

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

3
4 Maxwell L. Elliott^{1†}, Annchen R. Knodt^{1†}, David Ireland², Meriwether L. Morris¹, Richie Poulton²,
5 Sandhya Ramrakha², Maria L. Sison¹, Terrie E. Moffitt^{1,3-5}, Avshalom Caspi^{1,3-5}, Ahmad R.
6 Hariri^{1*}

7
8
9 *¹Department of Psychology & Neuroscience, Duke University, Box 104410, Durham, NC 27708,*
10 *USA*

11
12 *²Dunedin Multidisciplinary Health and Development Research Unit, Department of Psychology,*
13 *University of Otago, 163 Union St E, Dunedin, 9016, NZ*

14
15 *³Social, Genetic, & Developmental Psychiatry Research Centre, Institute of Psychiatry,*
16 *Psychology, & Neuroscience, King's College London, De Crespigny Park, Denmark Hill, London*
17 *SE5 8AF, UK*

18
19 *⁴Department of Psychiatry & Behavioral Sciences, Duke University School of Medicine, Durham,*
20 *NC 27708, USA*

21
22 *⁵Center for Genomic and Computational Biology, Duke University Box 90338, Durham, NC*
23 *27708, USA*

24
25 †These authors contributed equally to this work.

26
27 *Correspondence:

28 Ahmad R. Hariri, Ph.D.

29 Professor of Psychology and Neuroscience

30 Director, Laboratory of NeuroGenetics

31 Head, Cognition and Cognitive Neuroscience Training Program

32 Duke University

33 Durham, NC 27708, USA

34 Phone: (919) 684-8408

35 Email: ahmad.hariri@duke.edu

36
37 **Running head:** TASK-FMRI RELIABILITY NOVEL DATA AND META-ANALYSIS

38 Abstract

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

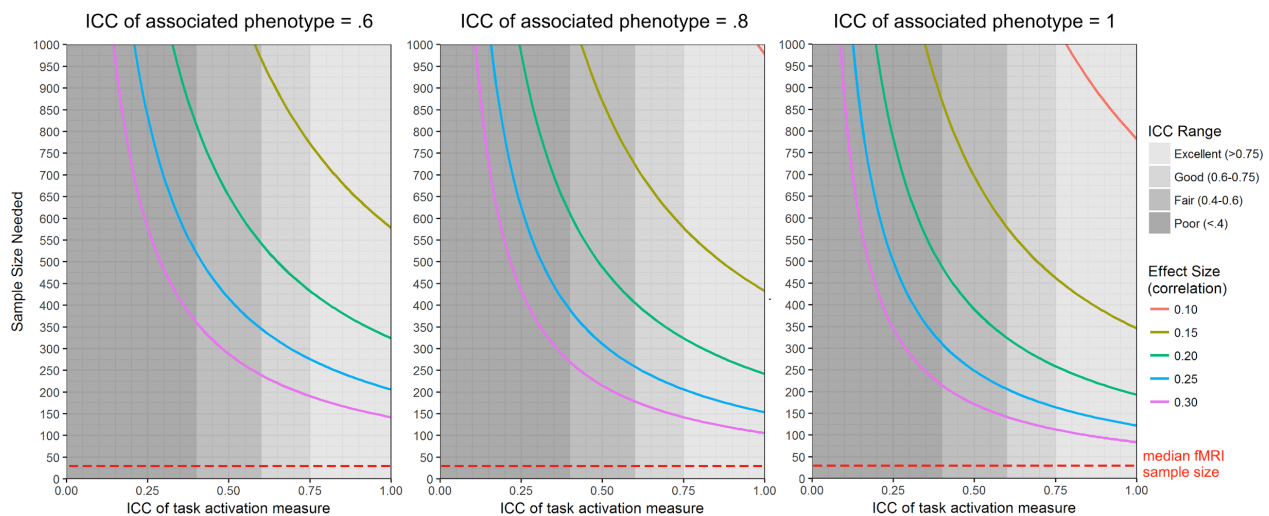
50 **Key words:** Neuroimaging, Individual Differences, Statistical Analysis, Cognitive Neuroscience

51 Introduction

52 Since functional magnetic resonance imaging (fMRI) was introduced in 1992 (Kwong et al., 1992),
53 scientists have had unprecedented ability to non-invasively observe brain activity in behaving humans. In
54 conventional fMRI, regional brain activity is estimated by measuring the blood oxygen level-dependent
55 (BOLD) signal which indexes changes in blood oxygenation associated with neural activity (Logothetis et
56 al., 2001). One of the most common forms of BOLD fMRI is based on tasks during which researchers
57 “map” brain activity associated with specific cognitive functions by contrasting the regional BOLD signal
58 during a control condition with the BOLD signal during a condition of interest. In this way, task-fMRI has
59 given neuroscientists unique insights into the brain basis of human behavior, from basic perception to
60 complex thought, and has given clinicians and mental-health researchers the opportunity to directly measure
61 dysfunction in the organ responsible for disorder.

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

78 The use of task-fMRI to study differences between people heralded the possibility that it could
79 offer a powerful tool for discovering biomarkers for brain disorders (Woo et al., 2017). Broadly, a
80 biomarker is a biological indicator often used for risk stratification, diagnosis, prognosis and evaluation of
81 treatment response. However, to be useful as a biomarker, an indicator must first be reliable. Reliability is
82 the ability of a measure to give consistent results under similar circumstances. It puts a limit on the
83 predictive utility, power, and validity of any measure (see **Box 1** and **Fig. 1**). In this way, reliability is
84 critical for both clinical applications and research practice. Measures with low reliability are unsuitable as
85 biomarkers and cannot predict clinical health outcomes. That is, if a measure is going to be used by
86 clinicians to predict the likelihood that a patient will develop an illness in the future, then the patient cannot
87 score randomly high on the measure at one assessment and low on the measure at the next assessment.



88 **Fig. 1.** The influence of task-fMRI test-retest reliability on sample size required for 80% power to detect
89 brain-behavior correlations of effect sizes commonly found in psychological research. Power curves are
90 calculated for three levels of reliability of the associated behavioral/clinical phenotype. The figure was
91 generated using the “pwr.r.test” function in R, with the value for “r” specified according to the attenuation
92 formula in Box 1. The figure emphasizes the impact of low reliability at the lower N range because most
93 fMRI studies are relatively small (median N = 28.5 (Poldrack et al., 2017)).

94
95
96
97 To progress toward a cumulative neuroscience of individual differences with clinical relevance we
98 must establish reliable brain measures. While the reliability of task-fMRI has previously been discussed
99 (Bennett & Miller, 2010; Herting et al., 2018), individual studies provide highly variable estimates, often
100 come from small test-retest samples employing a wide-variety of analytic methods, and sometimes reach

101 contradictory conclusions about the reliability of the same tasks (Manuck et al., 2007; Nord et al., 2017).
102 This leaves the overall reliability of task-fMRI, as well as the specific reliabilities of many of the most
103 commonly used fMRI tasks, largely unknown. An up-to-date, comprehensive review and meta-analysis of
104 the reliability of task-fMRI and an in-depth examination of the reliability of the most widely used task-
105 fMRI measures is needed. Here, we present evidence from two lines of analysis that point to the poor
106 reliability of commonly used task-fMRI measures. First, we conducted a meta-analysis of the test-retest
107 reliability of regional activation in task-fMRI. Second, in two recently collected datasets, we conducted
108 pre-registered analyses (https://sites.google.com/site/moffittcaspi/projects/home/projectlist/knodt_2019) of
109 the test-retest reliability of brain activation in *a priori* regions of interest across several commonly used
110 fMRI tasks.

111

112 **Methods**

113 **Meta-analytic Reliability of Task-fMRI**

114 We performed a systematic review and meta-analysis following PRISMA guidelines (see
115 Supplemental Fig. S1). We searched Google Scholar for peer reviewed articles written in English and
116 published on or before April 1, 2019 that included test-retest reliability estimates of task-fMRI activation.
117 We used the advanced search tool to find articles that include all of the terms “ICC,” “fmri,” and “retest”,
118 and at least one of the terms “ROI,” “ROIs,” “region of interest,” or “regions of interest.” This search yielded
119 1,170 articles.

120 ***Study Selection and Data Extraction.*** One author (MLM) screened all titles and abstracts before
121 the full texts were reviewed (by authors MLE and ARK). We included all original, peer-reviewed empirical
122 articles that reported test-retest reliability estimates for activation during a BOLD fMRI task. All ICCs
123 reported in the main text and supplement were eligible for inclusion. If ICCs were only depicted graphically
124 (e.g. bar graph), we did our best at judging the value from the graph. Voxel-wise ICCs that were only
125 depicted on brain maps were not included. For ICCs calculated based on more than 2 time points, we used
126 the average of the intervals as the value for interval (e.g. the average of the time between time points 1 and

127 2 and time points 2 and 3 for an ICC based on 3 time points). For articles that reported ICCs from sensitivity
128 analyses in addition to primary analyses on the same data (e.g. using different modeling strategies or
129 excluding certain subjects) we only included ICCs from the primary analysis. We did not include ICCs
130 from combinations of tasks. ICCs were excluded if they were from a longitudinal or intervention study that
131 was designed to assess change, if they did not report ICCs based on measurements from the same MRI
132 scanner and/or task, or if they reported reliability on something other than activation measures across
133 subjects (e.g., spatial extent of activation or multi-voxel patterns of activation within subjects).

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

138 ***Statistical Analyses.*** For most of the studies included, no standard error or confidence interval for
139 the ICC was reported. Therefore, in order to include as many estimates as possible in the meta-analysis, the
140 standard error of all ICCs was estimated using the Fisher r-to-Z transformation for ICC values (Chen et al.,
141 2018; McGraw & Wong, 1996).

142 A random-effects multilevel meta-analytic model was fit using tools from the metafor package in
143 R (“Metafor Package R Code for Meta-Analysis Examples,” 2019). In this model, ICCs and standard errors
144 were averaged within each unique sample, task, and test-retest interval (or “substudy”) within each article
145 (or “study”; (Borenstein et al., 2009)). For the results reported in the Main Article, the correlation between
146 ICCs in each substudy was assumed to be 1 so as to ensure that the meta-analytic weight for each substudy
147 was based solely on sample size rather than the number of ICCs reported. However, sensitivity analyses
148 revealed that this decision had very little impact on the overall result (see Supplemental Fig. S2). In the
149 meta-analytic model, substudies were nested within studies to account for the non-independence of ICCs
150 estimated within the same study. Meta-analytic summaries were estimated separately for substudies that
151 reported ICC values that had been thresholded (i.e., when studies calculated multiple ICCs, but only

152 reported values above a minimum threshold) because of the documented spurious inflation of effect sizes
153 that occur when only statistically significant estimates are reported (Kriegeskorte et al., 2009; Poldrack et
154 al., 2017; Vul et al., 2009; Yarkoni, 2009).

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

165

166 **Analyses of New Datasets**

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

172 The seven tasks employed in the HCP were designed to identify functionally relevant “nodes” in
173 the brain. These tasks included an “n-back” working memory / executive function task (targeting the
174 dorsolateral prefrontal cortex, or dlPFC (Drobyshevsky et al., 2006)), a “gambling” reward / incentive
175 processing task (targeting the ventral striatum (Delgado et al., 2000)), a motor mapping task consisting of
176 foot, hand, and tongue movements (targeting the motor cortex (Drobyshevsky et al., 2006)), an auditory

177 language task (targeting the anterior temporal lobe (Binder et al., 2011)), a social cognition / theory of mind
178 task (targeting the lateral fusiform gyrus, superior temporal sulcus, and other “social-network” regions
179 (Wheatley et al., 2007)), a relational processing / dimensional change detection task (targeting the
180 rostrolateral prefrontal cortex (R. Smith et al., 2007), or rIPFC), and a face-matching emotion processing
181 task (targeting the amygdala (Hariri et al., 2002)).

182 ***Dunedin Multidisciplinary Health and Development Study.*** The Dunedin Study is a longitudinal
183 investigation of health and behavior in a complete birth cohort of 1,037 individuals (91% of eligible births;
184 52% male) born between April 1972 and March 1973 in Dunedin, New Zealand (NZ) and followed to age
185 45 years (Poulton et al., 2015). Structural and functional neuroimaging data were collected between August
186 2016 and April 2019, when participants were 45 years old. In addition, 20 Study members completed the
187 entire scan protocol a second time (with a mean interval between scans of 79 days).

188 Functional MRI was collected during four tasks targeting neural “hubs” in four different domains:
189 a face-matching emotion processing task (targeting the amygdala (Hariri et al., 2002)), a Stroop executive
190 function task (targeting the dlPFC and the dorsal anterior cingulate cortex (Peterson et al., 1999)), a
191 monetary incentive delay reward task (targeting the ventral striatum (Knutson et al., 2000)), and a face-
192 name encoding episodic memory task (targeting the hippocampus (Zeineh et al., 2003)). See Supplemental
193 Methods for additional details, including fMRI pre-processing, for both datasets.

194 ***ROI Definition.*** Individual estimates of regional brain activity were extracted according to two
195 commonly used approaches. First, we extracted average values from *a priori* anatomically defined regions.
196 We identified the primary region of interest (ROI) for each task and extracted average BOLD signal change
197 estimates from all voxels within a corresponding bilateral anatomical mask.

198 Second, we used functionally defined regions based on group-level activation. Here, we generated
199 functional ROIs by drawing 5mm spheres around the group-level peak voxel within the target anatomical
200 ROI for each task (across all subjects and sessions). This is a commonly used strategy for capturing the
201 location of peak activation in each subject despite inter-subject variability in the exact location of the

202 activation. See Supplemental Materials for further details on ROI definition, overlays on the anatomical
203 template (Fig. S3), and peak voxel location (Table S2). We report analyses based on anatomically defined
204 ROIs in the Main Article and report sensitivity analyses using functional ROIs in the Supplement.

205 **Reliability Analysis.** Subject-level BOLD signal change estimates were extracted for each task,
206 ROI, and scanning session. Reliability was quantified using a 2-way mixed effects intraclass correlation
207 coefficient (ICC), with session modeled as a fixed effect, subject as a random effect, and test-retest interval
208 as an effect of no interest. This mixed effects model is referred to as ICC (3,1) by Shrout and Fleiss (1979),
209 and defined as:

$$210 \quad ICC(3,1) = (BMS - EMS) / (BMS + (k-1)*EMS)$$

211 where BMS = between-subjects mean square, EMS = error mean square, and k = number of
212 “raters,” or scanning sessions (in this case 2). We note that ICC (3,1) tracks the consistency of measures
213 between sessions rather than absolute agreement, and is commonly used in studies of task-fMRI test-retest
214 reliability due to the possibility of habituation to the stimuli over time (Plichta et al., 2012).

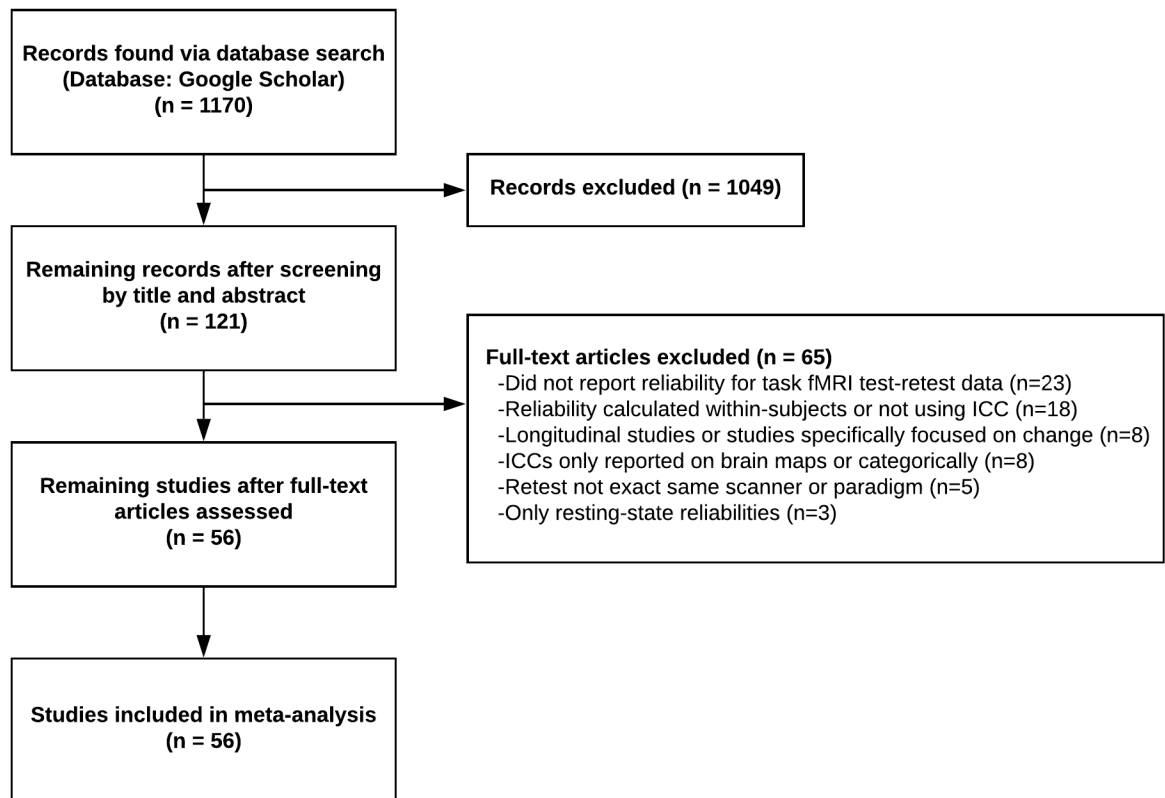
215 To test reliability for each task more generally, we calculated ICCs for all target ROIs across all 11
216 tasks. Since three of the tasks (the emotion, reward, and executive function tasks) were very similar across
217 the HCP and Dunedin Studies and targeted the same region, the same ROI was used for these tasks in both
218 studies, resulting in a total of eight unique target ROIs assessed for reliability. To further visualize global
219 patterns of reliability, we also calculated voxel-wise maps of ICC (3,1) using AFNI’s 3dICC_REML.R
220 function (Chen et al., 2013). Finally, to provide a benchmark for evaluating task-fMRI reliability, we
221 determined the test-retest reliability of three commonly used structural MRI measures: cortical thickness
222 and surface area for each of 360 parcels or ROIs (Glasser et al., 2016) as well as subcortical volume for 17
223 structures. These analyses were pre-registered
224 (https://sites.google.com/site/moffittcaspi/projects/home/projectlist/knodt_2019). Code and data for this
225 manuscript is available at
226 github.com/HaririLab/Publications/tree/master/ElliottKnodt2020PS_tfMRIreliability

227

228 Results

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

230 We identified 56 articles meeting criteria for inclusion in the meta-analysis, yielding 1,146 ICC
231 estimates derived from 1,088 unique participants across 90 distinct substudies employing 66 different task-
232 fMRI paradigms (**Fig. 2**). These articles were cited a total of 2,686 times, with an average of 48 citations
233 per article and 5.7 citations per article, per year. During the study-selection process, we discovered that
234 some analyses calculated many different ICCs (across multiple ROIs, contrasts, and tasks), but only
235 reported a subset of the estimated ICCs that were either statistically significant or reached a minimum ICC
236 threshold. This practice leads to inflated reliability estimates (Kriegeskorte et al., 2010, 2009; Poldrack et
237 al., 2017). Therefore, we performed separate analyses of data from un-thresholded and thresholded reports.



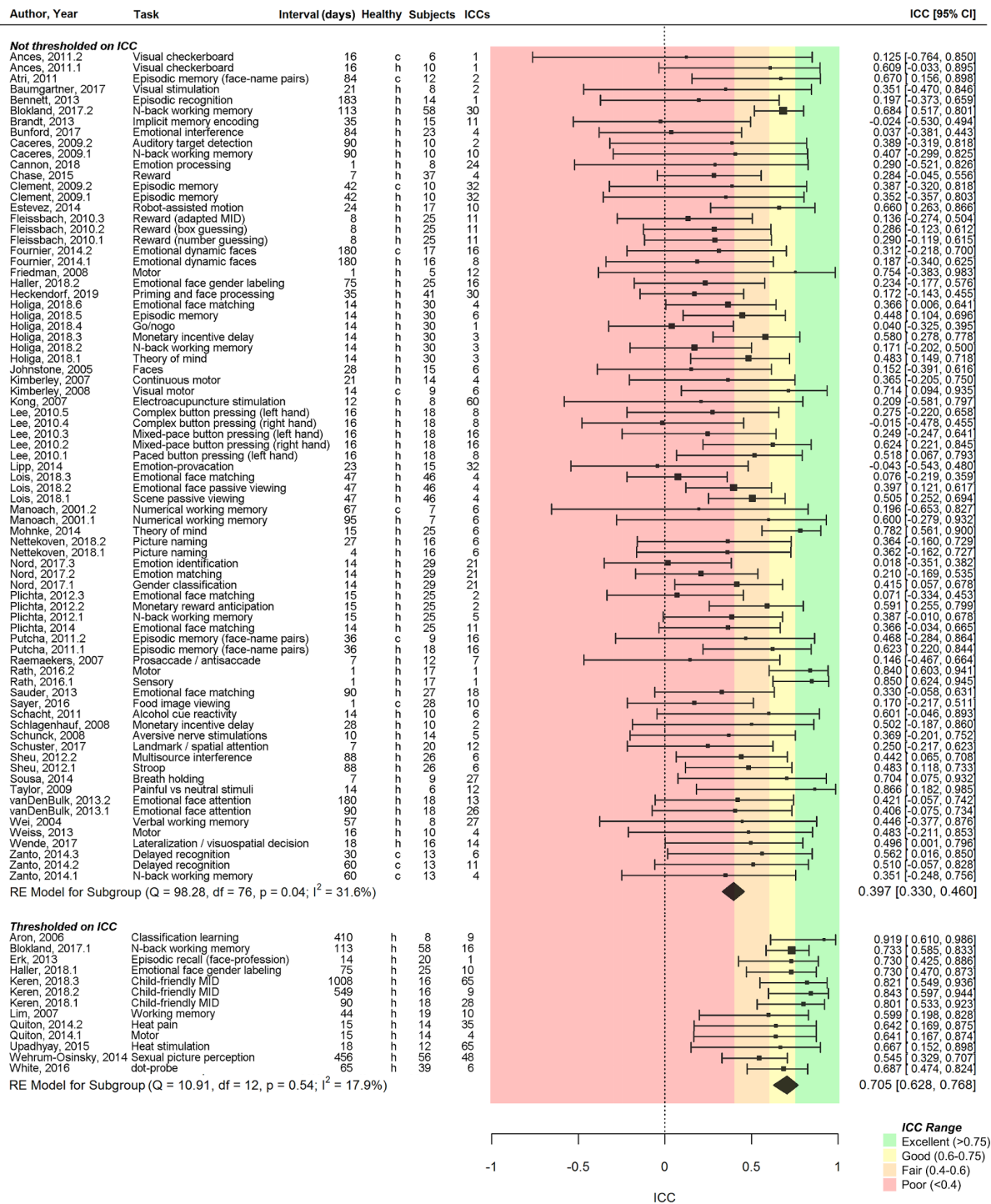
238

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

240

241

242 **Fig. 3** shows the test-retest reliability coefficients (ICCs) from 77 substudies reporting un-
243 thresholded values (average $N = 19.6$, median $N = 17$). 56% of the values fell into the range of what is
244 considered "poor" reliability (below .4), an additional 24% of the values fell into the range of what is
245 considered "fair" reliability (.4 - .6), and only 20% fell into the range of what is considered "good" (.6 - .75)
246 or "excellent" (above .75) reliability. A random effects meta-analysis revealed an average ICC of .397 (95%
247 CI, .330 - .460; $P < .001$), which is in the "poor" range (Cicchetti & Sparrow, 1981). There was evidence
248 of between-study heterogeneity ($I^2 = 31.6$; $P = 0.04$).



249
250
251
252
253
254
255
256
257
258

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. Each substudy is labelled as h if the sample in the study consisted of healthy controls or c if the study consisted of a clinical sample. Studies are split into two sub-groups. The first group of studies reported all ICCs that were calculated, thereby allowing for a relatively unbiased estimate of reliability. The second group of studies selected a subset of calculated ICCs based on the magnitude of the ICC or another non-independent statistic, and then only reported ICCs from that subset. This practice leads to inflated reliability estimates and therefore these studies were meta-analyzed separately to highlight this bias.

259 As expected, the meta-analysis of 13 substudies that only reported ICCs above a minimum
260 threshold (average $N = 24.2$, median $N = 18$) revealed a higher meta-analytic ICC of .705 (95% CI, .628 -
261 .768; $P < .001$; $I^2 = 17.9$). This estimate, which is 1.78 times the size of the estimate from un-thresholded
262 ICCs, is in the good range, suggesting that the practice of thresholding inflates estimates of reliability in
263 task-fMRI. There was no evidence of between-study heterogeneity ($I^2 = 17.9$; $P = 0.54$).

264 A moderator analysis of all substudies revealed significantly higher reliability for studies that
265 thresholded based on ICC ($Q_M = 6.531$, $df = 1$, $P = .010$; $\beta = .140$). In addition, ROIs located in the cortex
266 had significantly higher ICCs than those located in the subcortex ($Q_M = 114.476$, $df = 1$, $P < .001$; $\beta = .259$).
267 However, we did not find evidence that the meta-analytic estimate was moderated by task type, task design,
268 task length, test-retest interval, ROI type, sample type, or number of citations per year. Finally, we tested
269 for publication bias using the Egger random effects regression test (Egger et al., 1997) and found no
270 evidence for bias ($Z = .707$, $P = .480$).

271 The results of the meta-analysis were illuminating, but not without interpretive difficulty. First, the
272 reliability estimates came from a wide array of tasks and samples, so a single meta-analytical reliability
273 estimate could obscure truly reliable task-fMRI paradigms. Second, the studies used different (and some,
274 now outdated) scanners and different pre-processing and analysis pipelines, leaving open the possibility
275 that reliability has improved with more advanced technology and consistent practices. To address these
276 limitations and possibilities, we conducted pre-registered analyses of two new datasets, using state-of-the-
277 art scanners and practices to assess individual differences in commonly used tasks tapping a variety of
278 cognitive and affective functions.

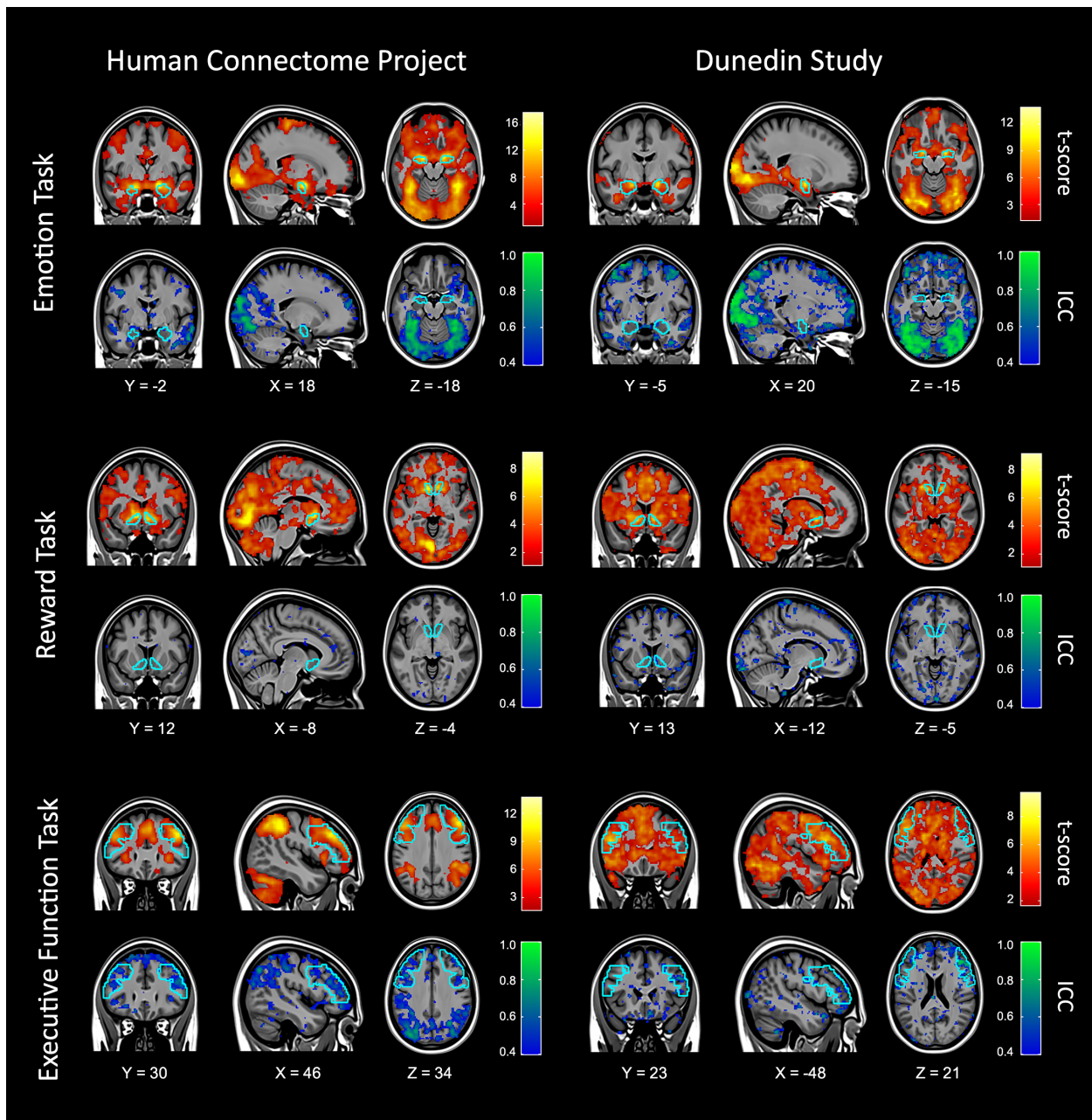
279

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

281 We evaluated test-retest reliabilities of activation in *a priori* regions of interest for 11 commonly
282 used fMRI tasks (see **Methods**). In the Human Connectome Project (HCP), 45 participants were scanned
283 twice using a custom 3T Siemens scanner, on average 140 days apart ($sd = 67.1$ days), using seven tasks

284 targeting emotion, reward, executive function, motor, language, social cognition, and relational processing.
285 This sample size was determined by the publicly available data in the HCP. In the Dunedin Study, 20
286 participants were scanned twice using a 3T Siemens Skyra, on average 79 days apart (sd = 10.3 days), using
287 four tasks targeting emotion, reward, executive control, and episodic memory. This sample size corresponds
288 to the average sample size used in the meta-analyzed studies. Three of the tasks were similar across the two
289 studies, allowing us to test the replicability of task-fMRI reliabilities. For each of the eight unique tasks
290 across the two studies, we identified the task's primary target region, resulting in a total of eight *a priori*
291 ROIs (see **Methods**).

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



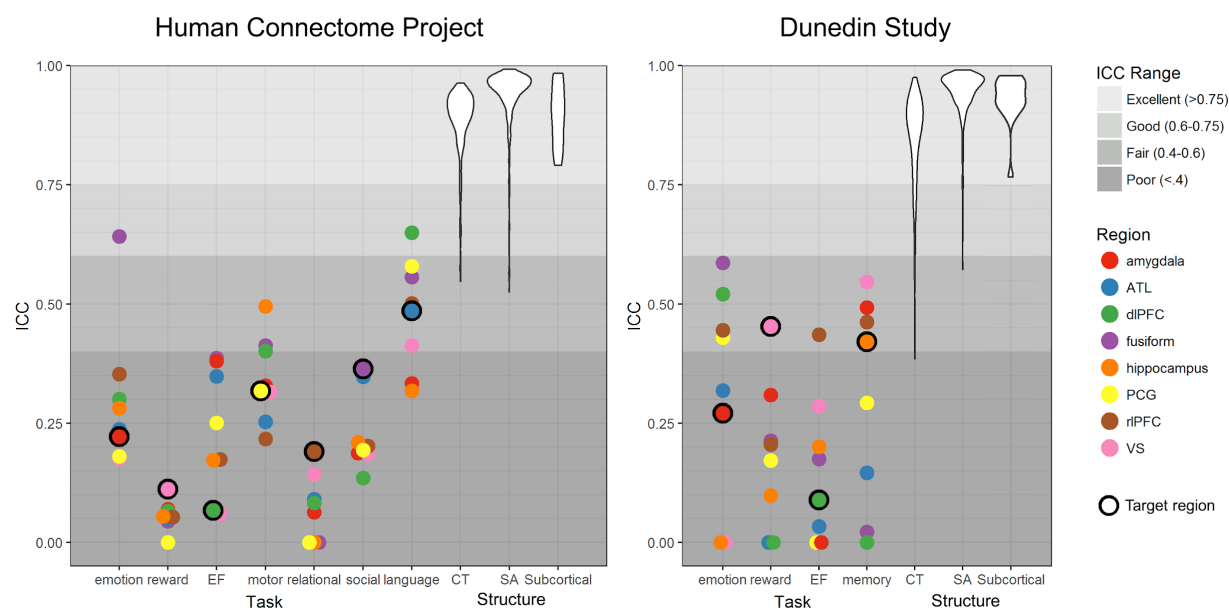
298
 299 **Fig. 4.** Whole-brain activation and reliability maps for three task-fMRI measures used in both the Human
 300 Connectome Project and Dunedin Study. For each task, a whole-brain activation map of the primary within-
 301 subject contrast (t-score) is displayed in warm colors (top) and a whole-brain map of the between-subjects
 302 reliability (ICC) is shown in cool colors (bottom). For each task, the target ROI is outlined in sky-blue. The
 303 activation maps are thresholded at $p < .05$ whole-brain corrected for multiple comparisons using threshold-
 304 free cluster enhancement (Smith & Nichols, 2009). The ICC maps are thresholded so that voxels with ICC
 305 $< .4$ are not colored. These images illustrate that despite robust within-subjects whole-brain activation
 306 produced by each task, there is poor between-subjects reliability in this activation, not only in the target
 307 ROI but across the whole-brain.

308
 309
 310 **Reliability of regional activation.** We investigated the reliability of task activation in both datasets

311 using four steps. First, we tested the reliability of activation in the target ROI for each task. Second, for

312 each task we also evaluated the reliability of activation in the other seven *a priori* ROIs. This was done to
 313 test if the reliability of target ROIs was higher than the reliability of activation in other (“non-target”) brain
 314 regions and to identify any tasks or regions with consistently high reliability. Third, we re-estimated
 315 reliability using activation in the left and right hemispheres separately to test if the estimated reliability was
 316 harmed by averaging across the hemispheres. Fourth, we tested if the reliability depended on whether ROIs
 317 were defined structurally (i.e., using an anatomical atlas) or functionally (i.e., using a set of voxels based
 318 on the location of peak activity). See Supplemental Fig. S5 for ICCs of behavior during each fMRI task.

319 **Reliability of regional activation in the Human Connectome Project.** First, as shown by the
 320 estimates circled in black in **Fig. 5**, across the seven fMRI tasks, activation in anatomically defined target
 321 ROIs had low reliability (mean ICC = .251; 95% CI, .142 - .360). Only the language processing task had
 322 greater than “poor” reliability (ICC = .485). None of the reliabilities entered the “good” range (ICC > .6).



323 **Fig. 5.** Test-retest reliabilities of region-wise activation measures in 11 commonly used task-fMRI
 324 paradigms (EF = executive function). For each task, ICCs were estimated for activation in the *a priori* target
 325 ROI (circled in black) and non-target ROIs selected from the other tasks. These plots show that task-fMRI
 326 measures of regional activation in both the Human Connectome Project and Dunedin Study are generally
 327 unreliable and the ROIs that are “targeted” by the task are rarely more reliable than non-target ROIs (ATL
 328 = anterior temporal lobe, dlPFC = dorsolateral prefrontal cortex, PCG = precentral gyrus, rIPFC =
 329 rostralateral prefrontal cortex, VS = ventral striatum). As a benchmark, ICCs of three common structural
 330 MRI measures (CT = Cortical Thickness, SA = Surface Area, and Subcortical Volume) are depicted as
 331 violin plots representing the distribution of ICCs for each of the 360 parcels for CT and SA, and 17
 332 subcortical structures for grey matter volume. Note that negative ICCs are set to 0 for visualization.

334
335
336 Second, the reliability of task activation in non-target ROIs was also low (**Fig. 5**; mean ICC = .239;
337 95% CI, .188 - .289), but not significantly lower than the reliability in target ROIs ($P = .474$).

338 Third, the reliability of task activation calculated from left and right ROIs separately resembled
339 estimates from averaged ROIs (mean left ICC = .207 in target ROIs and .196 in non-target ROIs, mean
340 right ICC = .259 in target ROIs and .236 in non-target ROIs; Supplemental Fig. S6).

341 Fourth, the reliability of task activation in functionally defined ROIs was also low (mean ICC =
342 .381; 95% CI, .317 - .446), with only the motor and social tasks exhibiting ICCs greater than .4 (ICCs =
343 .550 and .446 respectively; see Supplemental Fig. S6).

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

348 ***Reliability of regional activation in the Dunedin Study.*** First, as shown by the estimates circled in
349 black in **Fig. 5**, activation in the anatomically defined target ROI for each of the four tasks had low
350 reliability (mean ICC = .309; 95% CI, .145 - .472), with no ICCs reaching the "good" range ($ICC > .6$).

351 Second, the reliability of activation in the non-target ROIs was also low (**Fig. 5**; mean ICC = .193;
352 95% CI, .100 - .286), but not significantly lower than the reliability in target ROIs ($P = .140$).

353 Third, the reliability of task activation calculated for the left and right hemispheres separately was
354 similar to averaged ROIs (mean left ICC = .243 in target ROIs and .202 in non-target ROIs, mean right ICC
355 = .358 in target ROIs and .192 in non-target ROIs; Supplemental Fig. S6).

356 Fourth, functionally defined ROIs again did not meaningfully improve reliability (mean ICC =
357 .325; 95% CI, .197 - .453; see Supplemental Fig. S6).

358 ***Reliability of structural measures.*** To provide a benchmark for evaluating the test-retest reliability
359 of task-fMRI, we investigated the reliability of three commonly used structural MRI measures: cortical

360 thickness, surface area and subcortical grey matter volume. Consistent with prior evidence (Han et al., 2006;
361 Maclaren et al., 2014) that structural MRI phenotypes have excellent reliability (i.e., ICCs > .9), global and
362 regional structural MRI measures in the present samples demonstrated very high test-retest reliabilities (**Fig.**
363 **5**). For average cortical thickness, ICCs were .953 and .939 in the HCP and Dunedin Study datasets,
364 respectively. In the HCP, parcel-wise (i.e., regional) cortical thickness reliabilities averaged .886 (range
365 .547 - .964), with 100% crossing the "fair" threshold, 98.6% the "good" threshold, and 94.2% the "excellent"
366 threshold. In the Dunedin Study, parcel-wise cortical thickness reliabilities averaged .846 (range .385 -
367 .975), with 99.7% of ICCs above the "fair" threshold, 96.4% above "good", and 84.7% above "excellent."
368 For total surface area, ICCs were .999 and .996 in the HCP and Dunedin Study datasets, respectively. In
369 the HCP, parcel-wise surface area ICCs averaged .937 (range .526 - .992), with 100% crossing the "fair"
370 threshold, 98.9% crossing the "good" threshold, and 96.9% crossing the "excellent" threshold. In the
371 Dunedin Study, surface area ICCs averaged .942 (range .572 - .991), with 100% above the "fair" threshold,
372 99.7% above "good," and 98.1% above "excellent." For subcortical volumes, ICCs in the HCP averaged
373 .903 (range .791 - .984), with all ICCs above the "excellent" threshold. In the Dunedin Study, subcortical
374 volumes averaged .931 (range .767 - .979), with all ICCs above the "excellent" threshold. See Supplemental
375 Table S3 for reliabilities of each subcortical region evaluated.

376

377 Discussion

378 We found evidence that commonly used task-fMRI measures generally do not have the test-retest
379 reliability necessary for biomarker discovery or brain-behavior mapping. Our meta-analysis of task-fMRI
380 reliability revealed an average test-retest reliability coefficient of .397, which is below the minimum
381 required for good reliability (ICC = .6 (Cicchetti & Sparrow, 1981)) and far below the recommended cutoffs
382 for clinical application (ICC = .8) or individual-level interpretation (ICC = .9) (Guilford, 1946). Of course,
383 not all task-fMRI measures are the same, and it is not possible to assign a single reliability estimate to all

384 individual-difference measures gathered in fMRI research. However, we found little evidence that task type,
385 task length, or test-retest interval had an appreciable impact on the reliability of task-fMRI.

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

397

398 **The two disciplines of fMRI research**

399 Our results harken back to Lee Cronbach’s classic 1957 article in which he described the “two
400 disciplines of scientific psychology” (Cronbach, 1957). According to Cronbach, the “experimental”
401 discipline strives to uncover universal human traits and abilities through experimental control and group
402 averaging, whereas the “correlational” discipline strives to explain variation between people by measuring
403 how they differ from one another. A fundamental distinction between the two disciplines is how they treat
404 individual differences. For the experimental researcher, variation between people is error that must be
405 minimized to detect the largest experimental effect. For the correlational investigator, variation between
406 people is the primary unit of analysis and must be measured carefully to extract reliable individual
407 differences (Cronbach, 1957; Hedge et al., 2018).

408 Current task-fMRI paradigms are largely descended from the “experimental” discipline. Task-
409 fMRI paradigms are intentionally designed to reveal how the average human brain responds to provocation,
410 while minimizing between-subject variance. Paradigms that are able to elicit robust targeted brain activity

411 at the group-level are subsequently converted into tools for assessing individual differences. Within-subject
412 robustness is, then, often inappropriately invoked to suggest between-subject reliability, despite the fact
413 that reliable within-subject experimental effects at a group level can arise from unreliable between-subjects
414 measurements (Fröhner et al., 2019).

415 This reasoning is not unique to task-fMRI research. Behavioral measures that elicit robust within-
416 subject (i.e., group) effects have been shown to have low between-subjects reliability; for example, the
417 mean test-retest reliability of the Stroop Test (ICC = .45; (Hedge et al., 2018)) is strikingly similar to the
418 mean reliability of our task-fMRI meta-analysis (ICC = .397). Nor is it the case that MRI measures, or even
419 the BOLD signal itself, are inherently unreliable. Both structural MRI measures in our analyses (see Fig.
420 5), as well as measures of intrinsic functional connectivity estimated from long fMRI scans (Elliott et al.,
421 2019; Gratton et al., 2018), demonstrate high test-retest reliability. Thus, it is not the tool that is problematic
422 but rather the strategy of adopting tasks developed for experimental cognitive neuroscience that appear to
423 be poorly suited for reliably measuring differences in brain activation between people.

424

425 **Recommendations and Future Directions**

426 We next consider several avenues for maximizing the value of existing datasets as well as
427 improving the reliability of task-fMRI moving forward. We begin with recommendations that can be
428 implemented immediately (1, 2), before moving on to recommendations that will require additional data
429 collection and innovation (3, 4).

430

431 *1) Immediate opportunities for task-fMRI: from brain hotspots to whole-brain signatures*

432 Currently, the majority of task-fMRI measures are based on contrasts between conditions (i.e.,
433 change scores), extracted from ROIs. However, change scores will always have lower reliability than their
434 constituent measures (Hedge et al., 2018), and have been shown to undermine the reliability of task-fMRI
435 (Infantolino et al., 2018). However, contrast-based activation values extracted from ROIs represent only
436 one possible measure of individual differences that can be derived from task-fMRI data. For example,
437 several multivariate methods have been proposed to increase the reliability and predictive utility of task-

438 fMRI measures by exploiting the high dimensionality inherent in fMRI data (Dubois & Adolphs, 2016;
439 Yarkoni & Westfall, 2017). To name a few, the reliability of task-fMRI may be improved by developing
440 measures with latent variable models (Cooper et al., 2019), measuring individual differences in
441 representational spaces with multi-voxel pattern analysis (Norman et al., 2006), and training cross-validated
442 machine learning models that establish reliability through prediction of individual differences in
443 independent samples (Yarkoni & Westfall, 2017). In addition, in many already-collected datasets, task-
444 fMRI can be combined with resting-state fMRI data to produce reliable measures of intrinsic functional
445 connectivity (Elliott et al., 2019; Greene et al., 2018). Thus, there are multiple available approaches to
446 maximizing the value of existing task-fMRI datasets in the context of biomarker discovery and individual
447 differences research.

448

449 *2) Create a norm of reporting the reliability of task-fMRI measures*

450 The “replicability revolution” in psychological science (Nosek et al., 2015) provides a timely
451 example of how rapidly changing norms can shape research practices and standards. In just a few years,
452 practices to enhance replicability, like pre-registration of hypotheses and analytic strategies, have risen in
453 popularity (Nosek et al., 2018). We believe similar norms would be beneficial for task-fMRI in the context
454 of biomarker discovery and brain-behavior mapping. In particular, researchers should report the reliabilities
455 for all task-fMRI measures whenever they are used to study individual differences (Parsons et al., 2019).
456 In doing so, however, researchers need to ensure adequate power to evaluate test-retest reliability with
457 confidence. Given that correlations begin to stabilize with around 150 observations (Schönbrodt & Perugini,
458 2013), our confidence in knowing “the” reliability of any specific task will depend on collecting larger test-
459 retest datasets. We provide evidence that the task-fMRI literature generally has low reliability; however,
460 due to the relatively small size of each test-retest sample reported here, we urge readers to avoid making
461 strong conclusions about the reliability of specific fMRI tasks. In the pursuit of precise reliability estimates,
462 it will be important for researchers to collect larger test-retest samples, explore test-retest moderators (e.g.
463 test-retest interval) and avoid reporting inflated reliabilities that can arise from circular statistical analyses
464 (for detailed recommendations see (Kriegeskorte et al., 2010, 2009; Vul et al., 2009)).

465 Researchers can also provide evidence of between-subjects reliability in the form of internal
466 consistency. While test-retest reliability provides an estimate of stability over time that is suited for trait
467 and biomarker research, it is a conservative estimate that requires extra data collection and can be
468 undermined by habituation effects and rapid fluctuations (Hajcak et al., 2017). In some cases, internal
469 consistency will be more practical because it is cheaper, as it does not require additional data collection and
470 can be used in any situation where the task-fMRI measure of interest is comprised of multiple trials
471 (Streiner, 2003). Internal consistency is particularly well-suited for measures that are expected to change
472 rapidly and index transient psychological states (e.g., current emotions or thoughts). However, internal
473 consistency alone is not adequate for prognostic biomarkers. Establishing a norm of explicitly reporting
474 measurement reliability would increase the replicability of task-fMRI findings and accelerate biomarker
475 discovery.

476

477 *3) More data from more subjects*

478 Our ability to detect reliable individual differences using task-fMRI will depend, in part, on the
479 field embracing two complementary improvements to the status quo: 1) more subjects per study and 2)
480 more data per subject. It has been suggested that neuroscience is generally an underpowered enterprise, and
481 that small sample sizes undermine fMRI research in particular (Button et al., 2013; Szucs & Ioannidis,
482 2017). The results presented here suggest that this “power failure” may be further compounded by low
483 reliability in task-fMRI. The median sample size in fMRI research is 28.5 (Poldrack et al., 2017). However,
484 as shown in Fig. 1, task-fMRI measures with ICCs of .397 (the meta-analytic mean reliability) would
485 require $N > 214$ to achieve 80% power to detect brain-behavior correlations of .3, a moderate effect size
486 equal to the size of the largest replicated brain-behavior associations (Elliott et al., 2018; Nave et al., 2019).
487 For $r = .1$ (a small effect size common in psychological research (Funder & Ozer, 2019)), adequately
488 powered studies require $N > 2,000$. And, these calculations are actually best-case scenarios given that they
489 assume perfect reliability of the second “behavioral” variable (see Figure 1). Increasing the sample size of
490 task-fMRI studies and requiring power analyses that take into account unreliability represent a meaningful
491 way forward for boosting the replicability of individual differences research with task-fMRI.

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

507

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

509 Instead of continuing to adopt fMRI tasks from experimental studies emphasizing within-subjects
510 effects, we need to develop new tasks (and naturalistic stimuli) from the ground up with the goal of
511 optimizing their utility in individual differences research (i.e., between-subjects effects). Psychometrics
512 provides many tools and methods for developing reliable individual differences measures that have been
513 underutilized in task-fMRI development. For example, stimuli in task-fMRI could be selected based on
514 their ability to maximally distinguish groups of subjects or to elicit reliable between subject variance. As
515 noted in recommendation 1, psychometric tools for test construction could be adopted to optimize reliable
516 task-fMRI measures including item analysis, latent variable modelling, and internal-consistency measures
517 (Crocker & Algina, 2006).

518

519 Conclusion

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

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

529
530 Test-retest reliability is widely quantified using the intraclass correlation coefficient (ICC (Shrout
531 & Fleiss, 1979)). ICC can be thought of as the proportion of a measure's total variance that is accounted
532 for by variation between individuals. An ICC can take on values between -1 and 1, with values approaching
533 1 indicating nearly perfect stability of individual differences across test-retest measurements, and values at
534 or below 0 indicating no stability. Classical test theory states that all measures are made up of a true score
535 plus measurement error (Novick, 1965). The ICC is used to estimate the amount of reliable, true-score
536 variance present in an individual differences measure. When a measure is taken at two timepoints, the
537 variance in scores that is due to measurement error will consist of random noise and will fail to correlate
538 with itself across test-retest measurements. However, the variance in a score that is due to true score will
539 be stable and correlate with itself across timepoints (Crocker & Algina, 2006). Measures with ICC < .40
540 are thought to have "poor" reliability, those with ICCs between .40 - .60 "fair" reliability, .60 - .75 "good"
541 reliability, and > .75 "excellent" reliability. An ICC > .80 is considered a clinically required standard for
542 reliability in psychology (Cicchetti & Sparrow, 1981).

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

$$545 \quad r(A_{observed}, B_{observed}) = r(A_{true}, B_{true}) * \sqrt{reliability(A_{observed}) * reliability(B_{observed})}$$

546 Low reliability of a measure reduces statistical power and increases the sample size required to detect a
547 correlation with another measure. **Fig. 1** shows sample sizes required for 80% power to detect correlations
548 between a task-fMRI measure of individual differences in brain activation and a behavioral/clinical
549 phenotype, across a range of reliabilities of the task-fMRI measure and expected effect sizes. Power curves
550 are given for three levels of reliability of the hypothetical behavioral/clinical phenotype, where the first two
551 panels (behavioral ICC = .6 and .8) represent most typical scenarios.

552

553 Author Contributions

554 A.C., A.R.H., T.E.M, M.L.E., and A.R.K. conceived the study and data analysis plan. M.L.E.,
555 A.R.K., and M.L.S. prepared MRI data for analysis. M.L.M prepared data for meta-analysis. A.R.K.,
556 M.L.E., and M.L.S. conducted the analyses. M.L.E., A.R.K., A.C., A.R.H., and T.E.M. wrote the
557 manuscript. A.C., A.R.H., T.E.M., and R.P. designed, implemented, and/or oversaw the collection and
558 generation of the research protocol. S.R., D.I., and A.R.K. oversaw data collection. All authors discussed
559 the results and contributed to the revision of the manuscript.

560

561 Acknowledgments

562
563 Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium
564 (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH
565 Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell
566 Center for Systems Neuroscience at Washington University.

567 The Dunedin Study was approved by the NZ-HDEC (Health and Disability Ethics Committee).
568 The Dunedin Study is supported by NIA grants R01AG049789 and R01AG032282 and U.K. Medical
569 Research Council grant P005918. The Dunedin Multidisciplinary Health and Development Research Unit
570 is supported by the New Zealand Health Research Council and the New Zealand Ministry of Business,
571 Innovation and Employment (MBIE). MLE is supported by the National Science Foundation Graduate
572 Research Fellowship under Grant No. NSF DGE-1644868. Thanks to the members of the Advisory Board
573 for the Dunedin Neuroimaging Study. The authors would also like to thank Tim Strauman and Ryan
574 Bogdan for their feedback on an initial draft of this manuscript, as well as extensive feedback from peer
575 reviewers.

576 References

- 577 Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M.
578 F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M.,
579 Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z., ... WU-Minn HCP Consortium. (2013).
580 Function in the human connectome: task-fMRI and individual differences in behavior. *NeuroImage*,
581 *80*, 169–189.
- 582 Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance
583 imaging? *Annals of the New York Academy of Sciences*, *1191*, 133–155.
- 584 Binder, J. R., Gross, W. L., Allendorfer, J. B., Bonilha, L., Chapin, J., Edwards, J. C., Grabowski, T. J.,
585 Langfitt, J. T., Loring, D. W., Lowe, M. J., Koenig, K., Morgan, P. S., Ojemann, J. G., Rorden, C.,
586 Szaflarski, J. P., Tivarus, M. E., & Weaver, K. E. (2011). Mapping anterior temporal lobe language
587 areas with fMRI: a multicenter normative study. *NeuroImage*, *54*(2), 1465–1475.
- 588 Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-*
589 *Analysis*. <https://doi.org/10.1002/9780470743386>
- 590 Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M.
591 R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature*
592 *Reviews Neuroscience*, *14*(5), 365–376.
- 593 Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., & Cox, R. W. (2013). Linear mixed-effects modeling
594 approach to fMRI group analysis. *NeuroImage*, *73*, 176–190.
- 595 Chen, G., Taylor, P. A., Haller, S. P., Kircanski, K., Stoddard, J., Pine, D. S., Leibenluft, E., Brotman, M.
596 A., & Cox, R. W. (2018). Intra-class correlation: Improved modeling approaches and applications for
597 neuroimaging. *Human Brain Mapping*, *39*(3), 1187–1206.
- 598 Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of
599 specific items: applications to assessment of adaptive behavior. *American Journal of Mental*
600 *Deficiency*, *86*(2), 127–137.
- 601 Cooper, S. R., Jackson, J. J., Barch, D. M., & Braver, T. S. (2019). Neuroimaging of individual
602 differences: A latent variable modeling perspective. In *Neuroscience & Biobehavioral Reviews* (Vol.

- 603 98, pp. 29–46). <https://doi.org/10.1016/j.neubiorev.2018.12.022>
- 604 Crocker, L., & Algina, J. (2006). *Introduction to Classical and Modern Test Theory*. Wadsworth
605 Publishing Company.
- 606 Cronbach, L. J. (1957). The two disciplines of scientific psychology. In *American Psychologist* (Vol. 12,
607 Issue 11, pp. 671–684). <https://doi.org/10.1037/h0043943>
- 608 Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., & Fiez, J. A. (2000). Tracking the hemodynamic
609 responses to reward and punishment in the striatum. *Journal of Neurophysiology*, *84*(6), 3072–3077.
- 610 Drobyshevsky, A., Baumann, S. B., & Schneider, W. (2006). A rapid fMRI task battery for mapping of
611 visual, motor, cognitive, and emotional function. *NeuroImage*, *31*(2), 732–744.
- 612 Dubois, J., & Adolphs, R. (2016). Building a Science of Individual Differences from fMRI. *Trends in*
613 *Cognitive Sciences*, *20*(6), 425–443.
- 614 Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a
615 simple, graphical test. *BMJ*, *315*(7109), 629–634.
- 616 Elliott, M. L., Belsky, D. W., Anderson, K., Corcoran, D. L., Ge, T., Knodt, A., Prinz, J. A., Sugden, K.,
617 Williams, B., Ireland, D., Poulton, R., Caspi, A., Holmes, A., Moffitt, T., & Hariri, A. R. (2018). A
618 Polygenic Score for Higher Educational Attainment is Associated with Larger Brains. *Cerebral*
619 *Cortex*. <https://doi.org/10.1093/cercor/bhy219>
- 620 Elliott, M. L., Knodt, A. R., Cooke, M., Kim, M. J., Melzer, T. R., Keenan, R., Ireland, D., Ramrakha, S.,
621 Poulton, R., Caspi, A., Moffitt, T. E., & Hariri, A. R. (2019). General functional connectivity:
622 Shared features of resting-state and task fMRI drive reliable and heritable individual differences in
623 functional brain networks. *NeuroImage*, *189*, 516–532.
- 624 Fröhner, J. H., Teckentrup, V., Smolka, M. N., & Kroemer, N. B. (2019). Addressing the reliability
625 fallacy in fMRI: Similar group effects may arise from unreliable individual effects. *NeuroImage*,
626 *195*, 174–189.
- 627 Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and
628 Nonsense. In *Advances in Methods and Practices in Psychological Science* (Vol. 2, Issue 2, pp. 156–
629 168). <https://doi.org/10.1177/2515245919847202>

- 630 Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K.,
631 Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-
632 modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.
- 633 Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., Ortega, M.,
634 Hoyt-Drazen, C., Gratton, C., Sun, H., Hampton, J. M., Coalson, R. S., Nguyen, A. L., McDermott,
635 K. B., Shimony, J. S., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E., Nelson, S. M., & Dosenbach,
636 N. U. F. (2017). Precision Functional Mapping of Individual Human Brains. *Neuron*, *95*(4), 791–
637 807.e7.
- 638 Gratton, C., Laumann, T. O., Nielsen, A. N., Greene, D. J., Gordon, E. M., Gilmore, A. W., Nelson, S.
639 M., Coalson, R. S., Snyder, A. Z., Schlaggar, B. L., Dosenbach, N. U. F., & Petersen, S. E. (2018).
640 Functional Brain Networks Are Dominated by Stable Group and Individual Factors, Not Cognitive
641 or Daily Variation. *Neuron*, *98*(2), 439–452.e5.
- 642 Greene, A. S., Gao, S., Scheinost, D., & Constable, R. T. (2018). Task-induced brain state manipulation
643 improves prediction of individual traits. *Nature Communications*, *9*(1), 2807.
- 644 Guilford, J. P. (1946). New Standards For Test Evaluation. In *Educational and Psychological*
645 *Measurement* (Vol. 6, Issue 4, pp. 427–438). <https://doi.org/10.1177/001316444600600401>
- 646 Hajcak, G., Meyer, A., & Kotov, R. (2017). Psychometrics and the neuroscience of individual
647 differences: Internal consistency limits between-subjects effects. *Journal of Abnormal Psychology*,
648 *126*(6), 823–834.
- 649 Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert,
650 M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., & Fischl, B. (2006).
651 Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field
652 strength, scanner upgrade and manufacturer. *NeuroImage*, *32*(1), 180–194.
- 653 Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., & Weinberger, D. R. (2002). The amygdala response
654 to emotional stimuli: a comparison of faces and scenes. *NeuroImage*, *17*(1), 317–323.
- 655 Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not
656 produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186.

- 657 Herting, M. M., Gautam, P., Chen, Z., Mezher, A., & Vetter, N. C. (2018). Test-retest reliability of
658 longitudinal task-based fMRI: Implications for developmental studies. *Developmental Cognitive*
659 *Neuroscience*, 33, 17–26.
- 660 Infantolino, Z. P., Luking, K. R., Sauder, C. L., Curtin, J. J., & Hajcak, G. (2018). Robust is not
661 necessarily reliable: From within-subjects fMRI contrasts to between-subjects comparisons.
662 *NeuroImage*, 173, 146–152.
- 663 Knutson, B., Westdorp, A., Kaiser, E., & Hommer, D. (2000). fMRI visualization of brain activity during
664 a monetary incentive delay task. *NeuroImage*, 12(1), 20–27.
- 665 Kriegeskorte, N., Lindquist, M. A., Nichols, T. E., Poldrack, R. A., & Vul, E. (2010). Everything you
666 never wanted to know about circular analysis, but were afraid to ask. *Journal of Cerebral Blood*
667 *Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and*
668 *Metabolism*, 30(9), 1551–1557.
- 669 Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in
670 systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5), 535–540.
- 671 Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P.,
672 Kennedy, D. N., Hoppel, B. E., Cohen, M. S., & Turner, R. (1992). Dynamic magnetic resonance
673 imaging of human brain activity during primary sensory stimulation. *Proceedings of the National*
674 *Academy of Sciences of the United States of America*, 89(12), 5675–5679.
- 675 Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological
676 investigation of the basis of the fMRI signal. *Nature*, 412(6843), 150–157.
- 677 Maclaren, J., Han, Z., Vos, S. B., Fischbein, N., & Bammer, R. (2014). Reliability of brain volume
678 measurements: a test-retest dataset. *Scientific Data*, 1, 140037.
- 679 Manuck, S. B., Brown, S. M., Forbes, E. E., & Hariri, A. R. (2007). Temporal stability of individual
680 differences in amygdala reactivity. *The American Journal of Psychiatry*, 164(10), 1613–1614.
- 681 McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients.
682 In *Psychological Methods* (Vol. 1, Issue 1, pp. 30–46). <https://doi.org/10.1037//1082-989x.1.1.30>
- 683 Metafor Package R Code for Meta-Analysis Examples. (2019). In *Advanced Research Methods for the*

- 684 *Social and Behavioral Sciences* (pp. 365–367). <https://doi.org/10.1017/9781108349383.027>
- 685 Nave, G., Jung, W. H., Karlsson Linnér, R., Kable, J. W., & Koellinger, P. D. (2019). Are Bigger Brains
686 Smarter? Evidence From a Large-Scale Preregistered Study. *Psychological Science*, *30*(1), 43–54.
- 687 Nord, C. L., Gray, A., Charpentier, C. J., Robinson, O. J., & Roiser, J. P. (2017). Unreliability of putative
688 fMRI biomarkers during emotional face processing. *NeuroImage*, *156*, 119–127.
- 689 Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel
690 pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.
- 691 Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers,
692 C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R.,
693 Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). SCIENTIFIC
694 STANDARDS. Promoting an open research culture. *Science*, *348*(6242), 1422–1425.
- 695 Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution.
696 *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–
697 2606.
- 698 Novick, M. R. (1965). THE AXIOMS AND PRINCIPAL RESULTS OF CLASSICAL TEST THEORY.
699 In *ETS Research Bulletin Series* (Vol. 1965, Issue 1, pp. i – 31). [https://doi.org/10.1002/j.2333-
700 8504.1965.tb00132.x](https://doi.org/10.1002/j.2333-8504.1965.tb00132.x)
- 701 Nunnally, J. C. (1959). *Introduction to Psychological Measurement*.
- 702 Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of
703 Reporting the Reliability of Cognitive-Behavioral Measurements. In *Advances in Methods and
704 Practices in Psychological Science* (Vol. 2, Issue 4, pp. 378–395).
705 <https://doi.org/10.1177/2515245919879695>
- 706 Peterson, B. S., Skudlarski, P., Gatenby, J. C., Zhang, H., Anderson, A. W., & Gore, J. C. (1999). An
707 fMRI study of Stroop word-color interference: evidence for cingulate subregions subserving multiple
708 distributed attentional systems. *Biological Psychiatry*, *45*(10), 1237–1258.
- 709 Plichta, M. M., Schwarz, A. J., Grimm, O., Morgen, K., Mier, D., Haddad, L., Gerdes, A. B. M., Sauer,
710 C., Tost, H., Esslinger, C., Colman, P., Wilson, F., Kirsch, P., & Meyer-Lindenberg, A. (2012). Test-

- 711 retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. *NeuroImage*,
712 60(3), 1746–1758.
- 713 Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T.
714 E., Poline, J.-B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: towards transparent and
715 reproducible neuroimaging research. *Nature Reviews. Neuroscience*, 18(2), 115–126.
- 716 Poulton, R., Moffitt, T. E., & Silva, P. A. (2015). The Dunedin Multidisciplinary Health and
717 Development Study: overview of the first 40 years, with an eye to the future. *Social Psychiatry and*
718 *Psychiatric Epidemiology*, 50(5), 679–693.
- 719 Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research:
720 Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*,
721 10, 813.
- 722 Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? In *Journal of*
723 *Research in Personality* (Vol. 47, Issue 5, pp. 609–612). <https://doi.org/10.1016/j.jrp.2013.05.009>
- 724 Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability.
725 *Psychological Bulletin*, 86(2), 420–428.
- 726 Smith, R., Keramatian, K., & Christoff, K. (2007). Localizing the rostrolateral prefrontal cortex at the
727 individual level. *NeuroImage*, 36(4), 1387–1396.
- 728 Smith, S., & Nichols, T. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing,
729 threshold dependence and localisation in cluster inference. In *NeuroImage* (Vol. 44, Issue 1, pp. 83–
730 98). <https://doi.org/10.1016/j.neuroimage.2008.03.061>
- 731 Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal
732 consistency. *Journal of Personality Assessment*, 80(1), 99–103.
- 733 Swartz, J. R., Knodt, A. R., Radtke, S. R., & Hariri, A. R. (2015). A neural biomarker of psychological
734 vulnerability to future life stress. *Neuron*, 85(3), 505–511.
- 735 Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the
736 recent cognitive neuroscience and psychology literature. *PLoS Biology*, 15(3), e2000797.
- 737 Vanderwal, T., Eilbott, J., & Castellanos, F. X. (2018). Movies in the magnet: Naturalistic paradigms in

- 738 developmental functional neuroimaging. *Developmental Cognitive Neuroscience*, 100600.
- 739 Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., & WU-Minn
740 HCP Consortium. (2013). The WU-Minn Human Connectome Project: an overview. *NeuroImage*,
741 80, 62–79.
- 742 Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies
743 of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science: A Journal of*
744 *the Association for Psychological Science*, 4(3), 274–290.
- 745 Wheatley, T., Milleville, S. C., & Martin, A. (2007). Understanding animate agents: distinct roles for the
746 social network and mirror system. *Psychological Science*, 18(6), 469–474.
- 747 Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: brain
748 models in translational neuroimaging. *Nature Neuroscience*, 20(3), 365–377.
- 749 Yarkoni, T. (2009). Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical
750 Power-Commentary on Vul et al. (2009). *Perspectives on Psychological Science: A Journal of the*
751 *Association for Psychological Science*, 4(3), 294–298.
- 752 Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From
753 Machine Learning. *Perspectives on Psychological Science: A Journal of the Association for*
754 *Psychological Science*, 12(6), 1100–1122.
- 755 Zeineh, M. M., Engel, S. A., Thompson, P. M., & Bookheimer, S. Y. (2003). Dynamics of the
756 hippocampus during encoding and retrieval of face-name pairs. *Science*, 299(5606), 577–580.
- 757