Check for updates

Research Article

Psychological Science
1–15
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797620916786
www.psychologicalscience.org/PS

What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis

Maxwell L. Elliott[1], Annchen R. Knodt[1], David Ireland[2], Meriwether L. Morris[1], Richie Poulton[2], Sandhya Ramrakha[2], Maria L. Sison[1], Terrie E. Moffitt[1,3,4,5], Avshalom Caspi[1,3,4,5], and Ahmad R. Hariri[1]

[1]Department of Psychology & Neuroscience, Duke University; [2]Dunedin Multidisciplinary Health and Development Research Unit, Department of Psychology, University of Otago; [3]Social, Genetic, & Developmental Psychiatry Research Centre, Institute of Psychiatry, Psychology, & Neuroscience, King's College London; [4]Department of Psychiatry & Behavioral Sciences, Duke University School of Medicine; and [5]Center for Genomic and Computational Biology, Duke University

Abstract

Identifying brain biomarkers of disease risk is a growing priority in neuroscience. The ability to identify meaningful biomarkers is limited by measurement reliability; unreliable measures are unsuitable for predicting clinical outcomes. Measuring brain activity using task functional MRI (fMRI) is a major focus of biomarker development; however, the reliability of task fMRI has not been systematically evaluated. We present converging evidence demonstrating poor reliability of task-fMRI measures. First, a meta-analysis of 90 experiments (N = 1,008) revealed poor overall reliability—mean intraclass correlation coefficient (ICC) = .397. Second, the test-retest reliabilities of activity in a priori regions of interest across 11 common fMRI tasks collected by the Human Connectome Project (N = 45) and the Dunedin Study (N = 20) were poor (ICCs = .067–.485). Collectively, these findings demonstrate that common task-fMRI measures are not currently suitable for brain biomarker discovery or for individual-differences research. We review how this state of affairs came to be and highlight avenues for improving task-fMRI reliability.

Keywords

neuroimaging, individual differences, statistical analysis, cognitive neuroscience

Received 8/20/19; Revision accepted 2/1/20

Since functional MRI (fMRI) was introduced in 1992, scientists have had unprecedented ability to noninvasively observe human brain activity. In conventional fMRI, regional brain activity is estimated by measuring the blood-oxygen-level-dependent (BOLD) signal, which indexes changes in blood oxygenation associated with neural activity (Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001). In one of the most common forms of BOLD fMRI, researchers map brain activity associated with specific cognitive functions during certain tasks by contrasting the regional BOLD signal during a control condition with the BOLD signal during a condition of interest. In this way, task fMRI has given neuroscientists unique insights into the brain basis of human behavior, from basic perception to complex thought, and has given clinicians and mental-health researchers the opportunity to directly measure dysfunction in the organ responsible for disorder.

Corresponding Author:
Ahmad R. Hariri, Duke University, Department of Psychology & Neuroscience, Durham, NC 27708
E-mail: ahmad.hariri@duke.edu

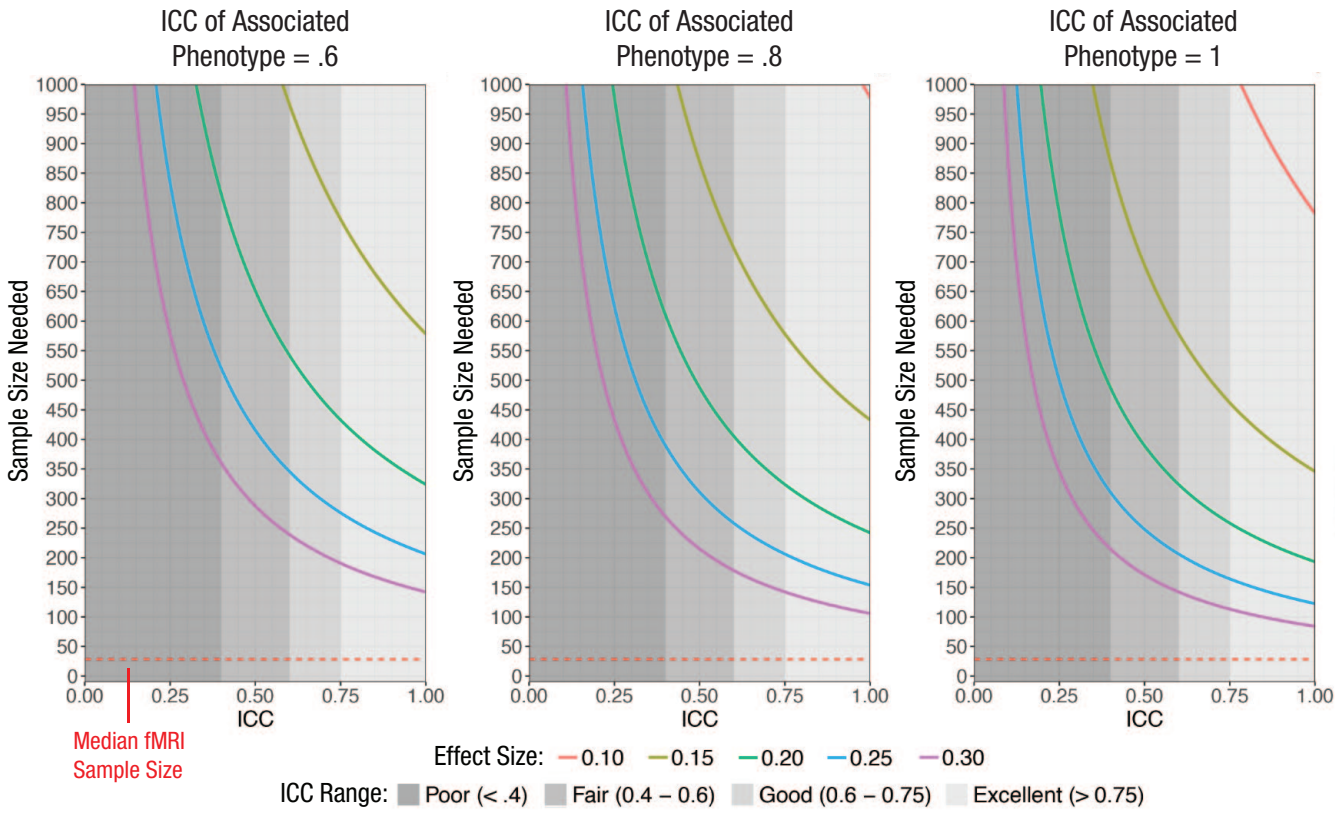

Fig. 1. The influence of task-functional MRI (fMRI) test-retest reliability on the sample size required for 80% power to detect brain–behavior correlations of effect sizes commonly found in psychological research. Power curves are shown for three levels of reliability of the associated behavioral or clinical phenotype. The figure was generated using the *pwr.r.test* function in R (Champely, 2018), with the value for *r* specified according to the attenuation formula in the Appendix. ICC = intraclass correlation coefficient.

Originally, task fMRI was primarily used to study functions supported by the average human brain. Researchers could measure within-subjects differences in activation between task and control conditions and average them across individuals to measure a group effect. To this end, fMRI tasks have been developed and optimized to elicit robust activation in a particular brain region of interest (ROI) or circuit when specific experimental conditions are contrasted. For example, increased amygdala activity is observed when people view emotional faces in comparison with geometric shapes, and increased ventral striatum activity is observed when people win money in comparison with when they lose it (Barch et al., 2013). The robust brain activity elicited using this within-subjects approach led researchers to use the same fMRI tasks to study between-subjects differences. The logic behind this strategy is straightforward: If a brain region activates during a task, then individual differences in the magnitude of that activation may contribute to individual differences in behavior as well as to any associated risk for disorder. Thus, if the amygdala is activated when people view threatening stimuli, then differences between people in the degree of amygdala activation should signal differences between them in threat sensitivity and related clinical phenomena, such as anxiety and depression (Swartz, Knodt, Radtke, & Hariri, 2015). In this way, fMRI was transformed from a tool for understanding how the average brain works to a tool for studying how the brains of individuals differ.

The use of task fMRI to study differences between people heralded the possibility that it could be a powerful tool for discovering biomarkers for brain disorders (Woo, Chang, Lindquist, & Wager, 2017). Broadly, a biomarker is a biological indicator often used for risk stratification, diagnosis, prognosis, and evaluation of treatment response. However, to be useful as a biomarker, an indicator must first be reliable. Reliability is the ability of a measure to give consistent results under similar circumstances. It puts a limit on the predictive utility, power, and validity of any measure (see Appendix and Fig. 1). Consequently, reliability is critical for both clinical applications and research practice. Measures with low reliability are unsuitable as biomarkers and cannot predict clinical health outcomes. That is, if a measure is going to be used by clinicians to predict the likelihood that a patient will develop an illness in the

future, then the patient cannot score randomly high on the measure at one assessment and low on the measure at the next assessment.

To progress toward a cumulative neuroscience of individual differences with clinical relevance, we must establish reliable brain measures. Although the reliability of task fMRI has previously been discussed (Bennett & Miller, 2010; Herting, Gautam, Chen, Mezher, & Vetter, 2018), individual studies provide highly variable estimates and often contain small test-retest samples and a wide variety of analytic methods. In addition, the authors of those studies may reach contradictory conclusions about the reliability of the same tasks (Manuck, Brown, Forbes, & Hariri, 2007; Nord, Gray, Charpentier, Robinson, & Roiser, 2017). This leaves the overall reliability of task fMRI, as well as the specific reliabilities of many of the most commonly used fMRI tasks, largely unknown. An up-to-date, comprehensive review and meta-analysis of the reliability of task fMRI and an in-depth examination of the reliability of the most widely used task-fMRI measures is needed. Here, we present evidence from two lines of analysis that point to the poor reliability of commonly used task-fMRI measures. First, we conducted a meta-analysis of the test-retest reliability of regional activation in task fMRI. Second, in two recently collected data sets, we analyzed the test-retest reliability of brain activation in a priori ROIs across several commonly used fMRI tasks (our design and analysis plans were posted prior to data analysis at https://sites.google.com/site/moffittcaspiprojects/home/projectlist/knodt_2019).

Method

Meta-analytic reliability of task fMRI

We performed a systematic review and meta-analysis following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (see Fig. S1 in the Supplemental Material available online). We searched Google Scholar for peer-reviewed articles written in English and published on or before April 1, 2019, that included test-retest reliability estimates of task-fMRI activation. We used the advanced search tool to find articles that included all of the terms "ICC" (i.e., intraclass correlation coefficient), "fMRI," and "retest" and at least one of the terms "ROI," "ROIs," "region of interest," or "regions of interest." This search yielded 1,170 articles.

Study selection and data extraction. One author (M. L. Morris) screened all titles and abstracts before the full texts were reviewed (by authors M. L. Elliott and A. R. Knodt). We included all original, peer-reviewed empirical articles that reported test-retest reliability estimates for activation during a BOLD fMRI task. All ICCs reported in the main text and the Supplemental Material were eligible for inclusion. If ICCs were depicted only graphically (e.g., in a bar graph), we did our best to judge the value from the graph. Voxelwise ICCs that were depicted only on brain maps were not included. For ICCs calculated on the basis of more than 2 time points, we used the average of the intervals as the value for the interval (e.g., the average of the time between Time Points 1 and 2 and Time Points 2 and 3 for an ICC based on three time points). For articles that reported ICCs from sensitivity analyses in addition to primary analyses on the same data (e.g. using different modeling strategies or excluding certain individuals) we included ICCs from only the primary analysis. We did not include ICCs from combinations of tasks. ICCs were excluded if they were from a longitudinal or intervention study that was designed to assess change, if they did not report ICCs based on measurements from the same MRI scanner or task, or if they reported reliability on something other than activation measures across individuals (e.g., spatial extent of activation or multivoxel patterns of activation within individuals).

Two authors (M. L. Elliott and A. R. Knodt) extracted data about sample characteristics (publication year, sample size, healthy vs. clinical), study design (test-retest interval, event-related vs. blocked, task length, and task type), and ICC reporting (thresholded vs. not thresholded). *Thresholding* occurs when studies calculate multiple ICCs but only report values above a minimum threshold. For each article, every reported ICC that met the above study-selection requirements was recorded.

Statistical analyses. For most of the studies included, no standard error or confidence interval (CI) for the ICC was reported. Therefore, in order to include as many estimates as possible in the meta-analysis, the standard error of all ICCs was estimated using the Fisher *r*-to-*Z* transformation for ICC values (Chen et al., 2018).

A random-effects multilevel meta-analytic model was fitted using tools from the *metafor* package in R (see Edlund & Nichols, 2019, appendix A). In this model, ICCs and standard errors were averaged within each unique sample, task, and test-retest interval (or *substudy*) was averaged within each article or study (Borenstein, Hedges, Higgins, & Rothstein, 2009). For the results reported here, the correlation between ICCs in each substudy was assumed to be 1 to ensure that the meta-analytic weight for each substudy was based solely on sample size rather than the number of ICCs reported. However, sensitivity analyses revealed that this decision had very little impact on the overall result (see Fig. S2 in the Supplemental Material). In the meta-analytic

model, substudies were nested within studies to account for the nonindependence of ICCs estimated within the same study. Meta-analytic summaries were estimated separately for substudies that reported ICC values that had been thresholded because of the documented spurious inflation of effect sizes that occurs when only statistically significant estimates are reported (Poldrack et al., 2017; Vul, Harris, Winkielman, & Pashler, 2009; Yarkoni, 2009).

To test for effects of moderators, we fitted a separate random-effects multilevel model to all 1,146 ICCs (i.e., without averaging within each substudy, because many substudies included ICCs with different values for one or more moderators). The moderators included were task length, task design (block vs. event-related), task type (e.g., emotion vs. executive control vs. reward), ROI type (e.g., structural or functional), ROI location (cortical vs. subcortical), sample type (healthy vs. clinical), retest interval, number of citations per year, and whether ICCs were thresholded on significance (see Table S1 in the Supplemental Material for descriptive statistics on all moderators tested). All moderators were simultaneously entered into the model as random effects. In the multilevel model, ICCs were nested within substudies, which were in turn nested within studies. This was done to account for the nonindependence of ICCs estimated within the same substudy, as well as the nonindependence of substudies conducted within the same study.

Analyses of new data sets

Human Connectome Project (HCP). The HCP is a publicly available data set that includes 1,206 participants with extensive structural and fMRI data (Van Essen et al., 2013). In addition, 45 participants completed the entire scan protocol a second time (with a mean interval between scans of approximately 140 days). All participants were free of current psychiatric or neurologic illness and were between 25 and 35 years of age.

The seven tasks employed in the HCP were designed to identify functionally relevant nodes in the brain. These tasks included an *n*-back working memory/executive-function task (targeting the dorsolateral prefrontal cortex, or dlPFC); a gambling-reward/incentive-processing task (targeting the ventral striatum); a motor-mapping task consisting of foot, hand, and tongue movements (targeting the motor cortex); an auditory language task (targeting the anterior temporal lobe); a social-cognition/theory-of-mind task (targeting the lateral fusiform gyrus, superior temporal sulcus, and other social-network regions); a relational-processing/dimensional-change-detection task (targeting the rostrolateral prefrontal cortex, or rlPFC); and a face-matching emotion-processing task (targeting the amygdala).

Dunedin Multidisciplinary Health and Development Study. The Dunedin Study is 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, and followed to age 45 years (Poulton, Moffitt, & Silva, 2015). Structural and functional neuroimaging data were collected between August 2016 and April 2019, when participants were 45 years old. In addition, 20 study members completed the entire scan protocol a second time (mean interval between scans = 79 days).

We collected fMRI during four tasks targeting neural hubs in four different domains: a face-matching emotion-processing task (targeting the amygdala), a Stroop executive-function task (targeting the dlPFC and the dorsal anterior cingulate cortex), a monetary-incentive delay-reward task (targeting the ventral striatum), and a face-name-encoding episodic-memory task (targeting the hippocampus). See Supplemental Methods in the Supplemental Material for additional details, including fMRI preprocessing, for both data sets.

ROI definition. Individual estimates of regional brain activity were extracted using two common approaches. First, we extracted average values from a priori anatomically defined regions. We identified the primary ROI for each task and extracted average BOLD signal-change estimates from all voxels within a corresponding bilateral anatomical mask.

Second, we used functionally defined regions based on group-level activation. Here, we generated functional ROIs by drawing 5-mm spheres around the group-level peak voxel within the target anatomical ROI for each task (across all individuals and sessions). This is a commonly used strategy for capturing the location of peak activation in each subject despite intersubject variability in the exact location of the activation. See the Supplemental Material for further details on ROI definition, overlays on the anatomical template (Fig. S3), and peak-voxel location (Table S2). Here, we report analyses based on anatomically defined ROIs, and we report sensitivity analyses based on functional ROIs in the Supplemental Material.

Reliability analysis. Subject-level BOLD signal-change estimates were extracted for each task, ROI, and scanning session. Reliability was quantified using a two-way mixed-effects ICC with session modeled as a fixed effect, subject as a random effect, and test-retest interval as an effect of no interest. This mixed-effects model is referred to as ICC (3,1)

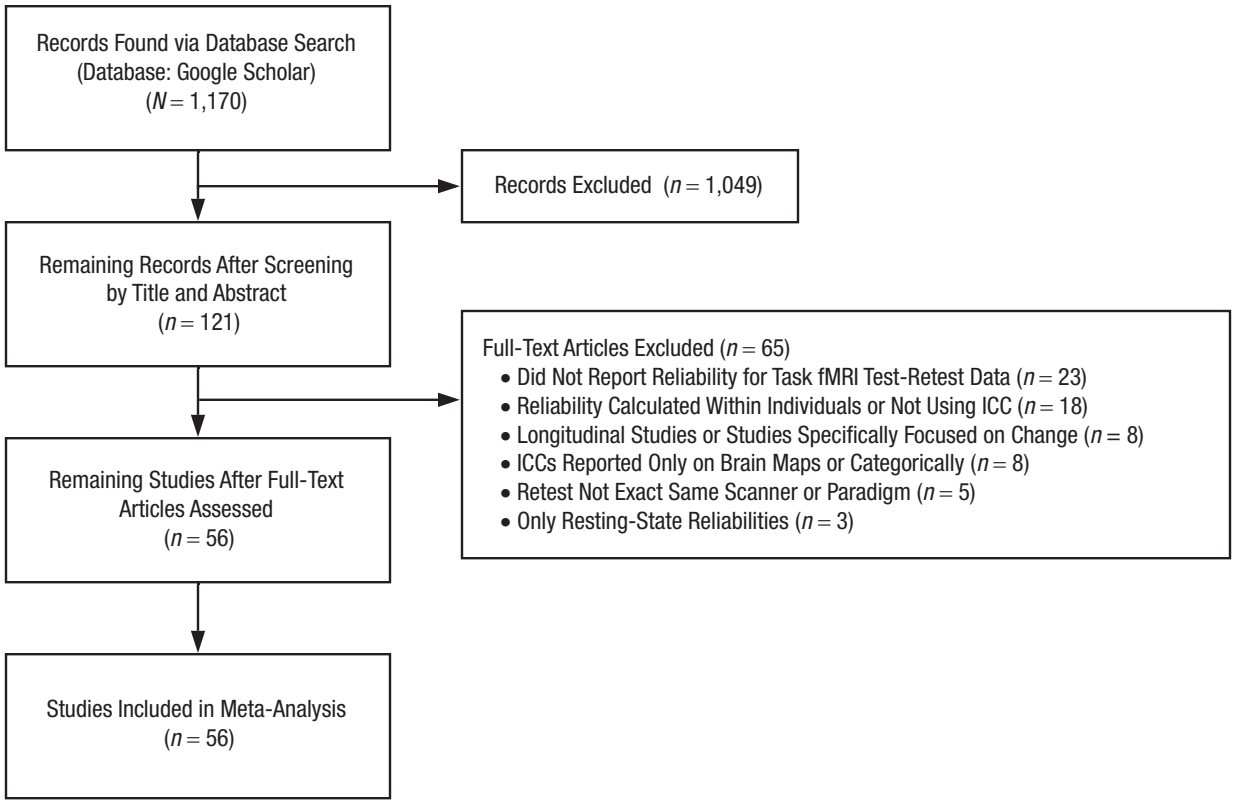

Fig. 2. Flow diagram for the systematic literature review and meta-analysis.

by Shrout and Fleiss (1979) and defined as ICC (3,1) = (BMS − EMS)/(BMS + (k − 1) × EMS), where BMS represents between-subjects mean square, EMS represents error mean square, and k is the number of raters or scanning sessions (in this case, two). We note that ICC (3,1) tracks the consistency of measures between sessions rather than absolute agreement and that it is commonly used in studies of task-fMRI test-retest reliability because of the possibility of habituation to the stimuli over time.

To test reliability for each task more generally, we calculated ICCs for all target ROIs across all 11 tasks. Because three of the tasks (the emotion, reward, and executive-function tasks) were very similar across the HCP and Dunedin studies and targeted the same region, the same ROI was used for these tasks in both studies, resulting in a total of eight unique target ROIs assessed for reliability. To further visualize global patterns of reliability, we also calculated voxelwise maps of ICC (3,1) using the 3dICC_REML.R function from Analysis of Functional NeuroImages (AFNI) software (Chen, Saad, Britton, Pine, & Cox, 2013). Finally, to provide a benchmark for evaluating task-fMRI reliability, we determined the test-retest reliability of three commonly used structural MRI measures: cortical thickness and surface area for each of 360 parcels or ROIs (Glasser et al., 2016), as well as gray-matter volume for 17 subcortical structures. Code and data for this study are available at github.com/HaririLab/Publications/tree/master/ElliottKnodt2020PS_tfMRIReliability.

Results

Reliability of individual differences in task fMRI: a systematic review and meta-analysis

We identified 56 articles meeting the criteria for inclusion in the meta-analysis, yielding 1,146 ICC estimates derived from 1,088 unique participants across 90 distinct substudies employing 66 different task-fMRI paradigms (Fig. 2). These articles were cited a total of 2,686 times, with an average of 48 citations per article and 5.7 citations per article per year. During the study-selection process, we discovered that some researchers calculated many different ICCs (across multiple ROIs, contrasts, and tasks) but reported only a subset of the estimated ICCs that were either statistically significant or reached a minimum ICC threshold. This practice leads to inflated reliability estimates (Kriegeskorte, Lindquist, Nichols, Poldrack, & Vul, 2010; Poldrack et al., 2017). Therefore, we performed separate analyses of data from unthresholded and thresholded reports.

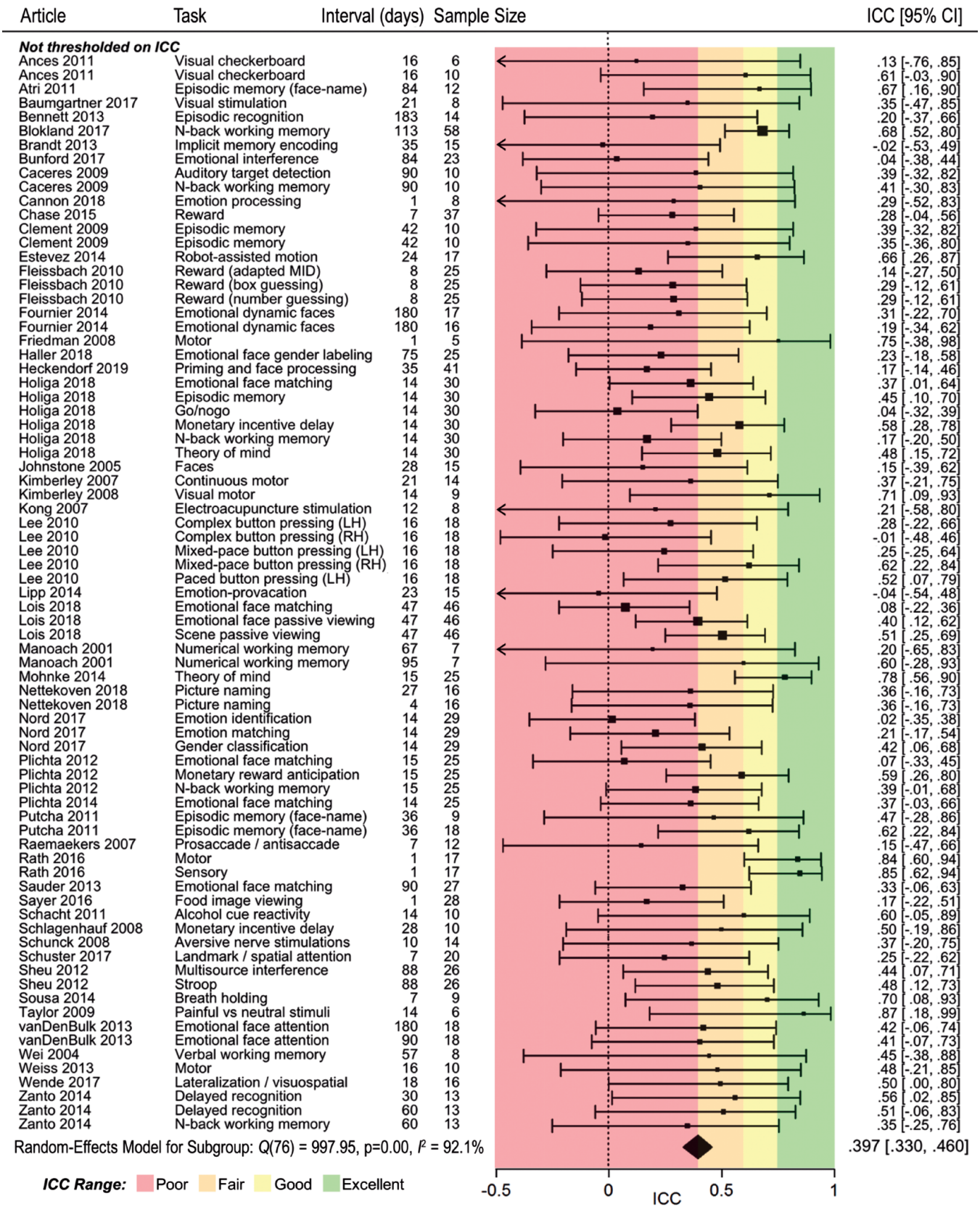

Fig. 3. *(continued on next page)*

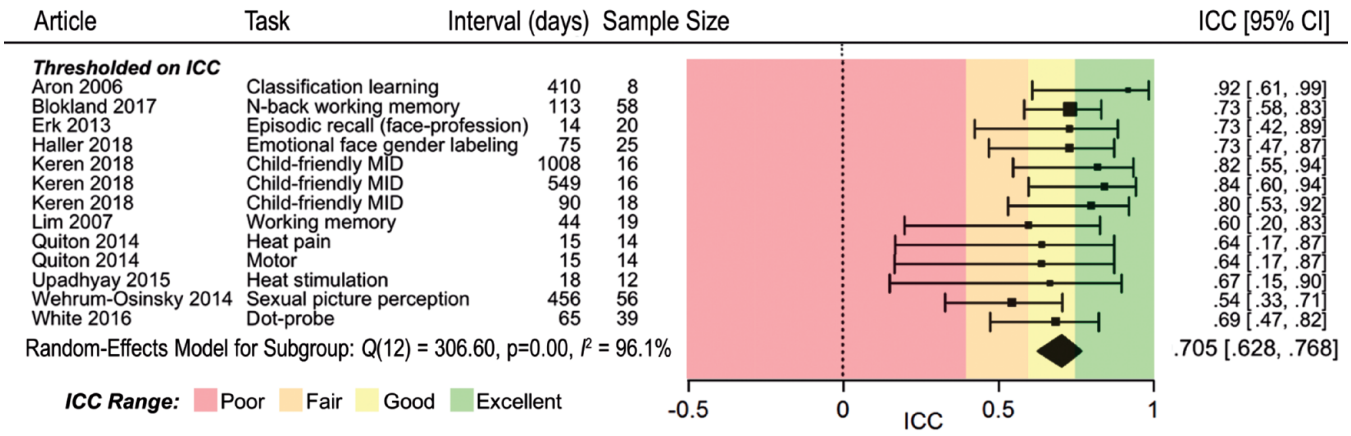

Fig. 3. Meta-analysis forest plot displaying the estimate of test-retest reliability for each task-functional MRI (fMRI) measure from all intraclass correlation coefficients (ICCs) reported in each study. The first column labels each article by the first author's last name and year of publication. References for all articles listed here are provided in the Supplemental Material available online. In the subject-type column, "h" indicates that the sample in the study consisted of healthy controls, and "c" indicates a clinical sample. Studies are split into two subgroups. In the first group of studies, authors reported all ICCs that were calculated, thereby allowing for a relatively unbiased estimate of reliability. In the second group of studies, authors selected a subset of calculated ICCs (on the basis of the magnitude of the ICC or another nonindependent statistic) and then reported ICCs only from that subset. This practice led to inflated reliability estimates, and therefore these studies were meta-analyzed separately to highlight this bias. Error bars indicate 95% confidence intervals (CIs). MID = monetary incentive delay; LH = left hand, RH = right hand.

Figure 3 shows the test-retest reliability coefficients (ICCs) from 77 substudies reporting unthresholded values (average $N = 19.6$, median $N = 17$). Fifty-six percent of the values fell in the range of poor reliability (below .4), an additional 24% of the values fell in the range of fair reliability (.4–.6), and only 20% fell in the range of good or excellent reliability (above .75). A random-effects meta-analysis revealed an average ICC of .397 (95% CI = [.330–.460], $p < .001$), which is in the poor range (Cicchetti & Sparrow, 1981). There was evidence of between-studies heterogeneity ($p = .04$, $I^2 = 31.6$).

As expected, the meta-analysis of 13 substudies that reported ICCs only above a minimum threshold (average $N = 24.2$, median $N = 18$) revealed a higher meta-analytic ICC of .705 (95% CI = [.628–.768], $p < .001$, $I^2 = 17.9$). This estimate, which is 1.78 times the size of the estimate from unthresholded ICCs, is in the good range, suggesting that the practice of thresholding inflates estimates of reliability in task fMRI. There was no evidence of between-studies heterogeneity ($p = .54$, $I^2 = 17.9$).

A moderator analysis of all substudies revealed significantly higher reliability for studies that were thresholded on the basis of the ICC, $Q_M(1) = 6.531$, $p = .010$, $\beta = 0.140$. In addition, ROIs located in the cortex had significantly higher ICCs than those located in the subcortex, $Q_M(1) = 114.476$, $p < .001$, $\beta = 0.259$. However, we did not find evidence that the meta-analytic estimate was moderated by task type, task design, task length, test-retest interval, ROI type, sample type, or number of citations per year. Finally, we tested for publication bias using the Egger random-effects regression test (Egger, Davey Smith, Schneider, & Minder, 1997) and found no evidence of bias ($Z = .707$, $p = .480$).

The results of the meta-analysis were illuminating, but interpreting them was not simple. First, the reliability estimates came from a wide array of tasks and samples, so a single meta-analytical reliability estimate could obscure truly reliable task-fMRI paradigms. Second, the studies used different (and some now outdated) scanners and different preprocessing and analysis pipelines, leaving open the possibility that reliability would be improved with more advanced technology and consistent practices. To address these limitations and possibilities, we analyzed two new data sets using state-of-the-art scanners and practices to assess individual differences in commonly used tasks that tap a variety of cognitive and affective functions.

Reliability of individual differences in task fMRI: analyses in two new data sets

We evaluated test-retest reliabilities of activation in a priori ROIs for 11 commonly used fMRI tasks (see the Method section). In the HCP, 45 participants were scanned twice using a custom 3T scanner (Siemens, Munich, Germany), 140 days apart on average (SD = 67.1 days), using seven tasks targeting emotion, reward, executive function, motor, language, social cognition, and relational processing. This sample size was determined by the publicly available data in the HCP. In the Dunedin Study, 20 participants were scanned twice using a 3T Siemens Skyra, 79 days apart on average ($SD = 10.3$ days), using four tasks targeting emotion, reward,

executive control, and episodic memory. This sample size corresponds to the average sample size used in the meta-analyzed studies. Three of the tasks were similar across the two studies, allowing us to test the replicability of task-fMRI reliabilities. For each of the eight unique tasks across the two studies, we identified the task's primary target region, resulting in a total of eight a priori ROIs (see the Method section).

Group-level activation. To ensure that the 11 tasks were implemented and processed correctly, we calculated the group-level activation in the target ROIs using the primary contrast of interest for each task (see Supplemental Methods in the Supplemental Material for details). These analyses revealed that each task elicited the expected robust activation in the target ROI at the group level (i.e., across all participants and sessions; see warm-colored maps in Figure 4 for the three tasks in common between the two studies and Fig. S4 in the Supplemental Material for the remaining tasks).

Reliability of regional activation. We investigated the reliability of task activation in both data sets using four steps. First, we tested the reliability of activation in the target ROI for each task. Second, we evaluated the reliability of activation in the other seven a priori ROIs for each task. This was done to test whether the reliability of target ROIs was higher than the reliability of activation in other (nontarget) brain regions and to identify any tasks or regions with consistently high reliability. Third, we reestimated reliability using activation in the left and right hemispheres separately to test whether the estimated reliability was harmed by averaging across the hemispheres. Fourth, we tested whether the reliability depended on whether ROIs were defined structurally (i.e., using an anatomical atlas) or functionally (i.e., using a set of voxels based on the location of peak activity). See Figure S5 in the Supplemental Material for ICCs of behavior during each fMRI task.

Reliability of regional activation in the HCP. First, as shown by the estimates circled in black in Figure 5, activation in anatomically defined target ROIs in the HCP had low reliability across the seven fMRI tasks (mean ICC = .251, 95% CI = [.142–.360]). Only the language-processing task had greater than poor reliability (ICC = .485). None of the reliabilities entered the good range (ICC > .6). Second, the reliability of task activation in nontarget ROIs was also low (Fig. 5; mean ICC = .239, 95% CI = [.188–.289]) but not significantly lower than the reliability in target ROIs (p = .474).

Third, the reliability of task activation calculated from left and right ROIs separately resembled estimates from averaged ROIs (mean left ICC = .207 in target ROIs and .196 in nontarget ROIs, mean right ICC = .259 in target ROIs and .236 in nontarget ROIs; see Fig. S6 in the Supplemental Material). Fourth, the reliability of task activation in functionally defined ROIs was also low (mean ICC = .381, 95% CI = [.317–.446]), with only the motor and social tasks exhibiting ICCs greater than .4 (ICCs = .550 and .446, respectively; see Fig. S6).

As an additional step, to account for the family structure present in the HCP, we reestimated reliability after removing one of each sibling/twin pair in the test-retest sample. Reliability in bilateral anatomical ROIs in the subsample of 26 unrelated individuals yielded reliabilities very similar to those in the overall sample (mean ICC = .301 in target ROIs and .218 in nontarget ROIs; Fig. S6).

Reliability of regional activation in the Dunedin Study. First, as shown by the estimates circled in black in Figure 5, activation in the anatomically defined target ROI in the Dunedin Study for each of the four tasks had low reliability (mean ICC = .309, 95% CI = [.145–.472]), with no ICCs reaching the good range (ICC > .6). Second, the reliability of activation in the nontarget ROIs was also low (Fig. 5; mean ICC = .193, 95% CI = [.100–.286]), but not significantly lower than the reliability in target ROIs (p = .140). Third, the reliability of task activation calculated for the left and right hemispheres separately was similar to the reliability in the averaged ROIs (mean left ICC = .243 in target ROIs and .202 in nontarget ROIs, mean right ICC = .358 in target ROIs and .192 in nontarget ROIs; see Fig. S6). Fourth, functionally defined ROIs again did not meaningfully improve reliability (mean ICC = .325, 95% CI = [.197–.453]; see Fig. S6).

Reliability of structural measures. To provide a benchmark for evaluating the test-retest reliability of task fMRI, we investigated the reliability of three commonly used structural MRI measures: cortical thickness, surface area, and subcortical gray-matter volume. Consistent with prior evidence (Han et al., 2006) that structural MRI phenotypes have excellent reliability (i.e., ICCs > .9), our results showed that global and regional structural MRI measures in the present samples demonstrated very high test-retest reliabilities (Fig. 5). For average cortical thickness, ICCs were .953 and .939 in the HCP and Dunedin Study data sets, respectively. In the HCP, parcel-wise (i.e., regional) cortical-thickness reliabilities averaged .886 (range = .547–.964), with 100% above the fair threshold, 98.6% above the good threshold, and 94.2% above the excellent threshold. In the Dunedin Study, parcel-wise cortical-thickness reliabilities averaged .846 (range = .385–.975), with 99.7% of ICCs above the fair threshold, 96.4% above the good threshold, and 84.7% above the excellent threshold. For total surface area, ICCs were .999

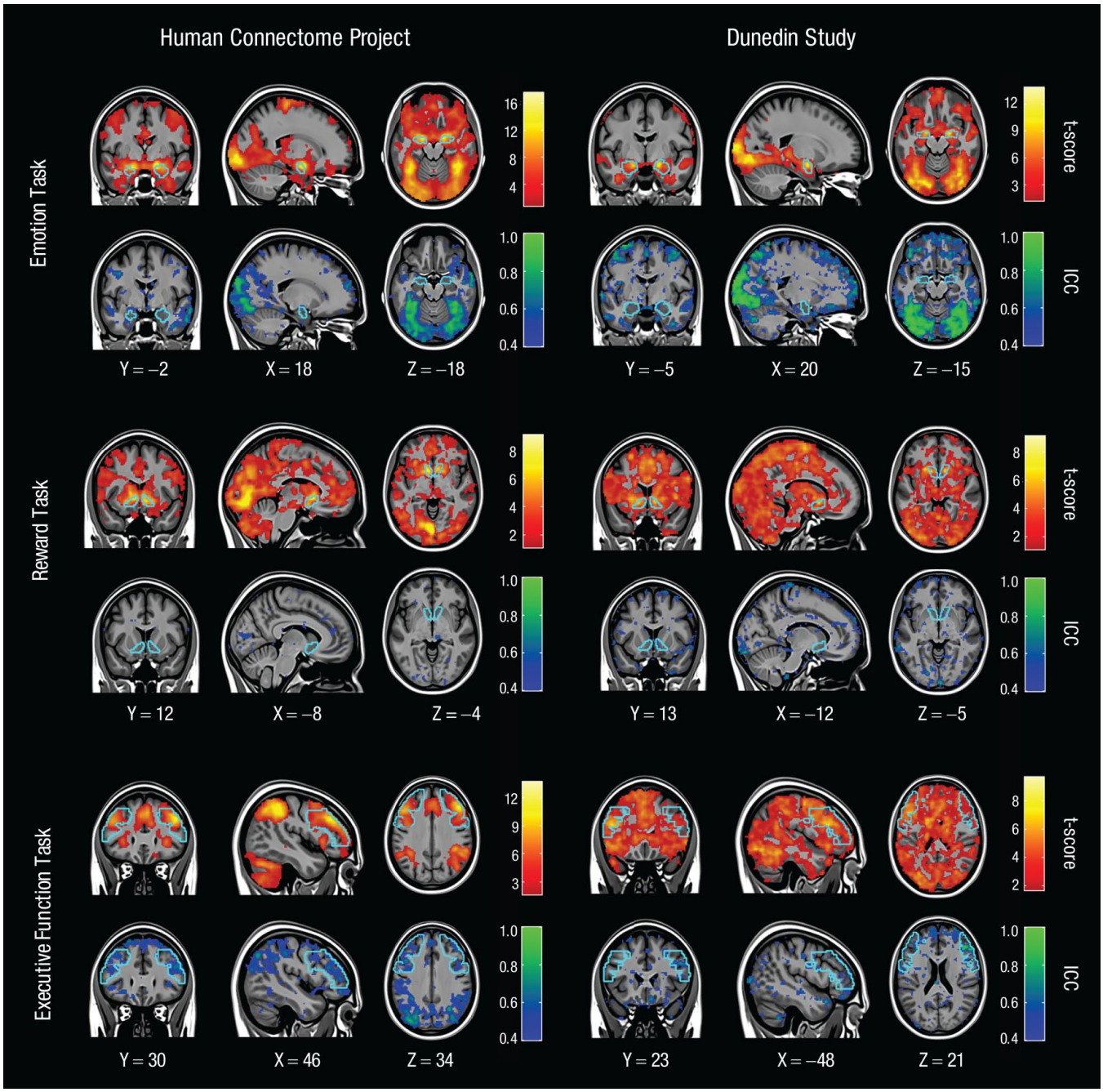

Fig. 4. Whole-brain activation and reliability maps for three task-functional MRI measures used in both the Human Connectome Project and the Dunedin Study. For each task, a whole-brain activation map of the primary within-subjects contrast (*t* score) is displayed in warm colors (top), and a whole-brain map of the between-subjects reliability (intraclass correlation coefficient, or ICC) is shown in cool colors (bottom). For each task, the target region of interest is outlined in sky blue. The activation maps are thresholded at $p < .05$ and are whole-brain corrected for multiple comparisons using threshold-free cluster enhancement. The ICC maps are thresholded so that voxels with ICCs of less than .4 are not colored. Values for X, Y, and Z are given in Montreal Neurological Institute coordinates.

and .996 in the HCP and Dunedin Study data sets, respectively. In the HCP, parcel-wise surface-area ICCs averaged .937 (range = .526–.992), with 100% above the fair threshold, 98.9% above the good threshold, and 96.9% above the excellent threshold. In the Dunedin Study, surface-area ICCs averaged .942 (range = .572–.991), with 100% above the fair threshold, 99.7% above the good threshold, and 98.1% above the excellent threshold. For subcortical volumes, ICCs in the HCP averaged .903 (range = .791–.984), with all ICCs above the excellent threshold. In the Dunedin Study, subcortical volumes averaged .931 (range = .767–.979), with all ICCs above the excellent

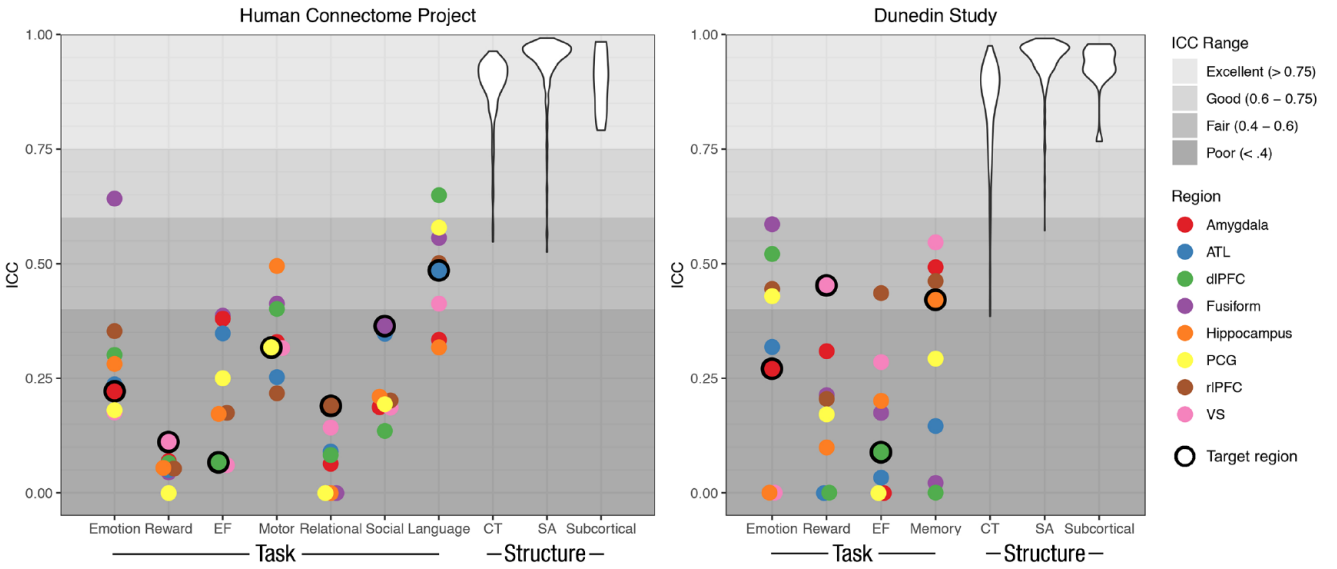

Fig. 5. Test-retest reliabilities of region-wise activation measures in 11 commonly used task-functional MRI paradigms and three common structural MRI measures, separately for the Human Connectome Project (left) and the Dunedin Study (right). For each task, intraclass correlation coefficients (ICCs) were estimated for activation in the a priori target region of interest (ROI; circled in black) and in nontarget ROIs selected from the other tasks. Nontarget ROIs were the anterior temporal lobe (ATL), dorsolateral prefrontal cortex (dlPFC), precentral gyrus (PCG), rostrolateral prefrontal cortex (rlPFC), and ventral striatum (VS). As a benchmark, ICCs of three common structural MRI measures—cortical thickness (CT), surface area (SA), and subcortical volume—are depicted as violin plots representing the distribution of ICCs for each of the 360 parcels for CT and SA and the 17 subcortical structures for gray-matter volume. Negative ICCs are set to 0 for purposes of visualization. EF = executive function.

threshold. See Table S3 in the Supplemental Material for an evaluation of the reliabilities of each subcortical region.

Discussion

We found evidence that commonly used task-fMRI measures generally do not have the test-retest reliability necessary for biomarker discovery or brain–behavior mapping. Our meta-analysis of task-fMRI reliability revealed an average test-retest reliability coefficient of .397, which is below the minimum required for good reliability (ICC = .6; Cicchetti & Sparrow, 1981) and far below the recommended cutoffs for clinical application (ICC = .8) or individual-level interpretation (ICC = .9). Of course, not all task-fMRI measures are the same, and it is not possible to assign a single reliability estimate to all individual-differences measures gathered in fMRI research. However, we found little evidence that task type, task length, or test-retest interval had an appreciable impact on the reliability of task fMRI.

We additionally evaluated the reliability of 11 commonly used task-fMRI measures in the HCP and the Dunedin Study. Unlike many of the studies included in our meta-analysis, these two studies were completed recently on modern scanners using cutting-edge acquisition parameters, up-to-date artifact reduction, and state-of-the-art preprocessing pipelines. Regardless, the average test-retest reliability was again poor (ICC = .228). In these analyses, we found no evidence that ROIs targeted by the task were more reliable than other, nontarget ROIs (mean ICC = .270 for target ROIs and .228 for nontarget ROIs) or that any specific task or target ROI consistently produced measures with high reliability. Of interest, the reliability estimate from these two studies was considerably smaller than the estimate from the meta-analysis (meta-analytic ICC = .397), possibly because preregistered analyses often yield smaller effect sizes than do analyses from publications without preregistration, which affords increased flexibility in analytic decision-making (Schäfer & Schwarz, 2019).

The two disciplines of fMRI research

Our results harken back to Lee Cronbach's classic 1957 article in which he described the "two disciplines of scientific psychology." According to Cronbach, the *experimental* discipline strives to uncover universal human traits and abilities through experimental control and group averaging, whereas the *correlational* discipline strives to explain variation between people by measuring how they differ from one another. A fundamental distinction between the two disciplines is how they treat individual differences. For the experimental researcher, variation between people is an error that

must be minimized to detect the largest experimental effect. For the correlational investigator, variation between people is the primary unit of analysis and must be measured carefully to extract reliable individual differences (Cronbach, 1957; Hedge, Powell, & Sumner, 2018).

Current task-fMRI paradigms are largely descended from the experimental discipline. Task-fMRI paradigms are intentionally designed to reveal how the average human brain responds to provocation, while minimizing between-subjects variance. Paradigms that are able to elicit robust targeted brain activity at the group level are subsequently converted into tools for assessing individual differences. Within-subjects robustness is, then, often inappropriately invoked to suggest between-subjects reliability, despite the fact that reliable within-subjects experimental effects at a group level can arise from unreliable between-subjects measurements (Fröhner, Teckentrup, Smolka, & Kroemer, 2019).

This reasoning is not unique to task-fMRI research. Behavioral measures that elicit robust within-subjects (i.e., group) effects have been shown to have low between-subjects reliability; for example, the mean test-retest reliability of the Stroop test (ICC = .45; Hedge et al., 2018) is strikingly similar to the mean reliability of our task-fMRI meta-analysis (ICC = .397). Nor is it the case that MRI measures, or even the BOLD signal itself, are inherently unreliable. Both structural MRI measures in our analyses (see Fig. 5), as well as measures of intrinsic functional connectivity estimated from long fMRI scans (Elliott et al., 2019; Gratton et al., 2018), demonstrate high test-retest reliability. Thus, it is not the tool that is problematic but the strategy of adopting tasks developed for experimental cognitive neuroscience that appear to be poorly suited for reliably measuring differences in brain activation between people.

Recommendations and future directions

We next consider several avenues for maximizing the value of existing data sets as well as improving the reliability of task fMRI moving forward. We begin with two recommendations that can be implemented immediately before moving on to two recommendations that will require additional data collection and innovation.

Immediate opportunities for task fMRI: from brain hot spots to whole-brain signatures. Currently, the majority of task-fMRI measures are based on contrasts between conditions (i.e., change scores) extracted from ROIs. However, change scores will always have lower reliability than their constituent measures (Hedge et al., 2018) and have been shown to undermine the reliability of task fMRI (Infantolino, Luking, Sauder, Curtin, & Hajcak, 2018). However, contrast-based activation values extracted from ROIs represent only one possible measure of individual differences that can be derived from task-fMRI data. For example, several multivariate methods have been proposed to increase the reliability and predictive utility of task-fMRI measures by exploiting the high dimensionality inherent in fMRI data (Dubois & Adolphs, 2016; Yarkoni & Westfall, 2017). To name a few, the reliability of task-fMRI may be improved by developing measures with latent-variable models (Cooper, Jackson, Barch, & Braver, 2019), measuring individual differences in representational spaces with multivoxel pattern analysis (Norman, Polyn, Detre, & Haxby, 2006), and training cross-validated machine-learning models that establish reliability through prediction of individual differences in independent samples (Yarkoni & Westfall, 2017). In addition, in many already-collected data sets, task fMRI can be combined with resting-state fMRI data to produce reliable measures of intrinsic functional connectivity (Elliott et al., 2019). Thus, there are multiple approaches available to maximize the value of existing task-fMRI data sets in the context of biomarker discovery and individual-differences research.

Create a norm of reporting the reliability of task-fMRI measures. The "replicability revolution" in psychological science provides a timely example of how rapidly changing norms can shape research practices and standards. In just a few years, practices to enhance replicability, such as preregistration of hypotheses and analytic strategies, have risen in popularity (Nosek, Ebersole, DeHaven, & Mellor, 2018). We believe similar norms would be beneficial for task fMRI in the context of biomarker discovery and brain–behavior mapping. In particular, researchers should report the reliabilities for all task-fMRI measures whenever they are used to study individual differences. In doing so, however, researchers need to ensure adequate power to evaluate test-retest reliability with confidence. Given that correlations begin to stabilize with around 150 observations (Schönbrodt & Perugini, 2013), our confidence in the reliability of any specific task will depend on collecting larger test-retest data sets. We provide evidence that the task-fMRI literature generally has low reliability; however, because of the relatively small size of each test-retest sample reported here, we urge readers to avoid drawing strong conclusions about the reliability of specific fMRI tasks. In the pursuit of precise reliability estimates, researchers must collect larger test-retest samples, explore test-retest moderators (e.g., the test-retest interval), and avoid reporting inflated reliabilities that can arise from circular statistical analyses (for detailed recommendations, see Kriegeskorte et al., 2010, and Vul et al., 2009).

Researchers can also provide evidence of between-subjects reliability in the form of internal consistency. Although test-retest reliability provides an estimate of

stability over time that is suited for trait and biomarker research, it is a conservative estimate that requires extra data collection and can be undermined by habituation effects and rapid fluctuations (Hajcak, Meyer, & Kotov, 2017). In some cases, internal consistency will be more practical because it is cheaper, as it does not require additional data collection and can be used in any situation in which the task-fMRI measure of interest involves multiple trials. Internal consistency is particularly well suited for measures that are expected to change rapidly and index transient psychological states (e.g., current emotions or thoughts). However, internal consistency alone is not adequate for prognostic biomarkers. Establishing a norm of explicitly reporting measurement reliability would increase the replicability of task-fMRI findings and accelerate biomarker discovery.

More data from more people. Our ability to detect reliable individual differences using task fMRI will depend, in part, on the field embracing two complementary improvements to the status quo: (a) more people per study and (b) more data per person. It has been suggested that neuroscience is generally an underpowered enterprise and that small sample sizes undermine fMRI research in particular (Button et al., 2013). The results presented here suggest that this "power failure" may be further compounded by low reliability in task fMRI. The median sample size in fMRI research is 28.5 (Poldrack et al., 2017). However, as shown in Figure 1, task-fMRI measures with ICCs of .397 (the meta-analytic mean reliability) would require a total sample of more than 214 to achieve 80% power to detect brain–behavior correlations of .3, a moderate effect size equal to the size of the largest replicated brain–behavior association (Elliott et al., 2018). For an r of .1 (a small effect size common in psychological research; Funder & Ozer, 2019), adequately powered studies require a total sample of more than 2,000. And these calculations are actually best-case scenarios, given that they assume perfect reliability of the second behavioral variable (see Fig. 1). Increasing the sample size of task-fMRI studies and requiring power analyses that take into account unreliability represent a meaningful way forward for boosting the replicability of individual-differences research with task fMRI.

Without substantially higher reliability, task-fMRI measures will fail to provide biomarkers that are meaningful on an individual level. One promising method to improve the reliability of fMRI is to collect more data per person. This approach has been shown to improve the reliability of functional connectivity (Elliott et al., 2019; Gratton et al., 2018), and preliminary efforts suggest this may be true for task fMRI as well (Gordon et al., 2017). Pragmatically, collecting additional fMRI data will be burdensome for participants, especially in children and clinical populations, where longer scan times often result in more data artifacts, particularly from increased motion. Naturalistic fMRI represents one potential solution to this challenge. In naturalistic fMRI, participants watch stimulus-rich movies during scanning instead of completing traditional cognitive neuroscience tasks. Initial efforts suggest that movie watching is highly engaging for participants, allows more data collection with less motion, and may even better elicit individual differences in brain activity by emphasizing ecological validity over experimental control (Vanderwal, Eilbott, & Castellanos, 2018). As the field launches large-scale neuroimaging studies (e.g., HCP, UK Biobank, and Adolescent Brain Cognitive Development) in the pursuit of brain biomarkers of disease risk, it is critical that we are confident in the psychometric properties of task-fMRI measures. This will require funders to advocate and support the collection of more data from more people.

Develop tasks from the ground up to optimize reliable and valid measurement. Instead of continuing to adopt fMRI tasks from experimental studies emphasizing within-subjects effects, we need to develop new tasks (and naturalistic stimuli) from the ground up with the goal of optimizing their utility in individual-differences research (i.e., between-subjects effects). Psychometrics provides many tools and methods for developing reliable individual-differences measures that have been underutilized in task-fMRI development. For example, stimuli in task fMRI could be selected on the basis of their ability to maximally distinguish groups of people or to elicit reliable between-subjects variance. As noted in the first recommendation, psychometric tools for test construction could be adopted to optimize reliable task-fMRI measures, including item analysis, latent variable modeling, and internal-consistency measures (Crocker & Algina, 2006).

Conclusion

A prominent goal of task-fMRI research has been to identify abnormal brain activity that could aid in the diagnosis, prognosis, and treatment of brain disorders. We find that commonly used task-fMRI measures lack the minimal reliability standards necessary for accomplishing this goal. Intentional design and optimization of task-fMRI paradigms are needed to measure reliable variation between individuals. As task-fMRI research faces the challenges of reproducibility and replicability, the importance of reliability must be stressed as well. In the age of individualized medicine and precision neuroscience, funding is needed for novel task-fMRI

research that embraces the psychometric rigor necessary to generate clinically actionable knowledge.

Appendix

Why is reliability critical for task-functional MRI (fMRI) research? Test-retest reliability is widely quantified using the intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979). The ICC can be thought of as the proportion of a measure's total variance that is accounted for by variation among individuals. An ICC can take on values between −1 and 1, with values close to 1 indicating nearly perfect stability of individual differences across test-retest measurements and values at or below 0 indicating no stability. Classical test theory states that all measures are made up of a true score plus measurement error (Novick, 1965). The ICC is used to estimate the amount of reliable, true-score variance present in an individual-differences measure. When a measure is taken at two time points, the variance in scores that is due to measurement error will consist of random noise and will fail to correlate with itself across test-retest measurements. However, the variance in a score that is due to the true score will be stable and correlate with itself across time points (Crocker & Algina, 2006). Measures with ICCs of less than .40 are thought to have poor reliability, between .40 and .60 fair reliability, .60 and .75 good reliability, and greater than .75 excellent reliability. An ICC greater than .80 is considered a clinically required standard for reliability in psychology (Cicchetti & Sparrow, 1981).

Reliability is critical for research because the correlation observed between two measures, A and B, is constrained by the square root of the product of each measure's reliability (Nunnally, 1959):

$$r(A_{\text{observed}}, B_{\text{observed}}) = r(A_{\text{true}}, B_{\text{true}}) \times \sqrt{\text{Reliability}\,(A_{\text{observed}}) \times \text{Reliability}(B_{\text{observed}})}\,.$$

Low reliability of a measure reduces statistical power and increases the sample size required to detect a correlation with another measure. Figure 1 shows sample sizes required for 80% power to detect correlations between a task-fMRI measure of individual differences in brain activation and a behavioral or clinical phenotype across a range of reliabilities of the task-fMRI measure and expected effect sizes. Power curves are given for three levels of reliability of the hypothetical behavioral or clinical phenotype. The first two panels (behavioral ICCs = .6 and .8) represent the most typical scenarios. The figure emphasizes the impact of low reliability at the lower *N* range because most fMRI studies are relatively small (median $N = 28.5$; Poldrack et al., 2017).

Transparency

Action Editor: Brent W. Roberts
Editor: D. Stephen Lindsay
Author Contributions

M. L. Elliott and A. R. Knodt contributed equally to this work and are joint first authors. A. Caspi, A. R. Hariri, T. E. Moffitt, M. L. Elliott, and A. R. Knodt conceived the study and data-analysis plan. M. L. Elliott, A. R. Knodt, and M. L. Sison prepared the MRI data for analysis. M. L. Morris prepared the data for meta-analysis. A. R. Knodt, M. L. Elliott, and M. L. Sison conducted the analyses. M. L. Elliott, A. R. Knodt, A. Caspi, A. R. Hariri, and T. E. Moffitt wrote the manuscript. A. Caspi, A. R. Hariri, T. E. Moffitt, and R. Poulton designed, implemented, and oversaw data collection and generation of the research protocol. S. Ramrakha, D. Ireland, and A. R. Knodt oversaw data collection. All authors discussed the results, contributed to the revision of the manuscript, and approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

The Dunedin Study is supported by National Institute on Aging (NIA) Grant No. R01AG049789, NIA Grant No. R01AG032282, and UK Medical Research Council Grant No. P005918. The Dunedin Multidisciplinary Health and Development Research Unit is supported by the New Zealand Health Research Council and the New Zealand Ministry of Business, Innovation and Employment (MBIE). M. L. Elliott is supported by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-1644868. The Human Connectome Project WU-Minn Consortium are funded by the 16 National Institutes of Health (NIH) Institutes and Centers that support the NIH Blueprint for Neuroscience Research and by the McDonnell Center for Systems Neuroscience at Washington University.

Open Practices

Data were provided in part by the Human Connectome Project WU-Minn Consortium (principal investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657). The Human Connectome Project data are publicly available at http://www.humanconnectomeproject.org/data/. Data from the Dunedin Study are not publicly available because the consent form at the last contact with these participants did not address the broad sharing of their data, nor the risks associated with broad sharing of these data. Because the population is small and participants' date of birth and location are well known, and because of the lack of information in the consent form, the institutional review board concluded that it is not appropriate to share these individual-level data collected through any National Institutes of Health-designated repository or other unrestricted open-access methods. Other investigators may contact the

principal investigator if interested in collaborating on a project that requires the use of individual-level data. R markdown code, results from Google Scholar searches, and intraclass correlation coefficients used for the meta-analysis are available on GitHub at github.com/HaririLab/Publications/tree/master/ElliottKnodt2020PS_tfMRIReliability. The design and analysis plans were posted on Google prior to data analysis (https://sites.google.com/site/moffittcaspiprojects/home/projectlist/knodt_2019).

ORCID iDs

Maxwell L. Elliott https://orcid.org/0000-0003-1083-6277
Avshalom Caspi https://orcid.org/0000-0003-0082-4600
Ahmad R. Hariri https://orcid.org/0000-0003-3052-9880

Acknowledgments

The Dunedin Study was approved by the New Zealand Health and Disability Ethics Committee. We thank the members of the Advisory Board for the Dunedin Neuroimaging Study. We also thank Tim Strauman and Ryan Bogdan for their feedback on an initial draft of this manuscript.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797620916786

References

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.

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, *1191*, 133–155.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: John Wiley. doi:10.1002/9780470743386

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376.

Champely, S. (2018). *Package 'pwr.'* Retrieved from http://cran.r-project.org/package=pwr

Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., & Cox, R. W. (2013). Linear mixed-effects modeling approach to FMRI group analysis. *NeuroImage*, *73*, 176–190.

Chen, G., Taylor, P. A., Haller, S. P., Kircanski, K., Stoddard, J., Pine, D. S., . . . Cox, R. W. (2018). Intraclass correlation: Improved modeling approaches and applications for neuroimaging. *Human Brain Mapping*, *39*, 1187–1206.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*, 127–137.

Cooper, S. R., Jackson, J. J., Barch, D. M., & Braver, T. S. (2019). Neuroimaging of individual differences: A latent variable modeling perspective. *Neuroscience & Biobehavioral Reviews*, *98*, 29–46. doi:10.1016/j.neubiorev.2018.12.022

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Wadsworth.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684. doi:10.1037/h0043943

Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI. *Trends in Cognitive Sciences*, *20*, 425–443.

Edlund, J. E., & Nichols, A. L. (Eds.). (2019). *Advanced research methods for the social and behavioral sciences*. Cambridge, England: Cambridge University Press.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634.

Elliott, M. L., Belsky, D. W., Anderson, K., Corcoran, D. L., Ge, T., Knodt, A., . . . Hariri, A. R. (2018). A polygenic score for higher educational attainment is associated with larger brains. *Cerebral Cortex*, *29*, 3496–3504. doi:10.1093/cercor/bhy219

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.

Fröhner, J. H., Teckentrup, V., Smolka, M. N., & Kroemer, N. B. (2019). Addressing the reliability fallacy in fMRI: Similar group effects may arise from unreliable individual effects. *NeuroImage*, *195*, 174–189.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*, 156–168. doi:10.1177/2515245919847202

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., . . . Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*, 171–178.

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., . . . Dosenbach, N. U. F. (2017). Precision functional mapping of individual human brains. *Neuron*, *95*, 791–807.

Gratton, C., Laumann, T. O., Nielsen, A. N., Greene, D. J., Gordon, E. M., Gilmore, A. W., . . . Petersen, S. E. (2018). Functional brain networks are dominated by stable group and individual factors, not cognitive or daily variation. *Neuron*, *98*, 439–452.

Hajcak, G., Meyer, A., & Kotov, R. (2017). Psychometrics and the neuroscience of individual differences: Internal consistency limits between-subjects effects. *Journal of Abnormal Psychology*, *126*, 823–834.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., . . . Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, *32*, 180–194.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reli-

able individual differences. *Behavior Research Methods*, *50*, 1166–1186.

Herting, M. M., Gautam, P., Chen, Z., Mezher, A., & Vetter, N. C. (2018). Test-retest reliability of longitudinal task-based fMRI: Implications for developmental studies. *Developmental Cognitive Neuroscience*, *33*, 17–26.

Infantolino, Z. P., Luking, K. R., Sauder, C. L., Curtin, J. J., & Hajcak, G. (2018). Robust is not necessarily reliable: From within-subjects fMRI contrasts to between-subjects comparisons. *NeuroImage*, *173*, 146–152.

Kriegeskorte, N., Lindquist, M. A., Nichols, T. E., Poldrack, R. A., & Vul, E. (2010). Everything you never wanted to know about circular analysis, but were afraid to ask. *Journal of Cerebral Blood Flow and Metabolism*, *30*, 1551–1557.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*, 150–157.

Manuck, S. B., Brown, S. M., Forbes, E. E., & Hariri, A. R. (2007). Temporal stability of individual differences in amygdala reactivity. *The American Journal of Psychiatry*, *164*, 1613–1614.

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.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*, 424–430.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, USA*, *115*, 2600–2606.

Novick, M. R. (1965). The axioms and principal results of classical test theory. *ETS Research Bulletin Series*, *1965*(1), i–31. doi:10.1002/j.2333-8504.1965.tb00132.x

Nunnally, J. C. (1959). *Introduction to psychological measurement*. New York, NY: McGraw-Hill.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., . . . Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*, 115–126.

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*, 679–693.

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, Article 813. doi:10.3389/fpsyg.2019.00813

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612. doi:10.1016/j.jrp.2013.05.009

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Swartz, J. R., Knodt, A. R., Radtke, S. R., & Hariri, A. R. (2015). A neural biomarker of psychological vulnerability to future life stress. *Neuron*, *85*, 505–511.

Vanderwal, T., Eilbott, J., & Castellanos, F. X. (2018). Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. *Developmental Cognitive Neuroscience*, *36*, Article 100600.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K., & WU-Minn HCP Consortium. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, *80*, 62–79.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience*, *20*, 365–377.

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—commentary on Vul et al. (2009). *Perspectives on Psychological Science*, *4*, 294–298.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122.