# Supplementary Information for

What is the test-retest reliability of common task-fMRI measures? New empirical evidence and a meta-analysis

Maxwell L. Elliott1†, Annchen R. Knodt1†, David Ireland2, Meriwether L. Morris1, Richie Poulton2, Sandhya Ramrakha2, Maria L. Sison1, Terrie E. Moffitt1,3-5, Avshalom Caspi1,3-5, Ahmad R. Hariri1\*

†These authors contributed equally to this work.

\*Correspondence: Ahmad Hariri

Email: ahmad.hariri@duke.edu

**This file includes:**

Supplementary Methods

Figs. S1 to S7

Tables S1 to S4

References for SI reference citations

## 

## Supplementary Methods

## Meta-analysis

We identified 56 articles meeting the inclusion criteria that were all included in the final meta-analysis [(Ances, Vaida, Ellis, & Buxton, 2011; Aron, Gluck, & Poldrack, 2006; Atri et al., 2011; Baumgartner et al., 2017; Bennett & Miller, 2013; Blokland et al., 2017; Brandt et al., 2013; Bunford, Kinney, Michael, & Klumpp, 2017; Caceres, Hall, Zelaya, Williams, & Mehta, 2009; Cannon, Cao, Mathalon, Gee, & on behalf of the NAPLS consortium, 2018; Chase et al., 2015; Clément & Belleville, 2009; Drew Sayer et al., 2016; Erk et al., 2013; Estévez et al., 2014; Fliessbach et al., 2010; Fournier, Chase, Almeida, & Phillips, 2014; Friedman et al., 2008; Haller et al., 2018; Heckendorf, Bakermans‐Kranenburg, van Ijzendoorn, & Huffmeijer, 2019; Holiga et al., 2018; Johnstone et al., 2005; Keren et al., 2018; Kimberley, Birkholz, Hancock, VonBank, & Werth, 2008; Kimberley, Khandekar, & Borich, 2008; Kong et al., 2007; Lee et al., 2010; Lim, Choo, & Chee, 2007; Lipp, Murphy, Wise, & Caseras, 2014; Lois, Kirsch, Sandner, Plichta, & Wessa, 2018; Manoach et al., 2001; Mohnke et al., 2014; Nettekoven, Reck, Goldbrunner, Grefkes, & Lucas, 2018; Nord, Gray, Charpentier, Robinson, & Roiser, 2017; Plichta et al., 2014, 2012; Putcha et al., 2011; Quiton, Keaser, Zhuo, Gullapalli, & Greenspan, 2014; Raemaekers et al., 2007; Rath et al., 2016; Sauder, Hajcak, Angstadt, & Phan, 2013; Schacht et al., 2011; Schlagenhauf et al., 2008; Schunck et al., 2008; Schuster et al., 2017; Sheu, Jennings, & Gianaros, 2012; Sousa, Vilela, & Figueiredo, 2014; Taylor & Davis, 2009; Upadhyay et al., 2015; van den Bulk et al., 2013; Wehrum-Osinsky et al., 2014; Weiss et al., 2013; Wei et al., 2004; Wende et al., 2017; White et al., 2016; Zanto, Pa, & Gazzaley, 2014)](https://paperpile.com/c/J9j3wd/ycDX+N2yu+CgHF+3Wfc+KLPT+d9Dw+ApcT+FPrF+vMr9+dLPF+j8i8+oZKM+TlMI+uJYs+akLA+1CQf+TyCM+FeD0+7QVi+254T+XYy5+5bqC+rdHz+fhcI+WMHV+HrRE+uXiU+Gfad+l37L+WFG6+llrx+mLaj+NcQe+nQ6B+6L2r+4Vy3+Rtd2+gzYp+aokv+w4g6+1PHK+IJCB+oOtx+AOWe+yoIi+1MfY+GX8g+wqAG+GtKD+aBKu+YKyr+ZRVe+TIUe+caeO+3m8W+i2ly).

## Datasets

***Human Connectome Project (HCP).***The MRI acquisition parameters and minimal preprocessing of these data have been described extensively elsewhere [(Glasser et al., 2013)](https://paperpile.com/c/J9j3wd/pEgz). Briefly, participants underwent extensive MRI measurement that included T1-weighted structural imaging and nearly two hours of fMRI scanning during resting-state and seven tasks. Task-fMRI is described extensively elsewhere [(Barch et al., 2013)](https://paperpile.com/c/J9j3wd/cLQb1). For our analyses we used the minimally preprocessed data in volumetric Montreal Neurological Institute (MNI) space (“fMRIVolume” pipeline).

Subjects who did not complete a task at one or both time points were removed from analysis for that task. Thus, the final test-retest dataset used in our reliability analyses included 45 subjects for the reward, motor, and executive function tasks; 44 subjects for the social task; 43 subjects for the language task, and 42 subjects for the emotion and relational tasks.

***Dunedin Multidisciplinary Health and Development Study****.* The Dunedin Studyis a longitudinal investigation of health and behavior in a complete birth cohort of 1,037 individuals (91% of eligible births; 52% male) born between April 1972 and March 1973 in Dunedin, New Zealand (NZ), and eligible based on residence in the province and participation in the first assessment at age 3. The cohort represents the full range of socioeconomic status on NZ’s South Island and matches the NZ National Health and Nutrition Survey on key health indicators (e.g., BMI, smoking, GP visits) [(Poulton, Moffitt, & Silva, 2015)](https://paperpile.com/c/J9j3wd/liJ9y). The cohort is primarily white (93%). Assessments were carried out at birth and ages 3, 5, 7, 9, 11, 13, 15, 18, 21, 26, 32, 38, and 45 years, when 94% of the 997 study members still alive took part, of whom 93% participated in MRI scanning.

The acquisition parameters and task-fMRI measures for the Dunedin Study have been described in detail elsewhere [(Elliott et al., 2019)](https://paperpile.com/c/J9j3wd/LmS0b). 20 study members completed the entire scan protocol a second time (mean days between scans = 79). These 20 study members made up the Dunedin test-retest sample used in this manuscript. All 20 study members completed all 4 fMRI tasks at both timepoints.

## Image Pre-processing

Minimal preprocessing was first applied to all data. For the HCP dataset this was done with the HCP minimal preprocessing pipeline [(Glasser et al., 2013)](https://paperpile.com/c/J9j3wd/pEgz). This includes correction for B0 distortion, realignment to the single-band reference image to correct for motion, registration to the subject’s structural scan, normalization to the 4D mean, brain masking, and non-linear warping to MNI space. All single-band reference images were visually inspected to ensure proper alignment to the anatomical image.

Similar “minimal preprocessing” steps were applied to the Dunedin Study dataset using custom processing scripts. Anatomical images for each subject were skull-stripped, intensity-normalized, and nonlinearly warped to a study-specific average template in the standard stereotactic space of the Montreal Neurological Institute template using the ANTs SyN registration algorithm [(Avants, Epstein, Grossman, & Gee, 2008; Klein et al., 2009)](https://paperpile.com/c/J9j3wd/Tt8w+2zcC). Time series images for each subject were despiked, slice-time-corrected, realigned to the first volume in the time series to correct for head motion using AFNI tools [(Cox, 1996)](https://paperpile.com/c/J9j3wd/2EEi), corrected for B0 distortions using SPM's fieldmap toolbox [(Jezzard & Balaban, 1995)](https://paperpile.com/c/J9j3wd/3V4B), coregistered to the anatomical image using FSL’s Boundary Based Registration [(Greve & Fischl, 2009)](https://paperpile.com/c/J9j3wd/2iAE), and spatially normalized into MNI space using the non-linear ANTs SyN warp from the anatomical image. All transformations were concatenated so that a single interpolation was performed. The 4D means from each normalized time series were visually inspected to ensure proper alignment to the anatomical template. Following “minimal preprocessing,” normalized time series from both datasets were smoothed to minimize noise and residual difference in gyral anatomy with a Gaussian filter, set at 6-mm full-width at half-maximum. Voxel-wise signal intensities were scaled to yield a time series mean of 100 for each voxel, allowing parameter estimates from the general linear model that was subsequently applied to be interpreted as percent signal change (PSC). The AFNI program 3dREMLfit (Cox, 1996) was used to fit a general linear model for first-level task-fMRI data analyses. Linear contrasts employing canonical hemodynamic response functions were used to estimate condition of interest effects for each task, while controlling for low frequency noise using a number of polynomial regressors appropriate for the length of each task. Contrasts of interest for each task were as follows in the HCP: “Faces > Shapes” for the emotion task, “Win > Loss” for the reward task, “2-back > 0-back” for the executive function N-back task, “Motor > Fixation” for the motor task, “Relation > Match” for the relational task, “Mental > Random” for the social task, and “Story > Math” for the language task; and in the Dunedin Study: “Faces > Shapes” for the emotion task, “Gain Anticipation > Neutral Anticipation” for the reward task, “Incongruent > Congruent” for the executive function Stroop task, and “Encoding > Distractor” for the episodic memory task. Volumes exhibiting excessive motion were identified with framewise displacement and standardized DVARS (Nichols, 2017; Power et al., 2014) thresholds specific to each dataset (0.39 mm and 4.9 for the HCP (Burgess et al., 2016) and 0.5 mm and 2.5 for the Dunedin Study, respectively) and subsequently censored from the GLM. To screen for signal dropout, for each subject we calculated the mean temporal SNR for each task within the target region and tested whether this value was an outlier (defined as < 3 standard deviations below the mean across subjects for the task). One subject in the Dunedin Study was identified as an outlier for tSNR in the reward task (ventral striatum); sensitivity analyses conducted without this subject resulted in a lower ICC for the task, leaving the conclusions unchanged.

Additionally, parcel-wise measures of cortical thickness and surface area in both datasets were extracted using the HCP’s “PostFreeSurfer” pipeline applied to the T1- and T2-weighted structural scans. Grey matter volumes for 17 regions of interest were extracted separately from the automatic segmentation (“aseg”) step of FreeSurfer version 6.0. FreeSurfer version 6.0 was used because the HCP FreeSurfer pipeline was optimized for the cortical surface, resulting in lower-quality segmentation of subcortical volumes in our dataset.

## Anatomical ROI Definition

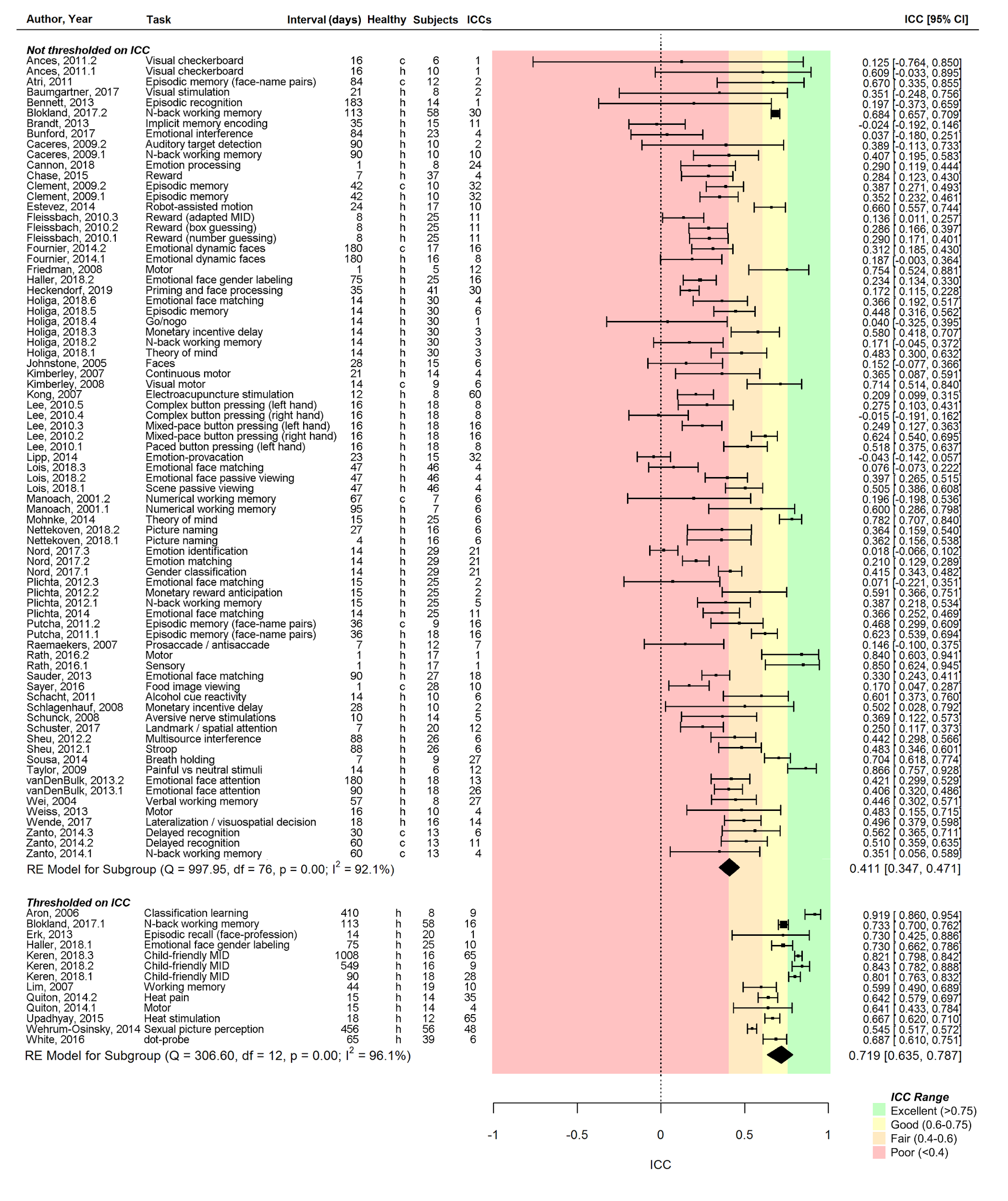
We identified the primary target region of interest (ROI) for each task for use in all analyses (see references for each task in the Methods section of the main text). For simplicity, we selected a single target region for the two tasks that are often used to target more than one region: 1) for the HCP social cognition task, we selected the fusiform gyrus over the other regions described in [(Wheatley, Milleville, & Martin, 2007)](https://paperpile.com/c/J9j3wd/c6GB) due to the presence of faces in other tasks, and 2) for the Dunedin executive function Stroop task, we selected the dorsolateral prefrontal cortex (dlPFC) over the dorsal anterior cingulate cortex to parallel the HCP executive function task.

Anatomical ROI masks for each ROI were selected based on alignment to the template as well as prior work for the respective task where applicable. We used anatomical definitions from the Automated Anatomical Labeling (AAL) atlas for the hippocampus and the fusiform gyrus (episodic memory and social cognition tasks), the Harvard Oxford Atlas (HOA) for the precentral gyrus and anterior temporal lobe (motor and language tasks; the latter consisted of the combined temporal pole, anterior superior temporal gyrus, anterior middle temporal gyrus, and anterior inferior temporal gyrus regions), the high-resolution amygdala template generated from 168 Human Connectome Project datasets (emotion tasks; Tyszka et al. 2016), and the Oxford-GSK-Imanova structural and connectivity striatal atlases distributed with FSL for the ventral striatum (reward tasks). The dlPFC (executive function task) was defined using the Brodmann Area Atlas provided by Wake Forest University PickAtlas (WFU PickAtlas; www.fmri.wfubmc.edu/downloads), with Brodmann Areas 9 and 46 combined, dilated by 2 voxels, and the medial aspect removed [(Tong et al., 2016)](https://paperpile.com/c/J9j3wd/Bs0R). The rostrolateral PFC (relational task) was defined as Brodmann Area 10 using the Brodmann Area Atlas provided with MRIcron (https://www.nitrc.org/projects/mricron). For our primary analyses, we combined the anatomical ROIs from the left and right hemisphere into a single bilateral mask. Further, all ROIs were masked to exclude non-grey matter voxels, using the grey matter probability map from the study-specific average template thresholded at 0.25.

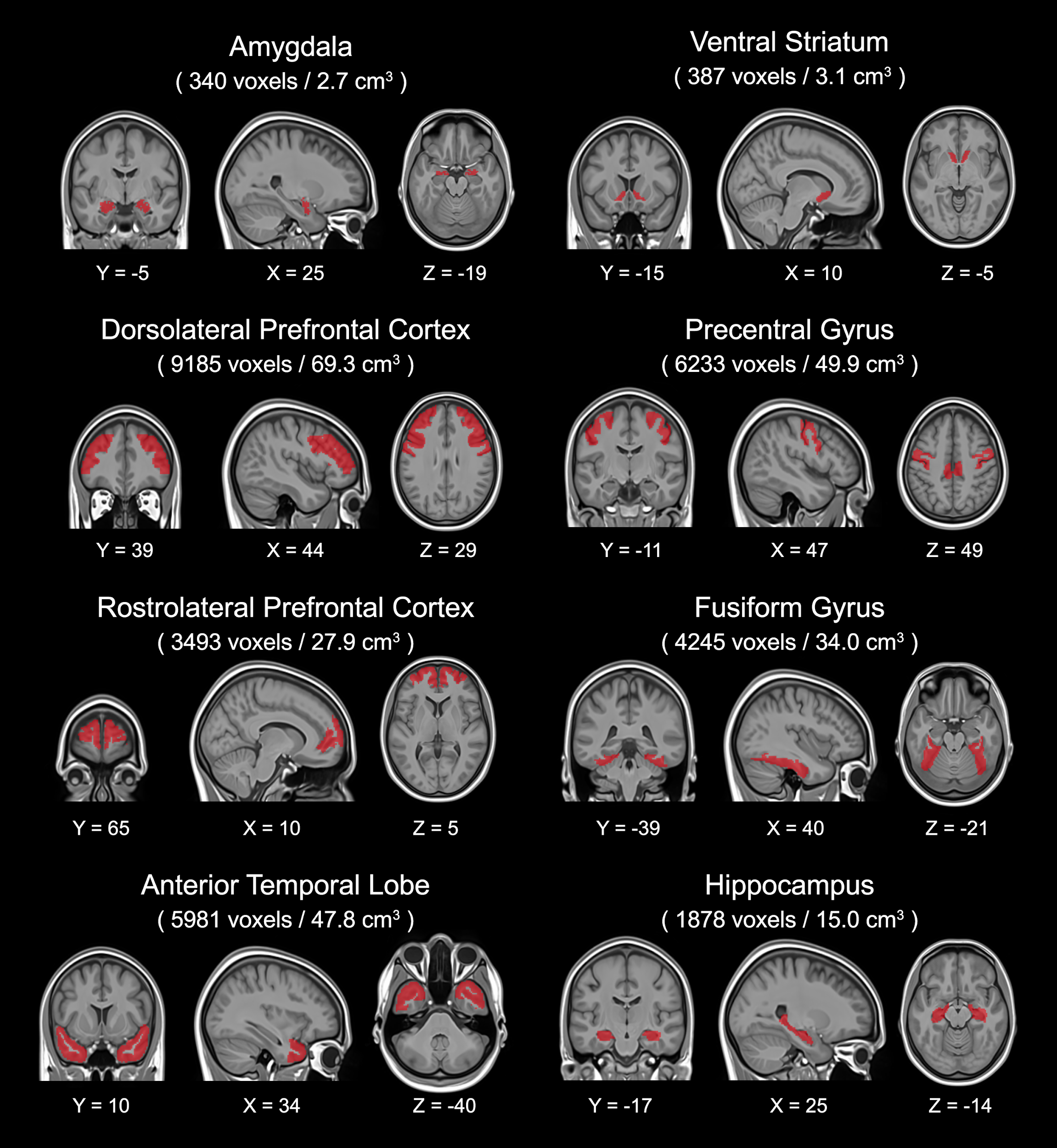
**Fig. S1.** PRISMA Checklist for systematic reviews and meta-analyses.

|  |  |  |  |
| --- | --- | --- | --- |
| **Section / topic** | **#** | **Checklist item** | **Reported on page #** |
| **TITLE** | | |  |
| Title | 1 | Identify the report as a systematic review, meta-analysis, or both. | 1 |
| **ABSTRACT** | | |  |
| Structured summary | 2 | Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number. | 2 |
| **INTRODUCTION** | | |  |
| Rationale | 3 | Describe the rationale for the review in the context of what is already known. | 3-5 |
| Objectives | 4 | Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS). | 4-5 |
| **METHODS** | | |  |
| Protocol and registration | 5 | Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number. | 5-6 |
| Eligibility criteria | 6 | Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale. | 5-6 |
| Information sources | 7 | Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched. | 5-6 |
| Search | 8 | Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated. | 5 |
| Study selection | 9 | State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis). | 5-6 |
| Data collection process | 10 | Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators. | 5-6 |
| Data items | 11 | List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made. | 6 |
| Risk of bias in individual studies | 12 | Describe the methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis. | 6-7 |
| Summary measures | 13 | State the principal summary measures (e.g., risk ratio, difference in means). | 5-6 |
| Synthesis of results | 14 | Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I2) for each meta-analysis. | 5-7 |
| Risk of bias across studies | 15 | Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies). | 5-7 |
| Additional analyses | 16 | Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified. | 6-7, S10 |
| **RESULTS** | | |  |
| Study selection | 17 | Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram. | Fig. 2 |
| Study characteristics | 18 | For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations. | Fig. 3, Table S1 |
| Risk of bias within studies | 19 | Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12). | 13, Fig S2 |
| Results of individual studies | 20 | For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot. | Fig. 3 |
| Synthesis of results | 21 | Present results of each meta-analysis done, including confidence intervals and measures of consistency. | 13, Fig. 3 |
| Risk of bias across studies | 22 | Present results of any assessment of risk of bias across studies (see Item 15). | Fig. S2 |
| Additional analysis | 23 | Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]). | 13, Fig. S2 |
| **DISCUSSION** | | |  |
| Summary of evidence | 24 | Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers). | 18-19 |
| Limitations | 25 | Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias). | 18-19 |
| Conclusions | 26 | Provide a general interpretation of the results in the context of other evidence, and implications for future research. | 24 |
| **FUNDING** | | |  |
| Funding | 27 | Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review. | 26 |

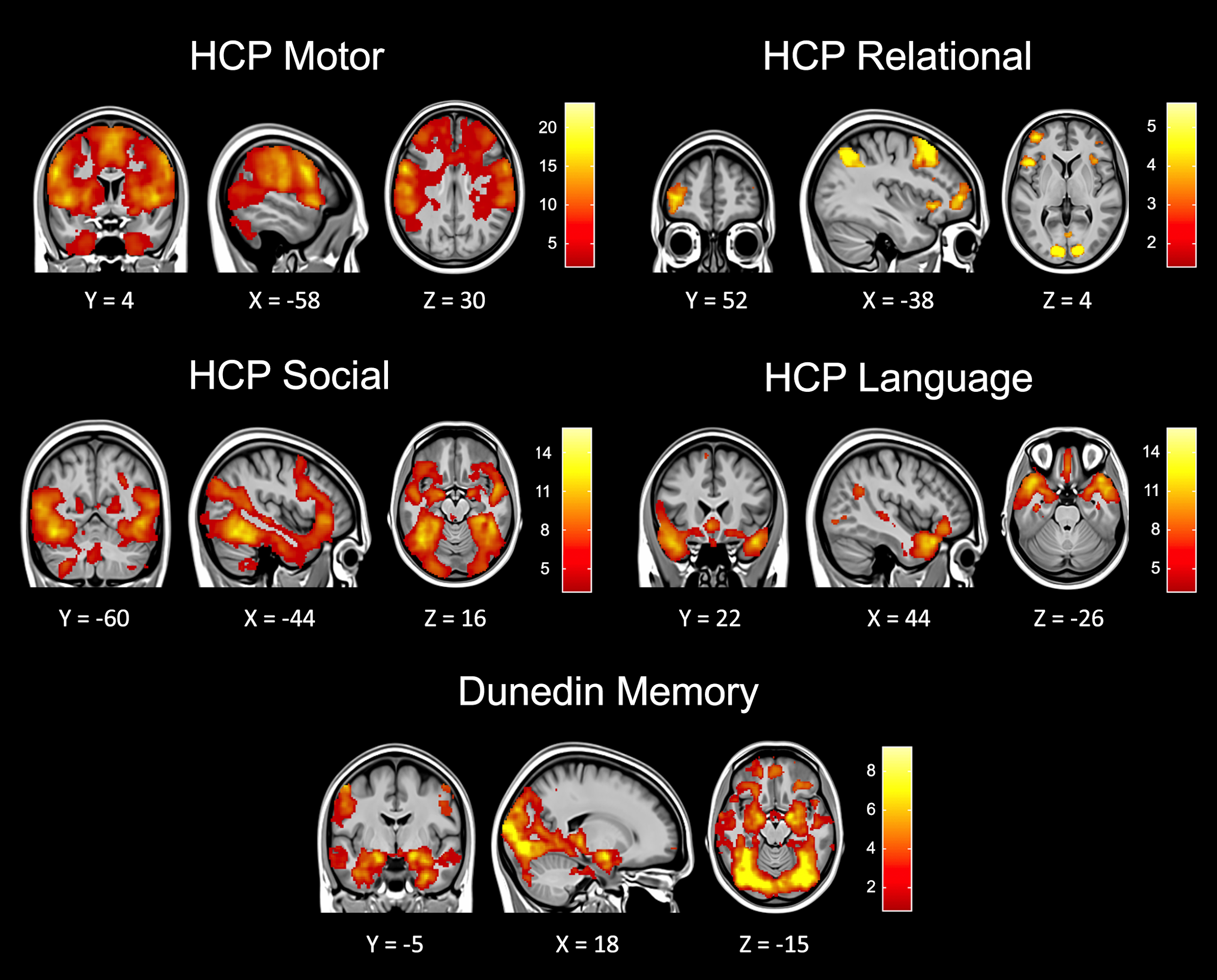
**Fig. S2.** Sensitivity analyses for the meta-analysis. Here the correlations between ICCs in each substudy were assumed to be 0, representing the extreme case where the ICCs within each substudy were completely independent.



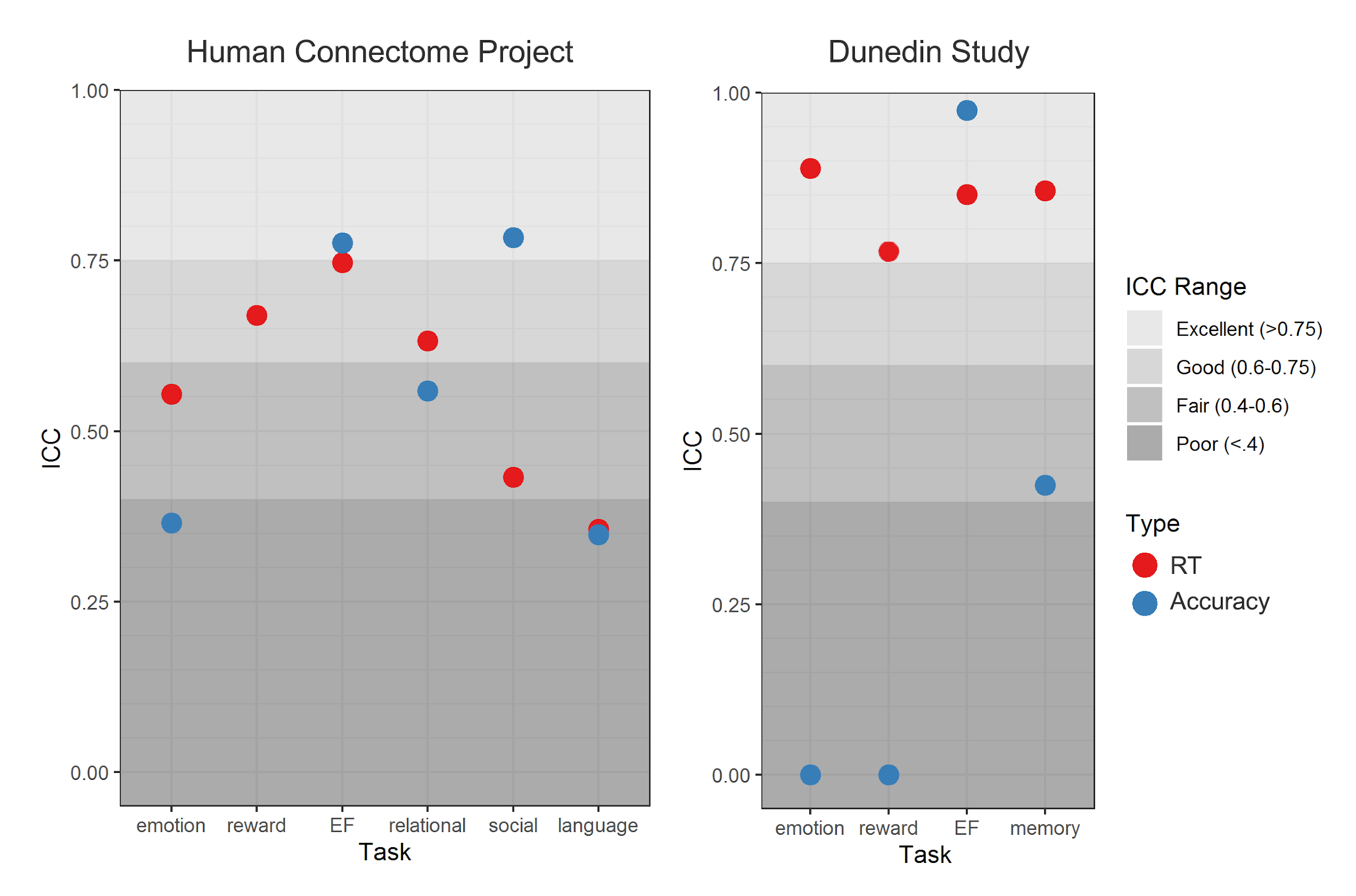
**Fig S3.** ROI masks (red) overlaid on the anatomical template.



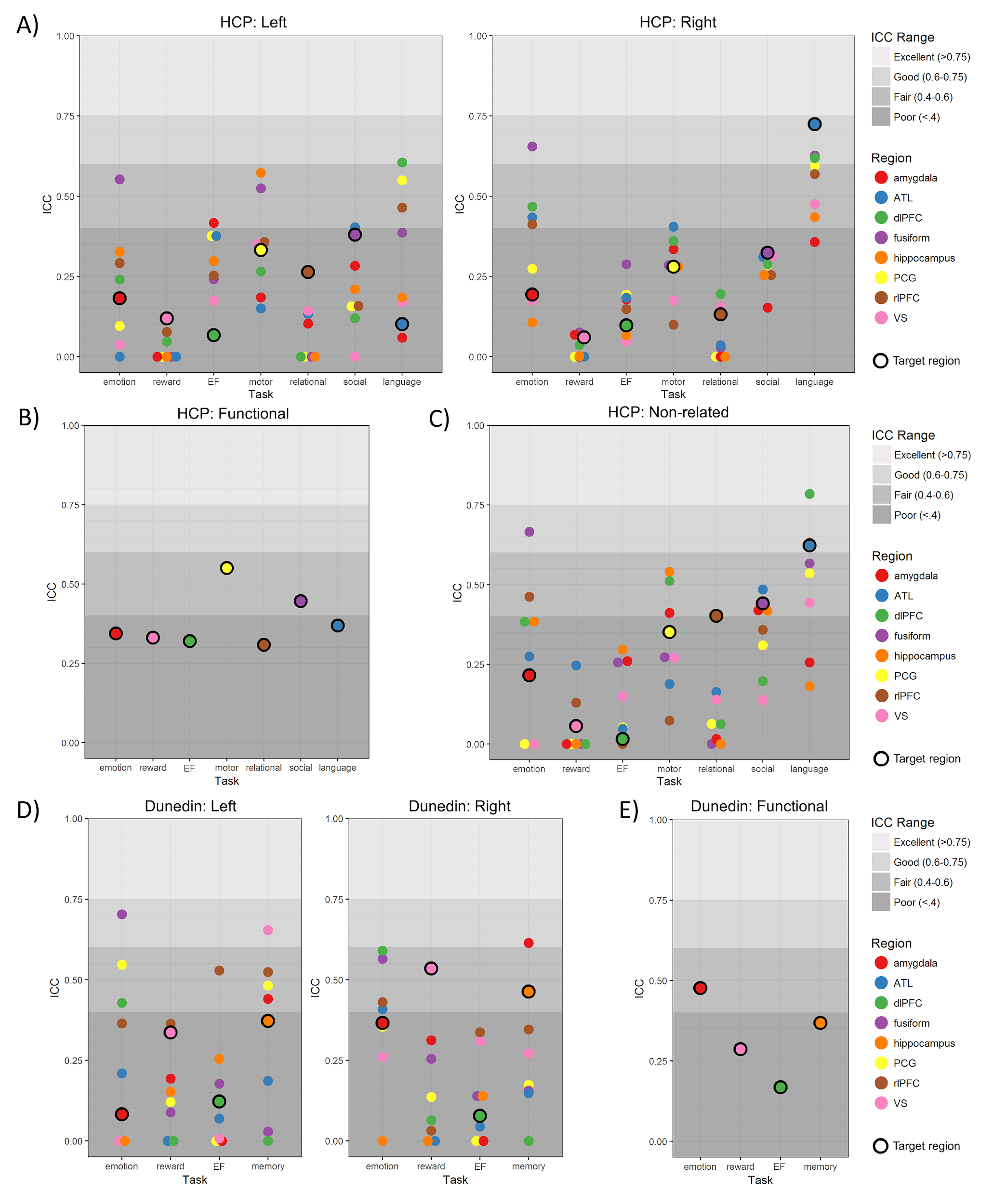
**Fig. S4.** Whole-brain activation maps for the five fMRI tasks not shown in Fig. 4 of the main text.



**Fig. S5.** ICCs for response time (RT) and accuracy from fMRI tasks. Each measure is given for the primary experimental condition of interest: the face-matching condition for the emotion tasks, the gain/reward condition for the reward tasks (note that accuracy is undefined for HCP’s task), the incongruent/2-back condition for the executive function (EF) tasks, the relation condition for the relational task, the theory of mind condition for the social task, the language condition for the language task, and the encoding condition for the memory task. The fMRI tasks in HCP and the Dunedin Study have widely variable ICCs. Of note, some tasks, like the emotion tasks, were designed to be very easy (i.e. have near ceiling accuracy); this limits between-subjects variance and leads to a low ICC. In this way, the low behavioral ICC was intended.



**Fig. S6.** Sensitivity analyses for test-retest reliability of region-wise activation measures in the Human Connectome Project (HCP) and the Dunedin Study. ICCs shown are calculated separately across hemispheres (A and D), within functionally defined ROIs (B and E), and in only the N=26 unrelated individuals in the HCP (C).



**Table S1.** Moderators tested in meta-analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Moderator** | **Level** | **Count** | **Mean (sd)** | **Min** | **Max** |
| *Number of citations per year* | - | - | 5.03 (4.38) | 0 | 23.7 |
| *Thresholded on ICC* | No | 840 | - | - | - |
| Yes | 306 | - | - | - |
| *Retest interval (days)* | - | - | 119.8 (242.3) | 1 | 1008 |
| *Sample type* | Healthy | 1036 | - | - | - |
| Clinical | 110 | - | - | - |
| *Task design* | Blocked | 859 | - | - | - |
| Event-related | 239 | - | - | - |
| Mixed | 48 | - | - | - |
| *Task length (minutes)* | - | - | 10.59 (6.45) | .9 | 31 |
| *Task type* | Emotion | 297 | - | - | - |
| Executive Control | 139 | - | - | - |
| Language | 12 | - | - | - |
| Memory | 134 | - | - | - |
| Motor | 124 | - | - | - |
| Pain | 177 | - | - | - |
| Reward | 210 | - | - | - |
| Sensory | 44 | - | - | - |
| Social | 9 | - | - | - |
| *ROI type* | Functional | 374 | - | - | - |
| ICC-based | 152 | - | - | - |
| Structural | 451 | - | - | - |
| Structural+Functional | 166 | - | - | - |
| Whole-brain | 3 | - | - | - |
| *ROI location* | Cortical | 744 | - | - | - |
| Subcortical | 399 | - | - | - |

**Table S2.** Activation in *a priori* anatomical target ROIs. Mean t-statistics are given for all voxels in each bilateral ROI, as well as left and right separately.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Human Connectome Project* | | | | |
| **Task (*ROI)*** | **Mean t**  **(bilateral)** | **Mean t**  **(left)** | **Mean t**  **(right)** | **Peak location**  **(x,y,z)** |
| Emotion *(Amygdala)* | 8.897404 | 9.118625 | 8.681328 | 18, -2, -18 |
| Reward *(Ventral Striatum)* | 5.721121 | 5.604154 | 5.831621 | -8, 12, -4 |
| Cognitive Control *(dlPFC)* | 3.264334 | 2.997859 | 3.519485 | 46, 30, 34 |
| Motor *(Precentral Gyrus)* | 7.594384 | 8.417101 | 6.747241 | -58, 4, 30 |
| Relational *(rlPFC)* | -0.23022 | -0.07973 | -0.36461 | -38, 52, 4 |
| Social *(Fusiform)* | 5.356837 | 5.977268 | 4.761796 | -44, -60, -16 |
| Language *(Anterior Temporal Lobe)* | 5.769964 | 6.041165 | 5.49485 | 44, 22, -26 |
| *Dunedin Study* | | | | |
| **Task *(ROI)*** | **Mean t**  **(bilateral)** | **Mean t**  **(left)** | **Mean t**  **(right)** | **Peak location**  **(x,y,z)** |
| Emotion *(Amygdala)* | 6.133734 | 5.877258 | 6.384245 | 18, -5, -15 |
| Reward *(Ventral Striatum)* | 4.20533 | 4.263221 | 4.150639 | 20, -5, -15 |
| Cognitive Control *(dlPFC)* | 3.163896 | 2.440572 | 3.884862 | -48, 23, 21 |
| Episodic Memory *(Hippocampus)* | 2.942571 | 2.794044 | 3.088901 | -12, 13, -5 |

## 

## 

**Table S3.** Test-retest reliabilities of grey matter volumes for 17 subcortical regions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Region*** | ***Human Connectome Project*** | | ***Dunedin Study*** | |
| ***Left*** | ***Right*** | ***Left*** | ***Right*** |
| Cerebellum | 0.982 | 0.983 | 0.891 | 0.94 |
| Thalamus | 0.889 | 0.875 | 0.919 | 0.971 |
| Caudate | 0.984 | 0.977 | 0.957 | 0.957 |
| Putamen | 0.934 | 0.979 | 0.917 | 0.973 |
| Pallidum | 0.791 | 0.815 | 0.901 | 0.922 |
| Hippocampus | 0.895 | 0.924 | 0.957 | 0.979 |
| Amygdala | 0.917 | 0.914 | 0.918 | 0.931 |
| Accumbens | 0.831 | 0.822 | 0.767 | 0.942 |
| Brainstem | 0.837 | | 0.979 | |

## References

[Ances, B., Vaida, F., Ellis, R., & Buxton, R. (2011). Test-retest stability of calibrated BOLD-fMRI in HIV- and HIV+ subjects. *NeuroImage*, *54*(3), 2156–2162.](http://paperpile.com/b/J9j3wd/uJYs)

[Aron, A. R., Gluck, M. A., & Poldrack, R. A. (2006). Long-term test–retest reliability of functional MRI in a classification learning task. *NeuroImage*, Vol. 29, pp. 1000–1006. https://doi.org/](http://paperpile.com/b/J9j3wd/N2yu)[10.1016/j.neuroimage.2005.08.010](http://dx.doi.org/10.1016/j.neuroimage.2005.08.010)

[Atri, A., O’Brien, J. L., Sreenivasan, A., Rastegar, S., Salisbury, S., DeLuca, A. N., … Sperling, R. A. (2011). Test-retest reliability of memory task functional magnetic resonance imaging in Alzheimer disease clinical trials. *Archives of Neurology*, *68*(5), 599–606.](http://paperpile.com/b/J9j3wd/1CQf)

[Avants, B., Epstein, C., Grossman, M., & Gee, J. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, Vol. 12, pp. 26–41. https://doi.org/](http://paperpile.com/b/J9j3wd/Tt8w)[10.1016/j.media.2007.06.004](http://dx.doi.org/10.1016/j.media.2007.06.004)

[Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., … WU-Minn HCP Consortium. (2013). Function in the human connectome: task-fMRI and individual differences in behavior. *NeuroImage*, *80*, 169–189.](http://paperpile.com/b/J9j3wd/cLQb1)

[Baumgartner, R., Cho, W., Coimbra, A., Chen, C., Wang, Z., Struyk, A., … Evelhoch, J. L. (2017). Evaluation of an fMRI USPIO-based assay in healthy human volunteers. *Journal of Magnetic Resonance Imaging: JMRI*, *46*(1), 124–133.](http://paperpile.com/b/J9j3wd/NcQe)

[Bennett, C. M., & Miller, M. B. (2013). fMRI reliability: influences of task and experimental design. *Cognitive, Affective & Behavioral Neuroscience*, *13*(4), 690–702.](http://paperpile.com/b/J9j3wd/254T)

[Blokland, G. A. M., Wallace, A. K., Hansell, N. K., Thompson, P. M., Hickie, I. B., Montgomery, G. W., … Wright, M. J. (2017). Genome-wide association study of working memory brain activation. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *115*, 98–111.](http://paperpile.com/b/J9j3wd/AOWe)

[Brandt, D. J., Sommer, J., Krach, S., Bedenbender, J., Kircher, T., Paulus, F. M., & Jansen, A. (2013). Test-Retest Reliability of fMRI Brain Activity during Memory Encoding. *Frontiers in Psychiatry / Frontiers Research Foundation*, *4*, 163.](http://paperpile.com/b/J9j3wd/CgHF)

[Bunford, N., Kinney, K. L., Michael, J., & Klumpp, H. (2017). Threat distractor and perceptual load modulate test-retest reliability of anterior cingulate cortex response. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *77*, 120–127.](http://paperpile.com/b/J9j3wd/llrx)

[Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C. R., & Mehta, M. A. (2009). Measuring fMRI reliability with the intra-class correlation coefficient. *NeuroImage*, *45*(3), 758–768.](http://paperpile.com/b/J9j3wd/dLPF)

[Cannon, T. D., Cao, H., Mathalon, D. H., Gee, D. G., & on behalf of the NAPLS consortium. (2018). Reliability of an fMRI paradigm for emotional processing in a multisite longitudinal study: Clarification and implications for statistical power. *Human Brain Mapping*, Vol. 39, pp. 599–601. https://doi.org/](http://paperpile.com/b/J9j3wd/IJCB)[10.1002/hbm.23875](http://dx.doi.org/10.1002/hbm.23875)

[Chase, H. W., Fournier, J. C., Greenberg, T., Almeida, J. R., Stiffler, R., Zevallos, C. R., … Phillips, M. L. (2015). Accounting for Dynamic Fluctuations across Time when Examining fMRI Test-Retest Reliability: Analysis of a Reward Paradigm in the EMBARC Study. *PloS One*, *10*(5), e0126326.](http://paperpile.com/b/J9j3wd/HrRE)

[Clément, F., & Belleville, S. (2009). Test-retest reliability of fMRI verbal episodic memory paradigms in healthy older adults and in persons with mild cognitive impairment. *Human Brain Mapping*, *30*(12), 4033–4047.](http://paperpile.com/b/J9j3wd/FPrF)

[Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, *29*(3), 162–173.](http://paperpile.com/b/J9j3wd/2EEi)

[Drew Sayer, R., Tamer, G. G., Jr, Chen, N., Tregellas, J. R., Cornier, M.-A., Kareken, D. A., … Campbell, W. W. (2016). Reproducibility assessment of brain responses to visual food stimuli in adults with overweight and obesity. *Obesity* , *24*(10), 2057–2063.](http://paperpile.com/b/J9j3wd/mLaj)

[Elliott, M. L., Knodt, A. R., Cooke, M., Kim, M. J., Melzer, T. R., Keenan, R., … Hariri, A. R. (2019). General functional connectivity: Shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. *NeuroImage*, *189*, 516–532.](http://paperpile.com/b/J9j3wd/LmS0b)

[Erk, S., Meyer-Lindenberg, A., Schmierer, P., Grimm, O., Tost, H., Mühleisen, T., … Walter, H. (2013). Functional impact of a recently identified quantitative trait locus for hippocampal volume with genome-wide support. *Translational Psychiatry*, *3*, e287.](http://paperpile.com/b/J9j3wd/6L2r)

[Estévez, N., Yu, N., Brügger, M., Villiger, M., Hepp-Reymond, M.-C., Riener, R., & Kollias, S. (2014). A Reliability Study on Brain Activation During Active and Passive Arm Movements Supported by an MRI-Compatible Robot. *Brain Topography*, Vol. 27, pp. 731–746. https://doi.org/](http://paperpile.com/b/J9j3wd/oOtx)[10.1007/s10548-014-0355-9](http://dx.doi.org/10.1007/s10548-014-0355-9)

[Fliessbach, K., Rohe, T., Linder, N. S., Trautner, P., Elger, C. E., & Weber, B. (2010). Retest reliability of reward-related BOLD signals. *NeuroImage*, *50*(3), 1168–1176.](http://paperpile.com/b/J9j3wd/akLA)

[Fournier, J. C., Chase, H. W., Almeida, J., & Phillips, M. L. (2014). Model specification and the reliability of fMRI results: implications for longitudinal neuroimaging studies in psychiatry. *PloS One*, *9*(8), e105169.](http://paperpile.com/b/J9j3wd/caeO)

[Friedman, L., Stern, H., Brown, G. G., Mathalon, D. H., Turner, J., Glover, G. H., … Potkin, S. G. (2008). Test-retest and between-site reliability in a multicenter fMRI study. *Human Brain Mapping*, Vol. 29, pp. 958–972. https://doi.org/](http://paperpile.com/b/J9j3wd/ZRVe)[10.1002/hbm.20440](http://dx.doi.org/10.1002/hbm.20440)

[Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., … WU-Minn HCP Consortium. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124.](http://paperpile.com/b/J9j3wd/pEgz)

[Greve, D. N., & Fischl, B. R. (2009). A Boundary-Based Cost Function for Within-Subject, Cross-Modal Registration. *NeuroImage*, Vol. 47, p. S100. https://doi.org/](http://paperpile.com/b/J9j3wd/2iAE)[10.1016/s1053-8119(09)70846-8](http://dx.doi.org/10.1016/s1053-8119(09)70846-8)

[Haller, S. P., Kircanski, K., Stoddard, J., White, L. K., Chen, G., Sharif-Askary, B., … Brotman, M. A. (2018). Reliability of neural activation and connectivity during implicit face emotion processing in youth. *Developmental Cognitive Neuroscience*, *31*, 67–73.](http://paperpile.com/b/J9j3wd/gzYp)

[Heckendorf, E., Bakermans‐Kranenburg, M. J., van Ijzendoorn, M. H., & Huffmeijer, R. (2019). Neural responses to children’s faces: Test–retest reliability of structural and functional MRI. *Brain and Behavior*, Vol. 9, p. e01192. https://doi.org/](http://paperpile.com/b/J9j3wd/3m8W)[10.1002/brb3.1192](http://dx.doi.org/10.1002/brb3.1192)

[Holiga, Š., Sambataro, F., Luzy, C., Greig, G., Sarkar, N., Renken, R. J., … Dukart, J. (2018). Test-retest reliability of task-based and resting-state blood oxygen level dependence and cerebral blood flow measures. *PloS One*, *13*(11), e0206583.](http://paperpile.com/b/J9j3wd/GX8g)

[Jezzard, P., & Balaban, R. S. (1995). Correction for geometric distortion in echo planar images from B0 field variations. *Magnetic Resonance in Medicine*, Vol. 34, pp. 65–73. https://doi.org/](http://paperpile.com/b/J9j3wd/3V4B)[10.1002/mrm.1910340111](http://dx.doi.org/10.1002/mrm.1910340111)

[Johnstone, T., Somerville, L. H., Alexander, A. L., Oakes, T. R., Davidson, R. J., Kalin, N. H., & Whalen, P. J. (2005). Stability of amygdala BOLD response to fearful faces over multiple scan sessions. *NeuroImage*, *25*(4), 1112–1123.](http://paperpile.com/b/J9j3wd/wqAG)

[Keren, H., Chen, G., Benson, B., Ernst, M., Leibenluft, E., Fox, N. A., … Stringaris, A. (2018). Is the encoding of Reward Prediction Error reliable during development? *NeuroImage*, *178*, 266–276.](http://paperpile.com/b/J9j3wd/aokv)

[Kimberley, T. J., Birkholz, D. D., Hancock, R. A., VonBank, S. M., & Werth, T. N. (2008). Reliability of fMRI during a continuous motor task: assessment of analysis techniques. *Journal of Neuroimaging: Official Journal of the American Society of Neuroimaging*, *18*(1), 18–27.](http://paperpile.com/b/J9j3wd/7QVi)

[Kimberley, T. J., Khandekar, G., & Borich, M. (2008). fMRI reliability in subjects with stroke. *Experimental Brain Research*, Vol. 186, pp. 183–190. https://doi.org/](http://paperpile.com/b/J9j3wd/j8i8)[10.1007/s00221-007-1221-8](http://dx.doi.org/10.1007/s00221-007-1221-8)

[Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M.-C., … Parsey, R. V. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, *46*(3), 786–802.](http://paperpile.com/b/J9j3wd/2zcC)

[Kong, J., Gollub, R. L., Megan Webb, J., Kong, J.-T., Vangel, M. G., & Kwong, K. (2007). Test–retest study of fMRI signal change evoked by electroacupuncture stimulation. *NeuroImage*, Vol. 34, pp. 1171–1181. https://doi.org/](http://paperpile.com/b/J9j3wd/ApcT)[10.1016/j.neuroimage.2006.10.019](http://dx.doi.org/10.1016/j.neuroimage.2006.10.019)

[Lee, J. N., Hsu, E. W., Rashkin, E., Thatcher, J. W., Kreitschitz, S., Gale, P., … Marchand, W. R. (2010). Reliability of fMRI motor tasks in structures of the corticostriatal circuitry: implications for future studies and circuit function. *NeuroImage*, *49*(2), 1282–1288.](http://paperpile.com/b/J9j3wd/5bqC)

[Lim, J., Choo, W.-C., & Chee, M. W. L. (2007). Reproducibility of Changes in Behaviour and fMRI Activation Associated with Sleep Deprivation in a Working Memory Task. *Sleep*, Vol. 30, pp. 61–70. https://doi.org/](http://paperpile.com/b/J9j3wd/uXiU)[10.1093/sleep/30.1.61](http://dx.doi.org/10.1093/sleep/30.1.61)

[Lipp, I., Murphy, K., Wise, R. G., & Caseras, X. (2014). Understanding the contribution of neural and physiological signal variation to the low repeatability of emotion-induced BOLD responses. *NeuroImage*, *86*, 335–342.](http://paperpile.com/b/J9j3wd/YKyr)

[Lois, G., Kirsch, P., Sandner, M., Plichta, M. M., & Wessa, M. (2018). Experimental and methodological factors affecting test-retest reliability of amygdala BOLD responses. *Psychophysiology*, *55*(12), e13220.](http://paperpile.com/b/J9j3wd/Gfad)

[Manoach, D. S., Halpern, E. F., Kramer, T. S., Chang, Y., Goff, D. C., Rauch, S. L., … Gollub, R. L. (2001). Test-retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. *The American Journal of Psychiatry*, *158*(6), 955–958.](http://paperpile.com/b/J9j3wd/3Wfc)

[Mohnke, S., Erk, S., Schnell, K., Schütz, C., Romanczuk-Seiferth, N., Grimm, O., … Walter, H. (2014). Further evidence for the impact of a genome-wide-supported psychosis risk variant in ZNF804A on the Theory of Mind Network. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *39*(5), 1196–1205.](http://paperpile.com/b/J9j3wd/w4g6)

[Nettekoven, C., Reck, N., Goldbrunner, R., Grefkes, C., & Lucas, C. W. (2018). Short- and long-term reliability of language fMRI. *NeuroImage*, Vol. 176, pp. 215–225. https://doi.org/](http://paperpile.com/b/J9j3wd/l37L)[10.1016/j.neuroimage.2018.04.050](http://dx.doi.org/10.1016/j.neuroimage.2018.04.050)

[Nord, C. L., Gray, A., Charpentier, C. J., Robinson, O. J., & Roiser, J. P. (2017). Unreliability of putative fMRI biomarkers during emotional face processing. *NeuroImage*, *156*, 119–127.](http://paperpile.com/b/J9j3wd/TIUe)

[Plichta, M. M., Grimm, O., Morgen, K., Mier, D., Sauer, C., Haddad, L., … Meyer-Lindenberg, A. (2014). Amygdala habituation: a reliable fMRI phenotype. *NeuroImage*, *103*, 383–390.](http://paperpile.com/b/J9j3wd/rdHz)

[Plichta, M. M., Schwarz, A. J., Grimm, O., Morgen, K., Mier, D., Haddad, L., … Meyer-Lindenberg, A. (2012). Test–retest reliability of evoked BOLD signals from a cognitive–emotive fMRI test battery. *NeuroImage*, Vol. 60, pp. 1746–1758. https://doi.org/](http://paperpile.com/b/J9j3wd/KLPT)[10.1016/j.neuroimage.2012.01.129](http://dx.doi.org/10.1016/j.neuroimage.2012.01.129)

[Poulton, R., Moffitt, T. E., & Silva, P. A. (2015). The Dunedin Multidisciplinary Health and Development Study: overview of the first 40 years, with an eye to the future. *Social Psychiatry and Psychiatric Epidemiology*, *50*(5), 679–693.](http://paperpile.com/b/J9j3wd/liJ9y)

[Putcha, D., O’Keefe, K., LaViolette, P., O’Brien, J., Greve, D., Rentz, D. M., … Sperling, R. (2011). Reliability of functional magnetic resonance imaging associative encoding memory paradigms in non-demented elderly adults. *Human Brain Mapping*, *32*(12), 2027–2044.](http://paperpile.com/b/J9j3wd/GtKD)

[Quiton, R. L., Keaser, M. L., Zhuo, J., Gullapalli, R. P., & Greenspan, J. D. (2014). Intersession reliability of fMRI activation for heat pain and motor tasks. *NeuroImage. Clinical*, *5*, 309–321.](http://paperpile.com/b/J9j3wd/fhcI)

[Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R. J. A., Kahn, R. S., & Ramsey, N. F. (2007). Test–retest reliability of fMRI activation during prosaccades and antisaccades. *NeuroImage*, Vol. 36, pp. 532–542. https://doi.org/](http://paperpile.com/b/J9j3wd/ycDX)[10.1016/j.neuroimage.2007.03.061](http://dx.doi.org/10.1016/j.neuroimage.2007.03.061)

[Rath, J., Wurnig, M., Fischmeister, F., Klinger, N., Höllinger, I., Geißler, A., … Beisteiner, R. (2016). Between- and within-site variability of fMRI localizations. *Human Brain Mapping*, *37*(6), 2151–2160.](http://paperpile.com/b/J9j3wd/1PHK)

[Sauder, C. L., Hajcak, G., Angstadt, M., & Phan, K. L. (2013). Test-retest reliability of amygdala response to emotional faces. *Psychophysiology*, *50*(11), 1147–1156.](http://paperpile.com/b/J9j3wd/TyCM)

[Schacht, J. P., Anton, R. F., Randall, P. K., Li, X., Henderson, S., & Myrick, H. (2011). Stability of fMRI striatal response to alcohol cues: a hierarchical linear modeling approach. *NeuroImage*, *56*(1), 61–68.](http://paperpile.com/b/J9j3wd/FeD0)

[Schlagenhauf, F., Juckel, G., Koslowski, M., Kahnt, T., Knutson, B., Dembler, T., … Heinz, A. (2008). Reward system activation in schizophrenic patients switched from typical neuroleptics to olanzapine. *Psychopharmacology*, *196*(4), 673–684.](http://paperpile.com/b/J9j3wd/yoIi)

[Schunck, T., Erb, G., Mathis, A., Jacob, N., Gilles, C., Namer, I. J., … Luthringer, R. (2008). Test–retest reliability of a functional MRI anticipatory anxiety paradigm in healthy volunteers. *Journal of Magnetic Resonance Imaging*, Vol. 27, pp. 459–468. https://doi.org/](http://paperpile.com/b/J9j3wd/vMr9)[10.1002/jmri.21237](http://dx.doi.org/10.1002/jmri.21237)

[Schuster, V., Herholz, P., Zimmermann, K. M., Westermann, S., Frässle, S., & Jansen, A. (2017). Comparison of fMRI paradigms assessing visuospatial processing: Robustness and reproducibility. *PloS One*, *12*(10), e0186344.](http://paperpile.com/b/J9j3wd/nQ6B)

[Sheu, L. K., Jennings, J. R., & Gianaros, P. J. (2012). Test-retest reliability of an fMRI paradigm for studies of cardiovascular reactivity. *Psychophysiology*, *49*(7), 873–884.](http://paperpile.com/b/J9j3wd/oZKM)

[Sousa, I., Vilela, P., & Figueiredo, P. (2014). Reproducibility of hypocapnic cerebrovascular reactivity measurements using BOLD fMRI in combination with a paced deep breathing task. *NeuroImage*, *98*, 31–41.](http://paperpile.com/b/J9j3wd/i2ly)

[Taylor, K. S., & Davis, K. D. (2009). Stability of tactile- and pain-related fMRI brain activations: an examination of threshold-dependent and threshold-independent methods. *Human Brain Mapping*, *30*(7), 1947–1962.](http://paperpile.com/b/J9j3wd/4Vy3)

[Tong, Y., Chen, Q., Nichols, T. E., Rasetti, R., Callicott, J. H., Berman, K. F., … Mattay, V. S. (2016). Seeking Optimal Region-Of-Interest (ROI) Single-Value Summary Measures for fMRI Studies in Imaging Genetics. *PloS One*, *11*(3), e0151391.](http://paperpile.com/b/J9j3wd/Bs0R)

[Upadhyay, J., Lemme, J., Anderson, J., Bleakman, D., Large, T., Evelhoch, J. L., … Becerra, L. (2015). Test–retest reliability of evoked heat stimulation BOLD fMRI. *Journal of Neuroscience Methods*, Vol. 253, pp. 38–46. https://doi.org/](http://paperpile.com/b/J9j3wd/d9Dw)[10.1016/j.jneumeth.2015.06.001](http://dx.doi.org/10.1016/j.jneumeth.2015.06.001)

[van den Bulk, B. G., Koolschijn, P. C. M. P., Meens, P. H. F., van Lang, N. D. J., van der Wee, N. J. A., Rombouts, S. A. R. B., … Crone, E. A. (2013). How stable is activation in the amygdala and prefrontal cortex in adolescence? A study of emotional face processing across three measurements. *Developmental Cognitive Neuroscience*, *4*, 65–76.](http://paperpile.com/b/J9j3wd/XYy5)

[Wehrum-Osinsky, S., Klucken, T., Kagerer, S., Walter, B., Hermann, A., & Stark, R. (2014). At the second glance: stability of neural responses toward visual sexual stimuli. *The Journal of Sexual Medicine*, *11*(11), 2720–2737.](http://paperpile.com/b/J9j3wd/1MfY)

[Weiss, C., Nettekoven, C., Rehme, A. K., Neuschmelting, V., Eisenbeis, A., Goldbrunner, R., & Grefkes, C. (2013). Mapping the hand, foot and face representations in the primary motor cortex — Retest reliability of neuronavigated TMS versus functional MRI. *NeuroImage*, Vol. 66, pp. 531–542. https://doi.org/](http://paperpile.com/b/J9j3wd/TlMI)[10.1016/j.neuroimage.2012.10.046](http://dx.doi.org/10.1016/j.neuroimage.2012.10.046)

[Wei, X., Yoo, S.-S., Dickey, C. C., Zou, K. H., Guttmann, C. R. G., & Panych, L. P. (2004). Functional MRI of auditory verbal working memory: long-term reproducibility analysis. *NeuroImage*, *21*(3), 1000–1008.](http://paperpile.com/b/J9j3wd/WMHV)

[Wende, K. C., Thiel, C., Sommer, J., Paulus, F. M., Krach, S., & Jansen, A. (2017). Mechanisms of hemispheric lateralization: A replication study. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *94*, 182–192.](http://paperpile.com/b/J9j3wd/Rtd2)

[Wheatley, T., Milleville, S. C., & Martin, A. (2007). Understanding animate agents: distinct roles for the social network and mirror system. *Psychological Science*, *18*(6), 469–474.](http://paperpile.com/b/J9j3wd/c6GB)

[White, L. K., Britton, J. C., Sequeira, S., Ronkin, E. G., Chen, G., Bar-Haim, Y., … Pine, D. S. (2016). Behavioral and neural stability of attention bias to threat in healthy adolescents. *NeuroImage*, *136*, 84–93.](http://paperpile.com/b/J9j3wd/aBKu)

[Zanto, T. P., Pa, J., & Gazzaley, A. (2014). Reliability measures of functional magnetic resonance imaging in a longitudinal evaluation of mild cognitive impairment. *NeuroImage*, Vol. 84, pp. 443–452. https://doi.org/](http://paperpile.com/b/J9j3wd/WFG6)[10.1016/j.neuroimage.2013.08.063](http://dx.doi.org/10.1016/j.neuroimage.2013.08.063)