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

David Dodell-Feder\textsuperscript{1,2}, Laura M. Tully\textsuperscript{3}, Emily Dudek\textsuperscript{1}, Christine I. Hooker\textsuperscript{4}

\textsuperscript{1}Department of Psychology, University of Rochester
\textsuperscript{2}Department of Neuroscience, University of Rochester Medical Center
\textsuperscript{3}Department of Psychiatry and Behavioral Sciences, UC Davis
\textsuperscript{4}Department of Psychiatry & Behavioral Sciences, Rush University Medical Center

Corresponding author: David Dodell-Feder, Department of Psychology, University of Rochester, 453 Meliora Hall, Rochester, NY 14627, d.dodell-feder@rochester.edu

Short title: Neural representation of mental states in SSD and FHR

\emph{This article has been accepted for publication in Social Cognitive and Affective Neuroscience Published by Oxford University Press.}
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 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 fMRI, 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-SSD, SSD showed reduced decoding accuracy for the category of mental states in right temporo-parietal junction (RTPJ)—which was related to false-belief accuracy—and 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.

Keywords: 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; Yang et al., 2016; Holt-Lunstad et al., 2017; Snyder-Mackler et al., 2020). Successfully navigating the social world and forming such connections hinges upon our ability to attribute and reason about the mental states (i.e., beliefs, desires, intentions) of others—a process 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 (SSD), which are associated with marked and persistent impairments on behavioral measures of ToM (Bora et al., 2009; Ventura et al., 2015). In support of the notion that ToM facilitates successful social interaction, 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 too are 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, medication); individuals at familial high-risk for schizophrenia (FHR)—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 are 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 SSD (van der Gaag, 2006; Hoffman, 2007; Selten et al., 2017).

Towards better understanding the nature of ToM impairment in SSD 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 right and left temporo-parietal junction (RTPJ, 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 versus 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, watching videos). In SSD, 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-SSD, SSD showed reduced ToM-related neural activity in MPFC, PC, and aspects of 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; Dodell-Feder, DeLisi, et al., 2014a; Villarreal et al., 2014; 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 versus MPFC) and nature (e.g., hyper- versus hypo-activation) of the abnormality in both SSD 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 SSD 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. And, further, that 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 SSD (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 are 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 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 SSD (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 SSD and FHR.

Thus, here, we evaluate whether and how the representational content of mental state information is disturbed in SSD 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 fMRI studies of the ToM network in SSD and FHR (Dodell-Feder, DeLisi, et al., 2014a; Dodell-Feder, Tully, et al., 2014). Both participant groups performed the false belief 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 SSD 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: (a) those describing outdated (i.e., false) beliefs, and (b) 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), 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 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, false belief 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 preregistered. The schizophrenia dataset included 38 participants between the ages of 18-58 years; 20 individuals with schizophrenia \((n=16, 80\%)\) or schizoaffective disorder \((n=4, 20\%; \text{hereafter, SSD})\) and 18 non-schizophrenia control participants \(\text{non-SSD}\) 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 et al. (Dodell-Feder, Tully, et al., 2014).
The familial high risk (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 with 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 20-35 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 \((n=9\text{ FHR, } n=2\text{ non-FHR})\). 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 et al. (Dodell-Feder, DeLisi, et al., 2014a).

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: (a) False Belief (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 (b) **False Physical Representation** (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 (5 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 3T 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 mm³). Functional data were acquired with a T2*-weighted echo-planar imaging sequence with parallel imaging (acceleration factor=2, 47 slices, voxel size=3 mm³, TR=2560 ms, TE=30 ms, flip angle=85°) for the SSD dataset, and a T2*-weighted echo-planar imaging sequence (40 slices, voxel size=3 mm³, TR=2560 ms, TE=30 ms, 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](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 (ART; 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.** Regions-of-interest (ROIs) were defined from an independent dataset reported in Dufour *et al.* (Dufour *et al.*, 2013) of 462 neurotypical participants who completed the false belief 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), left temporo-parietal junction (LTPJ; -48, -56, 22), precuneus (PC; 2, -56, 36), right superior temporal sulcus (RSTS; 58, -10, -14), and right temporo-parietal junction (RTPJ; 54, -52, 22; see Figure 1A). These regions have been demonstrated by meta-analysis to be most consistently recruited by the false belief 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.** False belief 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 et al. (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. (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 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 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 5 ROI tests conducted within-groups and between-groups. Effect sizes were calculated as Cohen’s d along with bias-corrected-and-accelerated (BCa) 95% 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 et al. (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 SSD; the correlation between voxels for FB in non-SSD; the correlation between voxels for FPR in SSD; the correlation between voxels for FPR in non-SSD—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 $p<.001$ and a cluster-wise FWE-corrected $p<.05$. 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, false belief 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 are accompanied by BCa 95% CIs generated from 10,000 bootstrap samples. We consider a finding to be unexpected under the null hypothesis when $q<.05$. Given that we find group differences in two ROIs, for task performance correlations, we corrected for 4 tests (2 ROIs x 2 conditions [FB, FPR]); for symptoms, we corrected for 6 tests (2 ROIs x 3 symptom categories [positive, negative, 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 et al. (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 1, Supplementary Figure 1). 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 2, Supplementary Figure 2).

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: \( t(30) = 1.16, p = .255, d = .35, 95\% \text{ CI } [-.31, 1.02] \), translation: \( t(25) = 1.06, p = .297, d = .26, 95\% \text{ CI } [-.40, .92] \). Further, neither mean translation nor rotation were
correlated with pattern discriminability in either ROI, DMPFC and translation $r(36) = .06$, 95% CI [-.27, .37], $p = .735$, DMPFC and rotation, $r(36) = .11$, 95% CI [-.22, .42], $p = .509$, RTPJ and translation $r(36) = .04$, 95% CI [-.29, .35], $p = .820$, RTPJ and rotation $r(36) = -.21$, 95% CI [-.50, .11], $p = .197$. 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 ANCOVAs, controlling for univariate activation. The group difference in DMPFC pattern discriminability was reduced to a trend level of significance, $F(1, 35) = 3.33$, $p = .077$, $\eta^2 = .08$, although the effect size based on the marginal means was medium in size, $d = .64$, with a 95% CI, [-.04, 1.31], largely overlapping with that of the non-adjusted model, [.18, 1.66]. When controlling for univariate activity, the impact of group on pattern discriminability in RTPJ remained statistically significant, $F(1, 35) = 7.54$, $p = .009$, $\eta^2 = .12$, with a large effect size, $d = .90$, 95% CI [.21, 1.60], similar in magnitude to the non-adjusted model, $d = 1.04$, 95% CI [.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 5). 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, $t(34) = 2.15$, $p = .039$, $d = -.70$, 95% CI [-1.37, -0.03], such that SSD exhibited higher pattern correlations than non-SSD (Table 5). Looking within each group, pattern correlations did not differ between FB and FPR in SSD, $t(19) = 1.60$, $p = .125$, $d_z = .36$, 95% CI [-.13, .85], as they did in non-SSD, $t(17) = 2.50$, $p = .023$, $d_z = .59$, 95% CI [.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, $Fs<=1.29, ps>=.264$.

The non-FHR and FHR groups showed classification accuracies above chance in all ROIs (Table 4, 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 3, 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 4, 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, and symptoms. On task performance, greater pattern discriminability was associated with better performance on the FB condition, $r(31)=.53, 95\% \text{ CI}[.17, .77], q=.007$, and this was not
moderated by group, $\beta = -.56$, 95% CI [-1.50, .37], $p = .229$. Task accuracy, either on the FB or FPR condition, was not associated with pattern discriminability in DMPFC, $r(31) = -.07$, and although we observed positive associations between FB accuracy and univariate activity in DMPFC, $r(31) = .37$, and RTPJ, $r(31) = .41$, at an uncorrected level ($p < .05$), 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, $r(31) = .18$, 95% CI [-.17, .50], $q = .547$, 95% CI of the correlation difference [.03, .65], meaning that increased pattern discriminability in RTPJ might specifically support false belief reasoning as opposed to reasoning about representations more generally. Further, the variance in FB accuracy accounted for by pattern discriminability, $\beta = .41$, 95% CI [.06, .76], $p = .023$, was above and beyond that accounted for by univariate activity, which was not associated with FB accuracy when taking multivariate pattern discriminability into account, $\beta = .19$, 95% CI [-.17, .54], $p = .295$.

On symptoms, we observed negative associations between RTPJ pattern discriminability and positive, $r(18) = -.52$, and disorganized symptoms, $r(18) = -.49$, at an uncorrected level ($p < .05$), but neither association survived FDR-correction. We also observed a positive association between DMPFC univariate activity and negative symptoms, $r(18) = .45$, 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 SSD and FHR have examined a single question: whether regions in the ToM network shows 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 SSD, FHR, and controls 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 SSD (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, SSD 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 SSD, 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 SSD versus non-SSD 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 SSD is due to SSD 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 SSD
vs non-SSD—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 SSD versus non-SSD. 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 SSD, 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 SSD (Kwapil, 1998; Tarbox and Pogue-Geile, 2008; Radua et al., 2018)—reduces 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 SSD 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 SSD, 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 intact involvement of these regions in the context of false belief reasoning, but not others involving 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 SSD. First, disrupted neural representation of mental state information may only occur in pre-SSD (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 SSD, 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? Towards addressing this question, we evaluated the associations between univariate activity, multivariate pattern discriminability, false belief 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 variance in FB accuracy above and beyond that explained by RTPJ
univariate activity suggesting particular importance of representational information for false
belief understanding. This finding is consistent with other work demonstrating that pattern
discriminability in RTPJ for intentional versus unintentional acts is associated with the extent to
which mental states are weighted when making moral judgments (Koster-Hale et al., 2013), and
other work in SSD 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 work 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 ($d=.92$ with power=$.80$), and large associations between brain and behavior ($r=.44$ for brain-task accuracy associations and $r=.61$ for brain-symptom associations with power=.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. Lastly, 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 SSD, 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 are not represented with sufficient distinguishing details or characteristics. This 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 performance on the FB task. In contrast, the ToM network in FHR can be characterized by altered involvement, but preserved 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.
Acknowledgments

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).

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.
References


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


Figure Legends

Figure 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. ROI = region-of-interest, DMPFC = dorsal medial prefrontal cortex, LTPJ = left temporo-parietal junction, PC = precuneus, RSTS = right superior temporal sulcus, RTPJ = right temporo-parietal junction.

Figure 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 $p<.001$ and a cluster-wise FWE-corrected $p<.05$. Colorbars depict $t$ values. R = right, L = left, A = anterior, P = posterior.
Table 1

Participant Characteristics

<table>
<thead>
<tr>
<th></th>
<th>SSD Dataset</th>
<th>FHR Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSD</td>
<td>Non-SSD</td>
</tr>
<tr>
<td><strong>n</strong></td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td><strong>Age, years</strong></td>
<td>38.8 (9.7)</td>
<td>32.4 (12.1)</td>
</tr>
<tr>
<td><strong>Sex, Male/Female (n)</strong></td>
<td>12/8</td>
<td>12/6</td>
</tr>
<tr>
<td><strong>Education, years</strong></td>
<td>15.0 (2.3)</td>
<td>14.2 (2.6)</td>
</tr>
<tr>
<td><strong>IQ</strong></td>
<td>108.7 (13.4)</td>
<td>107.4 (10.7)</td>
</tr>
</tbody>
</table>

**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<.001 |
| **Negative** | 1.9 (2.0) | 0.1 (0.2) | t(19)=4.03, p<.001 |
| **Disorganized** | 2.1 (1.6) | 0.4 (0.6) | t(24)=4.44, p<.001 |
| **General**  | 1.7 (1.7) | 0.2 (0.6) | t(22)=3.45, p=.002 |

*Note.* Values represent *M (SD)* unless otherwise noted. PANSS=Positive and Negative Syndrome Scale, SIPS=Structured Interview for Prodromal Syndromes, FB = False Belief, FPR = False Physical Representation.
## Table 2

**False Belief Task Performance**

<table>
<thead>
<tr>
<th></th>
<th>SSD Dataset</th>
<th>FHR Dataset</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSD</td>
<td>Non-SSD</td>
<td>Group</td>
<td>FHR</td>
<td>Non-FHR</td>
<td>Group</td>
</tr>
<tr>
<td>Accuracy, %</td>
<td></td>
<td></td>
<td>Difference</td>
<td></td>
<td></td>
<td>Difference</td>
</tr>
<tr>
<td>FB</td>
<td>74.8 (17.3)</td>
<td>80.6 (16.0)</td>
<td>t(31)=.99, $p=0.329$, $d^a=.34$ [-.38, 1.05]</td>
<td>88.5 (11.0)</td>
<td>90.1 (10.7)</td>
<td>t(35)=.47, $p=0.642$, $d=.15$ [-.50, .82]</td>
</tr>
<tr>
<td>FPR</td>
<td>79.9 (15.3)</td>
<td>82.2 (14.3)</td>
<td>t(31)=.45, $p=0.656$, $d=.16$ [-.57, .83]</td>
<td>86.7 (13.1)</td>
<td>90 (12.8)</td>
<td>t(35)=.78, $p=0.442$, $d=.26$ [-.45, .93]</td>
</tr>
<tr>
<td>Reaction Time, s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB</td>
<td>4.1 (0.6)</td>
<td>3.4 (0.6)</td>
<td>t(31)=3.35, $p=0.002$, $d=1.17$ [.35, 1.98]</td>
<td>3.3 (.6)</td>
<td>2.9 (.5)</td>
<td>t(34)=2.57, $p=0.015$, $d=.84$ [.19, 1.45]</td>
</tr>
<tr>
<td>FPR</td>
<td>3.7 (0.5)</td>
<td>3.5 (0.5)</td>
<td>t(31)=1.31, $p=0.201$, $d=.45$ [.26, 1.16]</td>
<td>3.1 (.5)</td>
<td>3.0 (.5)</td>
<td>t(34)=.91, $p=0.368$, $d=.30$ [-.41, .98]</td>
</tr>
</tbody>
</table>

*aCohen’s $d$ values with 95% bias-corrected-and-accelerated CIs derived from 10,000 bootstrap samples.*
Table 3

**ROI Classification Accuracy: SSD Dataset**

<table>
<thead>
<tr>
<th></th>
<th>SSD</th>
<th>Non-SSD</th>
<th>Group Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M [95% CI] Accuracy-Above-Chance %</td>
<td>One-Sample t-test</td>
<td>M [95% CI] Accuracy-Above-Chance %</td>
</tr>
<tr>
<td>DMPFC</td>
<td>6.3 [-4.4, 16.9]</td>
<td>$t(19)=1.23, q=.234$</td>
<td>27.8 [15.8, 39.8]</td>
</tr>
<tr>
<td>LTPJ</td>
<td>26.3 [16.6, 35.9]</td>
<td>$t(19)=5.69, q&lt;.001$</td>
<td>36.1 [25.5, 46.7]</td>
</tr>
<tr>
<td>PC</td>
<td>27.5 [16.8, 38.2]</td>
<td>$t(19)=5.40, q&lt;.001$</td>
<td>36.1 [25.5, 46.7]</td>
</tr>
<tr>
<td>RSTS</td>
<td>15.0 [1.6, 28.4]</td>
<td>$t(19)=2.35, q=.037$</td>
<td>29.2 [17.7, 40.7]</td>
</tr>
<tr>
<td>RTPJ</td>
<td>16.5 [2.4, 30.1]</td>
<td>$t(19)=2.46, q=.037$</td>
<td>41.7 [33.1, 50.2]</td>
</tr>
</tbody>
</table>
Table 4

ROI Classification Accuracy: FHR Dataset

<table>
<thead>
<tr>
<th></th>
<th>FHR</th>
<th>Non-FHR</th>
<th>Group Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M [95% CI] Accuracy-Above-Chance %</td>
<td>One-Sample t-test</td>
<td>M [95% CI] Accuracy-Above-Chance %</td>
</tr>
<tr>
<td>DMPFC</td>
<td>13.8 [2.7, 24.8]</td>
<td>$t(19)=1.23$, $q=.022$</td>
<td>23.7 [12.0, 35.4]</td>
</tr>
<tr>
<td>LTPJ</td>
<td>31.3 [20.6, 41.9]</td>
<td>$t(19)=6.14$, $q&lt;.001$</td>
<td>36.8 [27.5, 46.1]</td>
</tr>
<tr>
<td>PC</td>
<td>25.0 [10.8, 39.2]</td>
<td>$t(19)=3.68$, $q=.003$</td>
<td>36.8 [27.5, 46.1]</td>
</tr>
<tr>
<td>RSTS</td>
<td>17.5 [1.8, 33.2]</td>
<td>$t(19)=2.33$, $q=.031$</td>
<td>35.5 [25.4, 45.6]</td>
</tr>
<tr>
<td>RTPJ</td>
<td>30.0 [17.1, 42.9]</td>
<td>$t(19)=4.86$, $q&lt;.001$</td>
<td>31.6 [19.6, 43.5]</td>
</tr>
</tbody>
</table>
### Table 5

**ROI Pattern Correlations**

<table>
<thead>
<tr>
<th>ROI</th>
<th>Condition</th>
<th>SSD</th>
<th>Non-SSD</th>
<th>Group Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMPFC</td>
<td>FB</td>
<td>.98 (.54)</td>
<td>.97 (.52)</td>
<td><em>t</em>(36)=.05, <em>p</em>=.961, <em>d</em>=-.02, 95% CI [-.66, .66]</td>
</tr>
<tr>
<td></td>
<td>FPR</td>
<td>1.10 (.57)</td>
<td>1.03 (.63)</td>
<td><em>t</em>(35)=.37, <em>p</em>=.716, <em>d</em>=-.12, 95% CI [-.80, .56]</td>
</tr>
<tr>
<td>RTPJ</td>
<td>FB</td>
<td>1.25 (.45)</td>
<td>1.13 (.69)</td>
<td><em>t</em>(29)=.63, <em>p</em>=.533, <em>d</em>=-.21, 95% CI [-.83, .47]</td>
</tr>
<tr>
<td></td>
<td>FPR</td>
<td>1.04 (.54)</td>
<td>.63 (.63)</td>
<td><em>t</em>(34)=2.15, <em>p</em>=.039, <em>d</em>=-.70, 95% CI [-1.37, -.003]</td>
</tr>
</tbody>
</table>
### Table 6

**Whole-Brain Searchlight Analysis Group Differences**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Contrast</th>
<th>Region</th>
<th>MNI Coordinates $x$, $y$, $z$</th>
<th>Cluster Extent (voxels)</th>
<th>Cluster $p$ (FWE-corrected)</th>
<th>Peak t-value $(p&lt;.001)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD</td>
<td>Non-SSD&gt;SSD</td>
<td>RTPJ</td>
<td>66, -50, 22</td>
<td>165</td>
<td>&lt;.001</td>
<td>5.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Right Anterior Middle Temporal Gyrus</td>
<td>60, 3, -34</td>
<td>52</td>
<td>.032</td>
<td>4.71</td>
</tr>
<tr>
<td>SSD&gt;Non-SSD</td>
<td>Non-SSD&gt;SSD</td>
<td>No suprathreshold clusters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FHR</td>
<td>Non-FHR&gt;FHR</td>
<td>Right Superior/Middle Frontal Gyrus</td>
<td>26, 46, 16</td>
<td>270</td>
<td>&lt;.001</td>
<td>5.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Left Cerebellum/Left Fusiform Gyrus</td>
<td>-34, -80, -20</td>
<td>281</td>
<td>&lt;.001</td>
<td>4.51</td>
</tr>
<tr>
<td>FHR&gt;Non-FHR</td>
<td>No suprathreshold clusters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. Images were thresholded at a voxel-wise $p<.001$ and a cluster-wise FWE-corrected $p<.05$.  
