Moment-to-Moment Brain Signal Variability Reliably Predicts Psychiatric Treatment Outcome

Supplementary Information

This document includes:

1. Supplementary Methods
2. Supplementary Results
3. Supplementary Figures S1-S16
4. Supplementary Tables S1-S12
5. Supplementary References

SUPPLEMENTARY METHODS and MATERIALS

Recruitment of social anxiety disordered patients

Individuals answered online questionnaires on demographics, social anxiety, and depressive and insomnia symptoms as part of the screening. Eligible individuals were interviewed via telephone using (A) the full Mini-International Neuropsychiatric Interview (M.I.N.I.) version 7.0 (1) and (B) the social phobia and major depressive disorder sections of the Structured Clinical Interview for DSM-IV – Axis I Disorders (SCID-I) (2). Included patients were at least 18 years of age, had no neurological disorder, no concurrent psychological treatment, and if treated with a psychotropic medication, they agreed to maintain a stable dose at least 3 months before enrollment and during treatment in the current study. All participants also met magnetic resonance imaging (MRI) safety criteria (e.g., not pregnant, no ferromagnetic objects in the body). At screening, we excluded patients that were currently suffering from ongoing severe depression (as indexed by scoring >34 on the Montgomery Åsberg Depression Rating Scale, MADRS-S) (3), bipolar or psychotic disorders, alcohol or substance use disorders, or antisocial personality disorder. Further, patients that answered positive to any SAD comorbidity in the M.I.N.I. screening telephone interview were subject to a second face-to-face DSM-5 diagnostic interview. Twenty patients (44.4%, 20/45) had a concurrent psychiatric comorbidity, and four patients (8%, 4/45) were on a stable dosage of selective serotonin reuptake inhibitors (SSRIs), which did not change throughout the study period. Two patients had previously used beta blockers in social situations, but agreed not to use them during the study period. For further study and sample details, see Månsson et al., 2019 (4). See also Table S1 for a detailed summary of demographic and clinical status, and comorbid mental illness. Importantly, all patients that entered treatment also remained throughout the intervention, and took part in post-treatment assessments.

Table S1. Demographics, clinical status, concurrent psychotropic medications, and comorbid conditions in 45 social anxiety disordered patients.

Variable	**Patients, *n* = 45**	
Females, *n,* %	28	62.2
Age, average, ±SD	30.8	8.4
SAD duration in years mean, average, ±SD	17.0	10.0
Marital status, *n*, %		
Married/cohabiting with children	16	35.6
Married/cohabiting without children	10	22.2
Non-cohabiting partner	4	8.9
Single with children	3	6.7
Single without children	9	20.0
Other	3	6.7
Education, *n,* %		
Completed primary school	3	6.7
Completed secondary school	7	15.6
Completed vocational education	2	4.4
Ongoing university education	15	33.3
Completed university education	18	40.0
Concurrent psychotropic medications, *n,* %		
No concurrent medication	41	91.1
SSRIs	4	8.9
Psychiatric comorbidity (M.I.N.I and SCID), *n,* %		
No concurrent or previous psychiatric comorbidity	20	44.4
Previous depressive episode(s)	20	44.4
Current (and previous) depressive episodes	1	2.2
Current dysthymia	1	2.2
Previous panic disorder	3	6.7
Current panic disorder	1	2.2
Previous agoraphobia	1	2.2
Current generalized anxiety disorder	1	2.2

Abbreviations: SSRIs, selective serotonin-reuptake inhibitors; M.I.N.I., Mini International Neuropsychiatric Interview; SCID, Structured Clinical Interview for DSM; SAD, Social anxiety disorder;

Compliance to cognitive behavioral-therapy (CBT)

The patients undertook a weekly test with questions related to CBT and content of the module. To control for compliance, the patients had to give 100% correct responses on the multiple-choice questionnaire (with the possibility of redoing the test multiple times). After completion of the homework assignments and the multiple-choice quiz, the next module was made available to the patient.

Seven clinical psychologists served as therapists in the current study. The mean number (±SD) of years with experience working with CBT was 6.7±5.5 and the allocation of patients to the therapists was randomized. The mean (±SD) number of completed treatment

modules was 7.9 (±1.8) and 80% (36/45) of the patients completed at least 7 out of 9 modules. Further, at post-treatment, clinician rated their patient's compliance (i.e., none; to some degree; to a large extent) to exposure exercises (i.e., one key ingredient in this cognitive behavioral therapy) during the treatment.

Social anxiety, depressive, and insomnia outcomes

The LSAS-SR is a 48-item self-report questionnaire (each question consists of common social situations and the responder is asked to state both his/her anxiety and avoidance in these situations). LSAS-SR is a gold-standard questionnaire to assess treatment-related changes in social anxiety symptoms and the total score typically shows excellent test-retest reliability (i.e., $r = .83$) (5). LSAS-SR was the primary outcome measure of the current study and was administered at multiple times throughout the study period: screening (week 0), first (week 1) and second (week 9) baseline, and immediately after the treatment (week 18). To examine social anxiety as a general construct, secondary social anxiety measures were also collected, including the Social Interaction Anxiety Scale (SIAS) (6), the Social Phobia Scale (SPS) (6), and the Social Phobia Screening Questionnaire (SPSQ) (7). Further, post-treatment interviews were performed via telephone and included SCID-I on SAD (2) and M.I.N.I. on SAD (1), as well as the Clinical Global Impression-Improvement (CGI-I) scale (8). The interviews were performed by two external psychiatrists. Before the interview, the psychiatrists were informed about each patient's pre-treatment LSAS-SR score, but blind to any post-treatment self-reports. Depressive and insomnia symptoms were assessed using the Montgomery-Åsberg Depression Rating Scale, Self-reported version (MADRS-S) (3), and the Insomnia Severity Index (ISI) (9). Secondary social anxiety outcomes, and depressive and insomnia symptom questionnaires were administered at pre- and post-treatment.

Threats against internal and external validity

The current within-group design included multiple baseline assessments aimed at controlling for threats to internal validity (e.g., history, maturation, regression to the mean, instrumentation, testing). Threats like diffusion of treatment, demoralization and compensatory rivalry (see also ref 10) are not present here since we only have one group and all patients were offered an effective treatment. Although selection bias remains a potential threat against external validity, the average LSAS-SR value at pre-treatment screening in our study was 77, suggesting that our group of patients are well in line with what is seen in international treatment studies (ref 11; a network meta-analysis including 13.164 SAD patients that received psychiatric treatment, and reported that the median pre-treatment LSAS-SR was 78). In conclusion, our design protects against threats to internal validity and it is unlikely that we have a non-representative group of SAD patients.

Neuroimaging

MRI was performed twice for each patient (first and second baseline) and the two sessions were separated by eleven weeks (average number of days between sessions: 77.2±1.6). Also, the time of day each patient was scanned did not vary between the two baselines (average difference in time of day = 1.7±2.3 hours), nor did pre-scan session subjective sleepiness (Karolinska Sleepiness Scale; $B = 0.13$, BSE = 0.25, $Z = 0.53$, $p = .594$) (12).

In addition, brain scans were acquired half-way through the treatment (week 4), at post-treatment, and at 1-year follow-up, all of which will be analyzed in future work.

Magnetic resonance imaging

First, an anatomical T1-weighted image (fast spoiled gradient echo) was collected (180 slices, 1 mm thickness, field of view: 250 mm, voxel size: 0.5 × 0.5 × 1 mm^3) for each patient. Second, for resting-state and task, blood-oxygen-level-dependent (BOLD) contrast images were acquired using the following parameters: 30 ms echo time, 2000 ms repetition time; 80 degree flip-angle, field of view: 250 × 250 mm^3, matrix size: 96 × 96, and the voxel size was 1.95 × 1.95 × 3.90 mm. Thirty-seven slices with a thickness of 3.4 mm were acquired to capture the whole brain. Ten dummy scans were run before the image acquisition started to avoid signals resulting from progressive saturation.

Stimuli were presented on a computer screen and seen by the participant through a mirror attached to the head coil. Headphones and earplugs were used to reduce perception of scanner noise and cushions in the head coil reduced movement. Experienced MRI nurses were taking care of all participants.

Brain image preprocessing pipeline

fMRI data were preprocessed with both FSL5 (13,14) and SPM12. Pre-processing included motion-correction, initial bandpass filtering (.01–.10 Hz), and detrending (up to a cubic trend) using SPM12. We also utilized extended preprocessing steps to further reduce potential data artifacts (15–17) using FSL5. Specifically, we subsequently examined all functional volumes for artifacts via independent component analysis (ICA) within-run, within-person, as implemented in FSL/MELODIC (18). Noise components were identified according to several key criteria: A) Spiking (components dominated by abrupt time series spikes); B) Motion (prominent edge or "ringing" effects, sometimes [but not always] accompanied by large time series spikes); C) Susceptibility and flow artifacts (prominent air-tissue boundary or sinus activation; typically represents cardio/respiratory effects); D) White matter (WM) and ventricle activation (19); E) Low-frequency signal drift (20); F) High power in high-frequency ranges unlikely to represent neural activity (≥ 75% of total spectral power present above .10 Hz;); and G) Spatial distribution ("spotty" or "speckled" spatial pattern that appears scattered randomly across ≥ 25% of the brain, with few if any clusters (i.e., ~20 voxels at 3 × 3 × 3 mm voxel size).

Examples of these various components we typically deem to be noise can be found in previous work (21). By default, we utilized a conservative set of rejection criteria; if manual classification decisions were challenging due to mixing of "signal" and "noise" in a single component, we generally elected to keep such components. ICA components were reviewed by an experienced MRI research engineer and ambiguous noise/signal ICs were discussed within the research group to reach common decisions. Components identified as artifacts were then regressed from corresponding fMRI runs using the regfilt command in FSL. Finally, we registered functional images to participant-specific T1 images, and from T1 to 3 mm standard space (MNI 152) using FLIRT (affine). Finally, we masked the functional data with the GM tissue prior provided in FSL (probability > 0.37).

Voxel-wise estimation of brain signal variability

For resting state, SD_{BOLD} was computed across the entire denoised time series for each voxel. To calculate SD_{BOLD} for the socio-affective face task, we also performed a block normalization procedure to account for residual low frequency artifacts (as in previous work) (16). We first normalized all blocks for the socio-affective face task such that the overall 4D mean across voxels and blocks was 100, within-person. For each voxel, we then subtracted the block mean, concatenated across all task blocks, and computed voxel SD_{BOLD} across this concatenated time series (22).

We also sought to compare SD_{BOLD} results to a more typical mean-based measure of fMRI activity ($MEAN_{BOLD}$) during the socio-affective face task. Accordingly, we calculated $MEAN_{BOLD}$ by first expressing each within-block volume as percent change from the average of the ten preceding (fixation) block scans, calculating mean percent change within each block, and averaging across all face blocks (a typical method in the partial least squares, PLS data-analysis framework; see below for model details and implementation).

Statistical modeling

Outlier detection

The multivariate Mahalanobis distance measure was used to review and identify possible statistical outliers of the prediction models (as presented in the main manuscript, subheading *Cross-validation framework for brain and behavioral predictors of treatment outcome*) including all treatment outcome predictors (i.e., social anxiety change scores, as predicted by the pre-treatment Liebowitz social anxiety scale, self-report version (LSAS-SR), brain signal variability (task SD_{BOLD}, resting-state SD_{BOLD}), and average neural response (task $MEAN_{BOLD}$)). As shown in Figure S1, one patient was deemed an outlier and subsequently removed from testing at all levels of analysis in the current study.

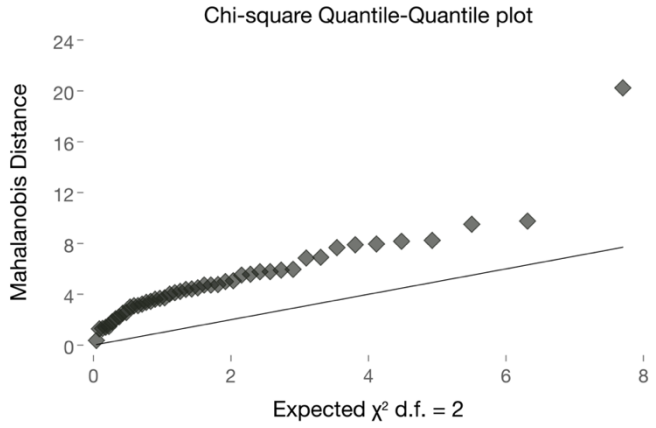

Figure S1. Detecting outliers. This figure demonstrates a quantile-to-quantile plot of the Mahalanobis distance measure. The measure is based on all initial predictors (i.e., task SD_{BOLD}, resting-state SD_{BOLD}, and pre-treatment LSAS-SR) and the LSAS-SR change score as outcome. The model includes all 46 patients initially included in the study and one observation was deemed as a multivariate statistical outlier and thus removed for further testing.

Treatment outcomes and predictors of clinical outcome

Longitudinal behavioral data were examined using repeated measure analyses with generalized estimating equations (default gaussian and exchangeable correlation structure) to calculate clinical outcomes across time (i.e., screening, first and second baseline, and post-treatment). Within-group Cohen's *d* effect sizes were calculated by dividing the mean difference with the standard deviation and correcting for the correlation between time-points (i.e., post-treatment vs screening). Simple and hierarchical linear regressions were used to regress pre-treatment predictor(s) and compliance to exposure exercises on clinical outcomes. Pre-treatment LSAS-SR as a predictor was compared with all brain-derived variables. Comparisons between predictive brain models were realized using multiple regression models and the adjusted R^2 values are reported throughout the paper. All correlations were performed in a parametric way (Pearson correlations).

Here, within PLS, we performed 1000 bootstraps, effectively picking random subsets of data for each of the 1000 iterations. For each subsample, we then re-calculate the PLS model linking brain (i.e., SD_{BOLD} or $MEAN_{BOLD}$) with LSAS-SR delta scores (using MATLAB 9.7.0.1190202 (R2019b; Natick, Massachusetts: The MathWorks Inc.; 2018).

There are a host of reasons to use PLS (23,24). For example, relative to standard GLM approaches, PLS is highly efficient, permitting the estimation of all brain weights in a single mathematical step via singular value decomposition (by definition, precluding the need for multiple comparison correction over voxels). PLS estimated "brain scores" also move fMRI-based models beyond the voxel (i.e., "indicator") level and into a psychometrically and statistically more desirable latent level. Further, PLS also utilizes assumption-free, non-parametric bootstrapped estimates (with replacement) of regional robustness that are not typical in the GLM community.

Partial least squares analysis at the first baseline

As described in detail in the main manuscript, the initial PLS models were based on a correlation matrix capturing the between-subject correlation of brain activity in each voxel and subject-wise delta total LSAS-SR score (post-treatment minus baseline 1, B1). The resulting voxel-wise BSRs for each model were thresholded at ±2 while excluding all clusters smaller than 20 voxels. The corresponding weights were applied to fMRI data recorded during the second baseline (measurement at B2). Below we display PLS models for task SD_{BOLD} (160 sec), task $MEAN_{BOLD}$ (160 sec), and resting-state SD_{BOLD} (340 sec) respectively (Tables S2-S4 and Figures S2-S4). As described in the main manuscript (subheading *Cross-validation framework for brain and behavioral predictors of treatment outcome*), the output from these initial behavioral PLS models were used in subsequent analyses to calculate the reliability-based brain scores.

We performed an additional series of PLS models to examine the influence of data volume on the strength and reliability of effects (i.e., for task-based SD_{BOLD} and $MEAN_{BOLD}$: first 40, first 80, and all 160 sec; for resting-state SD_{BOLD}: first 40, first 80, first 160, and all 340 sec). The same behavioral variable (the difference in total LSAS-SR score between B1 and post-treatment: delta Post-B1) was used in all of these models.

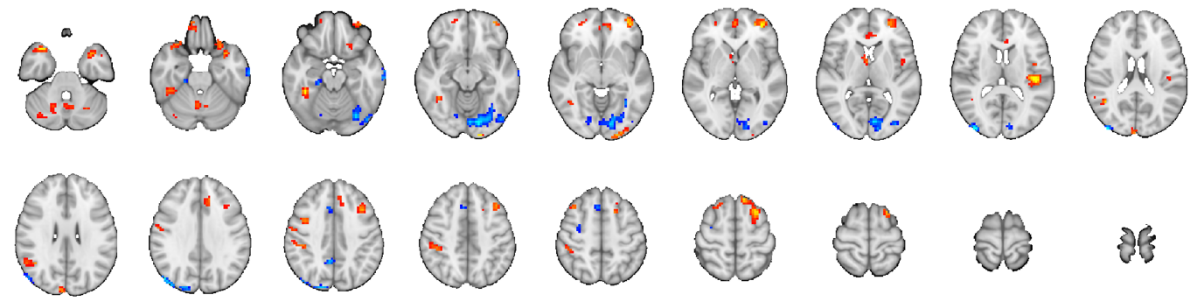

Figure S2. PLS-based task-related SD_{BOLD} at the first baseline. The figure demonstrates task-related variability associated with treatment outcome in 45 patients. Blue regions (negative BSRs) depict less variability, associated with better treatment outcome, whereas red/yellow (positive BSRs) represent high variability associated with better treatment outcome (color bar range ±2, ±4). Neural response pattern is thresholded at BSR ±2, with an extent threshold of 20 voxels, and the minimum distance between clusters is 10 mm. See also Table S2 for details. **Abbreviations:** BSR, Bootstrap ratio; SD_{BOLD}, Standard deviation of BOLD; BOLD, Blood-oxygen-level-dependent imaging; PLS, partial least squares;

Table S2. PLS-based task-related SD_{BOLD} at the first baseline. The table depicts task-related variability predicting treatment outcome in 45 patients. Negative BSRs depict less variability, predicting better treatment outcome, whereas positive BSRs represent high variability predicting better treatment outcome. Neural response pattern is thresholded at BSR ±2, with an extent threshold of 20 voxels, and the minimum distance between clusters is 10 mm.

Behavioral PLS	**MNI coordinates**				
Region	**x**	**y**	**z**	**BSR**	**k, voxels**
Middle Occipital Gyrus	-39	-90	18	-4.8833	54
Lingual Gyrus	15	-81	-12	-4.5411	453
ParaHippocampal Gyrus	-21	-27	-21	-4.5362	20
Superior Occipital Gyrus	-15	-90	36	-4.2837	92
Precuneus	-6	-51	39	-3.3357	24
Middle Temporal Gyrus	66	-18	-21	-3.3318	40
Posterior-Medial Frontal	-6	21	48	-3.1836	38
Lingual Gyrus	-15	-72	-9	-2.7278	23
Precentral Gyrus	-33	-6	57	-2.5941	21
Heschls Gyrus	45	-24	15	5.5041	96
Middle Orbital Gyrus	36	48	-3	5.3999	111
Inferior Temporal Gyrus	-39	-39	-18	4.7366	79
Temporal Pole	33	12	-27	4.5093	45
Temporal Pole	-33	18	-30	4.3716	90
Lingual Gyrus	18	-99	-9	4.3196	36
Superior Frontal Gyrus	27	18	63	4.2674	130
Angular Gyrus	-42	-54	24	4.1736	47
Medial Temporal Pole	45	9	-36	3.6766	25
Rolandic Operculum	45	-3	12	3.5904	21
Mid Cingulate Cortex	12	33	30	3.4495	45
Cerebelum VI	27	-60	-33	3.4281	104
Middle Frontal Gyrus	39	18	39	3.3858	62
Middle Frontal Gyrus	-33	21	57	3.1685	35
Thalamus	-3	0	3	3.1629	21
Cuneus	-6	-90	24	3.1627	32
Inferior Frontal Gyrus	24	18	-21	3.0661	36
Mid Orbital Gyrus	0	51	-3	3.0557	34
Precentral Gyrus	-39	3	39	3.0385	29
Postcentral Gyrus	-42	-33	42	-3.0383	68
Postcentral Gyrus	-57	-6	36	-2.9871	21
Superior Orbital Gyrus	-12	54	-24	-2.9376	25
Superior Orbital Gyrus	-27	51	-3	-2.8913	23
Cerebelum Crus 1	-36	-72	-27	-2.8403	21
Cerebellar Vermis 8	0	-66	-42	-2.7868	66
Anterior Cingulate Cortex	0	30	12	-2.6862	30
Dorsal Dentate Nucleus	-18	-60	-33	-2.6797	55

Abbreviations: BSR, Bootstrap ratio; SD_{BOLD}, Standard deviation of BOLD; BOLD, Blood-oxygen-level-dependent imaging; PLS, partial least squares;

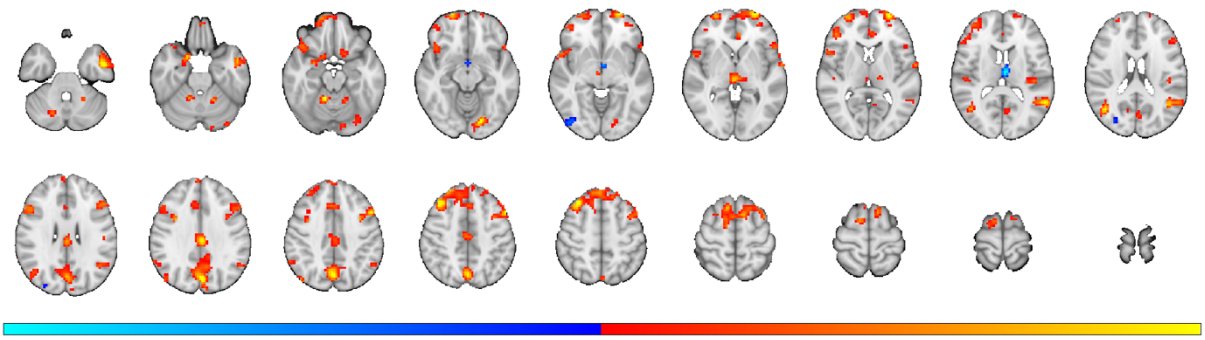

Figure S3. PLS-based task-related $MEAN_{BOLD}$ at the first baseline. The figure demonstrates task-related mean neural response associated with treatment outcome in 45 patients. Blue regions (negative BSRs) depict less activity, associated with better treatment outcome, whereas yellow (positive BSRs) represent more activity associated with better treatment outcome (color bar range ±2, ±4). Neural response pattern is thresholded at BSR ±2, with an extent threshold of 20 voxels, and the minimum distance between clusters is 10 mm. See also Table S3 for details. **Abbreviations:** BSR, Bootstrap ratio; $MEAN_{BOLD}$, Average BOLD activity; BOLD, Blood-oxygen-level-dependent imaging; PLS, partial least squares;

Table S3. PLS-based task-related $MEAN_{BOLD}$ at the first baseline. The table depicts task-related mean neural response predicting treatment outcome in 45 patients. Negative BSRs depict less activity, predicting better treatment outcome, whereas positive BSRs represent more activity predicting better treatment outcome. Neural response pattern is thresholded at BSR ±2, with an extent threshold of 20 voxels, and the minimum distance between clusters is 10 mm.

Behavioral PLS	**MNI coordinates**				
Region	**x**	**y**	**z**	**BSR**	**k, voxels**
Middle Temporal Gyrus	51	-3	-30	4.9126	154
Superior Orbital Gyrus	27	66	-3	4.8444	193
Middle Frontal Gyrus	51	15	42	4.4551	871
Cuneus	3	-75	33	4.4536	448
Middle Frontal Gyrus	-33	27	51	4.4255	211
Cerebelum IV-V	-12	-48	-21	4.199	39
Lingual Gyrus	18	-78	-12	4.186	117
Middle Temporal Gyrus	-42	-63	21	4.1779	111
Mid Cingulate Cortex	3	-24	33	4.0787	155
Middle Temporal Gyrus	57	-54	15	4.0547	182
Middle Frontal Gyrus	-27	63	3	3.8269	351
Thalamus	-3	-24	0	3.6117	82
ParaHippocampal Gyrus	-18	9	-24	3.6114	55
Superior Temporal Gyrus	63	-3	3	3.4839	37
Precentral Gyrus	-33	6	33	3.4639	31
Inferior Frontal Gyrus	-54	21	30	3.4046	59
Cerebelum IV-V	18	-51	-24	3.366	34
Cerebelum Crus 2	6	-72	-33	3.1282	36
SupraMarginal Gyrus	-63	-51	36	3.0459	26
Cerebelum VIII	24	-57	-48	3.0056	35
Temporal Pole	-45	24	-15	2.9605	134
Anterior Cingulate Cortex	6	39	0	2.9443	46
Superior Temporal Gyrus	-51	-21	12	2.9397	23
Cerebelum Crus 1	-21	-66	-33	2.9329	28
Cerebelum Crus 2	-39	-66	-39	2.8828	56

Heschls Gyrus	42	-24	15	2.8633	53
Superior Medial Gyrus	-3	57	39	2.7485	27
Inferior Frontal Gyrus	54	21	0	2.7409	40
Olfactory cortex	15	12	-18	2.6697	20
Middle Frontal Gyrus	33	51	0	2.4683	26
Thalamus	3	-12	15	-4.1941	44
Thalamus	3	-3	-9	-3.6736	20
Inferior Occipital Gyrus	-42	-81	-6	-2.7712	27
Middle Occipital Gyrus	-30	-78	21	-2.5005	20

Abbreviations: BSR, Bootstrap ratio; $MEAN_{BOLD}$, Average BOLD activity; BOLD, Blood-oxygen-level-dependent imaging; PLS, partial least squares;

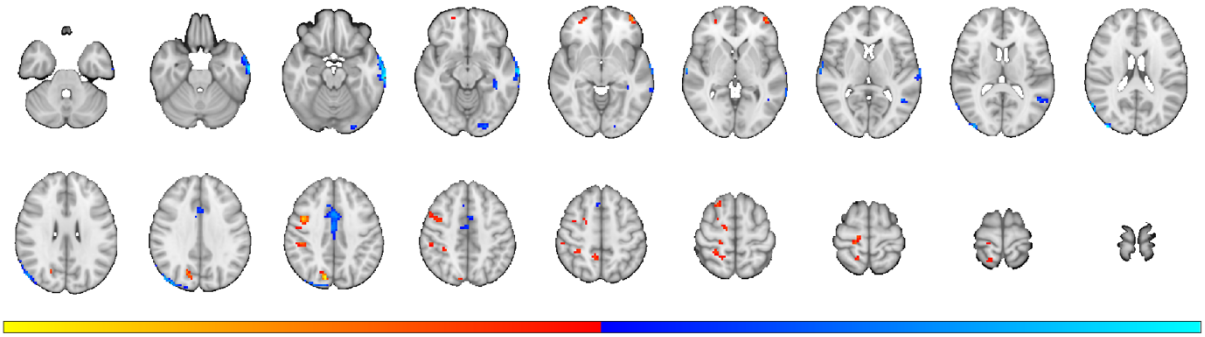

Figure S4. PLS-based resting-state SD_{BOLD} at the first baseline. The figure demonstrates resting-state neural variability associated with treatment outcome in 45 patients. Blue regions (negative BSRs) depict less variability, associated with better treatment outcome, whereas red/yellow (positive BSRs) represent high variability associated with better treatment outcome (color bar range ±2, ±4). Neural response pattern is thresholded at BSR ±2, with an extent threshold of 20 voxels, and the minimum distance between clusters is 10 mm. See also Table S4 for details. **Abbreviations:** BSR, Bootstrap ratio; SD_{BOLD}, Standard deviation of BOLD; BOLD, Blood-oxygen-level-dependent imaging; PLS, partial least squares;

Table S4. PLS-based resting-state SD_{BOLD} at the first baseline. The table depicts resting-state variability predicting treatment outcome in 45 patients. Negative BSRs depict less variability, predicting better treatment outcome, whereas positive BSRs represent high variability predicting better treatment outcome. Neural response pattern is thresholded at BSR ±2, with an extent threshold of 20 voxels, and the minimum distance between clusters is 10 mm.

Behavioral PLS	**MNI coordinates**				
Region	**x**	**y**	**z**	**BSR**	**k, voxels**
Middle Temporal Gyrus	69	-21	-18	-4.2981	199
Middle Occipital Gyrus	-39	-90	18	-4.2196	142
Mid Cingulate Cortex	0	-6	42	-4.1964	61
Mid Cingulate Cortex	3	6	42	-3.2932	92
Middle Temporal Gyrus	-66	-12	0	-3.2433	26
Lingual Gyrus	27	-90	-15	-3.2080	32
Middle Temporal Gyrus	45	-54	3	-3.1586	40
Fusiform Gyrus	39	-36	-12	-2.8847	23
Cuneus	-12	-75	39	4.0101	45
Middle Orbital Gyrus	42	54	-3	3.6862	24
Precentral Gyrus	-42	3	39	3.6293	72
Postcentral Gyrus	-18	-27	63	2.8923	27

Inferior Parietal Lobule	-45	-30	39	2.8885	113
Superior Orbital Gyrus	-27	51	-3	2.7853	20
Middle Frontal Gyrus	-24	21	60	2.6437	21
Middle Frontal Gyrus	-18	-6	57	2.4842	24

Abbreviations: BSR, Bootstrap ratio; SD_{BOLD}, Standard deviation of BOLD; BOLD, Blood-oxygen-level-dependent imaging; PLS, partial least squares;

Separate partial least squares analyses at the two baselines

In addition to the above-described PLS brain scores from the first baseline (i.e., used as input in the reliability-based cross-validation prediction models), we also performed control analyses by estimating two PLS models based on data from the SD_{BOLD} socio-affective task; one based solely on data from the first baseline assessment (same as Figure S2 and Table S2) and the other solely on data from the second baseline. The results of the two models and the correlation of their outputs can be found below (Figure S5 below). These clearly demonstrate that neuro-behavioral relationships as captured by PLS-based brain scores (separated by 11 weeks) are highly correlated (Figure S5A; Pearson's $r = .77$), that separate models of each baseline produce nearly identical effect sizes (brain × LSAS-SR change score correlations: $r_{B1} = .74$ (Figure S5B) and $r_{B2} = .76$ (Figure S5C), and that spatial patterns of these results show remarkable stability, especially in visual cortex (see Figure S5D and S5E).

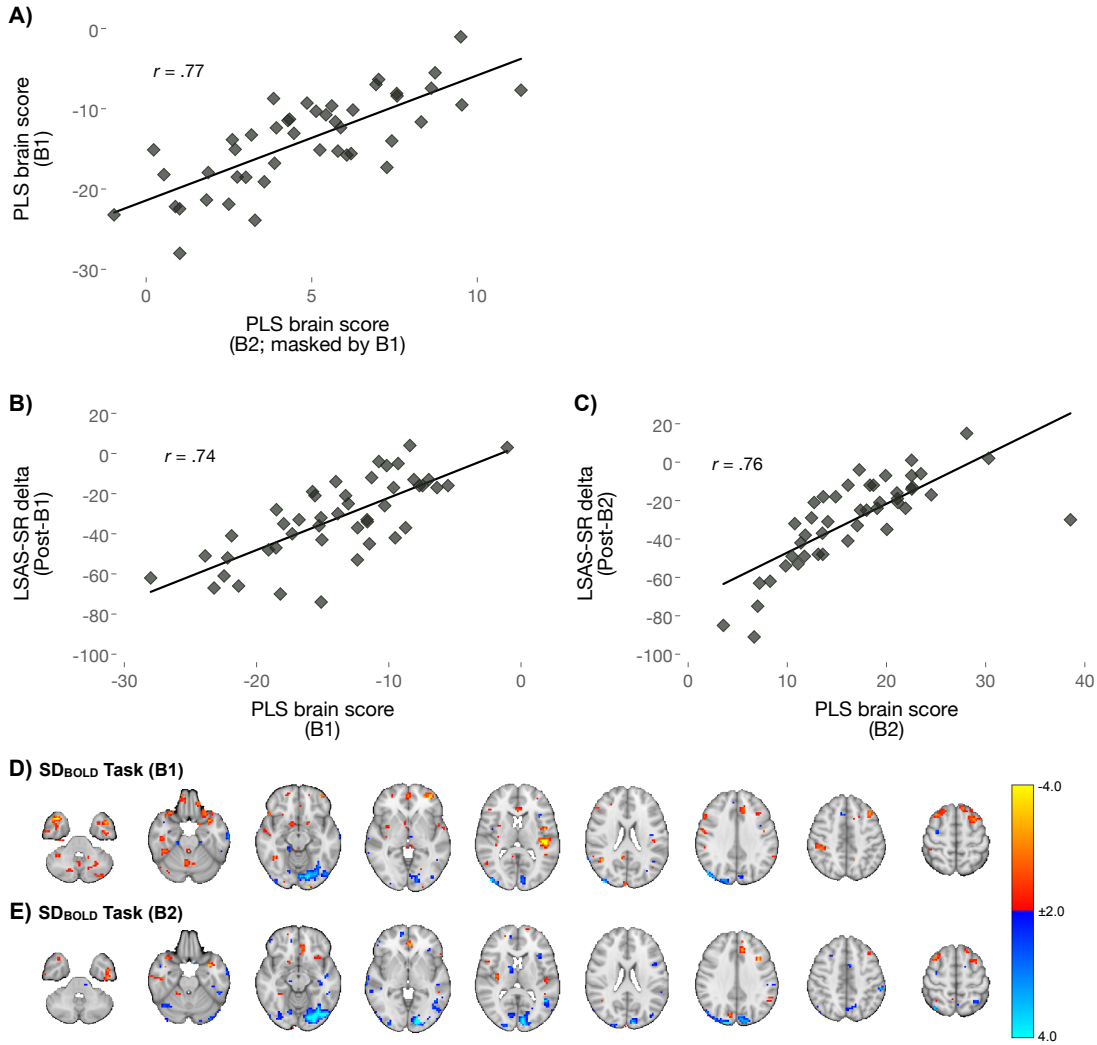

Figure S5 A) depicts the correlation between the first baseline (B1) PLS brain score, and the PLS brain score derived from the second baseline (B2; B2 data masked by B1), **B)** depicts the correlation between the PLS-based brain score derived from SD_{BOLD} task at B1, with the total LSAS-SR changes score (delta Post-B1), **C)** depicts the correlation between the PLS-based brain score derived from SD_{BOLD} task at B2, with the LSAS-SR changes score (delta Post-B2), **D)** displays the whole-brain voxel-wise map for the behavioral PLS model at B1, and **E)** displays the whole-brain voxel-wise map for the behavioral PLS model at B2. **Abbreviations:** PLS, partial least squares; B1, baseline 1; B2, baseline 2; LSAS-SR, Liebowitz social anxiety scale, self-report; Post, post-treatment;

Prediction accuracy estimation

To compare the predictive power of pre-treatment LSAS-SR and BOLD fMRI-derived (task SD_{BOLD}, task $MEAN_{BOLD}$, and resting-state SD_{BOLD}) variables, we calculated the Pearson correlation between predicted and observed LSAS-SR changes. To evaluate the significance of these correlations, we utilized a permutation approach. In brief, we permuted LSAS-SR change scores before data were divided into training and test folds, thus prior to model estimation and prediction of treatment outcome. Finally, predicted LSAS-SR change values were correlated with "observed" (i.e., shuffled) ones in each permutation. We repeated this procedure 1000 times and computed a permutation-based *P*-value by counting the number of permutations that resulted in a correlation between predicted and observed values that was higher than in the empirical (unshuffled) data. Additionally, 95% confidence intervals were estimated for all correlation coefficients using a bootstrap approach (1000 bootstraps).

Furthermore, and to offer a second metric for the relative comparison of predictors, we calculated the mean absolute scaled error (MASE) (25), which is defined as the ratio of the mean absolute prediction error to the mean absolute error of the one-step naïve forecast. While a MASE value of 1 represents equal predictive power of naive forecast and another predictor of interest, values below 1 depict a predominance of the predictor of interest where the improvement in prediction accuracy is 1–MASE %. MASE offers a scale-invariant measure of prediction accuracy and hence is directly comparable across different predictors regardless of their scale. Importantly, MASE penalizes over- and under-forecasting (i.e., too high vs. too low predicted scores, respectively) equally, rendering it a symmetric measure of prediction error. As other commonly used metrics of prediction accuracy do not offer scale independence or symmetry (e.g., root mean squared error, RMSE), MASE has been suggested as an ideal measure to compare the accuracy of different predictions (26).

Prediction model comparisons using multiple regressions

For all regression models aimed at comparing prediction model performance (i.e., outside of the cross-validation framework described in the main manuscript), we performed 1000 bootstraps (with replacement) to estimate bootstrapped 95% (normal-based) confidence intervals, as well as Monte Carlo permutation tests (1000 repetitions) for each variable. These regressions were performed using STATA (v15.1, STATA Corporation, College Station, TX, USA).

Data diagnostics

Social anxiety (delta LSAS-$SR_{POST-B1}$ and delta LSAS-$SR_{POST-B2}$) related brain scores on task SD_{BOLD}, resting-state SD_{BOLD}, and task $MEAN_{BOLD}$ were generated to predict treatment outcome. Reliability-based brain scores at the first and second baseline, and LSAS-SR change scores are displayed in Figure S6.

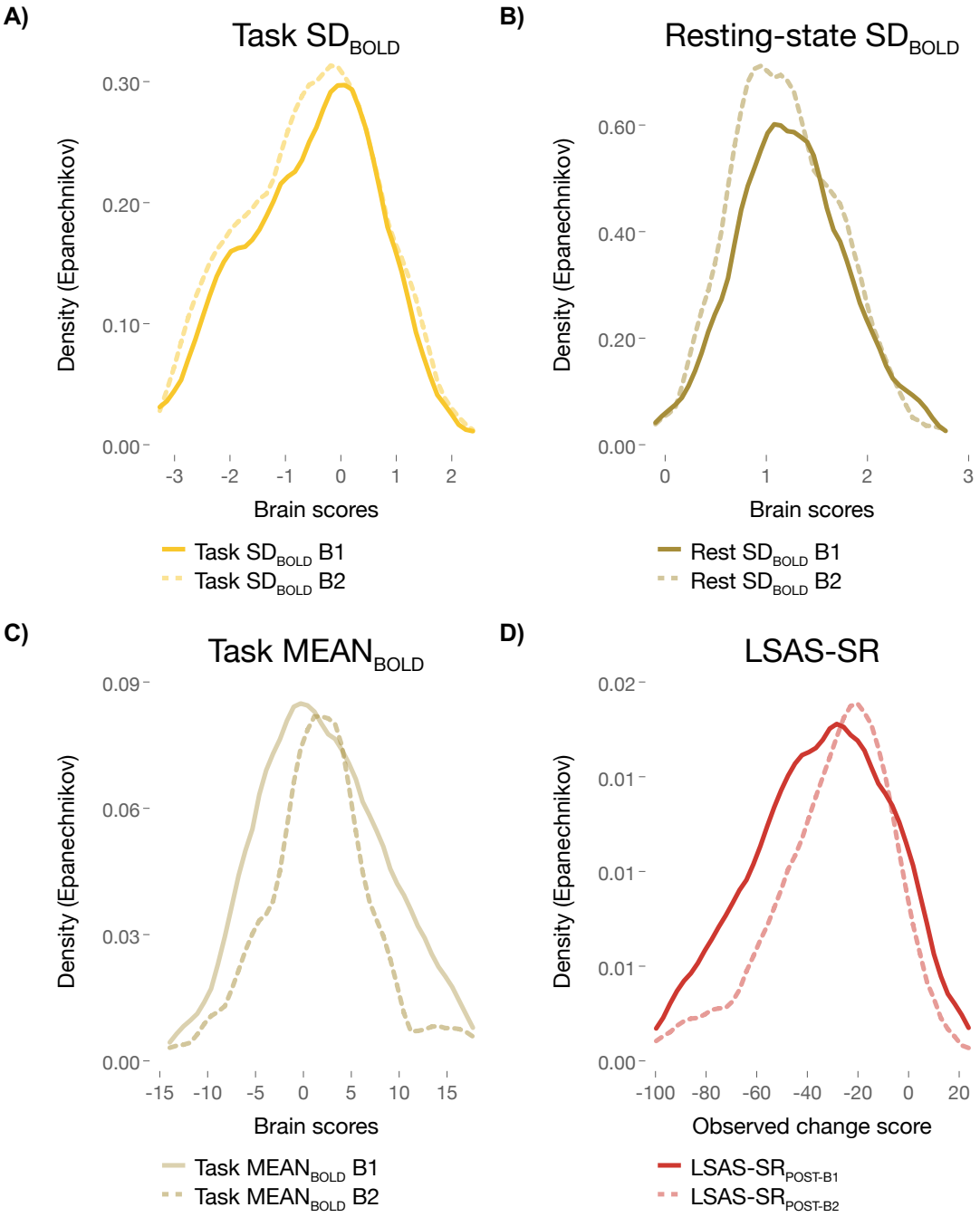

Figure S6. Brain score and behavioral data diagnostics. Brain scores at the first baseline (B1) and reliability-based brain scores at the second baseline (B2; for all details see the subheading *Reliability-based cross-validation framework for brain and behavioral prediction of treatment outcome* in the main manuscript) for **A)** task SD_{BOLD}, **B)** resting-state SD_{BOLD}, and **C)** task $MEAN_{BOLD}$. Each model includes the LSAS-SR change score (i.e., post-treatment minus baseline) and each model is based on 45 patients. **D)** displays the LSAS-SR change scores (Post-B1 and Post-B2). **Abbreviations:** LSAS-SR, Liebowitz social anxiety scale, self-report version; SD_{BOLD}, Standard deviation of BOLD; $MEAN_{BOLD}$, Average BOLD activity; BOLD, Blood-oxygen-level-dependent imaging; B1, Baseline 1; B2, Baseline 2; Post, Post-treatment;

Standard test-retest reliability estimation

Intraclass correlation coefficients (ICCs) based on the degree of consistency among measurements (C,1) were calculated to determine test-retest reliability on self-reported behavioral and brain-based variables (27) between the two baseline measurements (separated by 11 weeks). MATLAB 9.7.0.1190202 (R2019b; Natick, Massachusetts: The MathWorks Inc.; 2018) was used to compute ICCs (28). Between B1 and B2 measurements, ICCs were calculated on LSAS-SR (total score), and for each brain measure and data volume, on the reliability-based brain scores and on all voxels across the whole-brain (k = 51.609). Bootstrapped (× 1000 bootstraps) lower and upper bound 95% confidence intervals are reported. The ICCs were categorized as poor < 0.40, fair = 0.40 to 0.59, good = 0.60 to 0.74, or excellent ≥ 0.75 according to Cicchetti and Sparrow (29).

SUPPLEMENTARY RESULTS

Main and secondary treatment outcomes

As displayed in Table S5, main and secondary social anxiety outcomes, as well as depression and insomnia symptoms decreased significantly over the course of therapy.

Table S5. Main and secondary treatment outcomes. The table includes self-reported main and secondary outcomes from pre- to post-treatment for 45 patients. Group averages and 95% confidence intervals are displayed. Effect sizes represent Cohen's *d* between the first and last assessment of each measure.

	Pre-treatment assessments			**Post-treatment assessments**		
Self-reports	**Screening**	**Baseline 1**	**Baseline 2**	**Post-treatment**	**Pre to post change**	
	M (95% CI)	*M* (95% CI)	*M* (95% CI)	*M* (95% CI)	*M* difference (95% CI)	Effect size (95% CI)
LSAS-SR	77.29 (72.2, 82.4)	75.73 (70.5, 80.9)	73.33 (67.5, 79.1)	43.49 (37.1, 49.9)	-33.80 (-39.9, -27.7)*	1.62 (1.2, 2.1)
SIAS	55.18 (51.3, 59.1)	—	—	34.90 (31.1, 38.7)	-20.29 (-24.1, -16.4)*	1.49 (1.1, 1.9)
SPS	40.56 (36.4, 44.7)	—	—	19.60 (16.2, 23.0)	-20.96 (-24.9, -17.0)*	1.57 (1.1, 2.0)
SPSQ	34.09 (31.5, 36.7)	—	—	17.62 (15.1, 20.2)	-16.47 (-19.3, -13.7)*	1.86 (1.3, 2.4)
ISI	—	9.22 (7.6, 10.8)	—	6.98 (5.5, 8.4)	-2.24 (-3.7, -0.8)**	0.42 (0.1, 0.7)
MADRS-S	15.62 (14.0, 17.2)	13.07 (11.2, 15.0)	12.78 (10.6, 14.9)	9.04 (6.8, 11.3)	-6.58 (-8.4, -4.8)*	0.95 (0.6, 1.3)

*All permuted *ps* < 0.001; **Permuted *p* = .042;
Abbreviations: LSAS-SR, Liebowitz social anxiety scale, self-report version; SIAS, Social interaction anxiety scale; SPS, Social phobia scale; SPSQ, Social phobia screening questionnaire; ISI, Insomnia severity index; MADRS-S, Montgomery-Åsberg depression rating scale, self-report version;

Figure S7A displays the primary measure of social anxiety at each time-point (Screening, Baseline 1, Baseline 2, and Post-treatment), as well as the calculated changes scores (i.e., Post-treatment minus Baseline 1, and Post-treatment minus Baseline 2; Figure S7B). As expected, the two change scores were highly correlated: $r = .86$, $p < .001$; Figure S7C).

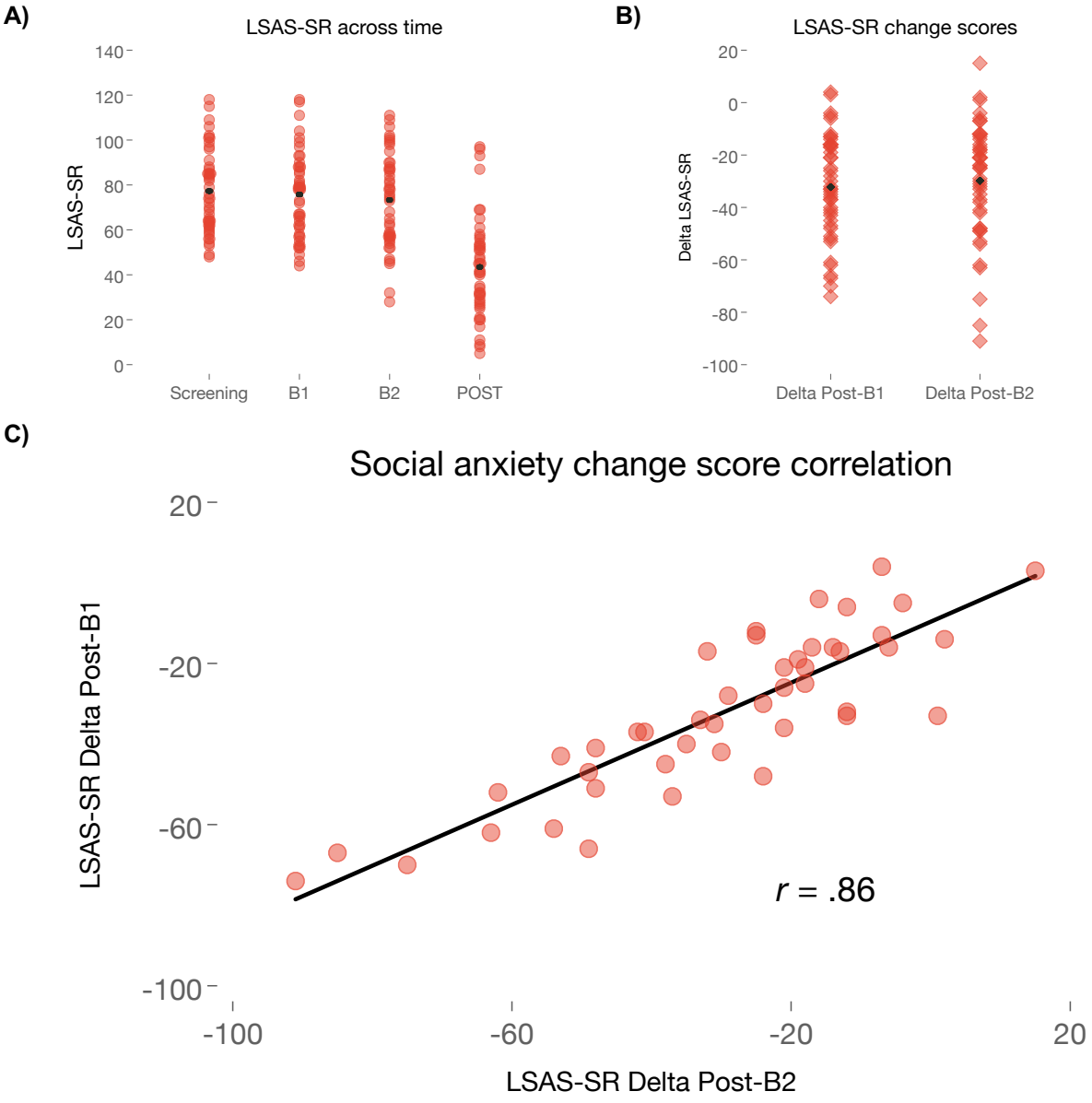

Figure S7. Main social anxiety treatment outcome. Figure **(A)** displays the main social anxiety outcome (Liebowitz social anxiety scale, self-report version, LSAS-SR) across all time-points, and **(B)** demonstrates LSAS-SR changes score for each individual and time-period. The black triangles denote the group average scores. **(C)** depicts the correlation between social anxiety change scores (LSAS-SR Post-treatment minus Baseline 1 *versus* LSAS-SR Post-treatment minus Baseline 2). All figures include 45 patients. **Abbreviations:** B1, Baseline 1; B2, Baseline 2; Post, Post-treatment;

Figure S8 displays the secondary self-reported measures of social anxiety and their respective changes scores (i.e., Post-treatment minus Baseline 1). Similarly, the secondary self-reported outcomes (change scores) on depressive and insomnia symptoms are displayed in Figure S8B.

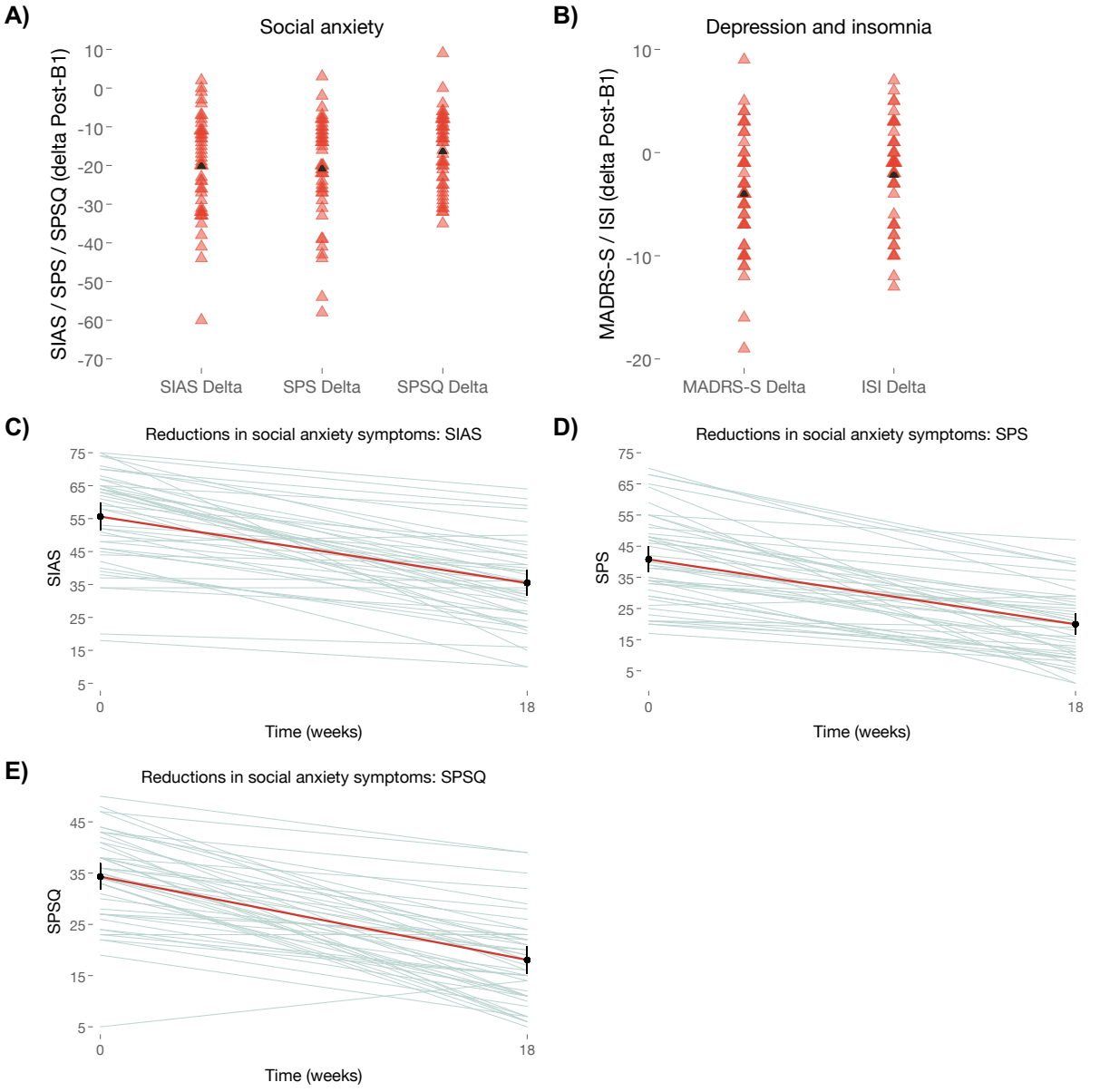

Figure S8. Secondary treatment outcomes. A) displays social anxiety changes scores (i.e., Post-treatment minus Baseline 1) with the black triangle denoting the group average change score, and **B)** depression and insomnia symptom change score (i.e., Post-treatment minus Baseline 1) with the group's average change score denoted with a black triangle. Also, individual slopes and group averages (solid red line), 95% CI (solid black error bars) are displayed for **C)** the Social interaction anxiety scale (SIAS), **D)** the Social phobia scale (SPS), and **E)** the Social phobia screening questionnaire (SPSQ). All figures include 45 patients. **Abbreviations:** B1, Baseline 1; B2, Baseline 2; Post, Post-treatment;

In addition to all self-reported measures, psychiatric interviews were completed at post-treatment including 45 patients. As displayed in Figure S9, according to the Clinical Global Impression-Improvement (CGI-I) assessment, a majority of the patients showed much or very much improvement after the treatment.

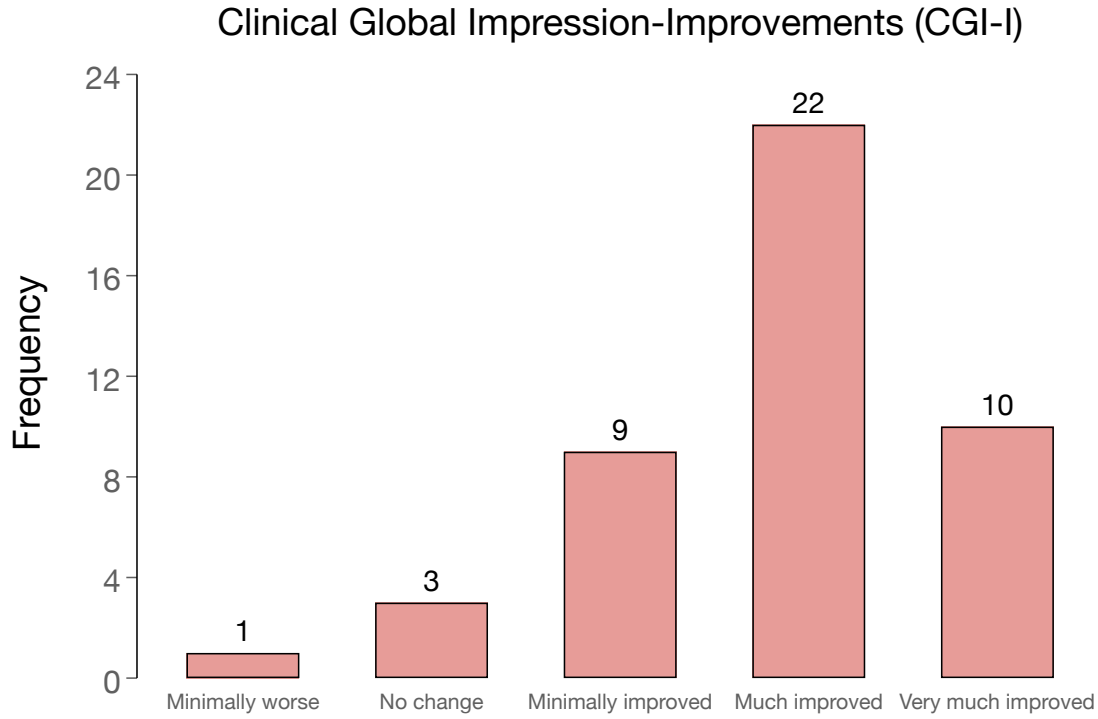

Figure S9. Clinical Global Impression-Improvements. The figure displays the clinician administrated CGI-I assessments at post-treatment (*n*=45). **Abbreviations:** CGI-I, Clinical global impression-improvement;

Despite the overall improvement, only a minority of patients were in probable remission at post-treatment, as indicated by a score less than 30 on LSAS-SR (28.9%, 13/45) or by being free from SAD according to DSM-5 criteria (17.8%, 8/45). Further, remission was determined by use of the DSM-5 criteria, as well as with the M.I.N.I. interview at post-treatment (see Table S6).

Table S6. Clinical interviews at post-treatment. Assessment of post-treatment remission status was conducted by use of structured clinical interviews of 45 patients.

	M.I.N.I.		**DSM-5**	
Indicating SAD diagnosis	*N*	*%*	*N*	*%*
No	24	53.3	8	17.8
Yes	21	46.7	37	82.2

Abbreviations: M.I.N.I., Mini International Neuropsychiatric Interview; DSM-5, The Diagnostic and Statistical Manual of Mental Disorders; SAD, Social anxiety disorder;

Treatment outcome and compliance to exposure exercise

Clinicians' ratings of exposure exercise compliance were related to larger reductions in LSAS-SR symptoms from the first baseline to post-treatment ($\beta = –0.30$, $p = .047$).

Treatment outcome predictions

Moment-to-moment brain signal variability during emotional face processing predicted social anxiety change scores with cognitive-behavioral therapy (CBT). As displayed in Table S7, task SD_{BOLD} predicted treatment outcome even after constraining the data volume to 80 or even 40 sec (see also Figure S10 demonstrating the consistency of spatial pattern across the three task SD_{BOLD} conditions). Table S7 also displays similar predictions and metrics for all other conditions (i.e., self-reported social anxiety, resting-state SD_{BOLD}, and task $MEAN_{BOLD}$).

Table S7. Pre-treatment predictors of treatment outcome. Zero-order treatment outcome predictions including 45 patients with social anxiety disorder.

				95% CI		
Pre-treatment predictors	**Data volume (items)**	**MASE**	***r***	**Lower**	**Higher**	**Permuted *p***
Behavioral, self-reports						
LSAS-SR at B1	48	0.65	.27	-.01	.56	.071
LSAS-SR at B2	48	0.58	.45	.20	.70	<.001
	Data volume (seconds)	**MASE**	***r***	**Lower**	**Higher**	**Permuted *p***
Brain, neural response						
Task SD_{BOLD}	160	0.54	.65	.51	.79	<.001
	80	0.53	.65	.52	.78	<.001
	40	0.53	.62	.45	.79	<.001
Resting-state SD_{BOLD}	340	0.55	.55	.35	.75	<.001
	160	0.63	.19	-.08	.46	.085
	80	0.62	.24	-.02	.51	.052
	40	0.64	.16	-.12	.45	.132
Task $MEAN_{BOLD}$	160	0.67	-.18	-.49	.12	.843
	80	0.67	-.28	-.58	.03	.962
	40	0.67	-.20	-.53	.12	.907

Abbreviations: MASE, Mean absolute scaled error; LSAS-SR, Liebowitz social anxiety scale, self-report version; B1, Baseline 1; B2, Baseline 2; SD_{BOLD}, Standard deviation of BOLD; $MEAN_{BOLD}$, Average BOLD activity; BOLD, Blood-oxygen-level-dependent imaging;

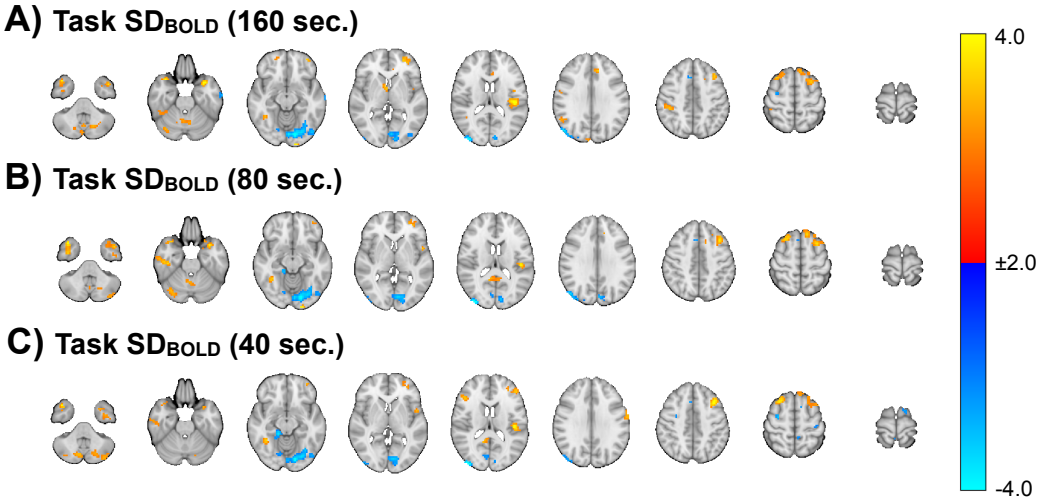

Figure S10. PLS-based task-related SD_{BOLD} at the first baseline. The figure demonstrates task-related variability associated with treatment outcome in 45 patients. Blue regions (negative BSRs) depict less variability, associated with better treatment outcome, whereas red/yellow (positive BSRs) represent high variability associated with better treatment outcome (color bar range ±2, ±4). Neural response pattern is thresholded at BSR ±2, with an extent threshold of 20 voxels, and the minimum distance between clusters is 10 mm. **A)** Representing the brain-behavioral association based on the full data volume model (80 TR/160 sec); and **B)** using half of the data volumes (40 TR/80 sec); and **C)** demonstrate this with only one-fourth of the data volumes (20 TR/40 sec). **Abbreviations**: SD_{BOLD}, Standard deviation of BOLD; BOLD, Blood-oxygen-level-dependent imaging; PLS, partial least squares; TR, Repetition time of the BOLD sequence;

Neither depression nor insomnia severity at pre-treatment predicted social anxiety treatment outcomes (all permuted *p*s>0.188). Similarly, treatment credibility ratings, demographics (as displayed in Table S1: age, sex, education, marital status), or psychiatric comorbidity (yes/no) did not predict treatment outcome (all permuted *p*s>0.071).

In Table S8, a multiple regression model is presented and includes all predictors of treatment outcome (i.e., actual LSAS-SR change score). This model includes all brain predictors when data volume remains constant across all conditions (i.e., 160 seconds stimuli duration), and shows that task-related neural variability outperforms all other potential predictors of treatment outcome. In a similar vein, Table S9 demonstrates a multiple regression model including all predictors of treatment outcome (i.e., actual LSAS-SR change score). This model includes unequal data volumes for the brain predictors. Specifically, task SD_{BOLD} and $MEAN_{BOLD}$ includes 160 seconds respectively, whereas the resting-state condition includes 340 seconds data. The model shows that the relatively shorter task-related neural variability remains the strongest treatment outcome predictor, and that the longer resting-state condition significantly contributed with some unique variance, see also Figure 4 in the main manuscript for details.

Table S8. Multiple pre-treatment predictors of treatment outcome: equal data volumes. Multiple prediction model including all fMRI conditions and self-reported social anxiety. Here, all conditions included equal data volumes (i.e., 160 sec). The regression includes 45 patients and the model's adjusted R^2 was 54%.

			95% CI			
Predictors	**Data volume (seconds)**	***β***	**Lower**	**Higher**	***Z***	**Permuted *p***
LSAS-SR at B2	—	.22	-.02	.46	1.81	.186
Task SD_{BOLD}	160	.61	.41	.82	5.85	<.001
Resting-state SD_{BOLD}	160	.26	.07	.46	2.63	.090
Task $MEAN_{BOLD}$	160	-.07	-.32	.17	0.60	.621

Abbreviations: fMRI, functional magnetic resonance imaging; SD_{BOLD}, Standard deviation of BOLD; $MEAN_{BOLD}$, Average BOLD activity; BOLD, Blood-oxygen-level-dependent imaging; Rest, Resting-state; B2, Baseline 2;

Table S9. Multiple pre-treatment predictors of treatment outcome: complete data volumes. Multiple prediction model including all fMRI conditions and self-reported social anxiety. Here, task-related fMRI conditions included 160 sec of data volumes, whereas the resting-state condition included 340 sec. The regression includes 45 patients and the model's adjusted R^2 was 57%.

			95% CI			
Predictors	**Data volume (seconds)**	***β***	**Lower**	**Higher**	***Z***	**Permuted *p***
LSAS-SR at B2	—	.29	.08	.50	2.65	.071
Task SD_{BOLD}	160	.41	.20	.62	3.83	.018
Resting-state SD_{BOLD}	340	.34	.09	.58	2.68	.039
Tast $MEAN_{BOLD}$	160	-.08	-.28	.12	0.78	.631

Abbreviations: fMRI, functional magnetic resonance imaging; SD_{BOLD}, Standard deviation of BOLD; $MEAN_{BOLD}$, Average BOLD activity; BOLD, Blood-oxygen-level-dependent imaging; Rest, Resting-state; B2, Baseline 2;

Thresholding, reliability and prediction accuracies

As reported in the main manuscript, we computed PLS-based neurobehavioral correlations between the first baseline blood-oxygen-level-dependent (BOLD) activity and reductions in LSAS-SR scores, and the resulting BSRs were thresholded at ± 2.0 ($p < .05$) while excluding all clusters smaller than 20 voxels. These thresholded maps were used to extract weights, and applied to MR data recorded at the second baseline measurement (11 weeks after the first baseline) to extract subject-specific brain scores. In addition to this procedure, here we also demonstrate model performance with variable BSR thresholding. Figure S11A) displays Pearson's *r*, B) mean absolute scaled error (MASE), and C) intraclass correlation coefficient (ICC) for BSR ratios at 0 (i.e., no threshold applied), 1, 2, and 3 for the main condition: task-related SD_{BOLD}. In general, Figure S11 demonstrates that predictive and reliability performance peaks at BSR level of 2, and deteriorates at an even stricter level. Interestingly, in a multiple prediction framework (i.e., including all potential pre-treatment brain predictors (task and resting-state SD_{BOLD}) and self-reported social anxiety), the prediction performance remains stable and accurate even without applying a threshold based on data from the first

baseline assessment (Pearson's $r_{ACT,PRED}$ = .69, MASE = 0.47, permuted p = .001, Figure S11). However, the model performance was optimal at a BSR of 2 (see also Figure S11A-B).

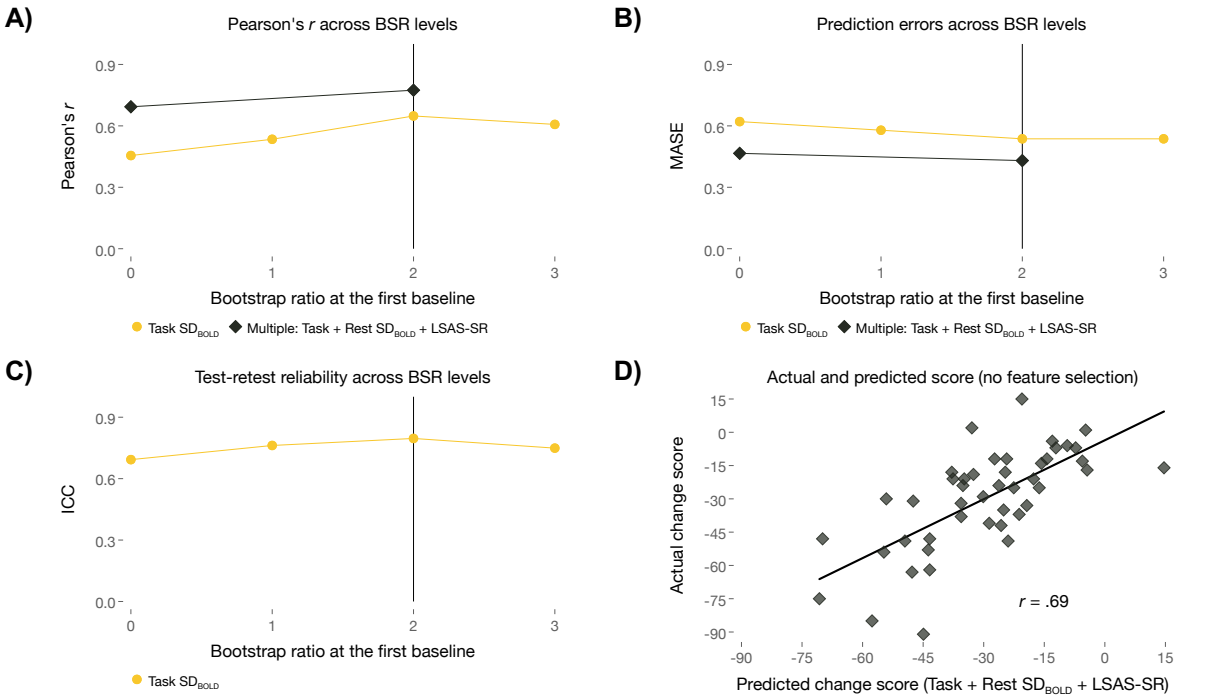

Figure S11. Prediction accuracies and test-retest reliabilities. A) depicts Pearson's $r_{ACTUAL,PREDICTED}$, **B)** MASE, and **C)** ICC across differently thresholded behavioral PLS models based on task-related variability (task SD_{BOLD}). The multiple prediction model (denoted with black diamonds) is based on the zero-order variables found to significantly predict treatment outcome (i.e., 160 sec task SD_{BOLD}, 340 sec resting-state SD_{BOLD}, and self-reported pre-treatment social anxiety at the second baseline), and in this forward prediction model a threshold was not set (i.e., no "feature selection") based on the outcome from the first baseline (see detailed description in the main manuscript). **D)** demonstrates the correlation between the actual social anxiety change score, and the predicted social anxiety change score based on multiple predictors (i.e., task SD_{BOLD}, resting-state SD_{BOLD}, and pre-treatment LSAS-SR). **Abbreviations:** MASE, Mean absolute scaled error; ICC, Intraclass correlation coefficient; BSR, Bootstrap ratio; PLS, Partial least squares; LSAS-SR, Liebowitz social anxiety scale, self-report version; SD_{BOLD}, Standard deviation of BOLD; BOLD, Blood-oxygen-level-dependent imaging;

Task-based signal variability prediction specifically generalizes across secondary social anxiety measures

As the task SD_{BOLD} condition was the strongest treatment outcome predictor, we investigated if this predictor also generalized across secondary social anxiety outcomes. To do so, we first estimated a single principal component analysis (PCA) representing all available secondary outcomes (i.e., self-reported SIAS, SPS, SPSQ, and the clinician-administered CGI-I; eigenvalue = 2.80, 70% explained variance and component loadings varied from .72 to .89). This component score correlated strongly with the task-based SD_{BOLD} predicted social anxiety change score used in our primary (LSAS-SR) analyses above (R^2 = 34%, permuted p < .001). Further, as Table S5 and Figure S8B depict, symptoms of depression (permuted p < .001) and insomnia (permuted p = .042) decreased over the course of therapy. However, the socio-affective face task SD_{BOLD} predicted social anxiety change score was not associated with reductions in depressive or insomnia symptoms (all permuted ps > .351).

Unique and shared neural variability between the socio-affective task and resting-state fMRI

Figure S12 displays the unique and overlapping brain regions comparing SD_{BOLD} for task and resting-state. Specifically, purple color depicts overlapping task- and resting-state related variability predicting treatment outcome. Blue color depicts unique variability within the task condition, and yellow color depicts unique variability as determined by the longer resting-state condition. In conclusion, the two conditions are largely independent and non-overlapping.

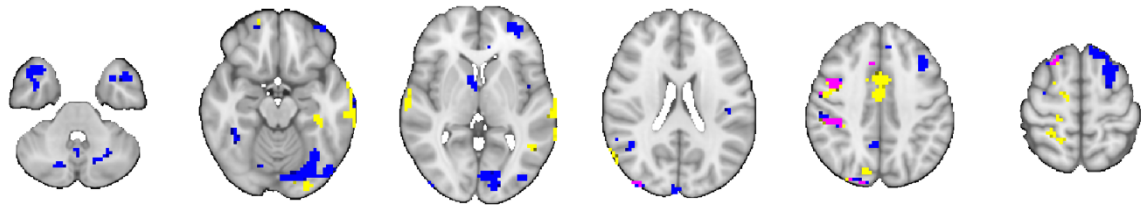

Figure S12. Task and resting-state SD_{BOLD} overlap at the first baseline. The figure demonstrates overlapping (purple) and unique neural activity for SD_{BOLD} task (160 sec condition; blue) and SD_{BOLD} resting-state (340 sec condition; yellow) associated with treatment outcome. The spatial pattern for each condition represents a binary mask and thus, direction (low/high variability) cannot be inferred. **Abbreviations:** SD_{BOLD}, standard deviation of BOLD; BOLD, Blood-oxygen-level-dependent imaging;

Test-retest reliability

The 11-week test-retest reliability (Baseline 1 vs Baseline 2) was excellent for the task-related SD_{BOLD} prediction model, and the reliability improved as a function of data volume. Specifically, the intraclass correlation coefficient ($ICC_{B1,B2}$) was lower for the shortest SD_{BOLD} condition (40 sec), and reliability was larger when 160 sec stimuli was included, see details in Table S10. When data volume was equated across conditions, resting-state SD_{BOLD} showed less reliability than task-related variability, and only when the full resting-state data volume was examined (340 sec, representing more than double the available task SD_{BOLD} data volume) did reliability improve to the level achieved by the 160 sec task SD_{BOLD} (resting-state $ICC_{B1,B2}$=0.81, CI=[0.74, 0.89]). In contrast, task-related $MEAN_{BOLD}$ showed poor reliability (all ICC's$_{B1,B2}$ ~0) across all conditions.

Table S10. Test-retest reliability of brain scores. Test-retest reliability of the brain scores included in each treatment outcome prediction model. Each ICC computation included 45 patients.

			95% CI	
Predictor	**Data volume (seconds)**	**ICC**	**Lower**	**Upper**
Task SD_{BOLD}	40	0.62	0.44	0.80
	80	0.78	0.66	0.89
	160	0.80	0.70	0.90
Resting-state SD_{BOLD}	40	0.36	0.11	0.62
	80	0.56	0.36	0.76
	160	0.67	0.52	0.82
	340	0.81	0.74	0.89
Task $MEAN_{BOLD}$	40	-0.15	-0.43	0.14
	80	-0.12	-0.38	0.14
	160	-0.05	-0.38	0.28

Abbreviations: ICC, Intraclass correlation coefficient; BOLD, Blood-oxygen level-dependent; LB, Lower bound confidence intervals; Avg, Average; SD_{BOLD}, Standard deviation of BOLD; $MEAN_{BOLD}$, Average BOLD activity; BOLD, Blood-oxygen-level-dependent imaging;

In addition to the reliability of the prediction models, voxel-wise ICC's are displayed in Table S11. In contrast to the previously presented reliability of the treatment outcome predictions, here, all voxels across the whole brain were included from each time-point (Baseline 1 and Baseline 2). Thus, these models represent a global measure of reliability across the whole-brain, rather than reliability of the condition (i.e., task or resting-state), or the treatment outcome prediction model.

Table S11. Test-retest reliability across all voxels. Test-retest reliability across all voxels in the whole-brain (k = 51.609). Average ICCs across 45 patients' whole-brain voxels are displayed below.

Predictor	Data volume (seconds)	Avg ICC	95% CI Lower	95% CI Upper
Task SD_{BOLD}	40	0.38	0.11	0.60
	80	0.46	0.20	0.65
	160	0.53	0.30	0.71
Resting-state SD_{BOLD}	40	0.35	0.08	0.58
	80	0.48	0.23	0.67
	160	0.58	0.36	0.74
	340	0.62	0.41	0.77
Task $MEAN_{BOLD}$	40	0.03	-0.26	0.31
	80	0.03	-0.25	0.31
	160	0.08	-0.21	0.35

Abbreviations: ICC, Intraclass correlation coefficient; LB, Lower bound confidence intervals; Avg, Average; SD_{BOLD}, Standard deviation of BOLD; $MEAN_{BOLD}$, Average BOLD activity; BOLD, Blood-oxygen-level-dependent imaging;

In addition to ICCs, below we also report Pearson r correlations between the first and second baseline. First, the average whole-brain voxel-wise correlation was r = .56 for the SD_{BOLD} socio-affective task condition, r = .63 for SD_{BOLD} resting-state, and r = .08 for the $MEAN_{BOLD}$ socio-affective task condition. Below, we display whole-brain voxel-wise Pearson's r (± 0.30; Figure S13A-C) and histograms of Pearson's r across all three conditions (Figure S13D).

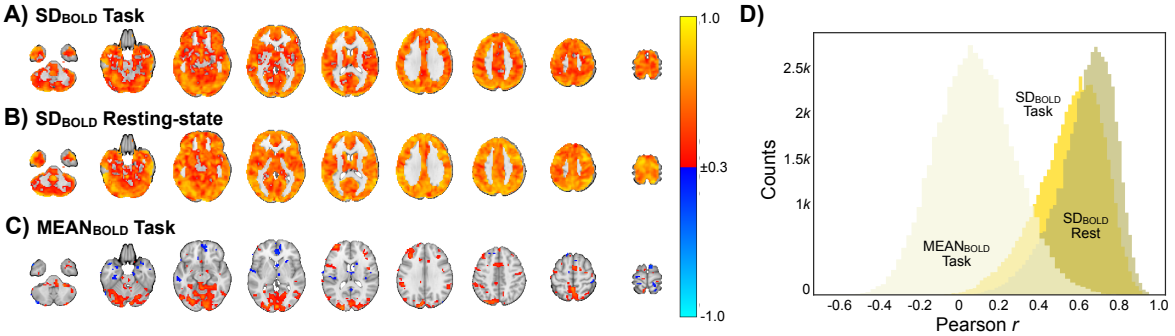

Figure S13. Whole-brain, voxel-wise correlations between baseline 1 and baseline 2. A) SD_{BOLD} socio-affective task condition. **B)** SD_{BOLD} Resting-state. **C)** $MEAN_{BOLD}$ socio-affective task condition. All maps are thresholded at r ± .30. **D)** Histograms of r values for each of the three conditions: $MEAN_{BOLD}$ socio-affective task condition (beige/cream color); SD_{BOLD} socio-affective task condition (yellow); SD_{BOLD} resting-state (khaki). **Abbreviations:** SD_{BOLD}, neural variability; $MEAN_{BOLD}$, conventional average neural response;

In addition to the previous whole-brain maps, we display cluster-wise voxel Pearson's *r* below (± 0.30; Figure S14). The clusters were defined by the behavioral PLS model (LSAS-SR post-treatment-baseline 1 delta score) with brain data from the first baseline (BSR > 2.0; extent threshold at 20 voxels).

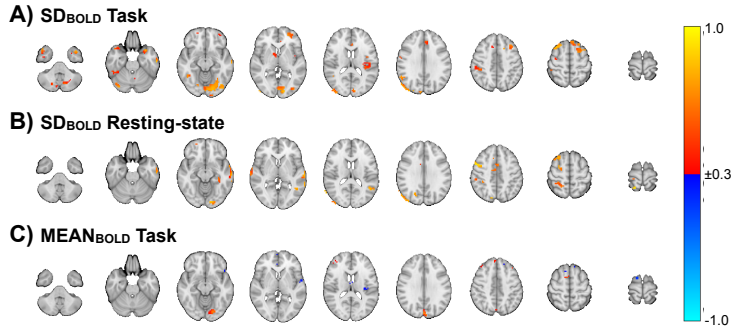

Figure S14. Cluster-level, voxel-wise correlations (Pearson *r*) between baseline 1 and baseline 2. A) SD_{BOLD} socio-affective task condition. **B)** SD_{BOLD} Resting-state. **C)** $MEAN_{BOLD}$ socio-affective task condition. All maps are thresholded at *r* ± .30. **Abbreviations:** SD_{BOLD}, neural variability; $MEAN_{BOLD}$, conventional average neural response;

We display the average Pearson's *r* for each cluster and condition in Figure S15. Each cluster was defined by a PLS prediction model with data from the first baseline (BSR ± 2; extent threshold at 20 voxels). As displayed in Figure S15C, the $MEAN_{BOLD}$ socio-affective task condition shows poor correlations between the first and second baseline for most of the clusters. In contrast, and as displayed by the diagonal red line in Figure S15A (SD_{BOLD} socio-affective task) and Figure S15B (SD_{BOLD} resting-state), each cluster shows relatively strong correlations between baselines for SD_{BOLD}.

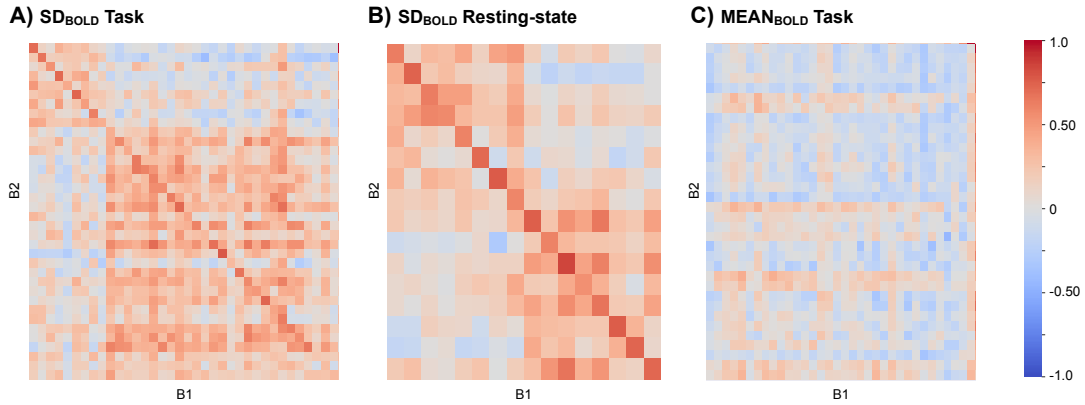

Figure S15. Cluster-wise correlations between baselines 1 and 2. A) SD_{BOLD} socio-affective task. **B)** SD_{BOLD} Resting-state. **C)** $MEAN_{BOLD}$ socio-affective task. **Abbreviations:** B1, baseline 1; B2, baseline 2; SD_{BOLD}, neural variability; $MEAN_{BOLD}$, conventional average neural response;

Control analyses

Additional control analyses were performed because of the apparent low test-retest reliability on the $MEAN_{BOLD}$ socio-affective task condition. Specifically, we used the general linear model (GLM) framework within SPM12 (www.fil.ion.ucl.ac.uk/spm/software/spm12/), and used the preprocessed and manually denoised images (same as we used for PLS analyses) as input in first level models (each baseline separately). Each subject's model included two onsets of socio-affective faces (with a duration of 80 seconds and fixation crosses as implicit baselines) and convolved with the canonical hemodynamic response function. Each subject's $MEAN_{BOLD}$ data were then implemented in a (second-level) one-sample *t*-test in order to examine average neural response to faces, relative to fixations.

The face > fixation contrast ($MEAN_{BOLD}$ at the first baseline; B1) revealed a topographical pattern commonly associated with face perception. Whole-brain peak-level voxels (family-wise error corrected (FWE) $p < .05$, $T > 5.25$) include the occipital lobe, fusiform gyrus, amygdala, hippocampus, and dorsomedial prefrontal cortex (see also Figure S16A and Table S12 below). Further, by use of the NeuroVault decoder (Gorgolewski et al., 2015), the following terms are associated with the found statistical map: occipital, v1, face, fusiform, visual cortex, early visual.

The same contrast (face > fixation) resulted in comparable group-level activation maps (FWE $p < .05$, T > 5.58) when data from the second baseline (B2) was used (see Figure S16B and Table S12 below). Despite overall convergence, B1 and B2 maps displayed considerable differences, especially in non-occipital regions (e.g., medial and dorsomedial prefrontal cortex, and amygdala). Unthresholded statistical maps are available online (https://neurovault.org/collections/9030/).

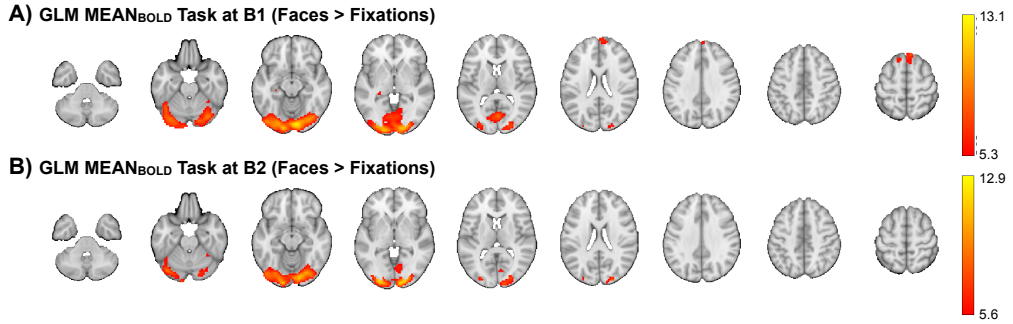

Figure S16. $MEAN_{BOLD}$ response to faces > fixations. A) depicts a whole-brain map ($T > 5.25$; FWE $p < .05$) demonstrating peak voxels for the $MEAN_{BOLD}$ condition at the first baseline scanning session, comparing faces to fixations as the implicit baseline. **B)** depicts a whole-brain map ($T > 5.58$; FWE $p < .05$) demonstrating peak voxels for the $MEAN_{BOLD}$ condition at the second baseline scanning session, comparing faces to fixations as the implicit baseline. See also Table S12 for details. **Abbreviations:** $MEAN_{BOLD}$: Conventional average neural response; FWE: Family-wise error corrected;

Table S12. $MEAN_{BOLD}$ response to faces, relative to fixations. The table displays voxel-wise whole-brain effect of larger $MEAN_{BOLD}$ responses to the task (faces) vs. fixation, at the first and second baseline. Voxels surviving whole-brain family-wise error correction (FWE $p < .05$) are reported.

Faces > Fixations (B1)

			MNI			
***p* (FWE)**	**mm^3**	***T***	***x***	***y***	***z***	**Anatomical location (AAL)**
<.001	75735	13.14	-12	-96	-6	Calcarine L
<.001	756	9.23	-27	-27	-6	Hippocampus L
<.001	2160	7.67	12	36	57	Frontal Superior Medial R
.001	1809	6.71	9	60	30	Frontal Superior Medial R
.002	594	6.31	-9	27	60	Frontal Superior Medial L
.002	189	6.30	-18	-6	-18	Amygdala L
.007	162	5.87	21	-30	-6	Hippocampus R
.019	189	5.52	36	-48	-24	Cerebelum 6 R
.024	216	5.45	-39	24	-18	Inferior Frontal Gyrus, Orbital L
.041	54	5.25	-48	9	-36	Temporal Inferior L

Faces > Fixations (B2)

			MNI			
***p* (FWE)**	**mm^3**	***T***	***x***	***y***	***z***	**Anatomical location (AAL)**
<.001	59157	12.71	-21	-93	-3	Middle Occipital Gyrus L
.012	81	5.58	-21	-30	-6	Hippocampus L

Abbreviations: L: Left; R: Right; FEW: Family-wise error corrected *P*; AAL: Automated anatomical labeling atlas; MNI: Montreal Neurological Institute coordinates; *k*: cluster size; $MEAN_{BOLD}$: conventional average neural response; B1: Baseline 1; B2: Baseline 2;

SUPPLEMENTARY REFERENCES

1. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, *et al.* (1998): The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 59 Suppl 20: 22-33;quiz 34-57.
2. First MB, Spitzer RL, Gibbon M, Williams JBW (2012): *Structured Clinical Interview for DSM-IV® Axis I Disorders (SCID-I), Clinician Version, Administration Booklet*. American Psychiatric Pub.
3. Svanborg P, Asberg M (2001): A comparison between the Beck Depression Inventory (BDI) and the self-rating version of the Montgomery Asberg Depression Rating Scale (MADRS). *J Affect Disord* 64: 203–216.
4. Månsson KNT, Lindqvist D, Yang LL, Svanborg C, Isung J, Nilsonne G, *et al.* (2019): Improvement in indices of cellular protection after psychological treatment for social anxiety disorder. *Transl Psychiatry* 9: 340.
5. Baker SL, Heinrichs N, Kim H-J, Hofmann SG (2002): The Liebowitz social anxiety scale as a self-report instrument: a preliminary psychometric analysis. *Behav Res Ther* 40: 701–715.
6. Mattick RP, Clarke JC (1998): Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behav Res Ther* 36: 455–470.
7. Furmark T, Tillfors M, Everz P, Marteinsdottir I, Gefvert O, Fredrikson M (1999): Social phobia in the general population: prevalence and sociodemographic profile. *Soc Psychiatry Psychiatr Epidemiol* 34: 416–424.
8. Zaider TI, Heimberg RG, Fresco DM, Schneier FR, Liebowitz MR (2003): Evaluation of the clinical global impression scale among individuals with social anxiety disorder. *Psychol Med* 33: 611–622.
9. Bastien CH, Vallières A, Morin CM (2001): Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep Med* 2: 297–307.
10. Wortman PM (1983): Evaluation Research: A Methodological Perspective. *Annu Rev Psychol* 34: 223–260.
11. Mayo-Wilson E, Dias S, Mavranezouli I, Kew K, Clark DM, Ades AE, Pilling S (2014): Psychological and pharmacological interventions for social anxiety disorder in adults: a systematic review and network meta-analysis. *Lancet Psychiatry* 1: 368–376.
12. Akerstedt T, Gillberg M (1990): Subjective and objective sleepiness in the active individual. *Int J Neurosci* 52: 29–37.
13. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM (2012): FSL. *Neuroimage* 62: 782–790.
14. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, *et al.* (2004): Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23 Suppl 1: S208-19.
15. Garrett DD, Kovacevic N, McIntosh AR, Grady CL (2011): The importance of being variable. *J Neurosci* 31: 4496–4503.
16. Garrett DD, Kovacevic N, McIntosh AR, Grady CL (2010): Blood oxygen level-dependent signal variability is more than just noise. *J Neurosci* 30: 4914–4921.
17. Garrett DD, Nagel IE, Preuschhof C, Burzynska AZ, Marchner J, Wiegert S, *et al.* (2015): Amphetamine modulates brain signal variability and working memory in younger and older adults. *Proc Natl Acad Sci U S A* 112: 7593–7598.
18. Beckmann CF, Smith SM (2004): Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans Med Imaging* 23: 137–152.

19. Birn RM (2012): The role of physiological noise in resting-state functional connectivity. *Neuroimage* 62: 864–870.
20. Smith AM, Lewis BK, Ruttimann UE, Ye FQ, Sinnwell TM, Yang Y, *et al.* (1999): Investigation of low frequency drift in fMRI signal. *Neuroimage* 9: 526–533.
21. Garrett DD, Kovacevic N, McIntosh AR, Grady CL (2013): The modulation of BOLD variability between cognitive states varies by age and processing speed. *Cereb Cortex* 23: 684–693.
22. Garrett DD, Grady CL, Hasher L (2010): Everyday memory compensation: the impact of cognitive reserve, subjective memory, and stress. *Psychol Aging* 25: 74–83.
23. McIntosh AR, Bookstein FL, Haxby JV, Grady CL (1996): Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage* 3: 143–157.
24. Krishnan A, Williams LJ, McIntosh AR, Abdi H (2011): Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. *Neuroimage* 56: 455–475.
25. Hyndman RJ, Koehler AB (2006): Another look at measures of forecast accuracy. *Int J Forecast* 22: 679–688.
26. Franses PH (2016): A note on the Mean Absolute Scaled Error. *Int J Forecast* 32: 20–22.
27. McGraw KO, Wong SP (1996): Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1: 30–46.
28. Salarian A (2016): Intraclass correlation coefficient (ICC). *MATLAB Cent File Exch*. Retrieved January 26, 2020, from https://de.mathworks.com/matlabcentral/fileexchange/22099-intraclass-correlation-coefficient-icc
29. Cicchetti DV, Sparrow SA (1981): Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic* 86: 127–137.