Poor test-retest reliability of task-fMRI: 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[1,3-5], Avshalom Caspi[1,3-5], Ahmad R. Hariri[1]

[1]*Department of Psychology & Neuroscience, Duke University Box 104410, Durham, NC 27708, USA*

[2]*Dunedin Multidisciplinary Health and Development Research Unit, Department of Psychology, University of Otago, 163 Union St E, Dunedin, 9016, NZ*

[3]*Social, Genetic, & Developmental Psychiatry Research Centre, Institute of Psychiatry, Psychology, & Neuroscience, King's College London, De Crespigny Park, Denmark Hill, London SE5 8AF, UK*

[4]*Department of Psychiatry & Behavioral Sciences, Duke University School of Medicine, Durham, NC 27708, USA*

[5]*Center for Genomic and Computational Biology, Duke University Box 90338, Durham, NC 27708, USA*

*These authors contributed equally to this work.

Correspondence:
Ahmad R. Hariri, Ph.D.
Professor of Psychology and Neuroscience
Director, Laboratory of NeuroGenetics
Head, Cognition and Cognitive Neuroscience Training Program
Duke University
Durham, NC 27708, USA
Phone: (919) 684-8408
Email: ahmad.hariri@duke.edu

Classification: Biological Sciences / Neuroscience & Social Sciences / Psychological and Cognitive Sciences

Key words: task-fMRI, reliability, biomarker, psychiatry, neurology, individual differences

Abstract

Identifying brain biomarkers of disease risk and treatment response is a growing priority in neuroscience. The ability to identify meaningful biomarkers is fundamentally limited by measurement reliability; measures that do not yield reliable values are unsuitable as biomarkers to predict clinical outcomes. Measuring brain activity using task-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 with 1,088 participants reporting 1,146 ICCs for task-fMRI revealed poor overall reliability (mean ICC = .397). Second, the test-retest reliabilities of activity in *a priori* regions of interest across 11 commonly used fMRI tasks collected in the Human Connectome Project and the Dunedin Longitudinal Study were poor (ICCs = .067 - .485). Collectively, these findings demonstrate that commonly used task-fMRI measures are not currently suitable for brain biomarker discovery or individual differences research in cognitive neuroscience (i.e., brain-behavior mapping). We review how this state of affairs came to be and consider several avenues for improving the reliability of task-fMRI.

Significance Statement

A biomarker with the potential to be useful in predicting clinical outcomes must yield values that are repeatable. We performed a comprehensive meta-analysis of the test-retest reliability of task-fMRI measures, which are widely adopted for biomarker discovery in neuroscience. We found that the meta-analytic reliability of task-fMRI was poor. We also investigated the reliability of many of the most commonly used task-fMRI measures in two datasets recently collected with cutting-edge scanners and methods. We found generally poor reliability for these task-fMRI measures. These findings indicate that many task-fMRI measures are not currently suitable for biomarker discovery or individual differences research in cognitive neuroscience.

Introduction

Since functional magnetic resonance imaging (fMRI) was introduced in 1992 (1), scientists have had unprecedented ability to non-invasively observe brain activity in behaving humans. In 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 (2). One of the most common forms of BOLD fMRI is based on tasks during which researchers "map" brain activity associated with specific cognitive functions by contrasting (i.e., subtracting) the regional BOLD signal during a control condition from the BOLD signal during a condition in which the brain is engaged in a task. In this way, task-fMRI has given neuroscientists unique insights into the brain basis of human behavior, from basic perception to complex thought (3–5), and has given neurologists and mental-health researchers the opportunity to directly identify dysfunction of the organ responsible for disorders: dementias and mental illnesses (6).

Originally, task-fMRI was primarily used to understand functions supported by the typical or average human brain by measuring within-subject differences in activation between task and control conditions, and averaging them together across subjects 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 subjects view threatening images in comparison with neutral images (7), and increased ventral striatum activity is observed when subjects win money in comparison to when they lose money (8). The robust brain activity elicited using this within-subjects approach led researchers to use the same fMRI tasks to study between-subject differences. The logic behind this strategy is straightforward and alluring: if a brain region activates during a task, then individual differences in the magnitude of that activation may contribute to individual differences in behavior and 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 phenomenon like anxiety and depression (9, 10). 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 offer a powerful approach to discovering biomarkers associated with both risk for disorders and response to treatments (6, 10). 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 **Box 1** and **Fig. 1**). In this way, reliability is critical for both clinical applications and research practice. Indicators that do not yield reliable values are unsuitable as biomarkers to predict clinical health outcomes. That is, if a test is going to be used by doctors to make a diagnosis, or to predict that a patient will develop an illness in the future, then the patient cannot score randomly high on the test at one assessment and low on the test at the next assessment.

To progress toward a cumulative neuroscience of individual differences with clinical relevance we must establish reliable brain measures. While the reliability of task-fMRI has previously been discussed (11), individual studies provide highly variable estimates, often come from small test-retest samples employing a wide-variety of analytic methods, and sometimes reach contradictory conclusions about the reliability of the same tasks (12, 13). 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 datasets, we conducted pre-registered analyses of the test-retest reliabilities of brain activation in *a priori* regions of interest across 11 commonly used fMRI tasks.

Results

Reliability of Individual Differences in Task-fMRI: A Systematic Review and Meta-analysis

We performed a systematic review and meta-analysis following PRISMA guidelines (see **Methods** and Supplemental Fig. S1). 56 articles met 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**). During the study-selection process, we discovered that some analyses calculated many different ICCs (across multiple ROIs, contrasts, and tasks), but only reported a subset of the estimated ICCs that were either statistically significant or reached a minimum ICC threshold. This practice, while usually well-intentioned, leads to inflated reliability estimates by capitalizing on chance with a circular approach. In this approach, ICCs are used twice in the same analysis: first, to select a subset of measures of interest, and a second time to report the reliability of those measures (14). Therefore, we performed separate analyses of data from un-thresholded and thresholded reports.

Fig. 3 shows the test-retest reliability coefficients (ICCs) from 77 substudies reporting un-thresholded values (average N = 19.6). 56% of the values fell into the range of what is considered "poor" reliability (below .4), an additional 24% of the values fell into the range of what is considered "fair" reliability (.4 - .6), and only 20% fell into the range of what is considered "good" (.6 - .75) or "excellent" (above .75) reliability. A random effects meta-analysis revealed an average ICC of .397 (95%

CI, .330 - .460; P < .001), which is in the "poor" range (15). There was evidence of between-study heterogeneity (I^2 = 31.6; P = 0.04).

As expected, the meta-analysis of 13 substudies that only reported ICCs above a minimum threshold (average N = 24.2) 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 un-thresholded 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-study heterogeneity (I^2 = 17.9; P = 0.54).

A moderator analysis of all substudies revealed significantly higher reliability for studies that thresholded based on ICC (Q_M = 6.531, df = 1, P = .010; β = .140). In addition, ROIs located in the cortex had significantly higher ICCs than those located in the subcortex (Q_M = 114.476, df = 1, P < .001; β = .259). However, we did not find evidence that the meta-analytic estimate was moderated by task type, task design (i.e., event-related versus blocked), task length, test-retest interval, ROI type (i.e., structural versus functional), sample type (i.e., healthy versus clinical), or number of citations per year. See Supplemental Table S1 for details on all moderators tested. Finally, we tested for publication bias using the Egger random effects regression test (16) and found no evidence for bias (Z = .707, P = .480; Supplemental Fig. S2).

The results of the meta-analysis were illuminating, but not without interpretive difficulty. 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 pre-processing and analysis pipelines, leaving open the possibility that reliability has improved with more advanced technology and consistent practices. To address these limitations and possibilities, we conducted pre-registered analyses of two new datasets,

using state-of-the-art scanners and practices to assess individual differences in commonly used tasks tapping a variety of cognitive and affective functions.

Reliability of Individual Differences in Task-fMRI: Pre-registered Analyses in Two New Datasets

We evaluated test-retest reliabilities of activation in *a priori* regions of interest for 11 commonly used fMRI tasks (see **Methods**). In the Human Connectome Project (HCP), 45 participants were scanned twice using a custom 3T Siemens scanner, on average 140 days apart (sd = 67.1 days), using seven tasks targeting emotion, reward, cognitive control, motor, language, social cognition, and relational processing. In the Dunedin Study, 20 participants were scanned twice using a 3T Siemens Skyra, on average 79 days apart (sd = 10.3 days), using four tasks targeting emotion, reward, cognitive control, and episodic memory. 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 **Methods**).

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 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 subjects and sessions; see warm-colored maps in **Fig. 4** for the three tasks in common between the two studies, and Supplemental Table S2 for statistics across all tasks).

Reliability of regional activation. We investigated the reliability of task activation in both datasets using four steps. First, we tested the reliability of activation in the target ROI for each task. Second, for each task we also evaluated the reliability of activation in the other seven *a priori* ROIs. This was done to test if the reliability of target ROIs was higher than the reliability of activation in other

("non-target") brain regions and to identify any tasks or regions with consistently high reliability. Third, we re-estimated reliability using activation in the left and right hemispheres separately to test if the estimated reliability was harmed by averaging across the hemispheres. Fourth, we tested if 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).

Reliability of regional activation in the Human Connectome Project. First, as shown by the estimates circled in black in **Fig. 5**, across the seven fMRI tasks, activation in anatomically defined target ROIs had low reliability (mean ICC = .246; 95% CI, .135 - .357). 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 non-target ROIs was also low (**Fig. 5**; mean ICC = .238; 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 = .191 in target ROIs and .194 in non-target ROIs, mean right ICC = .262 in target ROIs and .237 in non-target ROIs; Supplemental Fig. S3).

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 Supplemental Table S3 for all ICCs).

As an additional step, to account for the family structure present in the HCP, we re-estimated reliability after removing one of each sibling/twin pair in the test-retest sample. Reliability in bilateral anatomical ROIs in the subsample of N=26 unrelated individuals yielded reliabilities very similar to the overall sample (mean ICC = .281 in target ROIs and .217 in non-target ROIs; Supplemental Fig. S4).

Reliability of regional activation in the Dunedin Study. First, as shown by the estimates circled in black in **Fig. 5**, activation in the anatomically defined target ROI 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 non-target 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 averaged ROIs (mean left ICC = .243 in target ROIs and .202 in non-target ROIs, mean right ICC = .358 in target ROIs and .192 in non-target ROIs; Supplemental Fig. S5).

Fourth, functionally defined ROIs again did not meaningfully improve reliability (mean ICC = .325; 95% CI, .197 - .453; see Supplemental Table S4).

Reliability of structural measures. To provide a benchmark for evaluating task-fMRI, we investigated the reliability of two commonly used structural MRI measures: cortical thickness and surface area (17). Consistent with prior evidence (18, 19) that structural MRI phenotypes have excellent reliability (i.e., ICCs > .9), 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 datasets, respectively. In the HCP, parcel-wise (i.e., regional) cortical thickness reliabilities averaged .886 (range .547 - .964), with 100% crossing the "fair" threshold, 98.6% the "good" threshold, and 94.2% 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 "good", and 84.7% above “excellent.” For total surface area, ICCs were .999 and .996 in the HCP and Dunedin Study datasets, respectively. In the HCP, parcel-wise surface area ICCs averaged .937 (range .526 - .992), with 100% crossing the "fair" threshold, 98.9% crossing the "good" threshold, and

96.9% crossing the "excellent" threshold. In the Dunedin Study, surface area ICCs averaged .942 (range .572 - .991), with 100% above the "fair" threshold, 99.7% above "good", and 98.1% above "excellent".

Discussion

We found evidence that commonly used task-fMRI measures 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 (15)) and far below the recommended cutoffs for clinical application (ICC = .8) or individual-level interpretation (ICC = .9) (20). Of course, not all task-fMRI measures are the same, and it is not possible to assign a single reliability estimate to all individual-difference 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 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, non-target ROIs (mean ICC = .270 for target, .228 for non-target) 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 meta-analysis estimate (meta-analytic ICC = .397), possibly owing to the phenomenon that pre-registered analyses yield smaller effect sizes than past publications without pre-registration (21).

It has been suggested that neuroscience is an underpowered enterprise, and that small sample sizes undermine fMRI research, in particular (22, 23). The current results 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 (24). However, as shown in **Fig. 1,** task-fMRI measures with ICCs of .397 (the meta-analytic mean reliability) would require N > 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 associations (25, 26). For r = .1 (a small effect size common in psychological research (27)), adequately powered studies require N > 2,000. And, these calculations are actually best-case scenarios given that they assume perfect reliability of the second "behavioral" variable (see Supplemental Fig. S6 for power estimates with the measurement reliability consistent (i.e., lower) with most behavioral measures of interest).

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" (28). 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 error and needs to be minimized in order to detect the largest experimental effect. For the correlational investigator, variation between people is the primary unit of analysis and must be measured carefully in order to extract reliable individual differences (28, 29).

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-subject 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-subject robustness is, then, often inappropriately invoked to suggest between-subject reliability,

despite the fact that reliable within-subject experimental effects at a group level can arise from unreliable between-subjects measurements (30).

This reasoning is not unique to task-fMRI research. Behavioral measures that elicit robust group effects have been shown to have low between-subjects reliability; for example, the mean test-retest reliability of the Stroop Test (ICC = .45) (29) is strikingly similar to the mean reliability reported for the 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 (31, 32), demonstrate high test-retest reliability. Thus, it is not the tool that is problematic but rather the strategy of adopting tasks developed for experimental cognitive neuroscience; these appear to be poorly suited for reliably measuring differences in brain activation between people.

Recommendations and Future Directions

We next consider several avenues for improving the reliability of task-fMRI as well as maximizing the value of existing datasets. Some can be actioned now, whereas others will require innovation and development.

1) Immediate opportunities with previously collected task-fMRI data

Contrast-based activation values extracted from ROIs, while by far the most commonly reported in the literature, represent only one possible measure of individual differences that can be derived from fMRI data. For example, 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 (33, 34). To name a few, the reliability of task-fMRI may be improved by developing measures with latent variable models (35), measuring individual differences in representational spaces with multi-voxel pattern

analysis (36), and training cross-validated machine learning models that establish reliability through prediction of individual differences in independent samples (34). Further, instead of using task-fMRI to derive measures of contrast-based brain activation, task-fMRI data can be combined with resting-state fMRI data to produce reliable measures of intrinsic functional connectivity that have been shown to be better biomarkers of individual differences (32, 37). Thus, there are multiple actionable approaches to maximizing the value of existing task-fMRI datasets in the context of biomarker discovery and individual differences research.

2) Stop double-dipping when reporting reliability

"Double-dipping" arises from circular statistical reporting in which researchers report only the largest effect sizes after statistically thresholding a large number of noisy measures (14, 38). Double-dipping is being eradicated from task-fMRI research, but our meta-analysis found that it was still common in test-retest reliability studies (see **Fig. 3**). Studies that double-dipped reported reliability estimates that were on average 75% higher than those without double-dipping (ICC = .705 with double-dipping, ICC = .397 without). These spuriously inflated reliabilities are likely to mislead researchers into thinking that task-fMRI measures are reliable, and this conclusion may be compounded by small sample sizes (39). ROIs should be defined independently from ICC calculations and all ICCs from all brain regions that are investigated should be reported regardless of statistical significance.

3) Create a norm of reporting between-subjects reliability for all fMRI studies of individual differences

The "replicability revolution" in psychological science (40) provides a timely example of how rapidly changing norms can shape research practices and standards. In just a few years, practices to enhance replicability, like pre-registration of hypotheses and analytic strategies, have risen in popularity (41). We believe similar norms would be beneficial for task-fMRI in the context of biomarker discovery

and brain-behavior mapping, particularly the reporting of reliabilities for all task-fMRI measures that are used to study individual differences. Researchers can provide evidence in the form of between-subjects reliability such as test-retest or internal consistency. While 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 (42). 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 where the task-fMRI measure of interest is derived from repeated observations (43). 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, particularly when combined with large sample sizes, and accelerate biomarker discovery.

4) Develop tasks from the ground up to optimize reliable and valid measurement

As already mentioned, task-fMRI measures have been largely developed for experimental cognitive neuroscience where within-subjects effects are prioritized. Instead of adopting these measures, new tasks could be developed 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 that elicit brain activity that maximally distinguishes groups of subjects could be selected to maximize discriminant validity. Many psychometric tools for test construction could be adopted to create reliable task-fMRI measures including item analysis, latent variable modelling, and internal-consistency measures (44).

5) Be wary of difference scores (i.e., contrasts)

Change scores, which are produced by subtracting two measures, will always have lower reliability than their constituent measures (29). Currently, the majority of task-fMRI measures are based on contrasts between conditions (i.e., change scores), undermining their reliability (45). However, the widespread use of contrasts in task-fMRI is largely a vestige of experimental cognitive neuroscience. While experimental research aims to isolate cognitive processes through subtraction, there is no conceptual reason that individual differences research should use contrasts as the measure of interest. Instead, beta estimates from regressors of interest can be used directly. More specifically, measures can be developed for psychometric rigor by finding beta estimates that produce reliable variation between subjects, display internal consistency and, ultimately, construct validity (46).

6) Embrace ecological validity over experimental control

Individual differences in behavior, including psychopathology, arise from how the brain processes, perceives, and responds to the world. Tasks from cognitive neuroscience rarely approximate the richness of the human environment, instead preferring strict control over stimuli that "isolate" a single cognitive process. However, if the goal is to maximize reliable variation, individual differences may be better revealed when subjects are exposed to complex stimuli that elicit ecologically valid brain activity. One solution may be found in the growing field of "naturalistic fMRI," which surrenders experimental control by exposing individuals to rich audiovisual stimuli that contain complex social relationships, gripping emotional scenes, and even fear-inducing violence (47). While audio-visual stimuli can be hand-coded for variables of interest, there are now a number of tools for automatic feature extraction including object labelling, text analysis, sentiment analysis, and face detection (48). The field of naturalistic fMRI is growing in popularity and provides a frontier for fMRI researchers looking to develop more reliable measures of brain function (49).

Conclusion

A prominent goal of task-fMRI research has been to identify abnormal brain activity that could aid diagnosis, prognosis, and treatment of brain disorders. We find that commonly used task-fMRI measures lack minimal reliability standards necessary for accomplishing this goal. Intentional design and optimization of fMRI tasks are needed to measure reliable variation between individuals. As task-fMRI research faces the challenges of reproducibility and replicability, we draw attention to the importance of reliability as well. In the age of individualized medicine and precision neuroscience, task-fMRI research must embrace the psychometric rigor needed to generate clinically actionable knowledge.

Materials and Methods

Meta-analytic Reliability of Task-fMRI

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 include all of the terms “ICC,” “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. 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.

Study Selection and Data Extraction. One author (MM) screened all titles and abstracts before the full texts were reviewed (by authors MLE and ARK). We included all original, peer-reviewed empirical articles that reported test-retest reliability estimates for activation during a BOLD fMRI task. Articles (or in some cases, sets of ICCs within articles) were excluded if they had a test-retest interval of

less than one day, if the ICCs were from a longitudinal or experimental study that was designed to assess change, if they did not report ICCs based on measurements from the same MRI scanner and/or task, or if they reported reliability on something other than activation measures across subjects (e.g., spatial extent of activation or voxel-wise patterns of activation within subjects).

Two authors (MLE and ARK) extracted data about sample characteristics (study year, sample size, healthy versus clinical), study design (test-retest interval, event-related or blocked, task length, and task type), and ICC reporting (i.e., was the ICC thresholded?). For each article, every reported ICC meeting the above study-selection requirements was recorded.

Statistical Analyses. For most of the studies included, no standard error or confidence interval for the ICC was reported. Therefore, in order to include as many estimates as possible in the meta-analysis, we estimated the standard error of all ICCs using the fisher r-to-Z transformation for ICC values (50, 51).

A random-effects multilevel meta-analytic model was fit using tools from the metafor package in R (52). In this model, ICCs and standard errors were averaged within each unique sample, task, and test-retest interval (or "substudy") within each study (53). For the results reported in the Main Article, the correlation between ICCs in each substudy was assumed to be 1 so as 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 Supplemental Fig. S7). In the meta-analytic model, substudies were nested within studies to account for non-independence of ICCs estimated within the same study. Meta-analytic summaries were estimated separately for substudies that reported ICC values that had been thresholded (i.e., when studies calculated multiple ICCs, but only reported values above a minimum threshold) because of the documented spurious inflation of effect sizes that occur when only statistically significant estimates are reported (14, 38, 39).

To test for effects of moderators, a separate random-effects multilevel model was fit to all 1,146 ICCs (i.e., without averaging within each substudy, since many substudies included ICCs with different values for one or more moderators). To account for non-independence, ICCs were nested within substudies, which in turn were nested within studies.

Analyses of New Datasets

Human Connectome Project (HCP). This is a publicly available dataset that includes 1,206 participants with extensive structural and functional MRI (54). 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/cognitive control 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 (55)), a social cognition / theory of mind task (targeting the lateral fusiform gyrus, superior temporal sulcus, and other "social-network" regions (56)), a relational processing / dimensional change detection task (targeting the rostrolateral prefrontal cortex (57), or rlPFC), and an emotional processing face-matching 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 (NZ) and followed to age 45 years. 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 (with a mean interval between scans of 79 days).

Functional MRI was collected during four tasks targeting neural "hubs" in four different domains: an emotion processing face-matching task (targeting the amygdala), a cognitive control Stroop task (targeting the dlPFC and the dorsal anterior cingulate cortex), a monetary incentive delay reward task (targeting the ventral striatum), and an episodic memory face-name encoding task (targeting the hippocampus). See Supplemental Methods for additional details, including fMRI pre-processing, for both datasets.

ROI Definition

Individual estimates of regional brain activity were extracted according to two commonly used approaches. First, we extracted average values from *a priori* anatomically defined regions. We identified the primary region of interest (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 5mm spheres around the group-level peak voxel within the target anatomical ROI for each task (across all subjects and sessions). This is a commonly used strategy for capturing the location of peak activation in each subject despite inter-subject variability in the location of activation. See Supplemental Materials for further details on ROI definition and peak voxel location.

We report analyses based on anatomically defined ROIs in the Main Article and report sensitivity analyses using functional ROIs in the Supplement.

Reliability Analysis

Subject-level BOLD signal change estimates were extracted for each task, ROI, and scanning session. Reliability was quantified using a 2-way mixed effects intraclass correlation coefficient (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) by Shrout and Fleiss, and defined as:

$$ICC\ (3,1) = (BMS - EMS) / (BMS + (k-1)*EMS)$$

where BMS = between-subjects mean square, EMS = error mean square, and k = number of "raters," or scanning sessions (in this case 2). We note that ICC (3,1) tracks the consistency of measures between sessions rather than absolute agreement, and is commonly used in studies of task-fMRI test-retest reliability due to the possibility of habituation to the stimuli over time (58).

To test reliability for each task more generally, we calculated ICCs for all target ROIs across all 11 tasks. Since three of the tasks in each study are very similar and target the same region (the emotion, reward, and cognitive control tasks), this resulted in a total of eight ROIs assessed for reliability. To further visualize global patterns of reliability, we also calculated voxel-wise maps of ICC (3,1) using AFNI's 3dICC_REML.R function (59). Finally, to provide a benchmark for evaluating task-fMRI reliability, we determined the test-retest reliability of two commonly used structural MRI measures: cortical thickness and surface area for each of 360 parcels or ROIs (17).

Figures

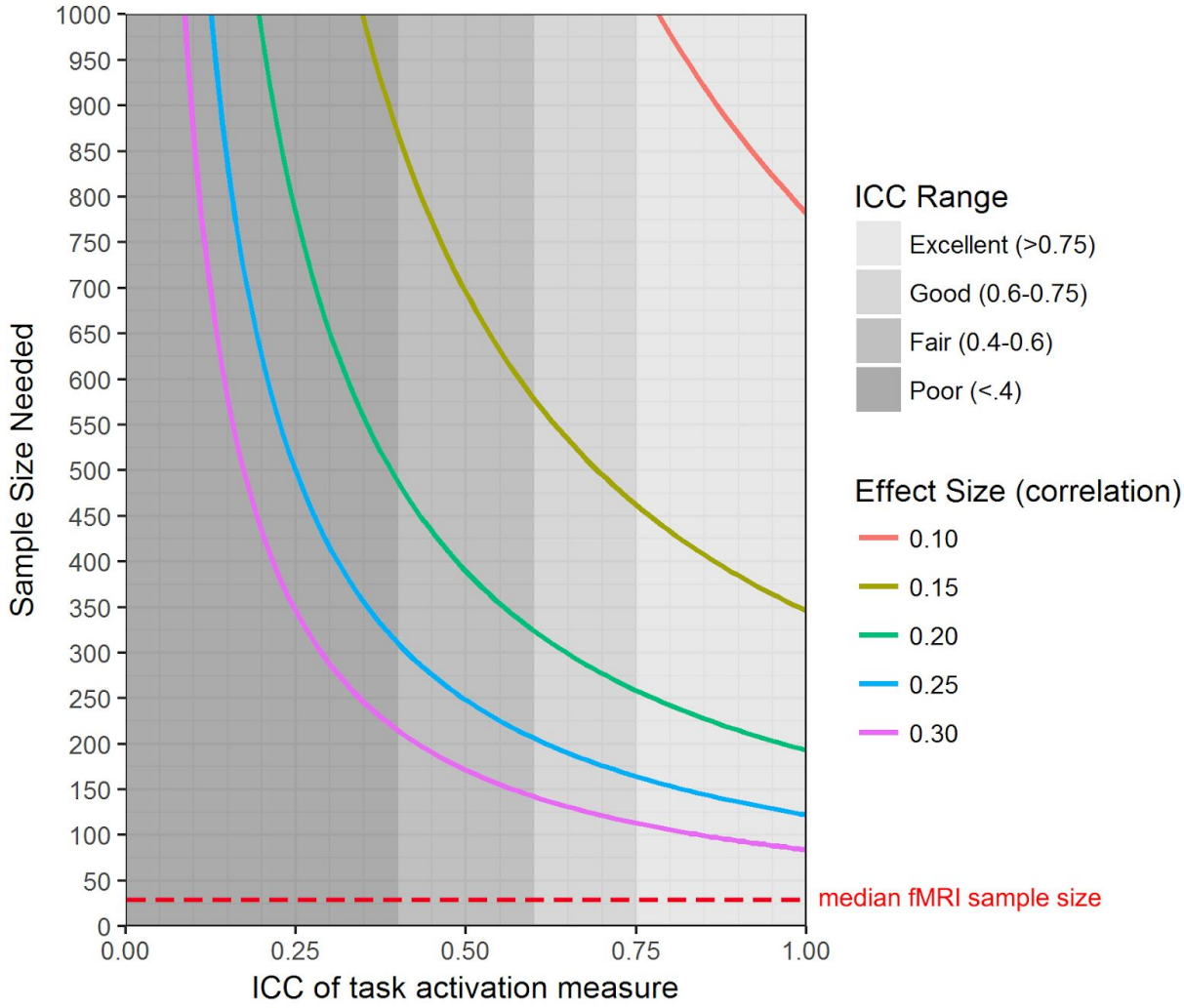

Fig. 1. The influence of task-fMRI test-retest reliability on sample size required for 80% power to detect brain-behavior correlations of effect sizes commonly found in psychological research. Perfect reliability for the behavioral/clinical measure is assumed (see Supplemental Fig. S6 for power curves calculated with less reliable behavioral/clinical measures). The figure was generated using the "pwr.r.test" function in R, with the value for "r" specified according to the attenuation formula in Box 1. The figure emphasizes the impact of low reliability at the lower N range because most fMRI studies are relatively small (median N = 28.5; (24))

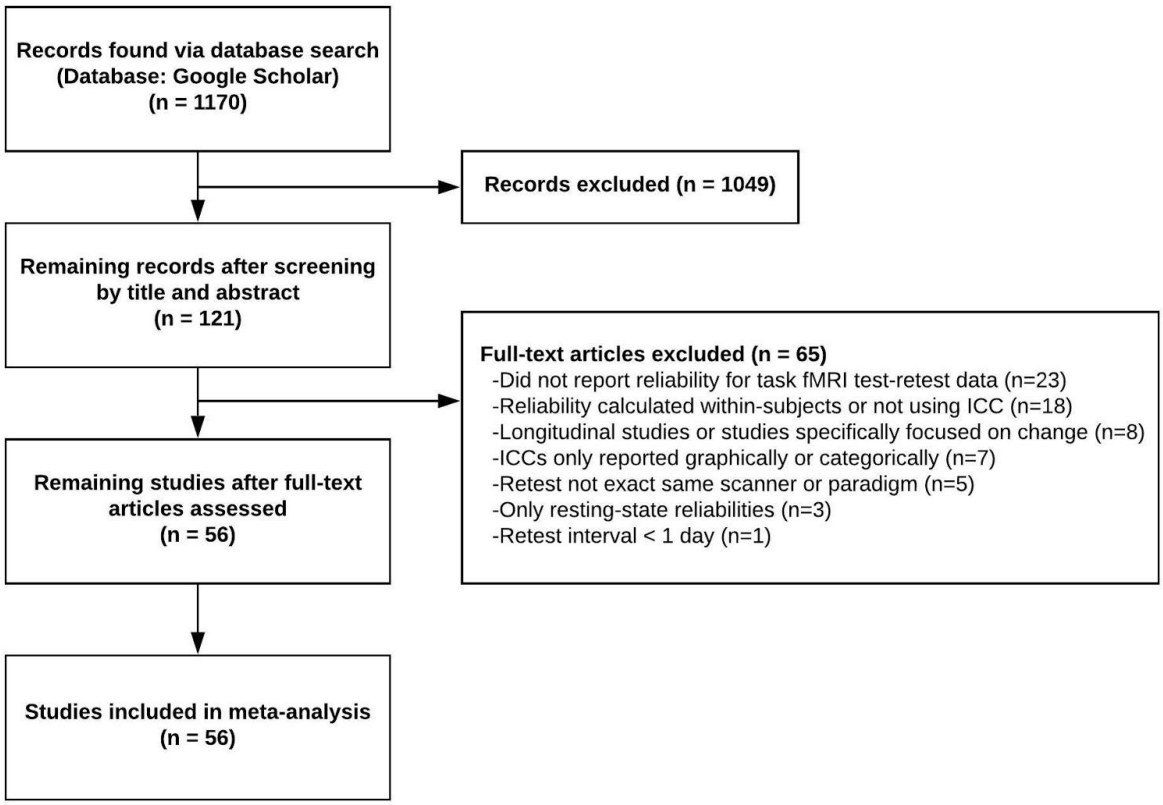

Fig. 2. Flow diagram for systematic literature review and meta-analysis.

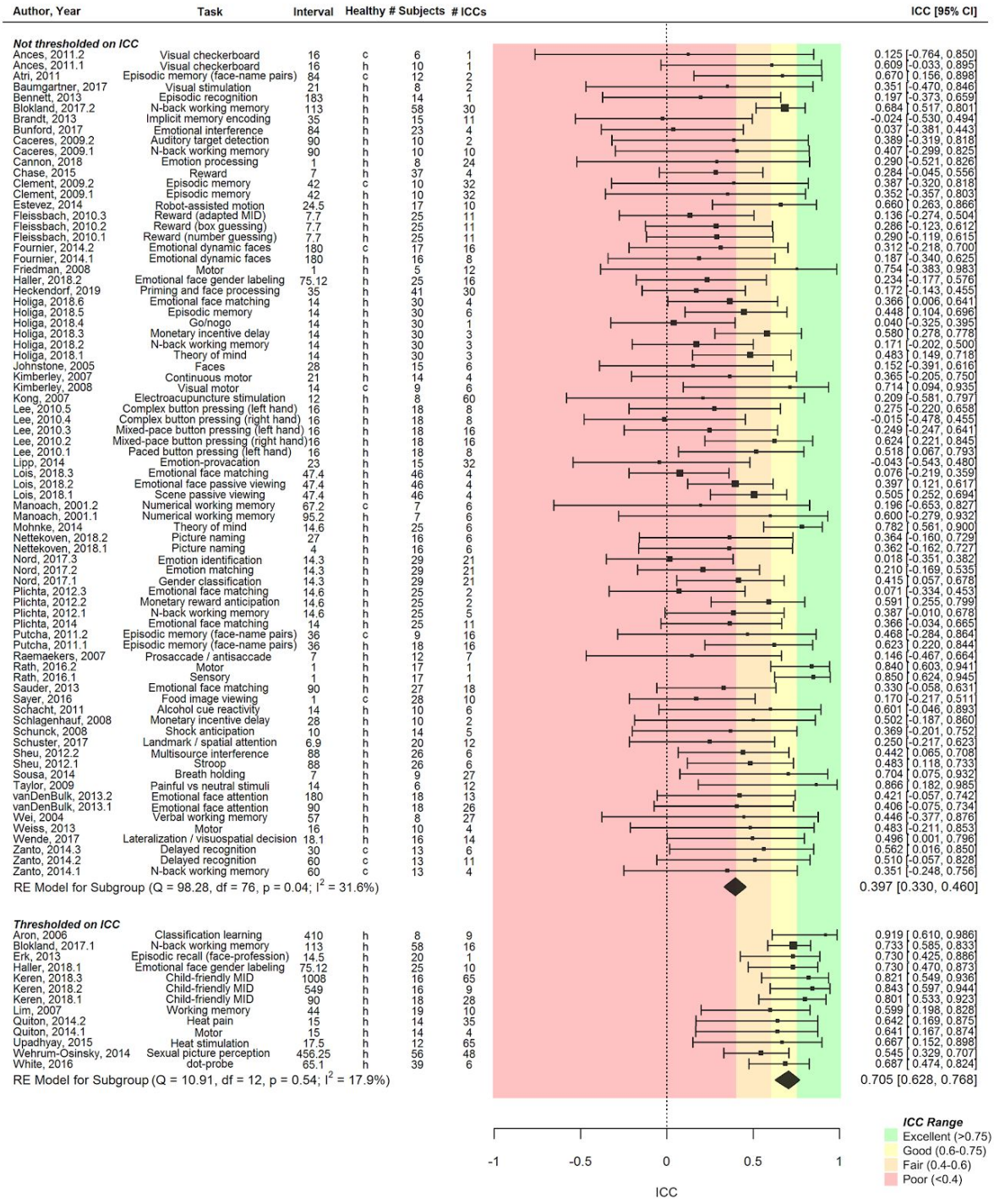

Fig. 3. Forest plot for the results of the meta-analysis of task-fMRI test-retest reliability. The forest plot displays the estimate of test-retest reliability of each task-fMRI measure from all ICCs reported in each study. Studies are split into two sub-groups. The first group of studies reported all investigated ICCs and did not "double-dip" by using a threshold for reporting, thereby allowing for a relatively unbiased estimate of reliability. The second group of studies excluded ICCs that were below a threshold. This double-dipping leads to inflated reliability estimates and therefore these studies were meta-analyzed separately to highlight this bias.

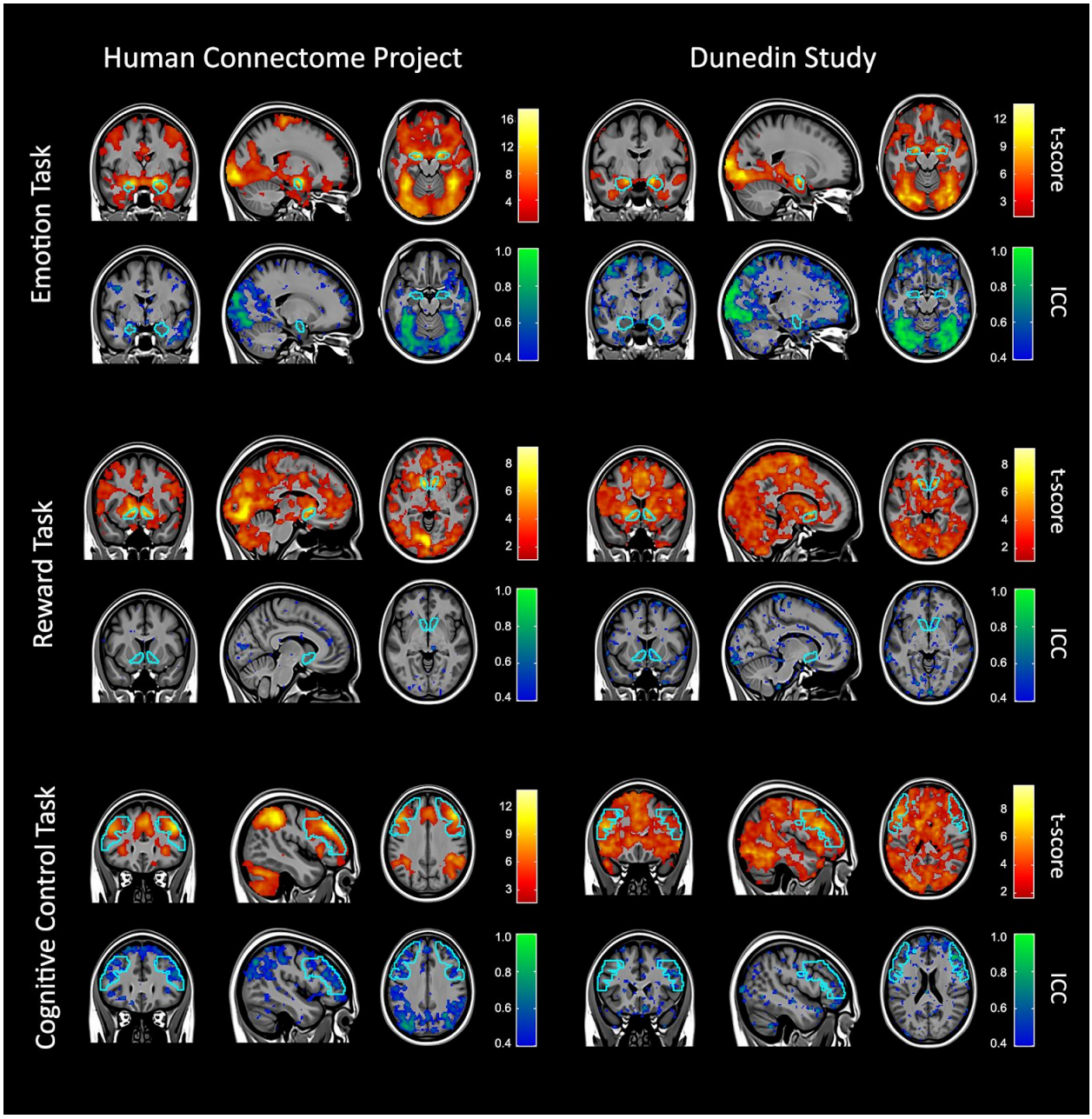

Fig. 4. Whole-brain activation and reliability maps for three task-fMRI measures used in both the Human Connectome Project and Dunedin Study. For each task, a whole-brain activation map of the primary within-subject contrast (t-score) is displayed in warm colors (top) and a whole-brain map of the between-subjects reliability (ICC) is shown in cool colors (bottom). For each task, the target ROI is outlined in sky-blue. These images illustrate that despite robust within-subjects whole-brain activation produced by each task, there is poor between-subjects reliability in this activation, not only in the target ROI but across the whole-brain.

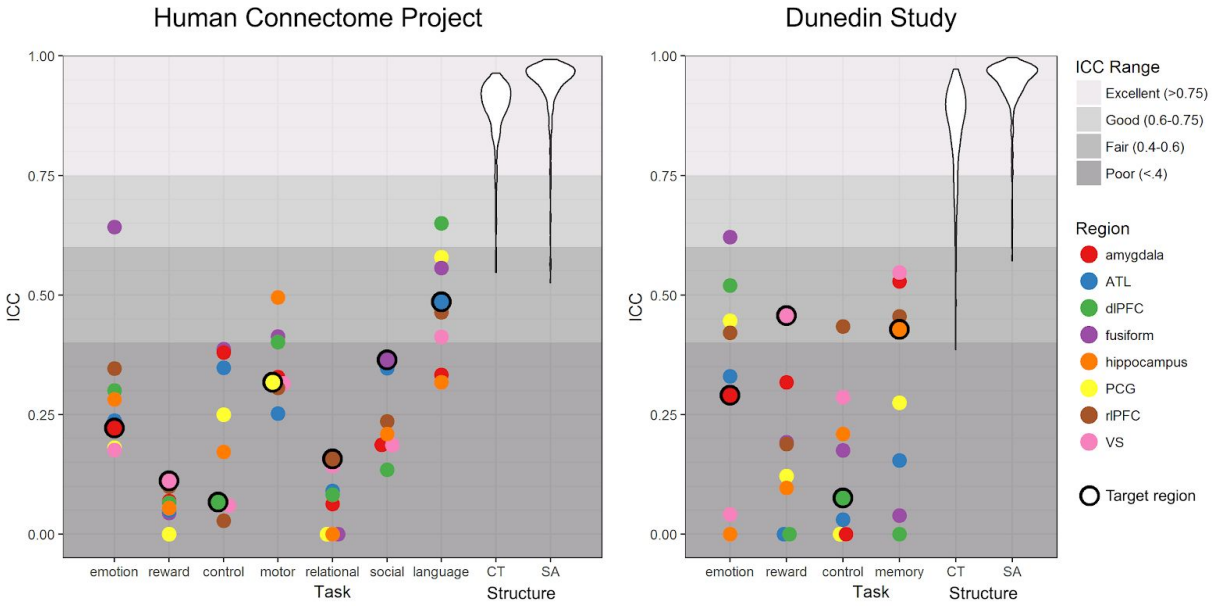

Fig. 5. Test-retest reliabilities of region-wise activation measures in 11 commonly used task-fMRI paradigms. For each task, ICCs were estimated for activation in the *a priori* target ROI (circled in black) and non-target ROIs selected from the other tasks. These plots show that task-fMRI measures of regional activation in both the Human Connectome Project and Dunedin Study are generally unreliable and the ROIs that are "targeted" by the task paradigm rarely are more reliable than non-target ROIs (ATL = anterior temporal lobe, dlPFC = dorsolateral prefrontal cortex, PCG = precentral gyrus, rlPFC = rostrolateral prefrontal cortex, VS = ventral striatum). As a benchmark, ICCs of two common structural MRI measures (CT = Cortical Thickness and SA = Surface Area) are depicted as violin plots representing the distribution of ICCs over 360 parcels (i.e., regions of interest). Note that negative ICCs are set to 0 for visualization.

Box 1: Why is reliability critical for task-fMRI research?

Test-retest reliability is widely quantified using the intraclass correlation coefficient (ICC (60)). ICC can be thought of as the proportion of a measure's total variance that is accounted for by variation between individuals. An ICC can take on values between -1 and 1, with values approaching 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 (61). 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 timepoints, 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 true score will be stable and correlate with itself across timepoints (44). Measures with ICC < .40 are thought to have "poor" reliability, those with ICCs between .40 - .60 "fair" reliability, .60 - .75 "good" reliability, and > .75 "excellent" reliability. An ICC > .80 is considered a clinically required standard for reliability in psychology (15).

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 (62):

$$r(A_{observed}, B_{observed}) = r(A_{true}, B_{true}) * \sqrt{reliability(A_{observed}) * reliability(B_{observed})}$$

Low reliability of a measure reduces statistical power and increases the sample size required to detect a correlation with another measure. **Fig. 1** shows sample sizes required for 80% power to detect correlations between a behavioral/clinical measure and a task-fMRI measure of individual differences in brain activation, across a range of reliabilities of the task-fMRI measure and expected effect sizes. This plot assumes perfect reliability of the hypothetical behavioral/clinical measure, thereby yielding best-case estimates about the impact of low reliability on statistical power (see Supplemental Fig. S6 for power curves calculated for less reliable behavioral/clinical measures).

Acknowledgments

Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

The Dunedin Study was approved by the NZ-HDEC (Health and Disability Ethics Committee). The Dunedin Study is supported by NIA grants R01AG049789 and R01AG032282 and U.K. Medical Research Council grant 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). MLE is supported by the National Science Foundation Graduate Research Fellowship under Grant No. NSF DGE-1644868. Thanks to the members of the Advisory Board for the Dunedin Neuroimaging Study. The authors would also like to thank Tim Strauman and Ryan Bogdan for their feedback on an initial draft of this manuscript.

References

1. Kwong KK, et al. (1992) Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc Natl Acad Sci U S A* 89(12):5675–5679.

2. Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A (2001) Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412(6843):150–157.

3. Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci* 17(11):4302–4311.

4. Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532(7600):453–458.

5. Saxe R, Kanwisher N (2003) People thinking about thinking peopleThe role of the temporo-parietal junction in "theory of mind." *NeuroImage* 19(4):1835–1842.

6. Matthews PM, Honey GD, Bullmore ET (2006) Applications of fMRI in translational medicine and clinical practice. *Nat Rev Neurosci* 7(9):732–744.

7. Hariri AR, Tessitore A, Mattay VS, Fera F, Weinberger DR (2002) The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage* 17(1):317–323.

8. Knutson B, Fong GW, Adams CM, Varner JL, Hommer D (2001) Dissociation of reward anticipation and outcome with event-related fMRI. *Neuroreport* 12(17):3683–3687.

9. Hariri AR (2002) Serotonin Transporter Genetic Variation and the Response of the Human Amygdala. *Science* 297(5580):400–403.

10. Swartz JR, Knodt AR, Radtke SR, Hariri AR (2015) A neural biomarker of psychological vulnerability to future life stress. *Neuron* 85(3):505–511.

11. Bennett CM, Miller MB (2010) How reliable are the results from functional magnetic resonance imaging? *Ann N Y Acad Sci* 1191:133–155.

12. Manuck SB, Brown SM, Forbes EE, Hariri AR (2007) Temporal stability of individual differences in amygdala reactivity. *Am J Psychiatry* 164(10):1613–1614.

13. Nord CL, Gray A, Charpentier CJ, Robinson OJ, Roiser JP (2017) Unreliability of putative fMRI biomarkers during emotional face processing. *Neuroimage* 156:119–127.

14. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci* 12(5):535–540.

15. Cicchetti DV, Sparrow SA (1981) Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic* 86(2):127–137.

16. Egger M, Davey Smith G, Schneider M, Minder C (1997) Bias in meta-analysis detected by a simple, graphical test. *BMJ* 315(7109):629–634.

17. Glasser MF, et al. (2016) A multi-modal parcellation of human cerebral cortex. *Nature*

536(7615):171–178.

18. Han X, et al. (2006) Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage* 32(1):180–194.

19. Maclaren J, Han Z, Vos SB, Fischbein N, Bammer R (2014) Reliability of brain volume measurements: a test-retest dataset. *Sci Data* 1:140037.

20. Guilford JP (1946) New Standards For Test Evaluation. *Educational and Psychological Measurement* 6(4):427–438.

21. Schäfer T, Schwarz MA (2019) The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Front Psychol* 10:813.

22. Button KS, et al. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14(5):365–376.

23. Szucs D, Ioannidis JPA (2017) Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol* 15(3):e2000797.

24. Poldrack RA, et al. (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci* 18(2):115–126.

25. Nave G, Jung WH, Karlsson Linnér R, Kable JW, Koellinger PD (2019) Are Bigger Brains Smarter? Evidence From a Large-Scale Preregistered Study. *Psychol Sci* 30(1):43–54.

26. Elliott ML, et al. (2018) A Polygenic Score for Higher Educational Attainment is Associated with Larger Brains. *Cereb Cortex*. doi:10.1093/cercor/bhy219.

27. Meyer GJ, et al. (2001) Psychological testing and psychological assessment. A review of evidence and issues. *Am Psychol* 56(2):128–165.

28. Cronbach LJ (1957) The two disciplines of scientific psychology. *American Psychologist* 12(11):671–684.

29. Hedge C, Powell G, Sumner P (2018) The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav Res Methods* 50(3):1166–1186.

30. Fröhner JH, Teckentrup V, Smolka MN, Kroemer NB (2019) Addressing the reliability fallacy in fMRI: Similar group effects may arise from unreliable individual effects. *Neuroimage* 195:174–189.

31. Gratton C, et al. (2018) Functional Brain Networks Are Dominated by Stable Group and Individual Factors, Not Cognitive or Daily Variation. *Neuron* 98(2):439–452.e5.

32. Elliott ML, et al. (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.

33. Dubois J, Adolphs R (2016) Building a Science of Individual Differences from fMRI. *Trends Cogn Sci* 20(6):425–443.

34. Yarkoni T, Westfall J (2017) Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspect Psychol Sci* 12(6):1100–1122.

35. Cooper SR, Jackson JJ, Barch DM, Braver TS (2019) Neuroimaging of individual differences: A latent variable modeling perspective. *Neuroscience & Biobehavioral Reviews* 98:29–46.

36. Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10(9):424–430.

37. Greene AS, Gao S, Scheinost D, Constable RT (2018) Task-induced brain state manipulation improves prediction of individual traits. *Nat Commun* 9(1):2807.

38. Vul E, Harris C, Winkielman P, Pashler H (2009) Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspect Psychol Sci* 4(3):274–290.

39. Yarkoni T (2009) Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power-Commentary on Vul et al. (2009). *Perspect Psychol Sci* 4(3):294–298.

40. Nosek BA, et al. (2015) SCIENTIFIC STANDARDS. Promoting an open research culture. *Science* 348(6242):1422–1425.

41. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT (2018) The preregistration revolution. *Proc Natl Acad Sci U S A* 115(11):2600–2606.

42. Hajcak G, Meyer A, Kotov R (2017) Psychometrics and the neuroscience of individual differences: Internal consistency limits between-subjects effects. *J Abnorm Psychol* 126(6):823–834.

43. Streiner DL (2003) Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J Pers Assess* 80(1):99–103.

44. Crocker L, Algina J (2006) *Introduction to Classical and Modern Test Theory* (Wadsworth Publishing Company).

45. Infantolino ZP, Luking KR, Sauder CL, Curtin JJ, Hajcak G (2018) Robust is not necessarily reliable: From within-subjects fMRI contrasts to between-subjects comparisons. *Neuroimage* 173:146–152.

46. Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. *Psychol Bull* 52(4):281–302.

47. Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R (2004) Intersubject synchronization of cortical activity during natural vision. *Science* 303(5664):1634–1640.

48. McNamara Q, De La Vega A, Yarkoni T (2017) Developing a Comprehensive Framework for Multimodal Feature Extraction. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*. doi:10.1145/3097983.3098075.

49. Vanderwal T, Eilbott J, Castellanos FX (2018) Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. *Dev Cogn Neurosci*:100600.

50. Chen G, et al. (2018) Intraclass correlation: Improved modeling approaches and applications for

neuroimaging. *Hum Brain Mapp* 39(3):1187–1206.

51. McGraw KO, Wong SP (1996) Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1(1):30–46.

52. Metafor Package R Code for Meta-Analysis Examples (2019) *Advanced Research Methods for the Social and Behavioral Sciences*:365–367.

53. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009) Introduction to Meta-Analysis. doi:10.1002/9780470743386.

54. Van Essen DC, et al. (2013) The WU-Minn Human Connectome Project: an overview. *Neuroimage* 80:62–79.

55. Binder JR, et al. (2011) Mapping anterior temporal lobe language areas with fMRI: a multicenter normative study. *Neuroimage* 54(2):1465–1475.

56. Wheatley T, Milleville SC, Martin A (2007) Understanding animate agents: distinct roles for the social network and mirror system. *Psychol Sci* 18(6):469–474.

57. Smith R, Keramatian K, Christoff K (2007) Localizing the rostrolateral prefrontal cortex at the individual level. *Neuroimage* 36(4):1387–1396.

58. Plichta MM, et al. (2012) Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. *Neuroimage* 60(3):1746–1758.

59. Chen G, Saad ZS, Britton JC, Pine DS, Cox RW (2013) Linear mixed-effects modeling approach to FMRI group analysis. *Neuroimage* 73:176–190.

60. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 86(2):420–428.

61. Novick MR (1965) THE AXIOMS AND PRINCIPAL RESULTS OF CLASSICAL TEST THEORY. *ETS Research Bulletin Series* 1965(1):i–31.

62. Nunnally JC (1959) *Introduction to Psychological Measurement*.