Online supplement for Hawco et al., "Separable and Replicable Neural Strategies During Social Brain Function in People With and Without Severe Mental Illness" Am J Psychiatry (doi: 10.1176/appi.ajp.2018.17091020)

Supplementary Table 1: SPINS Participant demographics and symptom scores across sites by Diagnosis

	site 1 (CMH)		site 2 (MRC)		site 3 (ZHH)	
	SSD only	**HC only**	**SSD only**	**HC only**	**SSD only**	**HC only**
Group	42	28	41	25	26	17
Sex (F:M)	11:31	14:14	10:31	9:16	13:13	8:09
Age	30.14 (8.33)	26.00 (7.05)	37.54 (10.93)	37.04 (11.14)	36.31 (8.68)	31.94 (9.24)
Education	13.44 (2.23)	15.64 (2.06)	12.90 (2.03)	15.76 (1.90)	13.42 (2.61)	14.94 (2.28)
BPRS	29.12 (6.63)	--	34.27 (8.01)	--	31.62 (8.96)	--
SANS diminished expression	0.84 (0.80)	--	1.36 (0.90)	--	1.03 (0.88)	--
SANS diminished motivation	1.50 (0.89)	--	2.26 (0.98)	--	1.76 (0.69)	--
Social Cognitive and Neurocognitive PCA Score	-0.38 (2.04)	1.89 (1.22)	-1.57 (2.45)	2.32 (1.01)	-2.08 (2.48)	1.35 (1.84)

Supplementary Table 2: PCA component loadings for the first component, calculated separately by cluster.

	Cluster 1	Cluster 2	Cluster 3
RMET	-0.26	0.31	0.30
RAD	-0.32	0.33	0.32
ER-40	0.26	-0.18	-0.29
TASIT 1	-0.26	0.28	0.32
TASIT 2	-0.31	0.30	0.33
TASIT 3	-0.34	0.32	0.33
processing speed	-0.30	0.30	0.28
attention / vigilance	-0.24	0.25	0.22
working memory	-0.32	0.32	0.28
verbal learning	-0.29	0.29	0.29
visual learning	-0.29	0.31	0.27
reasoning / problem solving	-0.26	0.23	0.20

Supplementary Table 3: Demographics and symptom scores across clusters in the replication sample

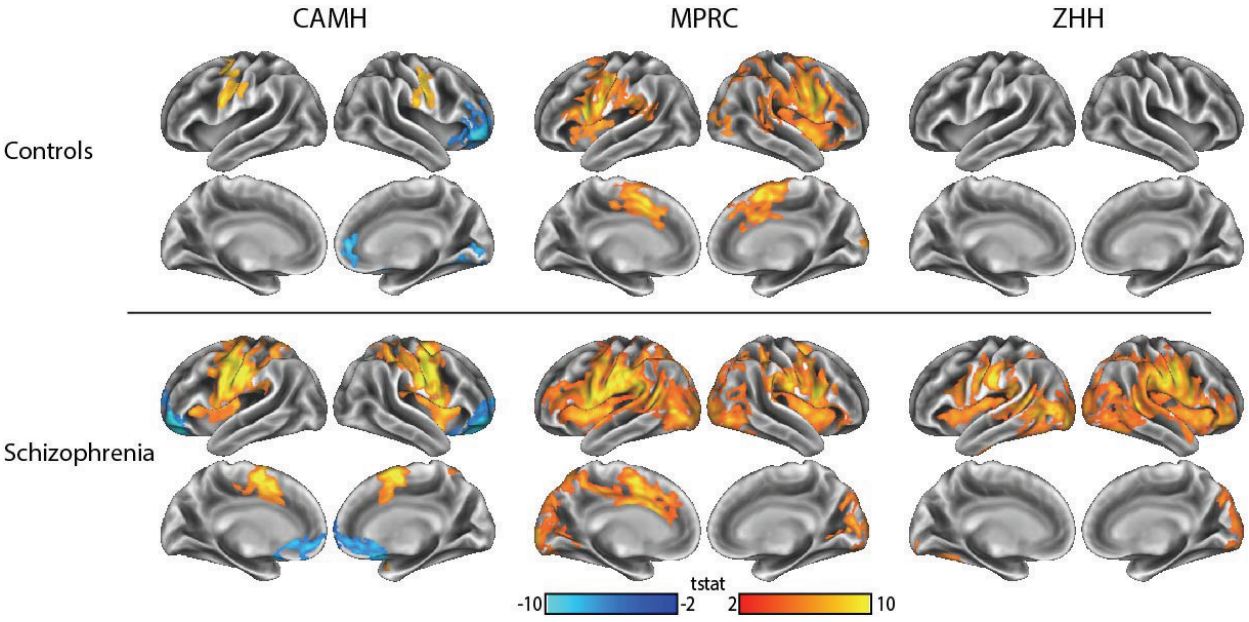
		Data by Cluster				
	Full Sample[1]	**cluster 1**	**cluster 2**	**cluster 3**	***P*[4]**	**effect size (η2)[4]**
Group (BP:HC:SZ)	37:38:32	10:12:7	16:13:17	9:12:7	0.66	-
Sex (F:M)	51:51	17:12	20:26	14:14	0.43	-
Age	31.25 (9.48)	31.72 (10.58)	30.50 (8.67)	32.79 (10.49)	0.61	0.001
Education	14.59 (2.15)	14.72 (2.43)	14.46 (2.09)	14.67 (2.00)	0.85	0.003
YMRS[2]	1.89 (2.19)	1.80 (2.57)	2.50 (2.28)	0.89 (1.17)	0.21	0.092
HDRS[3]	4.29 (2.31)	3.90 (2.23)	4.19 (2.51)	4.89 (2.15)	0.64	0.027
PANSS[3]	9.82 (4.55)	9.00 (4.19)	10.07 (4.36)	10.29 (5.31)	0.66	0.014

1. Clinical/demographic data not available for 5 participants (2 BD, 2 HC, 1 SSD)
2. Young Mania Rating Scale (YRMS) and Hamilton Depression Rating Scale (HDRS) administered to Bipolar Disorder cases only.
3. Positive and Negative Symptom Scale (PANSS) administered to SSD cases only. Scores represent the sum of items 1 to 7, negative symptom scores.
4. Statistical comparisons run across clusters only.

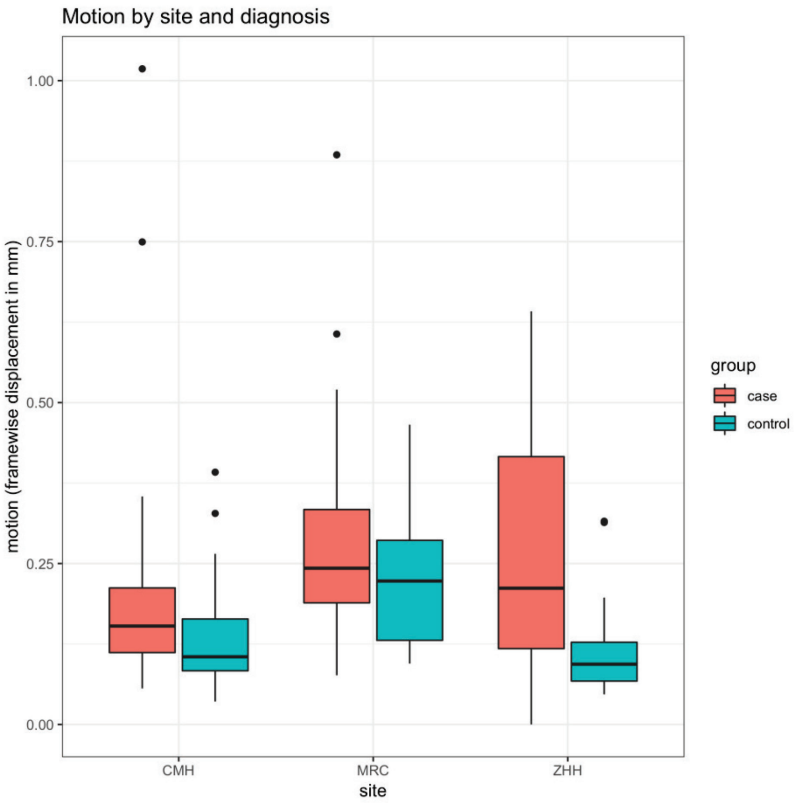
BD = euthymic bipolar disorder, HC = healthy controls, SZ = schizophrenia, F = female, M = male.

Supplementary Table 4: Descriptive statistics (mean and SD, z-scored based on all controls) for social cognitive and neurocognitive scores across sites and separated by diagnosis.

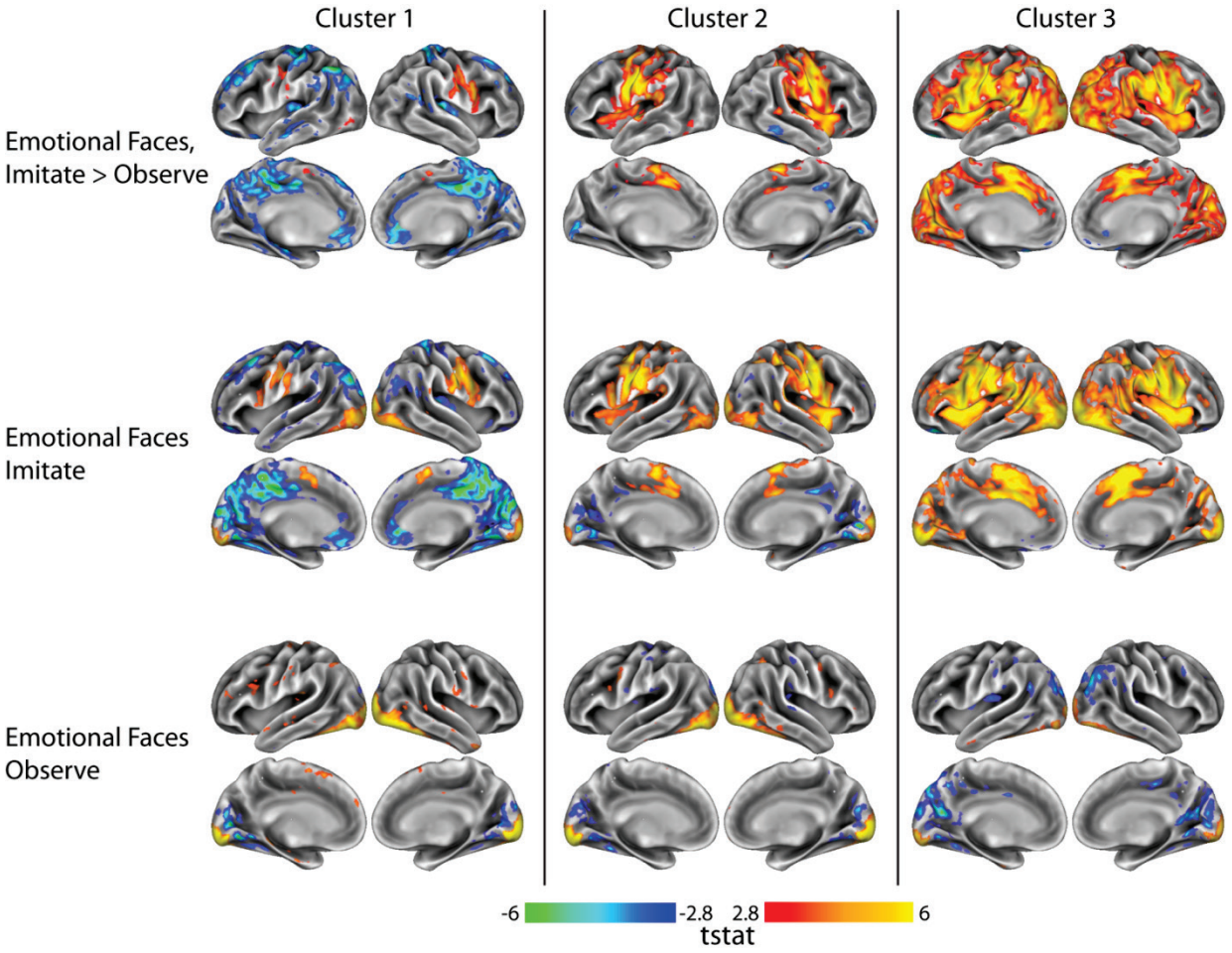
	SSD			HC		
	site 1 (CMH)	**site 2 (MRC)**	**site 3 (ZHH)**	**site 1 (CMH)**	**site 2 (MRC)**	**site 3 (ZHH)**
Social cognition						
RMET	-0.80 (1.43)	-1.58 (1.77)	-1.61 (1.47)	0.02 (0.82)	0.12 (0.83)	0.02 (1.13)
ER-40	-1.13 (1.92)	-3.37 (2.81)	-2.00 (2.64)	0.20 (0.86)	-0.31 (0.82)	0.41 (0.63)
TASIT 1	-0.23 (1.60)	-1.75 (1.74)	-1.83 (2.28)	0.26 (0.84)	0.17 (0.73)	-0.30 (0.91)
TASIT 2	-1.57 (1.91)	-2.59 (2.52)	-2.77 (2.59)	-0.09 (0.99)	0.29 (0.75)	-0.07 (1.00)
TASIT 3	-1.32 (1.06)	-1.94 (1.56)	-1.99 (1.48)	0.13 (0.72)	0.17 (0.84)	-0.06 (0.96)
Neuro cognition						
processing speed	-1.06 (1.26)	-1.30 (1.36)	-1.74 (1.28)	-0.13 (0.93)	0.16 (0.96)	-0.02 (1.18)
attention / vigilance	-0.96 (1.05)	-0.72 (1.23)	-1.24 (1.37)	-0.18 (0.93)	0.07 (1.18)	0.20 (0.83)
working memory	-1.10 (1.06)	-1.18 (1.15)	-1.54 (1.08)	0.02 (1.12)	0.25 (0.77)	-0.40 (1.01)
verbal learning	-0.98 (0.97)	-1.12 (0.97)	-1.57 (1.03)	-0.10 (1.11)	0.21 (0.86)	-0.15 (1.01)
visual learning	-0.73 (1.27)	-1.23 (1.24)	-1.52 (1.57)	0.06 (0.88)	0.19 (1.14)	-0.38 (0.91)
reasoning / problem solving	-0.66 (1.01)	-0.78 (1.14)	-1.05 (1.22)	-0.11 (0.95)	0.22 (1.04)	-0.13 (1.03)



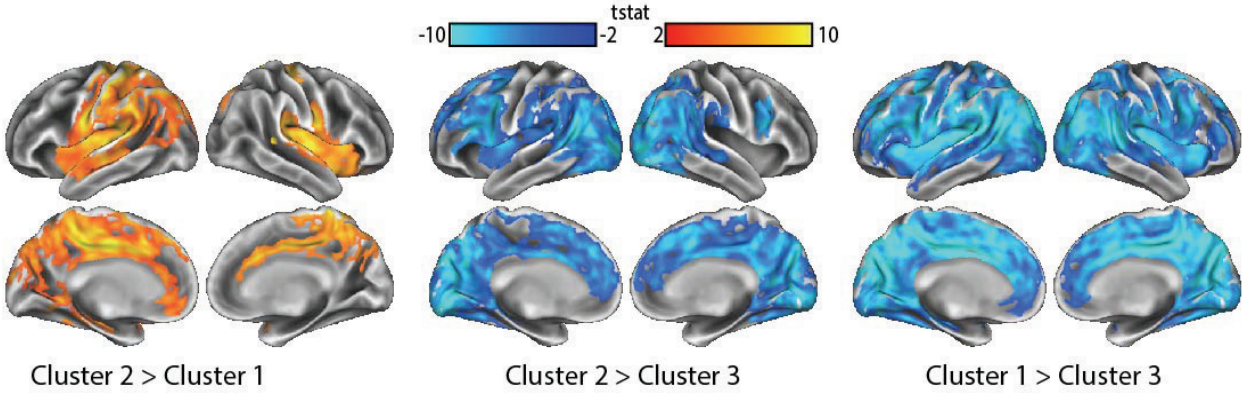
Supplementary Figure 1: Group activity patterns by site for HC and SSD; significant t-values are shown for all regions identified via threshold-free cluster enhancement as implemented in FSL's PALM function ($p < 0.05$ FEW corrected).



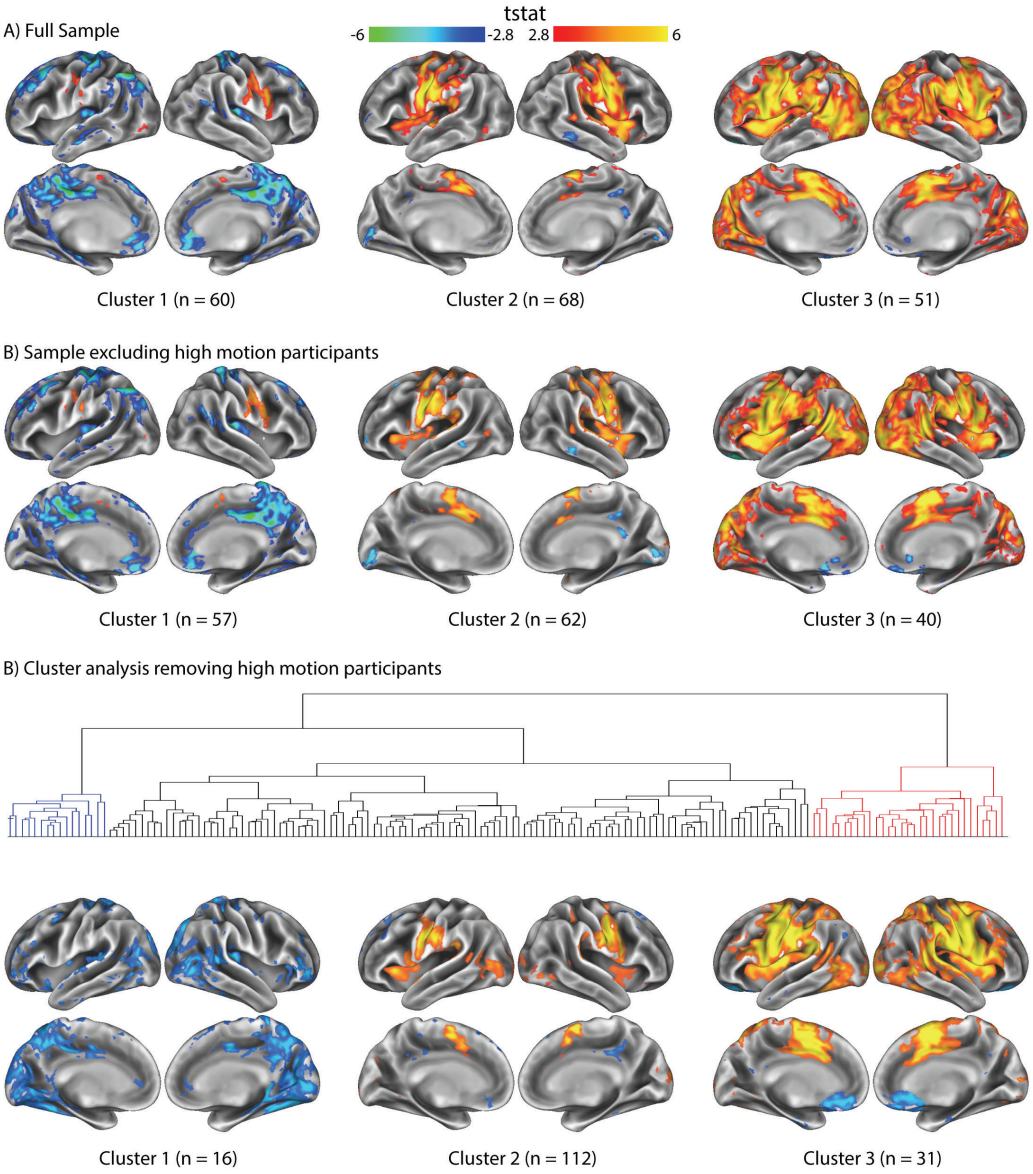
Supplementary Figure 2: Motion during the fMRI task, expressed as mean framewise displacement (FD), across sites and diagnostic groups.



Supplementary Figure 3: Second level main-effects analysis from each cluster for emotional faces during the Imitate and Observe runs, analysed using a group analysis in SPM12. Activity in the Imitate > Observe contrast can be driven by changes in either session (e.g. deactivation in the Imitate > Observe contrast can represent either decreased activity during imitation or increased activity during Observe). This data demonstrates that the different patterns in activity for Imitate > Observe across clusters are driven predominantly by the Imitate session.



Supplemental Figure 4: Group comparisons between clusters (two-sample t-tests, run in FSL-s PALM using TFCE. All t-stats shown are part of significant clusters. Blue colors represent the opposite pattern as the labeled group contrasts.



Supplementary Figure 5: A) original group maps with all participants included, for comparison. B) Reanalysis of group maps from each cluster, with high motion participants (mean FD > 0.4) removed. Cluster 3 had the largest number of high motion participants removed (n=11 excluded from cluster 3, while n=3 excluded from cluster 1 and n=6 excluded from cluster 2). C) Results of rerunning the hierarchical clustering analysis with high motion participants removed, showing a similar characterization of deactivators, typical activators, and hyper-activators. Group analyses were run using SPM12, $p < 0.005$ for display purposes.

Supplementary Appendix 1: Inclusion/exclusion criteria

Inclusion/Exclusion for the Multi-Site (SPINS study) sample: The CMH recruitment was conducted from the same start date until October, 2016. Participants were recruited separately at each of the three sites. All participants in the SSD group were individuals, who were currently receiving clinical treatment at one of the three sites, or had in the past, and had agreed to be contacted for future research studies. Participants in the HC group were recruited via each site's research registry, advertisements, or word-of-mouth. In order to be eligible for the study, all participants needed to be between the ages of 18 and 55 and be fluent in English.

Participants were excluded if they had: 1) Diagnosis of Mental Retardation (i.e. IQ<71); 2) Metal implants or a pace-maker that would preclude the MRI scan; 3) A history of a DSM-5 substance use disorder (other than tobacco) within the past six months; or a positive baseline urine drug screen; 4) Prior psychosurgery; 5) Type 1 diabetes mellitus (i.e., insulin-dependent diabetes mellitus with onset < 35 years of age and/or diabetes mellitus that has been complicated by a prior documented episode of ketoacidosis); 6) Acute or unstable medical illnesses (e.g. delirium, cancer, uncontrolled diabetes, decompensated cardiac, hepatic, renal or pulmonary disease, stroke (within past year), or myocardial infarction); 7) Significantly debilitating medical illnesses (e.g. encephalitis, aneurysms, tumors, or CNS infections); 8) Neurological disease associated with extrapyramidal signs and symptoms (e.g. Parkinson's disease); epilepsy, if the person has had one or more grand mal seizures in the past 18 months; history or physical signs of stroke; any diagnosis of a Central Nervous System (CNS) disorder; 9) History of head trauma resulting in loss of consciousness > 30 minutes that required medical attention; 10) Pregnancy; and 11) Suspected DSM-5 intellectual disability based upon clinical interview and psychosocial history.

Diagnostic status in all SSD and HC participant was confirmed via the SCID-IV-TR; however DSM-5 diagnostic criteria were applied. Healthy controls were excluded if they qualified for any current Axis I psychiatric disorder (according to the DSM-IV grouping for Axis 1), with the exception of adjustment disorder, phobic disorder, and past major depressive episode (over two years prior; presently unmedicated) or had a first degree relative with a history of psychosis. Cases were required to meet DSM-5 diagnostic criteria for schizophrenia, schizoaffective disorder, schizophreniform disorder, delusional disorder, or psychotic disorder not otherwise specified, and were included if clinically stable, as determined by no change in antipsychotic medication dose, or decrement in functioning, 30 days before study enrollment.

Inclusion/Exclusion for the independent, single site sample: Participants from the single-site sample were drawn from two study protocols. Protocol 1 included participants with Bipolar Disorder, while Protocol 2 included healthy control participants and individuals with schizophrenia spectrum disorder.

Protocol 1 (BD). In order to be eligible for the study, all participants needed to be between the ages of 18 and 49 and be fluent in English. Participants were excluded if they had a history of schizophrenia, schizoaffective, or other psychotic disorders, a history of substance use disorder within the past six months or a positive baseline urine toxicology drug screen (excluding tobacco), had electroconvulsive therapy within 6 months, had prior psychosurgery or a history of neurological disorder or head trauma, presented with acute, unstable, poorly controlled diabetes, or significantly debilitating medical illness. Participants were required to meet diagnostic criteria for bipolar I or II disorder as defined by the DSM-IV-TR as assessed via the SCID-IV, and were included if clinically euthymic for four weeks preceding study entry, as determined by both a

Hamilton Depression Rating Scale (HDRS)score of 17 or less and Young Mania Rating Scale (YRMS) scores of 10 or less at time of assessment.

Protocol 2 (HC and SSD). In order to be eligible for the study, all participants needed to be between the ages of 18 and 85 and be fluent in English. In order to be consistent with the primary SPINS sample, participants over age 55 were removed from further analysis. Participants were excluded if they had a diagnosis of mental retardation (IQ < 71), suspected intellectual disability, a history of substance use within the past six months or a positive baseline urine toxicology drug screen (excluding tobacco), had a history of neurological disorder or head trauma, presented with acute, unstable, or significantly debilitating medical illness. Healthy controls were excluded if they qualified for any current Axis I psychiatric disorder as determined by the Structured Clinical Interview for DSM (SCID-IV), with the exception of adjustment disorder and phobic disorder, or had a first degree relative with a history of psychosis. Cases were required to meet diagnostic criteria for schizophrenia (SCID-IV), schizoaffective disorder, schizophreniform disorder, delusional disorder, or psychotic disorder not otherwise specified, and were included if clinically stable, as determined by no change in antipsychotic medication dose, or decrement in functioning, 30 days before study enrollment.

Participant Training: Prior to the MRI scan, participants were introduced to the Imitate/Observe task. After an explanation of the task requirements (including explicit instructions to only move facial muscles but not move their head during imitation), participants were asked to lay flat on a bed, similar to the bed in the MRI. Study staff then presented a series of 10 faces, taken from the same stimulus bank but not included in the Imitate/Observe task, in a 'flip book' style held above the participant. The participant first practiced the Observe task, then

the Imitate task. If study staff noted that the participant had made any head movement, immediate feedback and instructions were given. The Imitate practice could be repeated if necessary to ensure participants could imitate the expressions without moving their head.

Supplementary Appendix 2: Social Cognitive Battery

The Penn Emotion Recognition Test ER40; [1]. The ER40 was administered to all participants to assess basic emotion recognition. Participants are randomly presented with 40 colour photographs of emotional faces and identify what emotion each face is showing from five options (happy, sad, anger, fear, and no emotion). Stimuli are balanced for emotion, as well as age, intensity of emotion, ethnicity, and gender for each emotion. The task is presented on a computerized, online platform, which outputs accuracy scores and median response times (http://webcnp.med.upenn.edu).

Reading the Mind in the Eyes Task RMET; [2]. A computerized version of the RMET was administered, including 36 trials presenting the eye region of black-and-white emotional faces. For each image, participants choose the most appropriate word to describe what the person is thinking or feeling from four choices (e.g., jealous, upset, panicked, arrogant). This task includes a mixture of somewhat basic and more complex mental states. Accuracy and reaction times were measured throughout.

The Relationships Across Domains (RAD) test [3] presents 25 written vignettes of 2-4 lines followed by three statements which describe the behaviour of the male-female dyad from each vignette in domains of social life different from that vignette. Participants indicate if the

behavior described in each statement is likely or unlikely to occur based on what was learned from the vignettes.

The Awareness of Social Inference Test - Revised TASIT-R; [4]. The TASIT involves viewing social video clips and is comprised of three subtests. TASIT 1, the Emotion Evaluation Test, includes 28 clips of actors portraying basic emotions. Participants choose one of seven labels (happy, surprised, sad, angry, anxious (fearful), revolted, or neutral) for each video to assess emotion recognition. In TASIT 2, Social Inference – Minimal, participants are shown 15 videos of sincere (5), simple sarcastic (5), and paradoxical sarcastic (5) interactions between two individuals. The actor's intentions can be inferred from the conversation, emotional expression, and other paralinguistic cues. Sincere exchanges are characterized by congruency between these elements, whereas for simple and paradoxical sarcastic exchanges, the meaning conveyed by the speaker conflicts with their paralinguistic cues (i.e., they mean the opposite of what they are saying). TASIT 3, Social Inference – Enriched, includes 16 clips of exchanges including lies (8) and sarcasm (8). These videos are similar to those in TASIT 2, but they also include enriched contextual information (i.e., a revealing camera shot or a prologue/epilogue). For lies, this additional information explicitly reveals if an individual is being deceptive, whereas paralinguistic cues must still be incorporated to make accurate inferences for sarcastic exchanges. For each of the TASIT 2 and 3 vignettes, participants are asked four questions assessing aspects of theory of mind. Total scores were calculated for TASIT 1, and subscores were calculated for each condition for TASIT 2 and TASIT 3.

1 Kohler, C. G., Bilker, W., Hagendoorn, M., Gur, R. E. & Gur, R. C. Emotion recognition deficit in schizophrenia: association with symptomatology and cognition. *Biological psychiatry* **48**, 127-136 (2000).

2 Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. & Plumb, I. The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of child psychology and psychiatry, and allied disciplines* **42**, 241-251 (2001).

3 Sergi, M. J. *et al.* Development of a measure of relationship perception in schizophrenia. *Psychiatry research* **166**, 54-62, doi:10.1016/j.psychres.2008.03.010 (2009).

4 McDonald, S., Flanagan, S. & Rollins, J. *The Awareness of Social Inference Test - Revised (TASIT-R)*. (Pearson Assessment, 2011).