes were calculated as Cohen's *d* along with bias-corrected-and-accelerated (BCa) 95% confidence intervals (CIs) generated from 10 000 bootstrap samples with the package *bootES* (Kirby and Gerlanc, 2013). We interpreted these effect sizes using conventional benchmarks (Cohen, 1988). To better understand the nature of group differences, we followed-up significant between group differences by evaluating the within-condition pattern correlations. Following Haxby (2001), we did this by splitting the data in half for each condition, calculating the beta value for each voxel within the ROI for each condition and, then, evaluating the correlation between betas in each voxel of the ROI for each condition. This analysis generated four values for each ROI—the correlation between voxels for FB in SSDs; the correlation between voxels for FB in non-SSDs; the correlation between voxels for FPR in SSDs; the correlation between voxels for FPR in non-SSDs—which were transformed using Fisher's *r*-to-*z* transformation, and then compared between groups using Welch's *t*-tests. Given that we used a different headcoil and acquisition parameters for the SSD and FHR dataset, we did not perform direct statistical comparisons between the SSD and FHR datasets for any analysis.

To investigate whether there were differences in classification accuracy in regions outside of the ToM network, we performed an exploratory whole-brain searchlight analysis using searchlights with a 4-voxel radius around the center voxel. Searchlights were passed through the whole-brain on a voxel-by-voxel basis, and classification was performed within each searchlight with the classification value (accuracy-minus-chance) being assigned to the center voxel. This created whole-brain classification maps for each participant representing the local information content around the center of each searchlight. These maps were analyzed at the group level by conducting one-sample *t*-tests within each group, and two-sample *t*-tests to compare local classification accuracy between groups. All images were thresholded at a voxel-wise $P < 0.001$ and a cluster-wise family-wise error (FWE)-corrected $P < 0.05$. Data were visualized with Surf Ice (https://www.nitrc.org/projects/surfice/).

Brain, behavior and symptom associations. To assess the behavioral and clinical impact of the neural measures, we evaluated the associations between univariate activity (using the FB-FPR contrast estimate for univariate activity), multivariate pattern information, FB task accuracy and symptoms. In order to reduce the number of tests and limit Type I error, we did this only in the dataset and ROIs in which we found group differences in either multivariate or univariate neural outcomes. All analyses were conducted using Pearson *r* correlations and were accompanied by BCa 95% CIs generated from 10 000 bootstrap samples. We consider a finding to be unexpected under the null hypothesis when $q < 0.05$. Given that we find group differences in two ROIs, for task performance correlations, we corrected for four tests (two ROIs × two conditions [FB, FPR]); for symptoms, we corrected for six tests (two ROIs × three symptom categories [positive, negative and disorganized]). We evaluated the association between the neural measures and task accuracy across all participants given that we did not expect the relation between brain and task performance to differ as a function of diagnostic status (e.g. Hawco *et al.*, 2019). For any association that was found to be unexpected under the null hypothesis, we evaluated whether the brain–task accuracy association was moderated by group by regressing task accuracy on the interaction of group and brain. For brain–symptom associations, we conducted these only within the clinical group because the PANSS was not administered to non-SSD participants. We conducted two follow-up analyses on associations that survived FDR-correction. First, we evaluated whether the brain–behavior association was specific to that behavioral variable (i.e. whether there was a difference brain–FB accuracy versus brain–FPR accuracy, and brain–positive symptom versus brain–negative symptom versus brain–disorganized symptom associations) by evaluating the 95% CI of the difference between the correlations using the method described in Zou (2007). Second, we evaluated the relative variance explained in the behavioral outcome by univariate activity versus multivariate pattern information with multiple linear regression.

Results

Univariate results

Results of all univariate analyses are reported in the Supplementary Materials, and the whole-brain analysis of the SSD dataset is reported in Dodell-Feder, Tully, *et al.* (2014). Briefly, ROI analysis revealed a group by condition interaction in DMPFC characterized by reduced neural activity for FB and FPR stories in SSD versus non-SSD (Supplementary Table S1, Supplementary Figure S1). Whole-brain analysis similarly revealed reduced neural activity for FB versus FPR in MPFC (Dodell-Feder, Tully, *et al.*, 2014). In contrast, ROI and whole-brain analysis revealed no differences in neural activity for FB versus FPR between FHR and non-FHR (Supplementary Table S2 and Supplementary Figure S2).

Multivariate results

Our main question concerned how mental state information was represented within ToM-related regions across the SSD and SSD-risk groups. First, we evaluated whether the ROIs distinguished between mental state and non-mental state information within each group by evaluating classification accuracy. Non-SSD participants showed above chance classification accuracy in all ROIs (Table 3, Figure 1B). SSD participants similarly showed above chance classification in all ROIs except for DMPFC. Comparing the classification accuracies between groups, the non-SSD group showed higher accuracy across all ROIs, with effect sizes ranging from small in PC to large in RTPJ (Figure 1C). The between-group difference in classification accuracy was unexpected under the null hypothesis in DMPFC and RTPJ. Given the sensitivity of MVPA analyses to movement, we evaluated whether group differences in movement might have been driving the differences in pattern discriminability. Neither translation nor rotation differed between the groups, rotation: $t(30) = 1.16$, $P = 0.255$, $d = 0.35$, 95% CI [−0.31, 1.02] and translation: $t(25) = 1.06$, $P = 0.297$, $d = 0.26$, 95% CI [−0.40, 0.92]. Further, mean translation and rotation were not correlated with pattern discriminability in either ROI, DMPFC and translation $r(36) = 0.06$, 95% CI [−0.27, 0.37], $P = 0.735$, DMPFC and rotation, $r(36) = 0.11$, 95% CI [−0.22, 0.42], $P = 0.509$, RTPJ and translation $r(36) = 0.04$, 95% CI [−0.29, 0.35], $P = 0.820$, RTPJ and rotation $r(36) = -0.21$, 95% CI [−0.50, 0.11], $P = 0.197$. Another possibility is that multivariate differences are being driven largely by differences in univariate activity. To address this possibility, we re-evaluated group differences with analysis of covariances (ANCOVAs), controlling for univariate activation. The group difference in DMPFC pattern discriminability was reduced to a trend level of significance, $F(1, 35) = 3.33$, $P = 0.077$, $\eta^2 = 0.08$, although the effect size based on the marginal means was medium in size, $d = 0.64$, with a 95% CI, [−0.04, 1.31], largely overlapping with that of the non-adjusted model, [0.18, 1.66]. When controlling for univariate activity, the impact of group on pattern discriminability in RTPJ remained statistically significant, $F(1, 35) = 7.54$, $P = 0.009$, $\eta^2 = 0.12$, with a large effect size, $d = 0.90$, 95% CI [0.21, 1.60], similar in magnitude to the non-adjusted model, $d = 1.04$, 95% CI [0.37, 1.69]. This suggests that univariate differences may be contributing to multivariate patterns differences in DMPFC, but not in RTPJ.

To better understand the source of the group difference in pattern discriminability, we evaluated group differences in within-condition pattern correlations. In DMPFC, pattern correlations values were similar and did not differ between groups for FB or FPR (Table 4). In RTPJ, pattern correlations values were similar and did not differ between groups for FB; however, we did observe a medium-sized difference in pattern correlations for FPR, $t(34) = 2.15$, $P = 0.039$, $d = -0.70$, 95% CI [−1.37, −0.003] such that SSD exhibited higher pattern correlations than non-SSD (Table 4). Looking within each group, pattern correlations did not differ between FB and FPR in SSD, $t(19) = 1.60$, $P = 0.125$, $d_z = 0.36$, 95% CI [−0.13, 0.85], as they did in non-SSD, $t(17) = 2.50$, $P = 0.023$, $d_z = 0.59$, 95% CI [0.04, 1.08], suggesting that SSD participants may be treating physical information like mental state information. One explanation for these findings is that group differences, or a lack thereof, may be attributable to noisier, less consistent patterns in SSD for either or both conditions. We tested for this possibility by examining homogeneity of variance in pattern correlations for each condition using Levene's test; however, no group differences emerged, $F\text{s} \leq 1.29$, $p\text{s} \geq 0.264$.

The non-FHR and FHR groups showed classification accuracies above chance in all ROIs (Table 5, Figure 1B). Although accuracy was higher in the non-FHR versus FHR group across all ROIs, with effect sizes ranging from trivially small in RTPJ to medium in RSTS (Figure 1C), none of these differences were unexpected under the null hypothesis.

Next, we used whole-brain exploratory searchlight analysis to address whether there were regions outside of the a priori ROIs that differed in classification accuracy as a function of group. In line with the ROI analysis, both the non-SSD and SSD group showed above chance classification in the ToM network, with the SSD group showing a smaller area of MPFC, located in the ventral aspect, that decoded condition (Supplementary Table S3, Figure 2A). A direct comparison of the groups revealed that compared to the SSD group, the non-SSD group showed greater classification accuracy in RTPJ as well as a region in anterior middle temporal gyrus (Table 6). There were no SSD > non-SSD classification differences.

Both the non-FHR and FHR group also showed above chance classification accuracy in the ToM network, with the FHR group showing a smaller area of MPFC that decoded condition (Supplementary Table S4, Figure 2B). Compared to FHR, non-FHR showed higher classification accuracy in a cluster spanning superior to middle frontal gyrus, and a cluster located primarily in left cerebellum that extended into fusiform gyrus (Table 6). There were no FHR > non-FHR classification differences.

Brain, behavior and symptom associations

To better understand the behavioral and clinical significance of the differences in DMPFC and RTPJ in the SSD dataset, we evaluated the associations between univariate activity (i.e. the FB-FPR contrast estimate), multivariate pattern information, task performance (Table 2) and symptoms. On task performance, greater pattern discriminability was associated with better performance on the FB condition, $r(31) = 0.53$, 95% CI [0.17, 0.77], $q = 0.007$, and this was not moderated by group,

Table 2. False belief task performance

	SSD dataset			FHR dataset		
	SSD	Non-SSD	Group difference	FHR	Non-FHR	Group difference
Accuracy, %						
FB	74.8 (17.3)	80.6 (16.0)	$t(31) = 0.99$, $P = 0.329$, $d^{a} = 0.34$ [−0.38, 1.05]	88.5 (11.0)	90.1 (10.7)	$t(35) = 0.47$, $P = 0.642$, $d = 0.15$ [−0.50, 0.82]
FPR	79.9 (15.3)	82.2 (14.3)	$t(31) = 0.45$, $P = 0.656$, $d = 0.16$ [−0.57, 0.83]	86.7 (13.1)	90 (12.8)	$t(35) = 0.78$, $P = 0.442$, $d = 0.26$ [−0.45, 0.93]
Reaction time, s						
FB	4.1 (0.6)	3.4 (0.6)	$t(31) = 3.35$, $P = 0.002$, $d = 1.17$ [0.35, 1.98]	3.3 (0.6)	2.9 (0.5)	$t(34) = 2.57$, $P = 0.015$, $d = 0.84$ [0.19, 1.45]
FPR	3.7 (0.5)	3.5 (0.5)	$t(31) = 1.31$, $P = 0.201$, $d = 0.45$ [=.26, 1.16]	3.1 (0.5)	3.0 (0.5)	$t(34) = 0.91$, $P = 0.368$, $d = 0.30$ [−0.41, 0.98]

[a] Cohen's d values with 95% bias-corrected-and-accelerated confidence intervals (CIs) derived from 10 000 bootstrap samples.

Table 3. ROI classification accuracy: SSD dataset

	SSD		Non-SSD		Group difference	
	M [95% CI] accuracy-above-chance %	One-sample t-test	M [95% CI] accuracy-above-chance %	One-sample t-test	Cohen's d [95% CI]	Two-sample t-test
DMPFC	6.3 [−4.4, 16.9]	$t(19) = 1.23$, $q = 0.234$	27.8 [15.8, 39.8]	$t(17) = 4.89$, $q < 0.001$	0.92 [0.18, 1.66]	$t(35) = 2.82$, $q = 0.019$
LTPJ	26.3 [16.6, 35.9]	$t(19) = 5.69$, $q < 0.001$	36.1 [25.5, 46.7]	$t(17) = 7.16$, $q < 0.001$	0.47 [−0.21, 1.19]	$t(35) = 1.44$, $q = 0.197$
PC	27.5 [16.8, 38.2]	$t(19) = 5.40$, $q < 0.001$	36.1 [25.5, 46.7]	$t(17) = 7.16$, $q < 0.001$	0.39 [−0.28, 1.07]	$t(36) = 1.20$, $q = 0.238$
RSTS	15.0 [1.6, 28.4]	$t(19) = 2.35$, $q = 0.037$	29.2 [17.7, 40.7]	$t(17) = 5.36$, $q < 0.001$	0.54 [−0.12, 1.25]	$t(36) = 1.69$, $q = 0.167$
RTPJ	16.5 [2.4, 30.1]	$t(19) = 2.46$, $q = 0.037$	41.7 [33.1, 50.2]	$t(17) = 10.31$, $q < 0.001$	1.04 [0.37, 1.69]	$t(31) = 3.28$, $q = 0.013$

Table 4. ROI pattern correlations

		M (SD) Fisher r-to-z value		
ROI	Condition	SSD	Non-SSD	Group difference
DMPFC				
	FB	0.98 (0.54)	0.97 (0.52)	$t(36) = 0.05$, $P = 0.961$, $d = -0.02$, 95% CI [−0.66, 0.66]
	FPR	1.10 (0.57)	1.03 (0.63)	$t(35) = 0.37$, $P = 0.716$, $d = -0.12$, 95% CI [−0.80, 0.56]
RTPJ				
	FB	1.25 (0.45)	1.13 (0.69)	$t(29) = 0.63$, $P = 0.533$, $d = -0.21$, 95% CI [−0.83, 0.47]
	FPR	1.04 (0.54)	0.63 (0.63)	$t(34) = 2.15$, $P = 0.039$, $d = -0.70$, 95% CI [−1.37, −0.003]

Table 5. ROI classification accuracy: FHR dataset

	FHR		Non-FHR		Group difference	
	M [95% CI] accuracy-above-chance %	One-sample t-test	M [95% CI] accuracy-above-chance %	One-sample t-test	Cohen's d [95% CI]	Two-sample t-test
DMPFC	13.8 [2.7, 24.8]	$t(19) = 1.23$, $q = 0.022$	23.7 [12.0, 35.4]	$t(18) = 4.26$, $q < 0.001$	0.42 [−0.26, 1.10]	$t(37) = 1.30$, $q = 0.339$
LTPJ	31.3 [20.6, 41.9]	$t(19) = 6.14$, $q < 0.001$	36.8 [27.5, 46.1]	$t(18) = 8.32$, $q < 0.001$	0.26 [−0.39, 0.88]	$t(37) = 0.83$, $q = 0.516$
PC	25.0 [10.8, 39.2]	$t(19) = 3.68$, $q = 0.003$	36.8 [27.5, 46.1]	$t(18) = 8.32$, $q < 0.001$	0.46 [−0.15, 1.03]	$t(32) = 1.46$, $q = 0.339$
RSTS	17.5 [1.8, 33.2]	$t(19) = 2.33$, $q = 0.031$	35.5 [25.4, 45.6]	$t(18) = 7.39$, $q < 0.001$	0.64 [0.01, 1.24]	$t(32) = 2.02$, $q = 0.257$
RTPJ	30.0 [17.1, 42.9]	$t(19) = 4.86$, $q < 0.001$	31.6 [19.6, 43.5]	$t(18) = 5.55$, $q < 0.001$	0.60 [−0.60, 0.70]	$t(37) = 0.19$, $q = 0.852$

$\beta = -0.56$, 95% CI [−1.50, 0.37], $P = 0.229$. Task accuracy, either on the FB or FPR condition, was not associated with pattern discriminability in DMPFC, $r(31) \leq 0.07$, and although we observed positive associations between FB accuracy and univariate activity in DMPFC, $r(31) = 0.37$ and RTPJ, $r(31) = 0.41$, at an uncorrected level ($P < 0.05$), neither association survived FDR correction. The association between RTPJ pattern discriminability and FB accuracy was larger in magnitude than the association between RTPJ pattern discriminability and FPR accuracy, $r(31) = 0.18$, 95% CI [−0.17, 0.50], $q = 0.547$, 95% CI of the correlation difference [0.03, 0.65], meaning that increased pattern discriminability in RTPJ might specifically support FB reasoning as opposed to reasoning about representations more generally. Further, the variance in FB accuracy accounted for by pattern discriminability, $\beta = 0.41$, 95% CI [0.06, 0.76], $P = 0.023$, was above and beyond that accounted for by univariate activity, which was not associated with FB accuracy when taking multivariate pattern discriminability into account, $\beta = 0.19$, 95% CI [−0.17, 0.54], $P = 0.295$.

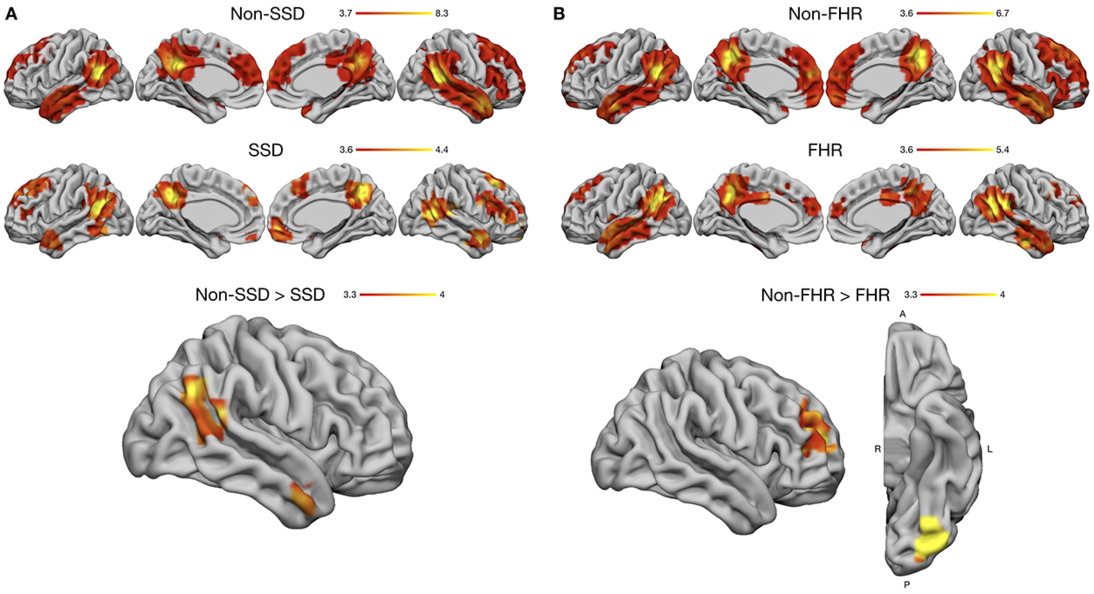

Fig. 2. Whole-brain exploratory searchlight analysis. SSD dataset (A) and FHR dataset (B). The top panels depict the Control group (non-SSD, non-FHR), the middle panel depicts the Case group (SSD, FHR), and the bottom panel depicts the control > case comparison. No differences were observed for case > control in either dataset. All images are thresholded at a voxel-wise $P < 0.001$ and a cluster-wise FWE-corrected $P < 0.05$. Color bars depict *t* values. R = right, L = left, A = anterior, P = posterior.

Table 6. Whole-brain searchlight analysis group differences

Dataset	Contrast	Region	MNI coordinates *x, y, x*	Cluster extent (voxels)	Cluster *P* (FWE-corrected)	Peak *t*-value ($P < 0.001$)
SSD	Non-SSD > SSD	RTPJ	66, −50, 22	165	<0.001	5.13
		Right anterior middle temporal gyrus	60, 3, −34	52	0.032	4.71
	SSD > Non-SSD	*No suprathreshold clusters*				
FHR	Non-FHR > FHR	Right superior/middle frontal gyrus	26, 46, 16	270	<0.001	5.44
		Left cerebellum/left fusiform gyrus	−34, −80, −20	281	<0.001	4.51
	FHR > Non-FHR	*No suprathreshold clusters*				

Images were thresholded at a voxel-wise $P < 0.001$ and a cluster-wise FWE-corrected $P < 0.05$.

On symptoms, we observed negative associations between RTPJ pattern discriminability and positive, $r(18) = -0.52$, and disorganized symptoms, $r(18) = -0.49$, at an uncorrected level ($P < 0.05$), but neither association survived FDR-correction. We also observed a positive association between DMPFC univariate activity and negative symptoms, $r(18) = 0.45$, at an uncorrected level, which too did not survive FDR-correction. All other associations were not unexpected under the null hypothesis.

Discussion

The majority of task-based fMRI research on ToM in SSDs and FHR have examined a single question: whether regions in the ToM network show aberrant of levels of involvement during mental state reasoning (i.e. hypo- or hyper-activation). Using univariate analysis, these studies have consistently shown that levels of activation between SSDs, FHR and control group are different (Kronbichler *et al.*, 2017; Kozhuharova *et al.*, 2020). However, there are other questions to be asked regarding how latent liability for an SSD and manifest illness impacts the functional properties of the ToM network. Answers to these other questions may shed more light on the neurobiological processes at work in the development of an SSD and the social difficulties it brings. MVPA, which characterizes regional representational content, may be useful in this regard as it has in other studies of neural function in SSDs (Yoon *et al.*, 2008).

Here, using MVPA, we assessed ToM-related activation patterns in a group of SSD and FHR participants who performed the same well-validated ToM task while undergoing fMRI. MVPA of ToM ROIs showed that compared to non-SSD participants, SSDs showed reduced classification accuracy in two regions of the ToM network thought to constitute a core mental state understanding network (Schurz *et al.*, 2014; Molenberghs *et al.*, 2016): DMPFC and RTPJ. In other words, the pattern of information in DMPFC and RTPJ for the category of mental states was less discriminable in SSD than it is in non-SSD participants, suggesting that in SSDs, mental information is not privileged in brain regions that are typically highly specialized for representing such information. Further, this difference could not clearly be explained by differences in univariate activation (although it reduced the multivariate difference in DMPFC to a trend level of significance) or because the SSD data were noisier (e.g. due to in-scanner motion). In RTPJ, the analysis of pattern correlations revealed a more highly stable, consistent response to FPR in SSDs versus non-SSDs that was equal in magnitude to the FB response. In consideration of this finding, one possibility is that the source of reduced pattern discriminability in SSDs is due to SSDs representing non-mental information in the same way they represent mental state information. If RTPJ treats physical information like mental information, this may help to explain reports of increased mind perception (Gray *et al.*, 2011; Raffard *et al.*, 2016), and 'hyper-ToM' (i.e. inappropriately ascribing mental states to others) in psychosis and certain psychotic-like experiences (i.e. paranoia, delusional ideation; Russell *et al.*, 2006; Fyfe *et al.*, 2008; Montag *et al.*, 2011; Clemmensen *et al.*, 2014). The fact that we observed more differences in multivariate pattern information versus univariate activity—in which we saw reduced activity for FB in SSDs vs non-SSDs—may reflect the fact that multivariate approaches afford better sensitivity at detecting differences as shown in other work (Kriegeskorte *et al.*, 2006; Raizada *et al.*, 2010).

Exploratory whole-brain searchlight analysis were partially consistent with these ROI findings revealing reduced decoding accuracy in RTPJ and RSTS in SSDs versus non-SSDs. These findings somewhat parallel those from studies of autism spectrum disorder, which is similarly characterized by marked ToM and social functioning deficits (Pinkham *et al.*, 2019). Specifically, studies of ASD have shown altered patterns of neural activity in RTPJ and MPFC during social cognitive tasks (Gilbert *et al.*, 2009; Koster-Hale *et al.*, 2013; Richardson *et al.*, 2020), suggesting that altered representation of mental state information may be a transdiagnostic marker of social dysfunction.

It remains an open question as to why in SSDs, DMPFC and RTPJ pattern discriminability would be reduced, and, as suggested by the pattern correlation analysis, why in RTPJ, physical information would be represented in a similar manner as mental information. One possibility is that early social skills deficits, social anhedonia and social withdrawal—characteristics that describe individuals who later develop SSDs (Kwapil, 1998; Tarbox and Pogue-Geile, 2008; Radua *et al.*, 2018)—reduce the quantity and quality of early social exposure in a way that prevents the specialization of ToM-related brain regions that occurs in typically developing youth (Saxe *et al.*, 2009; Gweon *et al.*, 2012; Bowman *et al.*, 2019). In partial support of this idea, pre-SSD individuals show progressive cortical thickness reductions in brain regions implicated in ToM, specifically MPFC and posterior temporal cortex (Cannon *et al.*, 2015), which may impact the functional specialization of these regions. This may also speak to why we saw a difference in DMPFC and RTPJ, and not other regions of the network; that is, in line with these other findings, the pathophysiology of SSDs may specifically affect these regions and the functional networks that they, in part, comprise, such as the default mode network which too is disrupted in at-risk groups (Dodell-Feder, DeLisi, *et al.*, 2014b; Karcher *et al.*, 2019).

In contrast to SSDs, we found no evidence of altered representational content for mental state information in FHR. Specifically, the activation patterns for the category of mental states in the a priori ROIs were equally discriminable in FHR as they were in non-FHR. The exploratory whole-brain searchlight analyses demonstrated reduced decoding accuracy in FHR compared to non-FHR in right superior to middle frontal gyrus and left cerebellum extending into fusiform gyrus. It is unclear what to make of these findings given that the role of these regions during mental state understanding is unknown. We note that we did not observe differences in the univariate analyses either, which is in contrast to prior work showing altered involvement of ToM brain regions (Kozhuharova *et al.*, 2020). This might reflect the intact involvement of these regions in the context of FB reasoning, but not other, more complex social scenarios (de Achával *et al.*, 2013; Dodell-Feder, DeLisi, *et al.*, 2014a; Mohnke *et al.*, 2016). There are several reasons as to why the MVPA findings in FHR diverge from those in SSDs. First, disrupted neural representation of mental state information may only occur in pre-SSDs (i.e. prodromal) and manifest illness, not simply in those at elevated risk due to a constitutional or acquired vulnerability factor. Second, there may exist differences in pattern discriminability in FHR that are simply smaller than those observed in SSDs, and we were underpowered to detect them. Third, FHR differences might have been obscured by a methodological difference between the FHR and SSD studies. Because of the small number of FHR participants and the methodological differences between the FHR and SSD studies, the between-sample differences should be interpreted with caution and replicated.

An important issue concerns the behavioral consequences of disrupted pattern information and/or activity, that is, to what extent do disruptions to representation or activation account for the social cognitive deficits and symptoms observed in the schizophrenia spectrum? Toward addressing this question, we evaluated the associations between univariate activity, multivariate pattern discriminability, FB task performance and symptoms. On task performance, we found that only RTPJ pattern discriminability in the SSD dataset was associated with FB accuracy and not task accuracy more generally. Further, pattern discriminability explained the variance in FB accuracy above and beyond that explained by RTPJ univariate activity, suggesting particular importance of representational information for FB understanding. This finding is consistent with work demonstrating that pattern discriminability in RTPJ for intentional *vs* unintentional acts is associated with the extent to which mental states are weighted when making moral judgments (Koster-Hale *et al.*, 2013) and with work in SSDs showing that pattern discriminability in regions recruited during visual object processing is associated with task performance (Yoon *et al.*, 2008). In contrast to task performance, the neural measures were not associated with symptoms at a corrected level.

Prior work on the representational content of ToM brain regions has shown that activity patterns in RTPJ and MPFC contain granular mental state information well beyond what was tested here, including the social impact of a mental state (i.e. the degree to which a mental state influences social relationships) (Tamir *et al.*, 2016; Thornton and Tamir, 2020), the epistemic context of a mental state (i.e. how a belief was formed and the justification for the belief) (Koster-Hale *et al.*, 2017), as well as affective states, their valence and the context in which an emotion occurs (Skerry and Saxe, 2015; Tamir *et al.*, 2016; Koster-Hale *et al.*, 2017; Thornton and Tamir, 2020). It has been suggested that representing these dimensions may facilitate social predictions (Tamir and Thornton, 2018; Thornton and Tamir, 2020), an idea supported by other works demonstrating that neural response in ToM brain regions can be characterized within a predictive coding framework (Carter *et al.*, 2012; Koster-Hale and Saxe, 2013; Tamir and Thornton, 2018; Thornton *et al.*, 2019; Park *et al.*, 2020; Richardson and Saxe, 2020). This raises the intriguing possibility that altered representation of mental state information in SSD contributes to difficulty in making social predictions (e.g. Sterzer *et al.*, 2018), which in turn contributes to social dysfunction. Because we did not test the integrity of information within the ROIs for these other dimensions nor their contribution to social prediction, these ideas are speculative. However, they would be worth addressing in future research.

Several limitations are notable. First, given the small sample sizes, we were adequately powered to detect large differences between groups ($d = 0.92$ with power $= 0.80$) and large associations between the brain and behavior ($r = 0.44$ for brain–task accuracy associations and $r = 0.61$ for brain–symptom associations with power $= 0.80$). This leaves open the strong possibility that there may be smaller yet clinically meaningful differences between groups or associations between brain and behavior that went undetected here, particularly in the FHR group. Second, we were unable to make direct statistical comparisons between SSD and FHR due to different acquisition parameters. Third, given that these data are cross-sectional, they can only suggest but not directly speak to how the ToM network changes from states of risk to manifest illness. Last, we tested how information is represented for only one specific context of mental state reasoning, without examining other dimensions within the category of mental states.

Conclusion

We find that in SSDs, core ToM brain regions demonstrate altered involvement and patterns of neural activity, including reduced discriminability between FB and FP, suggesting that the category of mental states is not represented with sufficient distinguishing details or characteristics. Thus, this may be partially driven by SSD representing physical information as they do mental information. The extent of altered pattern information in SSD carries functional implications as well as it was shown here to impact the performance on the FB task. In contrast, the ToM network in FHR can be characterized by altered involvement but preserved the informational content for the category of mental states. These data suggest that unlike aberrant involvement of these brain regions in ToM, which occurs in SSD risk states, the representation of mental states may be disrupted by the onset of illness and not before. This notion should be more directly evaluated in future work using larger samples with longitudinal paradigms.

Acknowledgements

The authors would like to thank Sarah Hope Lincoln, Beverly Pozuelos, Chinmanyi Tengshe, Nadia Lyyanage-Don, Todd Thompson, Emily Carol, Caitlin Carey, Gul Jabbar, Ashley Proal, Cheryl Best, Matthew Yung, Hong Yin, K. Juston Osborne, Erin Guty, Dost Ongur and Danielle Pfaff for assistance with subject recruitment and data collection. We would also like to thank the individuals who participated in these studies.

Funding

These studies were supported by Harvard University Research Funds (CIH), The William F Milton Fund (CIH), Sackler Scholar Programme in Psychobiology Fellowships (DDF, LMT), a training grant from the NIH Blueprint for Neuroscience Research (T90DA022759, DDF) and, indirectly, by a grant from the National Institute of Mental Health (1L30MH117569-01, DDF).

Conflict of interest

The authors declare no conflict of interest.

Supplementary data

Supplementary data are available at *SCAN* online.

References

Bora, E., Yucel, M., Pantelis, C. (2009). Theory of mind impairment in schizophrenia: meta-analysis. *Schizophrenia Research*, **109**, 1–9.

Bora, E., Pantelis, C. (2013). Theory of mind impairments in first-episode psychosis, individuals at ultra-high risk for psychosis and in first-degree relatives of schizophrenia: systematic review and meta-analysis. *Schizophrenia Research*, **144**, 31–6.

Bowman, L.C., Dodell-Feder, D., Saxe, R., *et al.* (2019). Continuity in the neural system supporting children's theory of mind development: longitudinal links between task-independent EEG and task-dependent fMRI. *Developmental Cognitive Neuroscience*, **40**, 100705.

Brainard, D.H. (1997). The psychophysics toolbox. *Spatial Vision*, **10**, 433–6.

Cannon, T.D., Chung, Y., He, G., *et al.* (2015). Progressive reduction in cortical thickness as psychosis develops: a multisite longitudinal neuroimaging study of youth at elevated clinical risk. *Biological Psychiatry*, **77**, 147–57.

Carter, R.M., Bowling, D.L., Reeck, C., *et al.* (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science*, **337**, 109–11.

Chang, C.-J., Chen, W.J., Liu, S.K., *et al.* (2002). Morbidity risk of psychiatric disorders among the first degree relatives of schizophrenia patients in Taiwan. *Schizophrenia Bulletin*, **28**, 379–92.

Clemmensen, L., van Os, J., Skovgaard, A.M., *et al.* (2014). Hyper-theory-of-mind in children with psychotic experiences. *PLoS One*, **9**, e113082.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, NJ: Erlbaum.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Coutanche, M.N., Thompson-Schill, S.L., Schultz, R.T. (2011). Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. *NeuroImage*, **57**, 113–23.

Couture, S.M., Penn, D.L., Roberts, D.L. (2006). The functional significance of social cognition in schizophrenia: a review. *Schizophrenia Bulletin*, **32**, S44–63.

Das, P., Lagopoulos, J., Coulston, C.M., *et al.* (2012). Mentalizing impairment in schizophrenia: a functional MRI study. *Schizophrenia Research*, **134**, 158–64.

de Achával, D., Villarreal, M.F., Costanzo, E.Y., *et al.* (2012). Decreased activity in right-hemisphere structures involved in social cognition in siblings discordant for schizophrenia. *Schizophrenia Research*, **134**, 171–9.

de Achával, D., Villarreal, M.F., Salles, A., *et al.* (2013). Activation of brain areas concerned with social cognition during moral decisions is abnormal in schizophrenia patients and unaffected siblings. *Journal of Psychiatric Research*, **47**, 774–82.

Dean, K., Stevens, H., Mortensen, P.B., *et al.* (2010). Full spectrum of psychiatric outcomes among offspring with parental history of mental disorder. *Archives of General Psychiatry*, **67**, 822.

Delacre, M., Lakens, D., Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of student's *t*-test. *International Review of Social Psychology*, **30**, 92.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., *et al.* (2011). fMRI item analysis in a theory of mind task. *NeuroImage*, **55**, 705–12.

Dodell-Feder, D., Tully, L.M., Lincoln, S.H., *et al.* (2014). The neural basis of theory of mind and its relationship to social functioning and social anhedonia in individuals with schizophrenia. *NeuroImage: Clinical*, **4**, 154–63.

Dodell-Feder, D., DeLisi, L.E., Hooker, C.I. (2014a). Neural disruption to theory of mind predicts daily social functioning in individuals at familial high-risk for schizophrenia. *Social Cognitive and Affective Neuroscience*, **9**, 1914–25.

Dodell-Feder, D., DeLisi, L.E., Hooker, C.I. (2014b). The relationship between default mode network connectivity and social functioning in individuals at familial high-risk for schizophrenia. *Schizophrenia Research*, **156**, 87–95.

Dufour, N., Redcay, E., Young, L., *et al.* (2013). Similar brain activation during false belief tasks in a large sample of adults with and without autism. *PLoS One*, **8**, e75468.

Erlenmeyer-Kimling, L. (1997). The New York High-Risk Project: prevalence and comorbidity of axis I disorders in offspring of schizophrenic parents at 25-year follow-up. *Archives of General Psychiatry*, **54**, 1096.

Faridi, K., Pawliuk, N., King, S., *et al.* (2009). Prevalence of psychotic and non-psychotic disorders in relatives of patients with a first episode psychosis. *Schizophrenia Research*, **114**, 57–63.

Fett, A.-K.J., Viechtbauer, W., Dominguez, M.-G., *et al.* (2011). The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: a meta-analysis. *Neuroscience and Biobehavioral Reviews*, **35**, 573–88.

First, M.B., Spitzer, R.L., Gibbon, M., *et al.* (2002). *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version*. New York, NY: Biometrics Research, New York State Psychiatric Institute.

Fyfe, S., Williams, C., Mason, O., *et al.* (2008). Apophenia, theory of mind and schizotypy: perceiving meaning and intentionality in randomness. *Cortex*, **44**, 1316–25.

Gilbert, S.J., Meuwese, J.D.I., Towgood, K.J., *et al.* (2009). Abnormal functional specialization within medial prefrontal cortex in high-functioning autism: a multi-voxel similarity analysis. *Brain*, **132**, 869–78.

Gottesman, I.I. (1991). *Schizophrenia Genesis: The Origins of Madness*. New York, NY: Freeman.

Gray, K., Jenkins, A.C., Heberlein, A.S., *et al.* (2011). Distortions of mind perception in psychopathology. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 477–9

Gweon, H., Dodell-Feder, D., Bedny, M., *et al.* (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts: behavioral and neural development in theory of mind. *Child Development*, **83**, 1853–68.

Hawco, C., Buchanan, R.W., Calarco, N., *et al.* (2019). Separable and replicable neural strategies during social brain function in people with and without severe mental illness. *American Journal of Psychiatry*, **176**, 521–30.

Haxby, J.V. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, **293**, 2425–30.

Haynes, J.-D., Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews: Neuroscience*, **7**, 523–34.

Hebart, M.N., Görgen, K., Haynes, J.-D. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, **8**, 88.

Hebart, M.N., Baker, C.I. (2018). Deconstructing multivariate decoding for the study of brain function. *NeuroImage*, **180**, 4–18.

Herold, R., Varga, E., Hajnal, A., *et al.* (2018). Altered neural activity during irony comprehension in unaffected first-degree relatives of schizophrenia patients—an fMRI study. *Frontiers in Psychology*, **8**, 2309.

Hoffman, R.E. (2007). A social deafferentation hypothesis for induction of active schizophrenia. *Schizophrenia Bulletin*, **33**, 1066–70.

Holt-Lunstad, J., Smith, T.B., Baker, M., *et al.* (2015). Loneliness and social isolation as risk factors for mortality: a meta-analytic review. *Perspectives on Psychological Science*, **10**, 227–37.

Holt-Lunstad, J., Robles, T.F., Sbarra, D.A. (2017). Advancing social connection as a public health priority in the United States. *American Psychologist*, **72**, 517–30.

Horan, W.P., Green, M.F., DeGroot, M., *et al.* (2012). Social cognition in schizophrenia, part 2: 12-month stability and prediction of functional outcome in first-episode patients. *Schizophrenia Bulletin*, **38**, 865–72.

House, J., Landis, K., Umberson, D. (1988). Social relationships and health. *Science*, **241**, 540–5.

Jáni, M., Kašpárek, T. (2018). Emotion recognition and theory of mind in schizophrenia: a meta-analysis of neuroimaging studies. *The World Journal of Biological Psychiatry*, **19**, S86–96.

Karcher, N.R., O'Brien, K.J., Kandala, S., *et al.* (2019). Resting-state functional connectivity and psychotic-like experiences in childhood: results from the adolescent brain cognitive development study. *Biological Psychiatry*, **86**, 7–15.

Kay, S.R., Fiszbein, A., Opler, L.A. (1987). The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, **13**, 261–76.

Kim, H.S., Shin, N.Y., Jang, J.H., *et al.* (2011). Social cognition and neurocognition as predictors of conversion to psychosis in individuals at ultra-high risk. *Schizophrenia Research*, **130**, 170–5.

Kirby, K.N., Gerlanc, D. (2013). BootES: an R package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods*, **45**, 905–27.

Kleiner, M., Brainard, D., Pelli, D. (2007). What's new in psychtoolbox-3? *Perception*, **36**, 1–16.

Koster-Hale, J., Saxe, R., Dungan, J., *et al.* (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 5648–53.

Koster-Hale, J., Richardson, H., Velez, N., *et al.* (2017). Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage*, **161**, 9–18.

Koster-Hale, J., Saxe, R. (2013). Theory of mind: a neural prediction problem. *Neuron*, **79**, 836–48.

Kozhuharova, P., Saviola, F., Ettinger, U., *et al.* (2020). Neural correlates of social cognition in populations at risk of psychosis: a systematic review. *Neuroscience and Biobehavioral Reviews*, **108**, 94–111.

Kriegeskorte, N., Goebel, R., Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 3863–8.

Kriegeskorte, N., Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, **38**, 649–62.

Kronbichler, L., Tschernegg, M., Martin, A.I., *et al.* (2017). Abnormal brain activation during theory of mind tasks in schizophrenia: a meta-analysis. *Schizophrenia Bulletin*, **43**, 1240–1250.

Kwapil, T.R. (1998). Social anhedonia as a predictor of the development of schizophrenia-spectrum disorders. *Journal of Abnormal Psychology*, **107**, 558–65.

Lavoie, M.-A., Plana, I., Bédard Lacroix, J., *et al.* (2013). Social cognition in first-degree relatives of people with schizophrenia: a meta-analysis. *Psychiatry Research*, **209**, 129–35.

Lee, J., Quintana, J., Nori, P., *et al.* (2011). Theory of mind in schizophrenia: exploring neural mechanisms of belief attribution. *Social Neuroscience*, **6**, 569–81.

Lee, J., Horan, W.P., Wynn, J.K., *et al.* (2016). Neural correlates of belief and emotion attribution in schizophrenia. *PLoS One*, **11**, e0165546.

Mar, R.A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology*, **62**, 103–34.

Marjoram, D., Job, D.E., Whalley, H.C., *et al.* (2006). A visual joke fMRI investigation into Theory of Mind and enhanced risk of schizophrenia. *NeuroImage*, **31**, 1850–8.

Maxwell, M.E. (1992). *Manual for the Family Interview for Genetic Studies (FIGS)*. Bethesda, MD: Clinical Neurogenetics Branch, National Institute of Mental Health.

Miller, T.J., McGlashan, T.H., Rosen, J.L., *et al.* (2003). Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability. *Schizophrenia Bulletin*, **29**, 703–15.

Mohnke, S., Erk, S., Schnell, K., *et al.* (2016). Theory of mind network activity is altered in subjects with familial liability for schizophrenia. *Social Cognitive and Affective Neuroscience*, **11**, 299–307.

Molenberghs, P., Johnson, H., Henry, J.D., *et al.* (2016). Understanding the minds of others: a neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews*, **65**, 276–91.

Montag, C., Dziobek, I., Richter, I.S., *et al.* (2011). Different aspects of theory of mind in paranoid schizophrenia: evidence from a video-based assessment. *Psychiatry Research*, **186**, 203–9.

Mur, M., Bandettini, P.A., Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, **4**, 101–9.

Norman, K.A., Polyn, S.M., Detre, G.J., *et al.* (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, **10**, 424–30.

Nurnberger, J.I. (1994). Diagnostic interview for genetic studies: rationale, unique features, and training. *Archives of General Psychiatry*, **51**, 849.

Op de Beeck, H.P. (2010). Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? *NeuroImage*, **49**, 1943–8.

Park, B., Fareri, D., Delgado, M., *et al.* (2020). The role of right temporo-parietal junction in processing social prediction error across relationship contexts. *Social Cognitive and Affective Neuroscience*, nsaa072.

Pinkham, A.E., Morrison, K.E., Penn, D.L., *et al.* (2019). Comprehensive comparison of social cognitive performance in autism spectrum disorder and schizophrenia. *Psychological Medicine*, **50**, 2557–2565.

Radua, J., Ramella-Cravaro, V., Ioannidis, J.P.A., *et al.* (2018). What causes psychosis? An umbrella review of risk and protective factors. *World Psychiatry*, **17**, 49–66.

Raffard, S., Bortolon, C., Khoramshahi, M., *et al.* (2016). Humanoid robots versus humans: how is emotional valence of facial expressions recognized by individuals with schizophrenia? An exploratory study. *Schizophrenia Research*, **176**, 506–13.

Raizada, R.D.S., Tsao, F.-M., Liu, H.-M., *et al.* (2010). Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: prediction of individual differences. *Cerebral Cortex*, **20**, 1–12.

Rasic, D., Hajek, T., Alda, M., *et al.* (2014). Risk of mental illness in offspring of parents with schizophrenia, bipolar disorder, and major depressive disorder: a meta-analysis of family high-risk studies. *Schizophrenia Bulletin*, **40**, 28–38.

Richardson, H., Gweon, H., Dodell-Feder, D., *et al.* (2020). Response patterns in the developing social brain are organized by social and emotion features and disrupted in children diagnosed with autism spectrum disorder. *Cortex*, **125**, 12–29.

Richardson, H., Saxe, R. (2020). Development of predictive responses in theory of mind brain regions. *Developmental Science*, **23**, e12863.

Russell, T.A., Reynaud, E., Herba, C., *et al.* (2006). Do you see what I see? Interpretations of intentional movement in schizophrenia. *Schizophrenia Research*, **81**, 101–11.

Saxe, R., Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in “theory of mind”. *NeuroImage*, **19**, 1835–42.

Saxe, R.R., Whitfield-Gabrieli, S., Scholz, J., *et al.* (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child Development*, **80**, 1197–209.

Schmidt, S.J., Mueller, D.R., Roder, V. (2011). Social cognition as a mediator variable between neurocognition and functional outcome in schizophrenia: empirical review and new results by structural equation modeling. *Schizophrenia Bulletin*, **37**, S41–54.

Schurz, M., Radua, J., Aichhorn, M., *et al.* (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, **42**, 9–34.

Selten, J.-P., Booij, J., Buwalda, B., *et al.* (2017). Biological mechanisms whereby social exclusion may contribute to the etiology of psychosis: a narrative review. *Schizophrenia Bulletin*, **43**, 287–292.

Skerry, A.E., Saxe, R. (2015). Neural representations of emotion are organized around abstract event features. *Current Biology*, **25**, 1945–54.

Snyder-Mackler, N., Burger, J.R., Gaydosh, L., *et al.* (2020). Social determinants of health and survival in humans and other animals. *Science*, **368**, eaax9553.

Sterzer, P., Adams, R.A., Fletcher, P., *et al.* (2018). The predictive coding account of psychosis. *Biological Psychiatry*, **84**, 634–43.

Tamir, D.I., Thornton, M.A., Contreras, J.M., *et al.* (2016). Neural evidence that three dimensions organize mental state representation: rationality, social impact, and valence. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 194–9

Tamir, D.I., Thornton, M.A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*, **22**, 201–12.

Tarbox, S.I., Pogue-Geile, M.F. (2008). Development of social functioning in preschizophrenia children and adolescents: a systematic review. *Psychological Bulletin*, **134**, 561–83.

Tarbox, S.I., Pogue-Geile, M.F. (2011). A multivariate perspective on schizotypy and familial association with schizophrenia: a review. *Clinical Psychology Review*, **31**, 1169–82.

Thornton, M.A., Weaverdyck, M.E., Tamir, D.I. (2019). The social brain automatically predicts others’ future mental states. *The Journal of Neuroscience*, **39**, 140–8.

Thornton, M.A., Tamir, D.I. (2020). People represent mental states in terms of rationality, social impact, and valence: validating the 3d Mind Model. *Cortex*, **125**, 44–59.

van der Gaag, M. (2006). A neuropsychiatric model of biological and psychological processes in the remission of delusions and auditory hallucinations. *Schizophrenia Bulletin*, **32**, S113–22.

Velthorst, E., Fett, A.-K.J., Reichenberg, A., *et al.* (2017). The 20-year longitudinal trajectories of social functioning in individuals with psychotic disorders. *American Journal of Psychiatry*, **174**, 1075–85.

Ventura, J., Ered, A., Gretchen-Doorly, D., *et al.* (2015). Theory of mind in the early course of schizophrenia: stability, symptom and neurocognitive correlates, and relationship with functioning. *Psychological Medicine*, **45**, 2031–43.

Villarreal, M.F., Drucaroff, L.J., Goldschmidt, M.G., *et al.* (2014). Pattern of brain activation during social cognitive tasks is related to social competence in siblings discordant for schizophrenia. *Journal of Psychiatric Research*, **56**, 120–9.

Walter, H., Ciaramidaro, A., Adenzato, M., *et al.* (2009). Dysfunction of the social brain in schizophrenia is modulated by intention type: an fMRI study. *Social Cognitive and Affective Neuroscience*, **4**, 166–76.

Wechsler, D. (2011). *WASI-II: Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: NCS Pearson.

Whitfield-Gabrieli, S., Nieto-Castanon, A., Ghosh, S. (2011). *Artifact Detection Tools (ART)* Cambridge, MA. https://www.nitrc.org/projects/artifact_detect/.

Yang, Y.C., Boen, C., Gerken, K., *et al.* (2016). Social relationships and physiological determinants of longevity across the human life span. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 578–83.

Yoon, J.H., Tamir, D., Minzenberg, M.J., *et al.* (2008). Multivariate pattern analysis of functional magnetic resonance imaging data reveals deficits in distributed representations in schizophrenia. *Biological Psychiatry*, **64**, 1035–41.

Zou, G.Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, **12**, 399–413.