

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

Impact of analytic decisions on test-retest reliability of individual and group estimates in functional magnetic resonance imaging: a multiverse analysis using the monetary incentive delay task

Michael I. Demidenko¹, Jeanette A. Mumford¹, Russell A. Poldrack¹

1. Department of Psychology, Stanford University, Stanford, United States

Correspondence concerning this article should be addressed to Michael Demidenko, Department of Psychology, Stanford University, 450 Serra Mall, Building 420, Stanford, CA 94305. E-mail: demidenm@stanford.edu

This is a preprint of the **Stage 2 Registered Report** that has been submitted for review to **Peer Community In: Registered Reports** on March 19th, 2024. An error was identified in the ABCD pipeline in April 2024. A correction was made May 2nd, 2024 (no interpretations changed) and resubmitted for review May 5th, 2024. Initial reviews were obtained June 17th, 2024. The Stage 2 revision was submitted on June 29th, 2024.

Abstract

Empirical studies reporting low test-retest reliability of individual blood oxygen-level dependent (BOLD) signal estimates in functional magnetic resonance imaging (fMRI) data have resurrected interest among cognitive neuroscientists in methods that may improve reliability in fMRI. Over the last decade, several individual studies have reported that modeling decisions, such as smoothing, motion correction and contrast selection, may improve estimates of test-retest reliability of BOLD signal estimates. However, it remains an empirical question whether certain analytic decisions *consistently* improve individual and group level reliability estimates in an fMRI task across multiple large, independent samples. This study used three independent samples (N s: 60, 81, 119) that collected the same task (Monetary Incentive Delay task) across two runs and two sessions to evaluate the effects of analytic decisions on the individual (intraclass correlation coefficient [ICC(3,1)]) and group (Jaccard/Spearman ρ) reliability estimates of BOLD activity of task fMRI data. The analytic decisions in this study vary across four categories: smoothing kernel (five options), motion correction (four options), task parameterizing (three options) and task contrasts (four options), totaling 240 different pipeline permutations. Across all 240 pipelines, the median ICC estimates are consistently low, with a maximum median ICC estimate of .43 - .55 across the three samples. The analytic decisions with the greatest impact on the median ICC and group similarity estimates are the *Implicit Baseline* contrast, Cue Model parameterization and a larger smoothing kernel. Using an *Implicit Baseline* in a contrast condition meaningfully increased group similarity and ICC estimates as compared to using the *Neutral* cue. This effect was largest for the Cue Model parameterization; however, improvements in reliability came at the cost of interpretability. This study illustrates that estimates of reliability in the MID task are consistently low and variable at small samples, and a higher test-retest reliability may not always improve interpretability of the estimated BOLD signal.

Keywords: Test-rest reliability, Intraclass Correlation, Jaccard Similarity, Functional Magnetic Resonance Imaging, Monetary Incentive Delay task, Individual Differences

Introduction

61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85

Reliability in functional magnetic resonance imaging (fMRI) is essential to individual differences research as well as for the development of clinical biomarkers. Unfortunately, numerous studies have demonstrated that reliability of individual estimates in fMRI is low (Elliott et al., 2020; Noble et al., 2019) and the reliability of group estimates in statistical maps is sensitive to varying analytical decisions made by researchers (Botvinik-Nezer et al., 2020)¹. Poor reliability can hamper validity in cognitive neuroscience research, reducing the ability to uncover brain-behavior effects (Hedge et al., 2018; Nikolaidis et al., 2022) and the ability to detect differences in distinct brain states and individual traits (Gell et al., 2023; Kragel et al., 2021). It remains to be seen whether certain analytic decisions *consistently* reduce individual and/or group reliability estimates of blood oxygen-level dependent (BOLD) activity across measurement occasions in univariate task fMRI analyses.

fMRI analysis involves a range of analytic decisions (Caballero-Gaudes & Reynolds, 2017; Soares et al., 2016) that can result in a vast number of statistical brain maps across which BOLD activity can vary subtly or substantially (Bowring et al., 2022; Carp, 2012; Li et al., 2021). Simple decisions, such as using different MNI template brains, can greatly affect the agreement between parameter estimates between two preprocessing pipelines (Li et al., 2021). Furthermore, the approach used to model a task design can also alter interpretations (Botvinik-Nezer et al., 2020). As a result of numerous arbitrary choices, preprocessing and task modeling decisions can significantly impact the reliability of voxel/region of interest (ROI) estimates (Dubois & Adolphs, 2016).

Different metrics of reliability provide quantitative indices of the consistency (or similarity) of estimates of BOLD activity in specific brain regions (or voxels) during fMRI task activation across repeated measurement occasions (Bennett & Miller, 2013). Researchers can quantify the consistency of two repeated measures in terms of estimated effects (continuous)

¹ Reliability of parameter estimates at the individual level and thresholded activation maps at the group level have previously been distinguished as “reliability” and “reproducibility” of BOLD activity, respectively (Bennett & Miller, 2013; Plichta et al., 2012; Zuo et al., 2014). We elect to refer to individual and group estimates as distinct forms of reliability and use ‘reproducibility’ to refer to a broader set of concepts describing various aspects of the ability to reproduce or generalize a research finding (e.g. Goodman et al. [2016]).

86 and/or the presence/absence of a significant effect (binary). In terms of the continuous effects,
87 reliability is an estimate of the consistency of the numerical representation of a measure (e.g.,
88 BOLD activity in the supplementary motor area during a finger tapping task [Witt et al., 2008])
89 of a mental process (e.g., index finger movement) across repeated measurement occasions within
90 an *individual* (e.g., task fMRI contrasts across two or more sessions, which can be hours, days or
91 weeks). This form of reliability is usually calculated using an intraclass correlation (ICC) at the
92 whole brain (i.e., voxel-wise) and/or ROI level. In terms of binary estimates of an effect,
93 reliability is an estimate of an experimental task's (e.g., finger tapping task [Witt et al., 2008])
94 ability to evoke statistically significant activation (above a pre-specified threshold) in the same
95 regions for *groups* of subjects for a specific condition (e.g., finger movement versus rest) across
96 measurement occasions (e.g., task fMRI contrasts across two or more scanning sessions). Binary
97 estimates of reliability are often calculated using Dice (Rombouts et al., 1998) or Jaccard's
98 similarity coefficients (Maitra, 2010). Together, these two forms of reliability reflect the
99 consistency (or agreement) in either the magnitude or the binary statistical significance of an
100 experimental effect occurring during task fMRI.

101 Traditionally, empirical studies have referred to the “robustness” of above-threshold
102 activation signals in group fMRI analyses as an implicit indicator of reliability of an fMRI task.
103 While a useful heuristic, Fröhner et al. (2019) argued that robustness across measurement
104 occasions only represents reliability of *group* (overall average) BOLD activity and does not
105 accurately represent *individual* variability in BOLD activity. In addition, thresholding is a
106 nonlinear operation that can result in substantial variability (Cohen & DuBois, 1999). When
107 quantifying reliability of BOLD activity in the brain, researchers often report an ICC or a
108 similarity coefficient for task fMRI (Bennett & Miller, 2013; Fröhner et al., 2019). The lack of
109 standardization makes it challenging to precisely quantify reliability, relative to individual
110 differences, and assess the impact of different fMRI analysis decisions on continuous and binary
111 estimates of reliability.

112 To date, several studies have examined the impact of analytic decisions, such as spatial
113 smoothing, motion correction and contrast modeling, on individual estimates of reliability of task
114 fMRI. Caceres et al. (2009, $n = 10$) found that an optimal smoothing kernel size of 8-10 FWHM
115 (full-width half-maximum) on a 1.5T scanner with 3.75mm voxels improved reliability. Results
116 regarding the impact of motion correction on reliability are mixed, with Gorgolewski et al.

117 (2013, $n = 11$) reporting a positive effect on reliability while Plichta et al. (2012, $n = 25$)
118 reporting no effect during a reward task and a negative effect during a faces and N-back task on
119 reliability. However, in a large, young sample, Kennedy et al. (2022, $n = 5,979 - 6,593$) reported
120 that excluding high motion subjects modestly improved reliability. Finally, Han et al. (2022, $n =$
121 $29 - 120$) and Kennedy et al. (2022, $n = 5,979 - 6,593$) reported that using an implicit baseline
122 for different tasks (e.g., rest phase during the task) rather than a neutral cue increased reliability
123 across measurement occasions. Some, but not all, of these findings are consistent with a previous
124 review of the fMRI reliability literature (Bennett & Miller, 2013), which suggests that motion,
125 spatial smoothing and task signal likely impacts reliability in task fMRI. However, differences in
126 modeling decisions across these studies leaves an important question unanswered: Are there
127 certain analytic decisions that *consistently* improve reliability (e.g., ICC) of neural activity for an
128 fMRI task across samples?

129 The ICC is a statistic adopted from behavioral research to estimate reliability of observed
130 scores across measurement occasions (Bartko, 1966; Fisher, 1934; Shrout & Fleiss, 1979;
131 Spearman, 1904). In the context of multi-session data, there are several ways to estimate an ICC,
132 but for typical univariate fMRI studies, two specific types (ICC[2,1] and ICC[3,1]) are
133 recommended (For a discussion, see Noble et al., 2021). As described elsewhere (Bennett &
134 Miller, 2013; Fisher, 1934), the ICC is similar to the product moment correlation. Unlike the
135 product moment correlation, which estimates separate means and variances between distinct
136 classes (e.g., age and height), the ICC estimates the mean and variances within a single class
137 (e.g., measure). For two or more variables from a single class, test-retest reliability estimates the
138 consistency (or agreement) of the observed scores across the measurement occasions. Using the
139 correlation coefficient as an example, if there are no differences in subjects' scores across two
140 measurement occasions, the correlation coefficient would be 1.0. However, if the measure is
141 affected by systematic and/or unsystematic error across measurement occasions, this would
142 impact the covariance between observed scores across subjects and decrease the linear
143 association between measures across the two occasions. Unlike the product moment correlation,
144 however, the ICC factors out measurement bias which reflects the reproducibility of observed
145 scores across measurement occasions (Liu et al., 2016). While the correlation between two
146 occasions ($\mathbf{A} = [1, 3, 6, 9, 12]$ & $\mathbf{B} = 3\mathbf{x}\mathbf{A} = [3, 9, 18, 27, 36]$) may be perfect ($r_{AB} = 1.0$), the
147 consistency in observed scores between the two measurement occasions would be lower

148 (ICC[3,1] = .60). In fMRI, the reliability of the BOLD signal may be impacted by biological
149 (e.g., differences in BOLD across brain region), analytic (e.g., task design and analytic
150 decisions), and participant-level factors (e.g., practice effects, motion, habituation and/or
151 development). These fluctuations, whether typical or atypical, may contribute to observed
152 differences and the reduced consistency in scores across measurement occasions, leading to
153 decreased estimates of reliability.

154 As discussed in prior work on fMRI reliability (Bennett & Miller, 2010, 2013; Caceres et
155 al., 2009; Chen et al., 2017; Herting et al., 2017; Noble et al., 2021), the ICC decomposes the
156 total variance of the data across all subjects and sessions into two key parts: *Between-subject* and
157 *Within-subject* variance (for statistical formulas and discussion of ICC, see Liljequist et al.,
158 [2019] and flowchart in McGraw & Wong [1996, p. 40]). The ICC estimate can be altered by
159 increasing the differences in BOLD activity between subjects (e.g., subjects differ more in
160 BOLD activity in index finger movements) and/or ensure that BOLD activity within subjects is
161 more similar across scans (e.g., BOLD activity in response to finger movements versus rest for
162 Subject A is consistent across Session 1 and Session 2). Some have argued that the low *between-*
163 *subject* variability may be a reason for low reliability of behavioral responses in experimental
164 tasks that are commonly used in fMRI (Hedge et al., 2018). However, there is little empirical
165 research on whether the culprit in the reportedly low reliability of fMRI signal across
166 measurement occasions is a *decreased between-subject* and/or an *increased within-subject*
167 variability. It also remains an open question whether certain analytic decisions differentially
168 impact the between/within subject variance and consistently improve reliability across different
169 samples with the same task. As it relates to prediction and global signal-to-noise ratio, evidence
170 from Churchill et al. (2015; $n = 25$) suggest that there are likely to be optimal preprocessing
171 pipelines; however, the degree to which these differ across datasets and individuals is currently
172 unknown.

173 The current study uses a multiverse (Steege et al., 2016) of analytic alternatives to
174 simultaneously evaluate the effects of analytic decisions on the continuous and binary reliability
175 estimates of neural activity in task fMRI in three samples. The three samples administered with
176 the comparable Monetary Incentive Delay (MID) task during fMRI across two runs and two
177 sessions. The purpose of multiple samples with the same task design is to evaluate the
178 consistency in findings across studies that vary in their sample populations and task design as

179 little evidence exists on the *consistency* of reliability estimates for the same task across
180 independent samples. **Aim 1** evaluates the effects of analytic decisions including task model
181 **smoothing, motion correction, parameterization (i.e., modeling) and task contrasts on** the impacts
182 on reliability, calculated using ICC(3,1) for individual [continuous] beta estimates and Jaccard's
183 similarity coefficient using significance thresholded group [binary] estimates ($p < .001$,
184 uncorrected) and Spearman correlation group [continuous] estimates. The decisions are noted in
185 **Table 1. Aim 1 Hypothesis** is that the highest produced ICC and similarity
186 coefficient/correlation is for the model decisions indicated by **blue** for A-D decisions in Table 1.
187 This, in part, is because the analytic strategy includes 1) motion correction techniques that limit
188 the number of noisy (high motion) subjects and reduce the number of degrees of freedom that are
189 lost due to censoring, 2) an optimal smoothing for the size of voxels, and 3) the highest
190 activation contrast from a task modeling phase that is relatively efficient. We hypothesize this to
191 be more so the case for the older (e.g., AHRB/MLS) than younger samples (e.g., ABCD) due to
192 changes occurring as a result of development (Herting et al., 2017; Noble et al., 2021). Due to
193 the lack of information regarding how the **between-subject variance (BS) and within-subject**
194 **variance (WS)** is impacted by analytic choices in task fMRI analyses, **Aim 2** evaluates the
195 change in **BS and WS** components. Due to the poor reliability of individual estimates in task
196 fMRI (Elliott et al., 2020), reported evidence of high between-subject variability in BOLD
197 activity (Turner et al., 2018), and limited evidence on changes in **BS and WS** variance
198 components in the MID task, we do not have a specific **Aim 2 Hypothesis**. Finally, seeing as the
199 ICC is, in some ways, similar to a moment product correlation (Bennett & Miller, 2010) which
200 stabilizes at larger sample sizes (Grady et al., 2020; Marek et al., 2022; Schönbrodt & Perugini,
201 2013), **Aim 3** evaluates at what sample the ICC stabilizes using the most optimal pipeline (e.g.,
202 highest median ICC) used in Aim 2. Stability of Jaccard coefficient group maps is not considered
203 in Aim 3 as these estimates are sensitive to significance thresholding. Using the evidence from
204 prior work on correlations (Grady et al., 2020; Schönbrodt & Perugini, 2013), the **Aim 3**
205 **Hypothesis** is that the ICC will stabilize a sample size between 150 to 500.

206

207 *Table 1.* Proposed Analytic Permutations: 360 Total
208 Modeling Combinations for MID task

First-level Pipeline Decisions	Options
--------------------------------	---------

A. Smoothing (FWHM)	
1. 1.5x voxel	ON / OFF
2. 2x voxel	ON / OFF
3. 2.5x voxel	ON / OFF
4. 3x voxel	ON / OFF
5. 3.5x voxel	ON / OFF
B. Motion Correction	
1. None	ON / OFF
2. Regress: Translation/Rotation (x,y,z) + Derivative (x,y,z)	ON / OFF
3. Regress: Regress: Translation/Rotation (x,y,z) + Derivative (x,y,z) + First 8 aCompCor Components	ON / OFF
4. Regress: Translation/Rotation (x,y,z) + Derivative (x,y,z) + First 8 aCompCor Components + Censor High Motion Volumes (FD \geq .9)	ON / OFF
#5. Regress: Translation/Rotation (x,y,z) + Derivative (x,y,z) + First 8 aCompCor Components, Exclude mean FD \geq .9	ON / OFF
#6. Regress: Translation/Rotation (x,y,z) + Derivative (x,y,z) + First 8 aCompCor Components + Censor High Motion Volumes, Exclude mean FD \geq .9	ON / OFF
C. Task Modeling	
1. MID: Cue Onset, Cue Duration only	ON / OFF
2. MID: Cue Onset, Cue + Fixation Duration	ON / OFF
3. MID: Fixation onset, Fixation Duration	ON / OFF
D. Task Contrasts	
1. MID: Big Win > Neutral	ON / OFF
2. MID: Big Win > Implicit	ON / OFF

3. MID: Small Win > Neutral	ON / OFF
4. MID: Small Win > Implicit	ON / OFF

209 **Blue text:** Model hypothesized to produce the highest test-retest
210 reliability; aCompCor: Anatomical Component Based Noise
211 Correction; MID: Monetary Incentive Delay task; FD: Framewise
212 displacement.
213 #Due to the lack of low motion subjects (zero mean FD <.90 in
214 2/3 samples), this decision was not included in the Stage 2
215 analyses, resulting in 240 analytic models.

216 Methods

217 To answer the questions proposed in Aim 1 and Aim 2, this study will require multiple
218 samples and tasks to obtain a comprehensive view of how analytic decisions impact group and
219 individual reliability metrics (Aim 1) and how **BS and WS** is impacted (Aim 2) across multiple
220 samples and similar MID task. We use three samples with subjects that have at least two
221 repeated sessions of data. To answer the question about the sample at which ICC stabilizes (Aim
222 3), we use the repeated session data from a large consortium sample.

223 The studies were selected based on two criteria. First, the goal is to derive group and
224 individual estimates of reliability using sample sizes that are larger than the reported median
225 sample size in fMRI research. The median reported sample size in fMRI is <30 subjects
226 (Poldrack et al., 2017; Szucs & Ioannidis, 2017). From the review of task fMRI reliability by
227 Bennet and Miller (2010), the median sample for individual (continuous) reliability is 10 subjects
228 (mean = 10.5 [range = 1 to 26]) and for group (binary) reliability is 9.5 subjects (mean = 11.2
229 [range = 4 to 45]). A recent review and analysis of task fMRI reliability suggests sample sizes
230 are increasing but remain lower than the median sample size in task fMRI, whereby the median
231 sample size for individual reliability in the meta-analysis are 18 subjects (mean = 26.4 [range = 5
232 to 467]) and the analyses are 45 & 20 subjects (Elliott et al., 2020). Second, the goal is to limit
233 the interaction between reliability estimates and unknown features of the data, such as the mental
234 processes, to get a sense of how the analytic pipeline impacts reliability estimates *consistently*
235 across a similar task design. Thus, the three samples described below exceed $N > 50$ and use a
236 nearly identical task that is known to evoke a strong BOLD response in specific brain regions to
237 achieve these two goals.

238 Participants²

239 Adolescent Brain Cognitive Development (ABCD) Study

240 The ABCD Study® is a longitudinal national study that was designed to study the change
241 in behavioral and biological measurements across development (Volkow et al., 2018). The focus
242 here is on the 4.0 brain imaging data that is released by the ABCD-BIDS Community Collection
243 (ABCC; Feczko et al. [2021]). As of February 2024, the ABCC data contains year 1
244 (approximately 11,000, participants Aged 9-10) and year 2 (approximately 7,000 participants,
245 Age 11-13) fMRI data. For Aims 1 and 2, we use a subsample of ABCD participants at the
246 University of Michigan site (site = 13) with maximum clean data available as this would be
247 sufficient to test the hypotheses and limit site and scanner effects. For Aim 3, we use a
248 subsample of $N = 2,000$ of the maximum clean data available from the ABCC sample and use an
249 adaptive design to answer at which N ICC stabilizes. To reduce the use of unnecessary
250 computational resources, the analyses are first performed in $N = 525$. If the difference between
251 average ICC estimate for interval N_i & N_{i-1} is $> .15$, the sample will be extended to $N = 1000$,
252 adding $N = 500$, until the plotted estimates are stable. As described elsewhere (Casey et al.,
253 2018), the study collected fMRI data during the StopSignal, Emotional N-back and MID tasks.
254 Reliability of consortium-derived region of interest level data for year 1 and year 2 has been
255 reported elsewhere (Kennedy et al., 2022). We expand on these findings by evaluating how
256 consistent these results are across studies and which analytic decisions impact estimates of
257 reliability. Here, we use the raw BOLD timeseries from the MID task as this is consistent with
258 the two other studies described below.

259 *Michigan Longitudinal Study (MLS)*

260 The MLS is a longitudinal study focused on the change in behavioral and biological
261 measurements across development. As described elsewhere (Martz et al., 2016; Zucker et al.,
262 2000), the MLS includes the Neuropsychological Risk cohort. The MLS Neuropsychological
263 Risk cohort contains year 1 (approximately 159 participants, Age 18-24) and year 2

² For the Stage 1 submission, the data for the different studies was not fully accessed, inspected, preprocessed or analyzed. Thus, the sample size approximations. The final N for each sample is expected to deviate from the approximated values because of complete data availability and quality control exclusions.

264 (approximately 150 participants, Age 20-26) fMRI data. The study collected fMRI data during
265 the affective word and MID tasks. Here, we use the raw BOLD data from the MID task as it is
266 consistent with the ABCD study and Adolescent Risk Behavior Study (described below).

267 *Adolescent Risk Behavior (AHRB) Study*

268 The AHRB study is a longitudinal study focused on the change in behavioral and
269 biological measurements across development. The AHRB study contains year 1 (approximately
270 108 participants, Age 17-20) and year 2 (approximately 66 participants, Age 19-22). The study
271 collected fMRI data during the Emotional Faces and MID tasks. Here, we use the raw BOLD
272 data from the MID task as it is consistent with the MLS and AHRB study.

273 FMRI Task, Data, Preprocessing

274 *FMRI Tasks*

275 Across the ABCD, AHRB and MLS studies, reward processing was measured using
276 comparable versions of the MID task. The MID task (Knutson et al., 2000) is used to model
277 BOLD signatures of the anticipation and receipt of monetary gains or losses. The MID task and
278 their nuanced differences across the ABCD, AHRB and MLS studies are described in
279 supplemental **Section 1.2**. The focus of the present work is on the anticipatory phase of the task.

280 *MRI Acquisition Details*

281 The acquisition details for the AHRB, ABCD and MLS datasets are summarized in
282 supplemental **Section 1.3 Table S2**.

283 *Data Quality Control and Preprocessing*

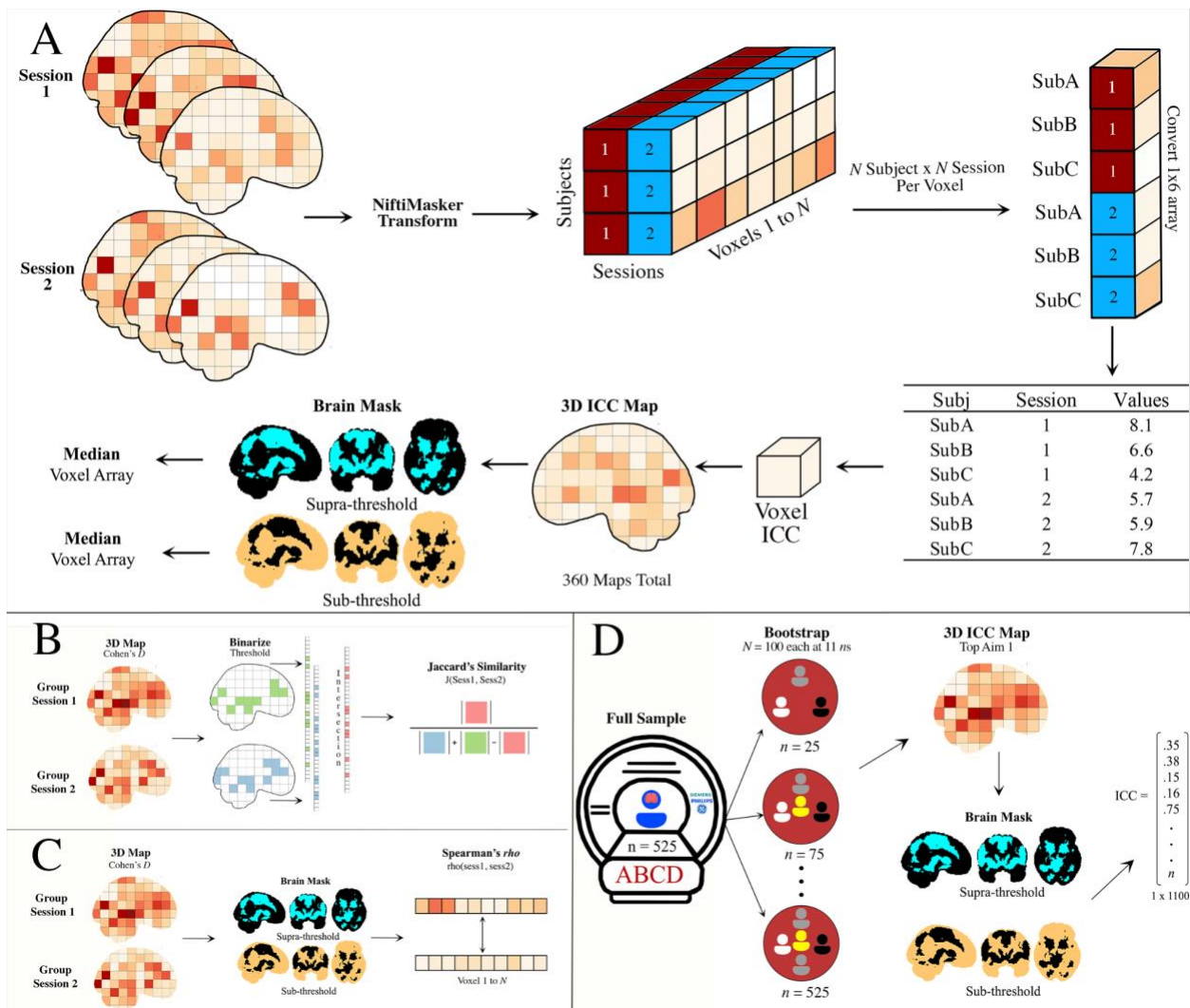
284 First, quantitative metrics reported from MRIQC version 23.1.0 (Esteban et al., 2023) for
285 the structural and BOLD data are evaluated to assess data quality and potentially problematic
286 subjects. Second, behavioral data were inspected to confirm that participants have the behavioral
287 data for each run and that participants performed at the targeted probe hit rate (e.g., at or near
288 60% overall probe hit rate, see supplemental **Section 1.2**). Then, structural and functional MRI
289 preprocessing is performed using fMRIPrep v23.1.4 (Esteban et al., 2022; RRID:SCR_016216),

290 which is based on Nipype 1.8.3 (Esteban, Markiewicz, Burns, et al., 2022; RRID:SCR_002502)
291 and the results are inspected to confirm no subjects' preprocessing steps failed.

292 Preprocessing between the ABCD, AHRB and MLS are held constant except for two
293 differences. First, the MLS datasets did not collect fieldmaps and the repetition time for MLS
294 (2000ms) is slower than the repetition time (800ms) in ABCD/AHRB. Therefore, fMRIPrep's
295 fieldmap-less distortion correction (SyN-SDC) is used to estimate and correct for fieldmap
296 distortions in MLS and slice-timing correction is applied *only* on the MLS data. For the ABCD
297 and AHRB data, fieldmap-less distortion correction is used *only* when a subject does not have
298 the necessary fieldmaps. Outside of these two exceptions, the preprocessing of the BIDS data
299 were preprocessed using identical pipelines. The complete preprocessing details are included in
300 supplemental **Section 1.4**

301 Analyses

302 This project is focused on the effects of analytic decisions on estimates of reliability
303 across (run/session) measurement occasions in task fMRI. As a reminder, reliability is the
304 estimate of how similar two measures (in this case, voxels for a given contrast from a fMRI 3D
305 volume) are in terms of estimated effects (continuous) and/or the presence/absence of a
306 significant effect (binary). We distinguish individual and group estimates in **Figure 1** and
307 describe the calculations below. For the continuous estimates of reliability described below, the
308 analyses will be performed separately on task voxels that exceed and do not exceed an *a priori*
309 specified threshold applied on the NeuroVault (Gorgolewski et al., 2015) meta-analysis
310 collection that comprises the anticipatory win phase across 15 whole brain maps for the MID
311 task (Wilson et al., 2018; Collection: 4258, Image ID: 68843). The *suprathreshold* task-positive
312 voxels are those that exceed the threshold ($z > 3.1$) and the *subthreshold* task voxels are those
313 that do not exceed the threshold ($z < 3.1$) in the map. We acknowledge that the threshold of $z =$
314 3.1 is arbitrary (uncorrected, p -value = .001) and that the voxels that fall below and above this
315 threshold may not be significantly different (Gelman & Stern, 2006). However, to constrain the
316 problem space this is a researcher's decision that is made in these analyses (Gelman & Loken,
317 2014; Simmons et al., 2011).



318

319 *Figure 1.* Diagram of (A) Continuous (individual), (B/C) binary/continuous (group) and (D)
 320 random subsampling of Estimates of Reliability across Measurement Occasions in 3D volumes
 321 of fMRI data.

322 Group = group average of activation; Sub = Subject; ICC = Intraclass Correlation; Supra- and Sub-threshold mask is
 323 > 3.1 of NeuroVault Vault Image ID #68843 (Collection #4258)

324 *Descriptive Statistics*

325 The mean, standard deviation, count and frequencies are reported for demographic
 326 variables from the ABCD, AHRB and MLS datasets. For ABCD, AHRB and MLS, participants
 327 self-reported on Age, Sex and Race/Ethnicity. ABCD: Sex is reported as sex at birth (Male,
 328 Female, Other, or Not Reported); Race/Ethnicity is reported on a 5-item scale: White, Black,
 329 Hispanic, Asian, Other. AHRB: Sex is reported as sex at birth (Male or Female); Race/Ethnicity
 330 is available on a 4-item scale: White, Non-Hispanic, Black, Non-Hispanic, Hispanic/Latinx,
 331 Other. MLS: Sex is reported as Sex at Birth; Race is available on an 8-item scale: Caucasian,

332 African American, Native American, Asian American, Filipino or Pacific Islander, Bi-Racial,
333 Hispanic-Caucasian, and Other.

334 Behavioral data from the MID task, such as the mean and distribution of probe hit rate
335 and mean response times (RT) across subjects, will be reported as supplemental information. The
336 task design is programmed to achieve a probe hit rate of approximately 60% for each subject. It
337 should be noted that the RT for the probe is not consistently collected across the ABCD, AHRB,
338 and MLS datasets.

339 *Impact of Analytic Decisions on Reliability in fMRI Data*

340 First-, second- and group-level analyses are performed using Python 3.9.7 and Nilearn
341 0.9.2 (Abraham et al., 2014). Details about these three analytic steps are described below and the
342 code is provided on Github. As listed in **Table 1** and described next, the analytic decisions will
343 be limited to the first-level analysis.

344 *Analytic Decisions:* For reasons described in the introduction, the focus of analytic
345 decisions in this paper will be on **four** categories: Smoothing, Motion Correction, Task Contrast
346 and Task Parametrization. As reported in empirical studies and meta-analyses of task fMRI
347 reliability (Bennett & Miller, 2010; Caceres et al., 2009), one way to improve reliability of fMRI
348 data is by increasing the signal-to-noise ratio in the BOLD data through different smoothing
349 kernels (Caceres et al., 2009), reducing motion effects in the fMRI data (Gorgolewski et al.,
350 2013; Kennedy et al., 2022) and using task designs/contrasts that evoke increased neural activity
351 (Han et al., 2022; Kennedy et al., 2022). These analytic decisions are described in greater detail
352 in supplemental **Section 1.1**.

353 *Within-run Analysis:* A general linear model (GLM) is fit using Nilearn (e.g.,
354 *FirstLevelModel*) to estimate the response to task-relevant conditions in the BOLD timeseries for
355 each participant/voxel. The BOLD timeseries are masked and spatially smoothed using specified
356 full-width half-maximum (FWHM) Gaussian kernel options (see ‘Smoothing’ in **Table 1**) and
357 the timeseries are prewhitened using an ‘ar1’ noise model. A GLM is fit (using *FirstLevelModel*)
358 for a design matrix that includes the 15 task-relevant regressors (see task details in supplemental
359 **Section 1.2**) and a set of nuisance regressors. Depending on the decision criteria (see ‘Motion
360 Correction’ in **Table 1**), nuisance regressors may include, for example, **A**) estimated translation
361 and rotation (+ derivatives) of head motion or **A** + first eight aCompCor noise components and

362 the corresponding cosine regressors for high pass filtering (with a cutoff of 128 seconds) that are
363 calculated by fMRIPrep (see preprocessing of functional data). Task regressors are convolved
364 with the SPM hemodynamic response function (HRF). The resulting beta estimates from the
365 GLM, for each individual subject and run, are used to compute four contrasts for the MID task
366 (see ‘Task Contrasts’ in **Table 1**).

367 *Within-session Analysis:* Per subject, each study collected two runs for each of two
368 sessions. For each of the four contrast types, the beta and variances estimates from the two MID
369 runs for each subject are averaged using Nilearn’s precision-weighted fixed effects model (i.e.,
370 *compute_fixed_effects*).

371 *Group-level Analysis (within-session):* The MID task weighted fixed effects contrast files
372 are used in a group-level mixed effect model (i.e., Nilearn’s *SecondLevelModel*) to average the
373 within-subject estimates across subjects. These group maps are used as measures of the average
374 activation patterns during the MID task in each of the studies across each of the four contrast
375 types within each session.

376 The resulting individual and group maps from the four contrasts are used in calculating
377 two different estimates of reliability (described in detail below). First, the resulting *within-run*
378 *analysis* maps (i.e., for each run) are used for the continuous estimate of reliability *within* each
379 session (i.e., reliability across runs). Then, the resulting *within-session analysis* maps, computed
380 from the weighted fixed effects model, are used in the continuous estimate of reliability *between*
381 the two sessions. Due to the temporal difference within and between sessions, the reliability
382 within sessions would be hypothesized to be greater than between sessions. The resulting group-
383 level analysis maps are used in the binary estimate of reliability *between* sessions.

384 *Estimate of Reliability for Continuous Outcomes: Intraclass Correlation*

385 Reliability for continuous outcomes at the individual level is estimated using ICC. The
386 ICC is an estimate of between-subject and within-subject variance that summarizes how similar
387 the signal intensities are for a given voxel from a 3D volume across sessions. As described in
388 Liljequist et al. (2019), there are several versions of the ICC, which vary in whether the subjects
389 and sessions are considered to be fixed (e.g., ICC[1]), subjects are considered to be random and
390 sessions are considered to be fixed (e.g., consistency, estimated via ICC[3,1]) or the subjects and
391 sessions are considered to be random (e.g., agreement, estimated via ICC[2,1]). In the case of

392 these analyses, we assume that subjects are random but do not assume that sessions are random
393 for two reasons. First, in the case of reliability of runs within a session, the runs are administered
394 in a fixed manner and the state of the participant cannot be assumed to be random for each.
395 Second, in the case of reliability across sessions, during the follow-up session subjects have
396 experienced the MRI environment and the task design in the scanner. In this case, again, it is
397 difficult to assume that sessions are in fact random as the practice and session effects may be
398 present. Thus, we estimate the consistency (ICC[3,1]) of the signal intensity for a given voxel
399 across measurement occasions.

400 Several packages exist to calculate ICC and Jaccard/Dice coefficients. For example,
401 *ICC_rep_anova & Similarity* in Python (Gorgolewski et al., 2011), *fmreli* in MATLAB (Fröhner
402 et al., 2019) and *3dICC* in AFNI (Chen et al., 2017). However, these packages are either a)
403 limited to a specific ICC calculation (e.g., ICC[3,1]), b) not easy to integrate into reproducible
404 python code (e.g., *fmreli*), c) do not include similarity calculations (e.g., *3dICC*), or do not return
405 information about between-subject, within-subject and between-measure variance components.
406 Thus, to have the flexibility to estimate ICC(1), ICC(2,1) and ICC(3,1), Dice and Jaccard
407 similarity coefficients and Spearman correlations simultaneously, we wrote and released an
408 open-source Python package with reliability and similarity functions that works on 3D NifTi
409 fMRI images.

410 The *PyReliMRI* v2.1.0 (Demidenko, Mumford & Poldrack, 2024) Python package is used
411 to calculate continuous estimates of reliability. *PyReliMRI* implements a voxel-wise ICC
412 calculation (e.g., *voxelwise_icc*) for 3D NifTI images between runs and/or between sessions (see
413 the ICC example in study flowchart, **Figure 1A**). The function takes in a list of lists (e.g., list of
414 session 1 and list of session 2) of ordered paths to the preprocessed data [in MNI space] for
415 session 1 (or run 1) and session 2 (or run 2) subjects, and a binary [MNI space] brain mask. The
416 package is flexible to take in more than 2 sessions (or runs). An ICC type option (e.g., 'icc_1',
417 'icc_2' or 'icc_3') indicates the type of ICC estimate that is calculated across the voxels within
418 the masked 3D volume. The function returns a dictionary with five separate 3D volumes
419 containing the voxel-wise (1) ICC estimate, (2) lower bound ICC, (3) upper bound ICC, (4)
420 **Between-subject variance (BS) and (5) Within-subject variance (WS) and, in case of ICC(2,1),**
421 **(5) Between-measure variance, or the measurement additive bias.** Like the ICC & 95%
422 confidence calculation in the *pingouin* package (Vallat, 2018), the ICC confidence interval in

423 *PyReliMRI* is calculated using the *f*-statistic (Bonett, 2002) to reduce the computation time
424 compared to using bootstrapped estimates.

425

$$426 \quad ICC(3,1) = \frac{MSBS - MSError}{MSBS + MSError} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_v^2} \quad \text{Equation 1}$$

427

428

429 *Aim 1a*: evaluated the effect of analytic decisions (see **Table 1**; **Figure 1A**) on the
430 ICC(3,1) (equation 1 for two measurement occasions) for individual [continuous] estimates of
431 voxel activity across the ABCD, AHRB and MLS studies. The parameters in Equation 1 are:
432 *MSBS* is the Mean Squared Between-subject Error and *MSError* is the Mean Squared Error. As
433 described in Liljequist et al. (2019), the differences in the numerator is the between-subject
434 variance (σ_r^2) and the denominator is the sum of the between-subject variance (σ_r^2) and the
435 within-subject variance (or noise, [σ_v^2]). For each study, *voxelwise_icc* within the *brain_icc.py*
436 script is used to estimate the voxel-wise ICC(3,1) for between run and between session reliability
437 across the 360 model permutations. First, voxel-wise average and standard deviation from the
438 resulting ICCs for the 360 model permutations are reported in two 3D volumes. Second, the
439 range and distribution of median ICCs across each study (three) and analytic decision category
440 (four) are plotted across suprathreshold task-positive and subthreshold ICCs using Rainclouds
441 (Allen et al., 2019) and the median and standard deviation are reported in a table. Third, to
442 visualize the ordered median ICCs across the 360 model permutations for suprathreshold task-
443 positive and subthreshold ICCs, specification curve analyses are used (Simonsohn et al., 2020).
444 Specifically, results across the 360 model permutations are reported using a specification curve
445 to represent the range of estimated effects across the variable permutations. This consists of two
446 panels: Panel A represents the *ordered* median ICC coefficients and the associated 95%
447 confidence interval (across samples) colored based on no significance (gray), negative (red) or
448 positive (blue) significance from the Null (Null here is 0) and Panel B represents the analytic
449 decisions from each of the four categories (see **Table 1**) that produced the median ICC estimates.
450 The median ICC estimates from the 360 models are reported separately for suprathreshold task-
451 positive and subthreshold activation (the specification curve for all ICC estimates for
452 suprathreshold task-positive and subthreshold activation are provided as supplemental
453 information). Finally, to evaluate the effect of the analytic decisions on the median ICC,

454 hierarchical linear modeling (HLM) is performed as implemented in the *lmer()* function from the
455 *lme4* R package (Bates et al., 2020). HLM is used to regress the median ICC on the [four]
456 analytic decisions as fixed effects with a random intercept model is fit (Matuschek et al.,
457 2017) for samples across the suprathreshold task-positive and subthreshold maps. Multiple
458 comparisons corrections are applied using the Tukey adjustment as implemented in the *emmeans*
459 package (Lenth et al., 2023). For these HLM models, the interpretation focuses on the
460 significant, non-zero effect of an independent variable (e.g., smoothing) on the dependent
461 variable (e.g., median ICC) while the remaining independent variables are assumed to be zero.

462 *Aim 2:* evaluated the change in between- and within-subject variance across the analytic
463 model permutations. Similar to Aim 1 (**Figure 1A**), *voxelwise_icc* within the *brain_icc.py* script
464 is used to estimate the **BS and WS** across the 360 model permutations. The range and
465 distribution of median **BS and WS** across each study and analytic decision category are plotted
466 across suprathreshold task-positive and subthreshold **BS/WS** using Rainclouds. Then, two
467 separate specification curve analyses report the *ordered* median **BS and WS** coefficients in one
468 panel and the analytic decisions that produced the **BS and WS** estimates in a second panel
469 separately for suprathreshold task-positive and subthreshold activation. Finally, like Aim 1, two
470 HLMs are used to regress the median **BS and median WS** on the [four] analytic decisions as
471 fixed effects with a random intercept only for sample across the suprathreshold task-positive and
472 subthreshold maps. Multiple comparisons corrections are applied using the Tukey adjustment.
473 Like Aim 1, the interpretation focuses on the significant, non-zero effect of an independent
474 variable (e.g., smoothing) on the dependent variable (e.g., median **BS or median WS**) while the
475 remaining independent variables are assumed to be zero.

476 *Aim 3:* evaluated the sample size at which the ICC stabilizes (**Figure 1D**). The chosen
477 pipeline is based on the highest median ICC across the studies for the suprathreshold task-
478 positive mask from Aim 1a and is rerun for the ABCD sample. Based on this pipeline, the first-
479 level analysis steps are repeated for $N = 525$ from the $N = 2000$ subsample for only the ABCD
480 data. Then, *voxelwise_icc* within the *brain_icc.py* script is used to derive estimates of the median
481 ICC, **BS and WS** for the between runs (e.g., measurement occasions) reliability across randomly
482 sampled subjects for 25 to 525 subjects in intervals of 50. Similar to the methods in Liu et al.
483 (2023), 100 iterations are performed at each N (with replacement) and the median ICC, the
484 associated **BS and WS** estimates are retained from *voxelwise_icc*. The average and 95%

485 confidence interval for the estimates across the 100 iterations is plotted for each interval of N
486 with the y-axis representing the median ICC and x-axis representing N . The plotted values will
487 be used to infer change and stability in the estimated median ICCs and variance components
488 across the sample size. If stability is not achieved by $N = 500$, the sample is extended to $N =$
489 1,000 and the analyses are repeated.

490 *Estimate of Reliability: Jaccard Coefficient for Binary & Spearman Correlation for Continuous* 491 *Outcomes*

492 The estimate of reliability for group analyses is estimated using the Jaccard Similarity for
493 binary and Spearman correlation for continuous outcomes. The estimates are used to evaluate
494 how the MID task evokes BOLD activation above a pre-specified threshold ($p < .001$) in the
495 same voxels for *groups* of subjects across measurement occasions (run/session) in the ABCD,
496 AHRB and MLS studies.

497 The *PyReliMRI* package is used. *PyReliMRI* calculates the similarity between two 3D
498 volumes using a Jaccard's coefficient which, in short, is the intersection divided by the union
499 between two binary images (see **Figure 1B**) or the Spearman correlation, which is ranked
500 correlation between two continuous variables (see **Figure 1C**). The Jaccard coefficient ranges
501 from 0 to 1, whereby higher values reflect greater similarity between two images. Like the
502 product-moment correlation, the Spearman correlation ranges from -1 to 1, whereby values >0
503 indicate a positive association between images and values <0 indicate a negative association
504 between images. The function (i.e., *image_similarity*) takes in the paths for MNI *image file1* and
505 *image file2*, a specified MNI mask and integer (i.e., z-stat/t-stat) at which to threshold the image.
506 The images are masked (if a mask is provided), thresholded at the specified integer (if a
507 threshold is provided) and the resulting images are binarized per user's input (i.e., if threshold =
508 0, the resulting similarity = 1). Based on the specified similarity metric, the resulting estimates
509 are similarity (e.g., Dice/Jaccard) or correlation coefficient (e.g., Spearman) between the two 3D
510 NIfTI images. For similarity between 2+ NIfTI images, *pairwise_similarity* is used. Similar to
511 *image_similarity*, *pairwise_similarity* takes in paths for an MNI mask, a threshold integer for the
512 3D volumes and the similarity type. Unlike *image_similarity*, *pairwise_similarity* allows for a
513 list (2+) of paths pointing to 3D volumes and creates pairwise-combinations across the image
514 paths between which to estimate similarity. The function returns the similarity coefficient in a

515 dataframe with the resulting similarity (or correlation coefficient) and the image label (e.g.,
516 basename of the provided path for given volume).

517

$$518 \quad J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad \text{Equation 2}$$

519

$$520 \quad \text{Spearman Correlation}_{A,B} = \frac{6\sum d_i^2}{n(n^2-1)} \quad \text{Equation 3}$$

521

522 *Aim 1b*: evaluated the effect of analytic decisions (see Table 1) in the Jaccard's similarity
523 coefficient (Equation 2; **Figure 1B**) and Spearman correlation (Equation 3; **Figure 1C**) using
524 the group binary & continuous estimates. In Equation 2, $J(A, B)$ is the
525 similarity coefficient between A (session 1) and B (session 2). This is
526 derived from intersection, $|A \cap B|$, which represents the elements
527 that are common to both A and B divided by the union, $|A \cup B|$, or the
528 elements that are both in A and/or B. In Equation 3, the Spearman Rank Coefficient, as
529 implemented in Scipy stats using *spearmanr* (Virtanen et al., 2020), is ranked correlation
530 between unthresholded images A and B, whereby $\sum d^2$ is the sum of squared differences between
531 ranked values in session A and B, normalized by $(n * (n^2 - 1))$.

532 Since the Jaccard similarity coefficient is sensitive to thresholding and sample size
533 (Bennett & Miller, 2010), in *Aim 1b* an equal sample size (e.g., $N \sim 60^3$) is chosen for each study
534 to compare how the similarity between sessions varies across studies. For all 360 pipelines, a
535 group-level (average) activation map is estimated for each session. In the case of the Jaccard
536 coefficient, the group maps are thresholded at $p < .001$. In the case of the Spearman coefficient,
537 the group maps are masked using a suprathreshold task-positive map from NeuroVault
538 (<https://identifiers.org/neurovault.collection:4258>; Image ID: 68843). Then, the paths for the
539 pipelines and sessions are called using the *pairwise_similarity* within the *similarity.py* script. The
540 resulting coefficients report the similarity between analytic pipelines and sessions for each study.
541 For each study, the coefficients are plotted to reflect the distribution and range of coefficients.

³ At Stage 1 the sample was based on an approximation. During Stage 2, we realized it would be more effective to take advantage of the complete available data by using standardized effect Cohen's *d* maps.

542 Both Jaccard's and Spearman correlation are reported separately. Like Aim 1a & Aim 2, two
543 HLMs are used to regress the Jaccard coefficients and Spearman correlation on the [four]
544 analytic decisions nested within study. Multiple comparisons corrections are applied using the
545 Tukey adjustment.

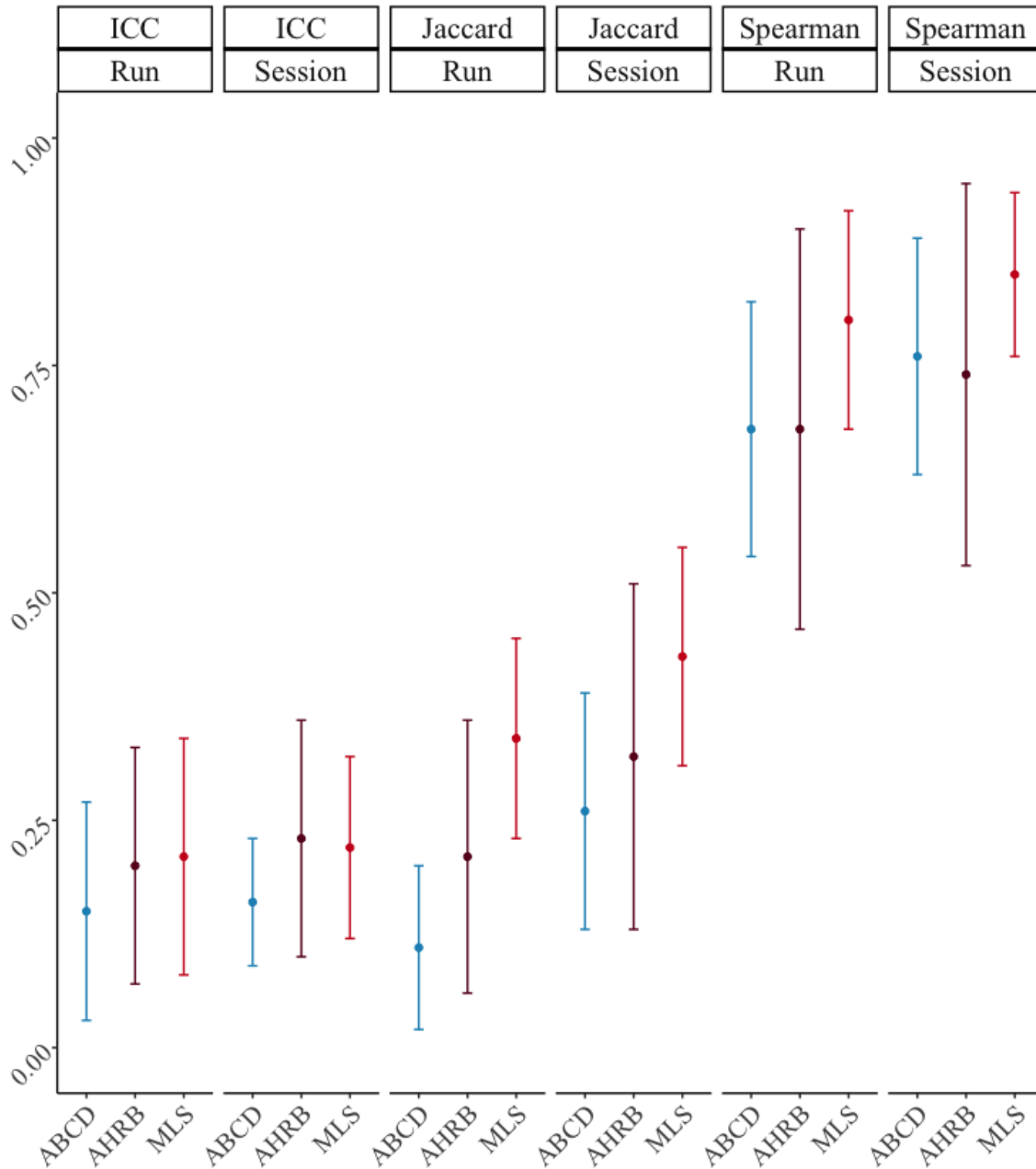
546 Results

547 Given the breadth of the analyses (see **Table 2**), the results in the main text focus on the
548 Session 1 between-run individual- and group-level reliability estimates for the supra-threshold
549 mask. Differences are briefly noted for between-session reliability estimates and sub-threshold
550 models and are reported in detail in the supplemental materials.

551 As permitted, aggregate and individual subjects' data are made publicly available on
552 NeuroVault (Gorgolewski et al., 2015) and/or OpenNeuro (Markiewicz et al., 2021). The
553 complete set of group-level and ICC maps are publicly available on Neurovault for ABCD (6180
554 images; <https://identifiers.org/neurovault.collection:17171>), AHRB (2400 images;
555 <https://identifiers.org/neurovault.collection:16605>) and MLS (2400 images;
556 <https://identifiers.org/neurovault.collection:16606>). For each run and session, the BIDS input
557 data and derivations for MRIQC v23.1.0 and fMRIPrep v23.1.4 are available on OpenNeuro for
558 AHRB (Demidenko, Huntley, et al., 2024) and MLS (Demidenko, Klaus, et al., 2024). Since the
559 ABCD data are governed by a strict data use agreement (March 2024), the processed data will be
560 made publicly available via the NDA at a later date as part of the ABCC release. The final code
561 for all analyses is publicly available on Github
562 (https://github.com/demidenm/Multiverse_Reliability⁴).

563 In the supplemental information of the Stage 1 submission, we stated that we would
564 adjust the smoothing weight for the MLS as its voxel size, 4 mm anisotropic, would result in
565 greater inherent smoothness of the data than ABCD/AHRB samples (2.4 mm isotropic voxel). A
566 weight of .50 was applied to the smoothing kernels of the MLS data. This resulted in 3.6, 4.8,
567 6.0, 7.2 and 8.4 mm smoothing kernels for the AHRB/ABCD data and 3.0, 4.0, 5.0, 6.0 and
568 7.0mm smoothing kernels for the MLS data (**Figure S4**). In the results, the MLS ordinal values
569 are relabeled to map onto the values used for AHRB/ABCD for reporting purposes.

⁴ Will revise with final Zenodo citation prior to Stage 2 acceptance.



570
 571 *Figure 2. Session 1 Between-runs and Between-sessions: Mean +/- 1 Standard Deviation (SD) of*
 572 *Supra-threshold median Intraclass Correlation Coefficient (ICC), Jaccard and Spearman*
 573 *Similarity Coefficients from 240 analytic models across ABCD, AHRB and MLS Samples.*
 574 *Note: Estimates in supplemental Table S5*

575 Deviations from Stage 1 Registered Report

576 There are one moderate and two minor deviations from the Stage 1 Registered Report

577 (<https://doi.org/10.17605/OSF.IO/NQGEH>). First, fieldmap-less distortion correction is not

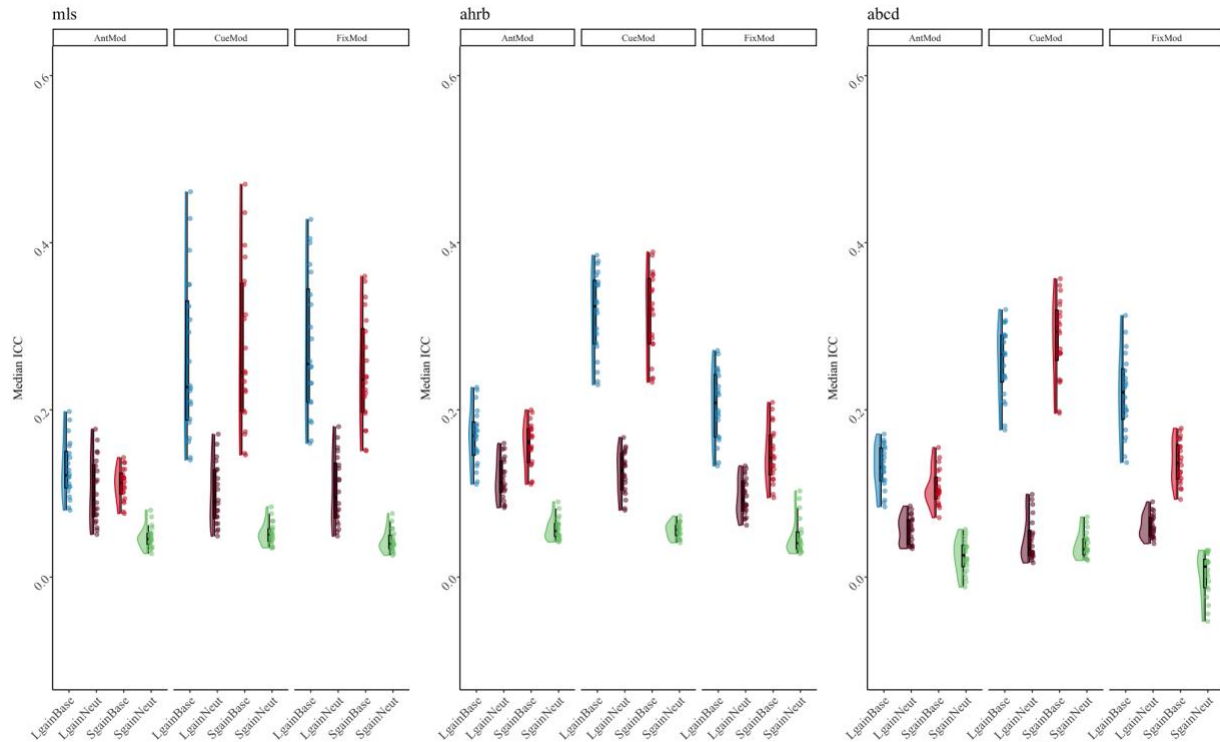
578 applied on the MLS data because the data were collected using spiral acquisition. The ABCC
579 data selects a single fieldmap within a session to apply on *all* of the functional runs, so subjects
580 without a fieldmap folder are excluded and fieldmap-less distortion correction is not used on the
581 ABCD data. In AHRB, fieldmap-less distortion correction was used for only *one* subject.
582 Second, in Aim 1b we proposed to use thresholded images (e.g., $p < .001$, approx. $t > 3.2$) to
583 estimate the Jaccard/Spearman similarity between the model permutations for the estimated
584 group maps. However, this statistic is arbitrarily sensitive to differences in the number of model
585 permutations when subjects are excluded in cases of failed preprocessing features, such
586 aCompCor mask errors. To improve the interpretability of the similarity estimates across
587 analyses with different numbers of included observations (see supplemental **Figure S3**), we
588 converted all t -statistic group maps to Cohen's d effect size maps using the formula: $\frac{t\text{-statistic}}{\sqrt{N}}$.
589 Cohen's $d = .40$ is used as the alternative threshold for Aim 1b as for pre-registered $N \sim 60$ a
590 conversion of $t\text{-statistic} = 3.2$ would be near this threshold. Third, the analyses proposed to
591 evaluate 360 analytic decisions across the three samples. However, no subjects in the final
592 AHRB and MLS samples exceeded mean FD = .9 so it was not possible to perform Motion
593 option 5 (Motion option 3 + exclude mean FD $\geq .9$) or Motion option 6 (Motion option 4 +
594 exclude mean FD $\geq .9$). As a result, the model permutations are restricted to 240 permutations (5
595 = FWHM, 6 \rightarrow 4 = Motion; 3 = Model Parameterization; 4 = Contrasts) with relevant data
596 across the three samples and are the focus of the below analyses.

597 Descriptive Statistics

598 The final sample for Aim 1 and Aim 2 for ABCD, AHRB and MLS samples (mean FD <
599 .90) from the University of Michigan site that had two runs for at least two sessions, had
600 behavioral data, and passed QC are N s 119, 60 and 81, respectively. For $N = 15$ subjects in the
601 ABCD sample aCompCor ROIs failed, but otherwise the data passed QC and so these subjects
602 were not excluded in Motion option3 and option4 models that include the top-8 aCompCor
603 components as regressors. The final random subsample from the Baseline ABCD data for Aim 3
604 is $N = 525$.

605 Demographic information across the three samples for Aim 1 and Aim 2 (ABCD = 119;
606 AHRB = 60; MLS = 81) are reported in supplemental **Table S4**. The average number of days
607 between sessions is largest for the MLS sample (1090 days), followed by ABCD (747 days) and
608 AHRB (419 days; **Figure S5**). On average, mean FD was higher in the ABCD sample versus the
609 AHRB and MLS samples (**Figure S6; Table S5**). The samples also differed on average response
610 probe accuracy (%), whereby on average MLS participants had a higher and faster probe
611 response accuracy than ABCD and AHRB samples.

612 The estimated model efficiency, defined as $Efficiency = \frac{1}{c(X'X)^{-1}c'}$, varied as a
613 function of Model Parameterization and Contrast types across the three samples (see **Figure S7**).
614 The Anticipation Model (i.e., onset times locked to Cue onset and duration the combined
615 duration of Cue and Fixation cross) was consistently estimated to be the most efficient model
616 across the three samples for the *Large Gain* versus *Neutral* and *Small Gain* versus *Neutral*
617 contrasts.

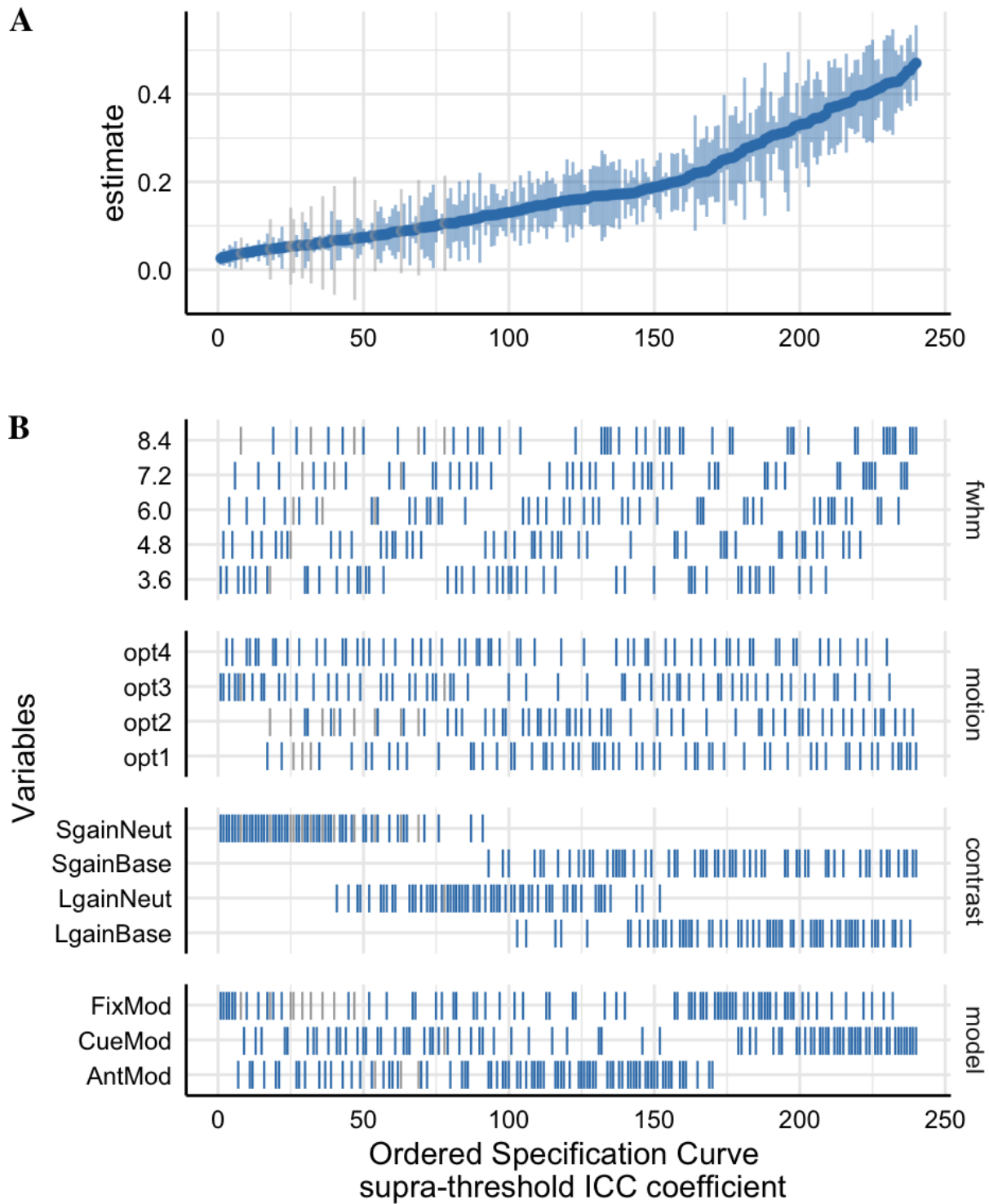


618
 619 **Figure 3.** Supra-threshold Median ICC Session 1 between-run reliability estimates for Contrast
 620 (con) and Model Parameterization analytic options across the ABCD, AHRB and MLS samples.
 621 Complete distribution across four analytic options in supplemental **Figure S9**.

622 Aim 1a: Effect of analytic decisions on median ICC estimates for individual
 623 continuous maps

624 Aim 1a proposed to evaluate the estimated individual map similarity between
 625 measurement occasions (runs/sessions) using the ICC(3,1) across 240 pipeline permutations. In
 626 **Table S5 (Figure 2)**, the median between-run Session 1 ICCs are slightly lower than the
 627 between-session ICCs (between-run: ABCD = .11 [range: -.04 - .43]; AHRB = .18 [range: .00 -
 628 .52]; MLS = .18 [range: .04 - .55]; between-session: ABCD = .15 [range: .03 - .34]; AHRB = .21
 629 [range: .04 - .53]; MLS = .21 [range: .06 - .47]). The mean and standard deviation of the 3D
 630 volumes across the 240 analytic decisions are reported in supplemental **Figure S8**. Across the
 631 three samples, a consistent pattern is observed, whereby the regions with the highest ICCs, on
 632 average, are within the visual and motor regions. Notably, the lowest ICCs, on average, are
 633 within the ventricles and white matter. The supra-threshold distribution of the median estimates
 634 across the four model options and three samples are reported in **Figure 3** and the specification

635 curve of the median ICC estimates are reported in Figure 4. Note, the sub-threshold reported in
 636 supplemental **Figure S10**.



637 **Figure 4.** The supra-threshold Specification Curve of the Session 1 Between-run Median ICC
 638 estimates across 240 pipeline permutations for the ABCD, AHRB and MLS samples. Full length
 639 of estimates reported in **Figure S11**.
 640

641 A. The distribution of the point estimate (average) and distribution (error bars) across the three samples. B. The
642 model options (four) associated with each estimate.

643
644 The effects reported in **Figure 3** and **Figure 4** illustrate that the largest differences in the
645 median ICC estimate is associated with model parameterization and the contrast type. Even
646 though the Anticipation Model ('AntModel') has the highest estimated contrast efficiency within
647 each sample, contrary to our hypothesis the highest median ICC is associated with the Cue
648 Model ('CueMod') in which the onset and duration are locked to the cue stimulus. However,
649 using an interaction to probe the distributions in **Figure 3**, *post hoc* analyses suggest the Cue
650 Model finding is largely driven by the *Implicit Baseline* contrasts (see Aim 1b) and the plot of
651 the Model Parameterization-by-Contrast in supplemental **Figure S12** suggests negligible
652 differences between Model Parameterization for the contrast of the *Neutral* contrasts.

653 Independent of model parameterization and consistent with our hypothesis and previous
654 reports in the task fMRI literature (Han et al., 2022; Kennedy et al., 2022), the highest median
655 ICC is consistently observed for the *Large Gain* versus *Implicit Baseline* contrast. In line with
656 the reported estimates in **Figure 3** and **Figure 4**, the HLM model for the supra-threshold mask
657 shows a significant association between different FWHM, Motion, Model Parameterization and
658 Contrasts model options compared to their respective reference values (**Table 3**). Specifically,
659 the median ICC estimates increased with larger smoothing kernels and decreased with more
660 stringent motion correction. Additionally, primarily driven by the *Implicit Baseline* conditions,
661 median ICC for the 'CueMod' and 'FixMod' increased in comparison to the 'AntMod' (see
662 interaction plot in **Figure S12**). Last, median ICC decreased in comparison to the *Large Gain*
663 versus *Implicit Baseline* contrast. For example, the contrast *Large Gain* versus *Neutral* has an
664 median ICC that is .17 lower, on average, compared to the *Implicit Baseline* contrast when
665 holding other decisions constant (see marginal means comparisons in supplemental **Table S6**).

666 While most parameters are significant in **Table 3**, the effects vary in their relative importance in
667 the model. The variability in the median ICC estimate across 240 pipelines and three samples is
668 best explained by contrast (marginal ΔR^2 : .55) and model parameterization (marginal ΔR^2 : .10).
669 FWHM and motion had a smaller impact on ΔR^2 , .03 and .03 respectively. In fact, including
670 aCompCor components (Motion option 3) and aCompCor components + censoring high motion
671 volumes (Motion option 4) is associated with a slight decrease in the median ICC estimate as
672 compared to no motion correction (Motion option 1), $b = -.05$ and $b = -.05$, respectively. A
673 similar finding is observed for the sub-threshold mask, whereby the contrast (ΔR^2 : .56) and
674 model parameterization (ΔR^2 : .10) decision had a larger impact on ΔR^2 than the FWHM (ΔR^2 :
675 .04) or motion (ΔR^2 : .02) decisions (see **Figure S14**; **Table S7**). In general, the voxelwise
676 distribution of ICC estimates tends to be higher for the supra-threshold mask than the sub-
677 threshold masks (see supplemental Figure S14). Interpretations are generally consistent for
678 between-session median ICC estimates across the 240 pipeline permutations (see **Table S9** and
679 **Figure S18, S19**).

680 We had hypothesized that the ICC estimates in the older samples (AHRB/MLS) would
681 meaningfully differ from the younger sample (ABCD). Overall, ICC estimates were higher in the
682 older than younger sample for *between-run*, $t(497.2) = 5.53$, $p < .001$, $d = .43$, and *between-*
683 *session*, $t(669.9) = 9.57$, $p < .001$, $d = .66$.

684 **Table 3. Hierarchical Linear Model: (A) Linear associations between the analytic decisions and**
685 **the Session 1 between-run median Intraclass Correlation Coefficient (ICC[3,1]), Between-**
686 **subject (BS) and Within-subject variance (WS) from supra-threshold mask and (B) the impact of**
687 **the analytic category on the marginal R².**

A. HLM Estimates for Supra-threshold Mask												
<i>Predictors</i>	Median ICC(3,1)			Median BS			Median WS					
	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>			
(Intercept)	.23	.20 – .26	<.001	.27	.18 – .35	<.001	.91	.72 – 1.10	<.001			
Reference [3.6]												
fwhm [4.8]	.02	.01 – .04	.003	-.03	-.06 – .00	.09	-.23	-.28 – -.18	<.001			
fwhm [6.0]	.04	.03 – .06	<.001	-.04	-.07 – -.01	.003	-.36	-.41 – -.31	<.001			
fwhm [7.2]	.06	.04 – .07	<.001	-.06	-.09 – -.03	<.001	-.44	-.49 – -.39	<.001			
fwhm [8.4]	.07	.05 – .08	<.001	-.07	-.10 – -.04	<.001	-.49	-.54 – -.44	<.001			
Reference [opt1]												
motion [opt2]	-.01	-.03 – .00	.07	-.04	-.06 – -.01	.01	-.14	-.18 – -.09	<.001			
motion [opt3]	-.05	-.06 – -.04	<.001	-.10	-.13 – -.08	<.001	-.23	-.28 – -.19	<.001			
motion [opt4]	-.05	-.06 – -.03	<.001	-.10	-.13 – -.08	<.001	-.24	-.28 – -.20	<.001			
Reference [AntMod]												
model [CueMod]	.10	.09 – .11	<.001	.15	.13 – .17	<.001	.26	.23 – .30	<.001			
model [FixMod]	.05	.04 – .06	<.001	.12	.10 – .14	<.001	.27	.23 – .31	<.001			
Reference [LgainBase]												
con [LgainNeut]	-.17	-.18 – -.16	<.001	-.22	-.25 – -.19	<.001	-.28	-.32 – -.23	<.001			
con [SgainBase]	-.02	-.04 – -.01	<.001	-.02	-.05 – .00	.09	.00	-.04 – .05	.93			
con [SgainNeut]	-.23	-.24 – -.22	<.001	-.24	-.27 – -.21	<.001	-.31	-.35 – -.26	<.001			
B. Analytic Category Model Impact												
Comparison	χ^2	New			χ^2	New			χ^2	New		
		Orig	R2	R2		ΔR^2	Orig	R2		R2	ΔR^2	Orig
[Full] vs [New - fwhm]	95	.72	.69	.03	25	.47	.45	.02	384	.52	.31	.21
[Full] vs [New - motion]	81	.72	.69	.03	81	.47	.42	.05	138	.52	.46	.06

[Full] vs [New - model]	263	.72	.62	.10	162	.47	.37	.10	221	.52	.42	.10
[Full] vs [New - con]	864	.72	.17	.55	397	.47	.17	.30	285	.52	.38	.14

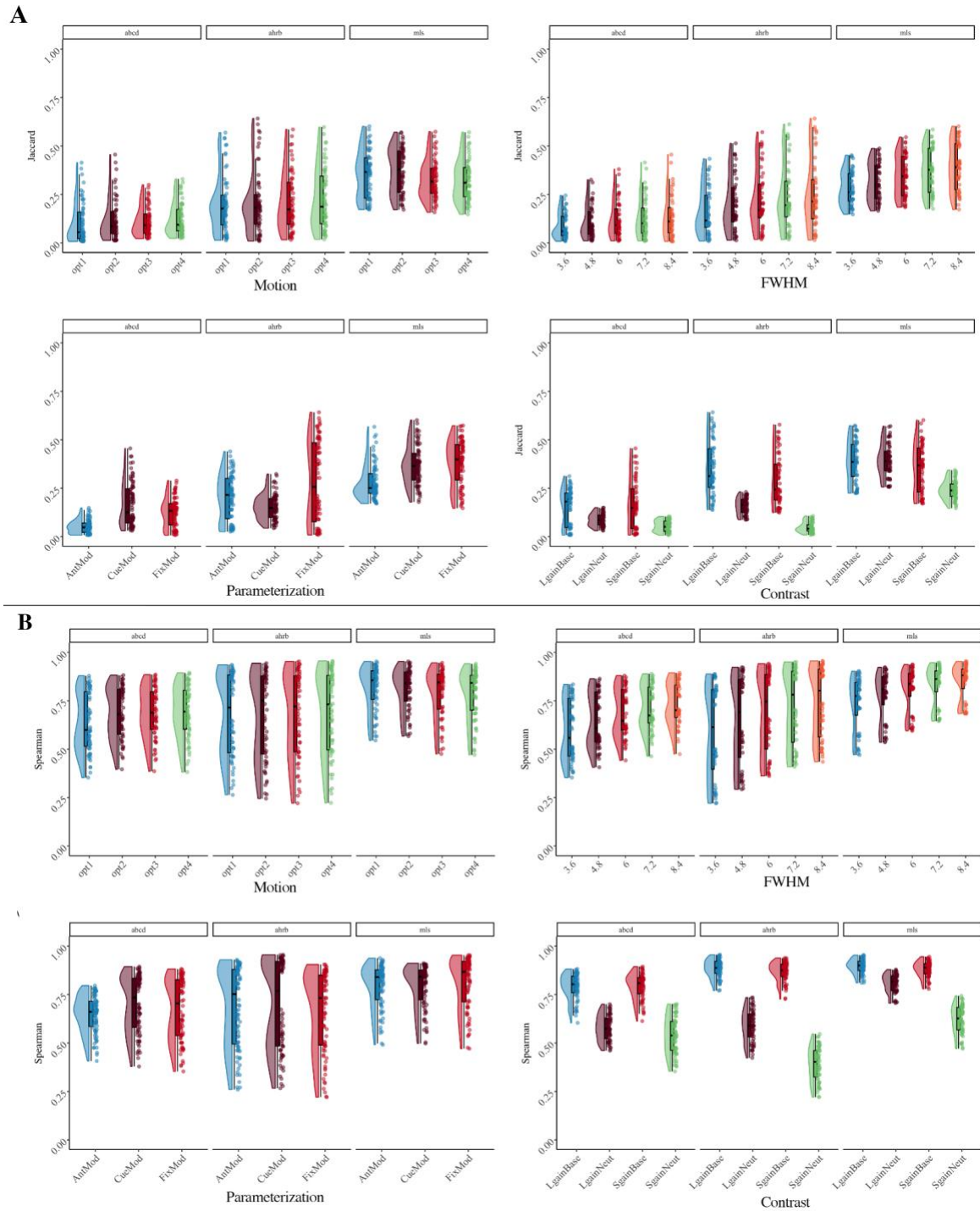
688 **Summary of Findings for Aim 1a:**

689 Overall, between-run ICCs are slightly lower than between-session ICCs. Across the
690 three samples, the highest ICCs, on average, are within visual and motor areas and the lowest
691 ICCs are within the ventricles and white matter. In Table 1, it was hypothesized that the optimal
692 analytic decisions would be: FWHM Smoothing 2.5x the voxel size, Motion correction that
693 includes translation/rotation, their derivatives, the first 8 aCompCor components and exclusion
694 of > .90 mFD subjects, the anticipation Model Parameterization, and Contrast *Large Gain* >
695 *Implicit Baseline*. Contrary to registered hypotheses: (1) smoothing had a small but linear effect
696 on ICC estimates, whereby the largest median ICC was for the largest FWHM smoothing kernel
697 (3.5x voxel size); (2) Motion correction had minimal and negative impact on median ICCs in
698 case of more rigorous corrections; and (3) the Cue and Fixation Models had higher estimated
699 median ICCs than the Anticipation model. *Post hoc* analyses illustrated Model Parameterization
700 is largely driven by the Implicit Baseline contrast, as Model Parameterization has a negligible
701 impact on between condition contrasts. Consistent with registered hypotheses, the *Large Gain*
702 versus *Implicit Baseline* had the highest estimated median ICC. Contrary to registered
703 hypotheses, there was little evidence to suggest that analytic decisions differentially impacted
704 estimated median ICCs between developmental samples (e.g., oldest MLS/AHRB versus
705 younger ABCD data). Finally, the older samples (AHRB/MLS) had higher between- and
706 between-session estimated ICCs than the younger sample (ABCD).

707 **Aim 1b: Effect of analytic decisions on Jaccard (binary) and Spearman**
708 **(continuous) similarity estimates of group maps**

709 Aim 1b proposed to evaluate the estimated group map similarity between measurement
710 occasions (runs/sessions) using a Jaccard similarity for thresholded binary maps and a Spearman
711 similarity for continuous measures across the 240 pipeline permutations. The distribution of the
712 estimates across the four model options and three samples are reported in **Figure 5** for Jaccard
713 and supra-threshold Spearman similarity. The specification curve of the Session 1 between-run

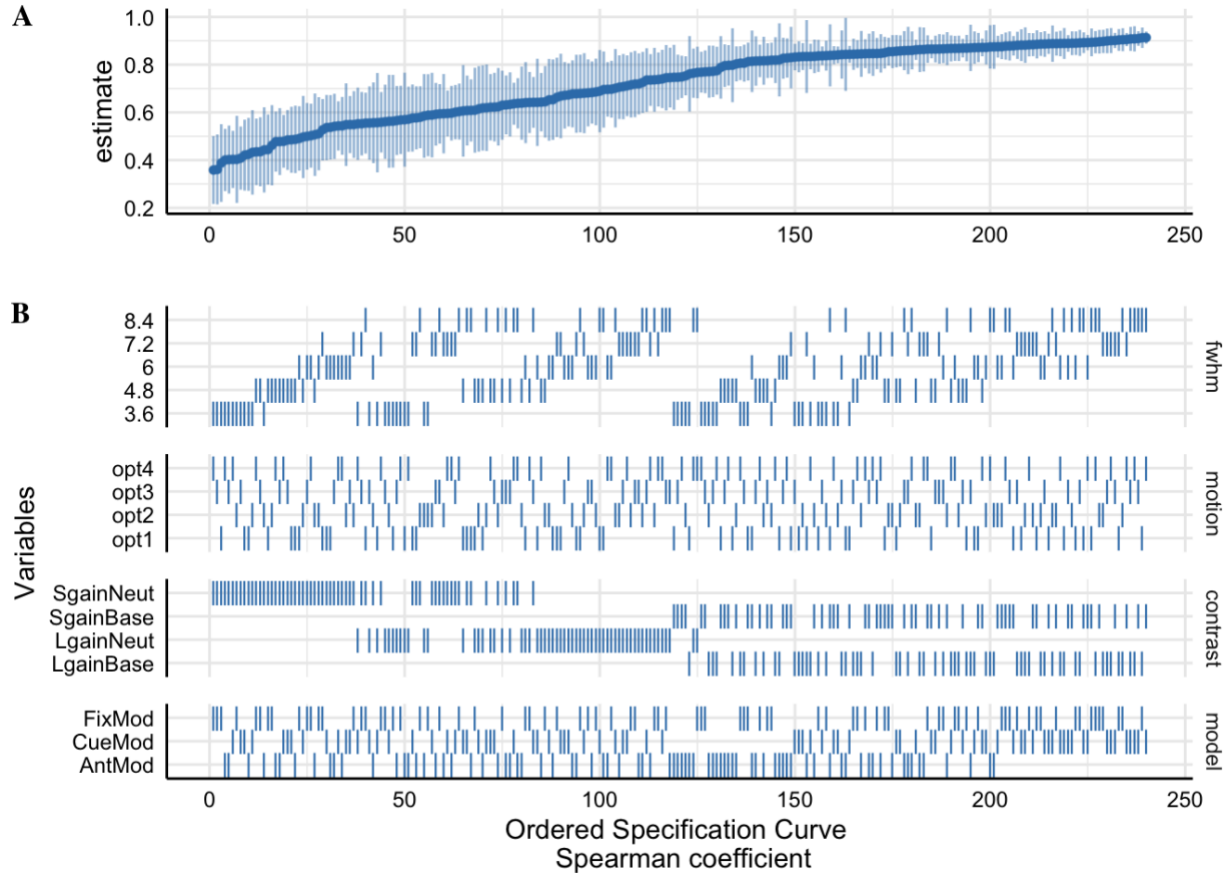
714 estimates are reported in **Figure 6** for Spearman similarity (see **Figure S21** for Jaccard). Based
715 on the group-level Cohen's *d* maps, there is a high similarity between the *Small Gain* and *Large*
716 *Gain* versus *Implicit Baseline* (and *Large Gain*) contrasts that appears to be driven by the
717 *Implicit Baseline* condition and high similarity between Cue and Fixation models (see **Figure**
718 **S22**).



719

720 **Figure 5. (A) Jaccard and (B) supra-threshold Spearman Session 1 Between-run similarity**
 721 **estimates across [Four] analytic options for between-run reliability across the ABCD, AHRB and**
 722 **MLS samples.**

723



724
 725 **Figure 6.** The supra-threshold Specification Curve of the *Session 1 Between-run Spearman*
 726 *similarity* estimates across 240 pipeline permutations for the ABCD, AHRB and MLS samples.
 727 A. The distribution of the point estimate (average) and distribution (error bars) across the three samples. B. The
 728 model options (four) associated with each estimate.

729
 730 Similar to Aim 1a (**Table S5; Figure 2**), on average the Session 1 between-run supra-
 731 threshold Spearman similarity is slightly lower than the supra-threshold between-session
 732 Spearman similarity (between-run: ABCD = .68 [range: .35 - .89]; AHRB = .73 [range: .22 -
 733 .96]; MLS = .84 [range: .47 - .96]; between-session: ABCD = .80 [range: .40 - .94]; AHRB = .82
 734 [range: .32 - .97]; MLS = .87 [range: .59 - .97]). A similar trend is observed for the Jaccard
 735 Similarity coefficient. The effects reported in **Figure 5** illustrate that the analytic categories have
 736 unique impacts on the estimated Jaccard and supra-threshold Spearman coefficients. While the
 737 Jaccard coefficient varies most across contrast and model parameterization options (**Figure 5A**),
 738 the Spearman similarity varies most across FWHM and contrast type (**Figure 5B**). The
 739 specification curve for the Spearman similarity coefficients illustrate a near ceiling similarity for
 740 estimates **at the upper tail of the estimates** and little variability across **the three** samples (**Figure**
 741 **6**). The HLM estimates indicate that a change from 3.6 to 8.4 FWHM results in a $b = .08$

742 increase in Jaccard similarity and a $b = .13$ increase in Spearman similarity. Furthermore, the
 743 change from the contrast *Large Gain* versus *Implicit Baseline* to *Large Gain* versus *Neutral*
 744 results in a $b = -.09$ decrease in Jaccard Similarity and a $b = -.20$ decrease in Spearman
 745 similarity. While most parameters are significant in **Table 4**, the effects vary in relative
 746 importance in the model. The variability in the estimated coefficients across 240 pipelines and
 747 three samples is best explained by Contrast (marginal ΔR^2 : .21) and model parameterization
 748 (marginal ΔR^2 : .05) for Jaccard similarity coefficient, and Contrast (marginal ΔR^2 : .66) and
 749 FWHM (marginal ΔR^2 : .08) for supra-threshold Spearman similarity coefficient. Surprisingly,
 750 the motion regressor options had a near-zero impact on the variability on both Jaccard and
 751 Spearman similarity coefficients. Similar to Aim 1a, *post hoc* analyses illustrate an interaction
 752 between Contrasts and Model Parameterization (**Figure S23**), whereby the largest driver of
 753 Model Parameterization differences in the Spearman *rho* similarity is as a function of the
 754 contrasts included the *Implicit Baseline*.

755 **Table 4. Hierarchical Linear Model: (A) Linear associations between the analytic decisions and**
 756 **the Jaccard and Spearman supra-threshold mask Session 1 between-run similarity and (B) the**
 757 **impact of the analytic category on the marginal R².**

A. HLM Group-map Estimates						
	Jaccard			Spearman		
<i>Predictors</i>	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>
(Intercept)	.20	.09 – .31	<.001	.76	.69 – .83	<.001
Reference [3.6]						
fwhm [4.8]	.03	.01 – .05	.004	.05	.04 – .07	<.001
fwhm [6.0]	.05	.03 – .07	<.001	.09	.07 – .10	<.001
fwhm [7.2]	.07	.05 – .09	<.001	.11	.10 – .13	<.001
fwhm [8.4]	.08	.06 – .10	<.001	.13	.12 – .15	<.001
Reference [opt1]						
motion [opt2]	.01	-.00 – .03	.13	.01	-.00 – .03	.05
motion [opt3]	.00	-.02 – .02	.85	.01	-.00 – .02	.20
motion [opt4]	.00	-.01 – .02	.69	.01	-.00 – .03	.08
Reference [AntMod]						
model [CueMod]	.05	.04 – .07	<.001	.02	.01 – .03	<.001

model [FixMod]	.08	.07 – .10	<.001	.01	-.00 – .02	.18
Reference [LgainBase]						
con [LgainNeut]	-.09	-.10 – -.07	<.001	-.20	-.21 – -.18	<.001
con [SgainBase]	-.03	-.05 – -.01	.001	-.01	-.02 – .00	.17
con [SgainNeut]	-.18	-.20 – -.16	<.001	-.34	-.35 – -.32	<.001

B. Analytic Category Model Impact

Comparison	χ^2	Orig R2	New R2	ΔR^2	χ^2	Orig R2	New R2	ΔR^2
[Full] vs [New - fwhm]	78	.30	.26	.04	292	.74	.66	.08
[Full] vs [New - motion]	3	.30	.30	.00	5	.74	.74	.00
[Full] vs [New - model]	104	.30	.25	.05	14	.74	.73	.01
[Full] vs [New - con]	348	.30	.09	.21	1205	.74	.08	.66

758

759 The group-level maps indicate a notable difference in contrasts using the *Neutral* and

760 *Implicit Baseline* conditions (NeuroVault ABCD:

761 <https://identifiers.org/neurovault.collection:17171> AHRB:

762 <https://identifiers.org/neurovault.collection:16605>

763 ; MLS: <https://identifiers.org/neurovault.collection:16606>). As **Figure S22** shows, the *Large*

764 *Gain* versus *Neutral* contrast reflects a qualitatively comparable activation map across Cue,

765 Fixation and Anticipation Models. On the other hand, the *Large Gain* versus *Implicit Baseline*

766 contrast differs across models, where the most notable pattern is that the Cue model is negative

767 of the Fixation model across the samples. Specifically, in ABCD, AHRB and MLS there is

768 increased negative activity in the insular, visual, motor and visual areas, in the Cue Model, and

769 this pattern is mostly **opposite of the** Fixation Model. Meanwhile, in the Anticipation model there

770 is high positive activity in the dorsal striatal, SMA and Insular regions. This reflects the variable

771 meanings of *Implicit Baseline* across the models. The relative symmetry between the Cue and

772 Fixation models is consistent with the fact that each serves as the B_0 in the models, e.g.,

773 $B_{1[Condition\ A,Cue]} - B_{0[All\ Fixation + Probe\ Phase]}$ and $B_{1[Condition\ A,fixation]} -$

774 $B_{0[All\ Cue + Probe\ Phase]}$. The Anticipation model is more variable as it is contrasted with a more

775 narrow phase of the task, e.g., $B_{1[Condition\ A,Cue+Fixation]} - B_{0[Probe\ Phase]}$.

776 Summary of Findings for Aim 1b:

777 Similar to Aim 1a, on average, the supra-threshold Session 1 between-run Spearman and
778 Jaccard similarity is slightly lower between-session similarity. Spearman similarity meaningfully
779 differed across Contrast, Model Parametrization and Smoothing, and it is near the ceiling for the
780 upper tail of the Spearman similarity estimates. Like Aim 1a, Model Parametrization is driven by
781 the Implicit Baseline. Finally, mean-based group activity maps illustrate that the Cue and
782 Fixation models are opposite of each other when the contrast is a between condition and implicit
783 baseline comparison.

784 Aim 2: Effect of analytic decisions on median BS/WS estimates from individual
785 continuous maps

786 Aim 2 proposed to evaluate the changes in the Between-subject variance (BS) and
787 Within-subject variance (WS) components that differentially relate to the ICC(3,1) across the
788 240 workflow permutations. The supra- and sub-threshold distributions across the four model
789 options and three samples are reported in supplemental Figure S24 & S25 and specification
790 curves for BS in supplemental Figure S28 and WS in supplemental Figure S29. The HLM
791 estimates (Table 3) suggest that the Implicit Baseline contrasts increase BS variance and more
792 stringent motion correction decrease BS variance, and Implicit Baseline contrasts and larger
793 smoothing kernels reduce WS variance. The variability in the estimated BS coefficients across
794 240 pipelines and three samples is best explained by Contrast (ΔR^2 : .30), model parameterization
795 (ΔR^2 : .10) and then motion (ΔR^2 : .04). The variability in the estimated WS coefficients across
796 240 pipelines and three samples is best explained by FWHM (ΔR^2 : .21), Contrast (ΔR^2 : .14) and
797 then model parameterization (ΔR^2 : .10). A comparable trend is observed in the between-session
798 estimates (Table S9), with the exception of Contrast selection explaining more variability (ΔR^2 :
799 .26) than FWHM (ΔR^2 : .16). We avoid interpreting the sub-threshold mask as it includes regions
800 that are high-noise (e.g., white matter and ventricles) and drop-out areas (e.g. cerebellar and
801 medial orbital frontal cortex) which exaggerates the BS and WS components.

802 Aim 3: Stability of the ICC, BS and WS Components across Sample Size

803 As expected, based on sampling theory which demonstrates that variability decreases as a
804 function of the square root of N , the variability in estimates decreased as N increased.
805 Specifically, the bootstrapped estimates for the median ICC, BS and WS change slowly at higher
806 intervals of N (Figure 7). In *post hoc* comparisons of whole brain voxelwise ICC maps, the
807 largest variability occurs below $N = 275$. As reported in supplemental Figure S36, at $N = 25$ the
808 minimum and maximum median whole brain ICC maps have a wider voxelwise distribution of
809 ICC values which are notably different (Cohen's $d = 1.9$). With increasing N , Cohen's d of the
810 whole brain voxelwise distributions between the minimum and maximum 3D ICC maps narrows,
811 $d = 1.4$ at $N = 225$ and $d = 1.0$ at $N = 525$, respectively.

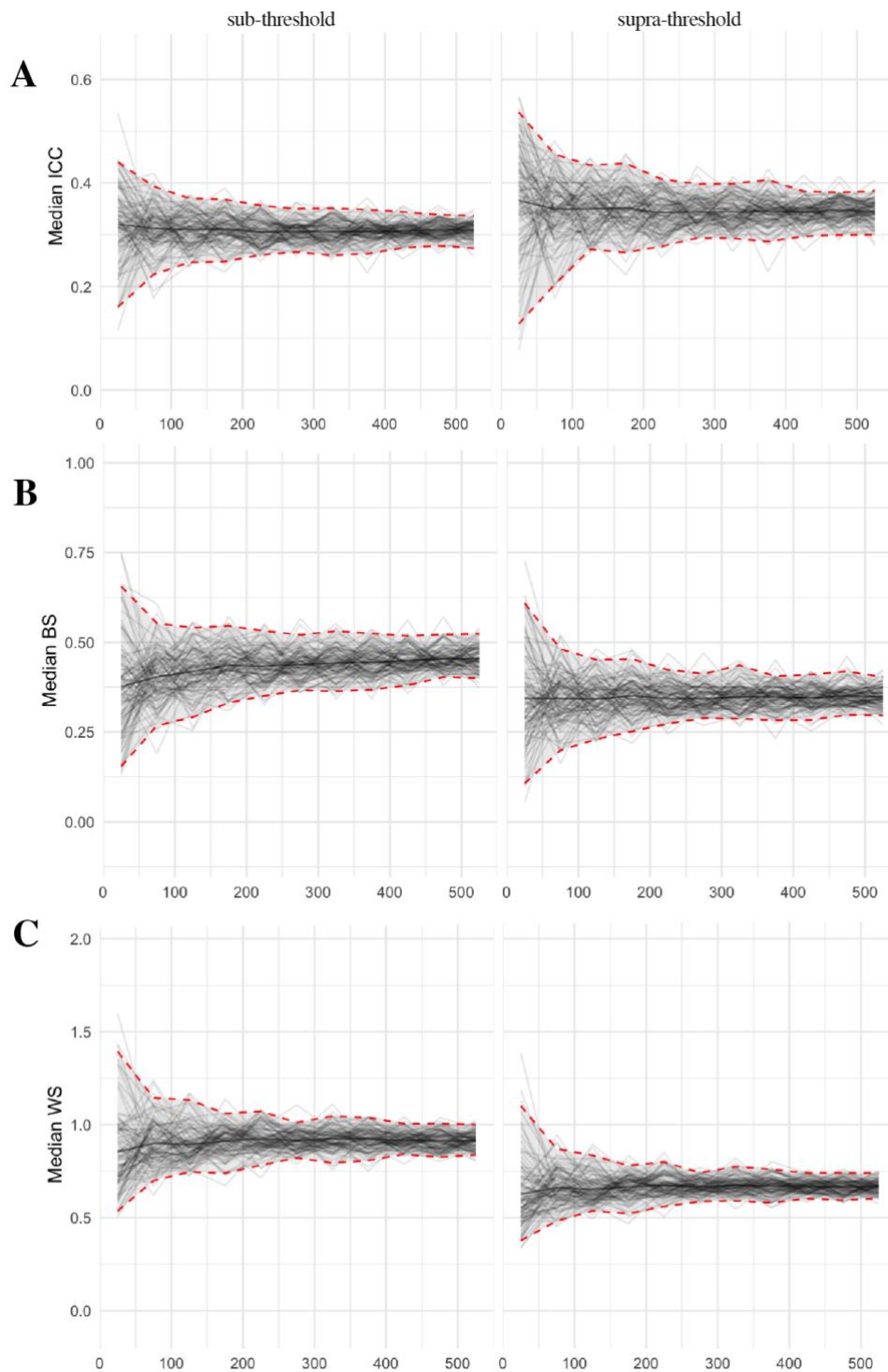


Figure 7. Changes in the Supra- & Supra-threshold Median Intraclass Correlation (ICC), Between-subject variance (BS) and Within-subject variance (WS) estimate in the ABCD sample for N 25 to 525 with 100 bootstraps at each N

Note: Based on the top model from Figure 2: *Small Gain vs Implicit Baseline Contrast*, 'CueMod' Model, Motion option 1 and FWHM 8.4.

835 Post Hoc Analyses

836 An exploratory set of analyses were performed to evaluate 1) the effect of analytic
837 decisions on ICC for the Left and Right Nucleus Accumbens and 2) the association between
838 voxelwise Cohen's d estimates at the group-level and the voxelwise ICC maps. These are
839 reported in supplemental **section 2.6.**

840 Discussion

841 Understanding the analytic decisions that may consistently increase individual- and/or
842 group-level reliability estimates has implications for the study of individual differences using
843 fMRI. The current study expands on previous work by simultaneously evaluating the effects of
844 smoothing, motion correction, task parameterization and contrast selection on the continuous and
845 binary reliability estimates of BOLD activity during the MID task for run- and session-level data
846 across three independent samples. The five major findings are: (1) The ICC(3,1) test-retest
847 reliability estimates in the MID task are consistently low; (2) Group-level estimates of reliability
848 are higher than individual [ICC] estimates; (3) Contrast selection and Model Parameterization
849 have the largest impact on median ICC estimates, and Smoothing and Contrast selection has the
850 largest impact on similarity estimates; however, gains in reliability across different contrasts
851 comes at the cost of interpretability and may differ; (4) Motion correction strategies in these
852 analyses did not meaningfully improve individual or group similarity estimates and, in some
853 cases, *reduced* estimates of reliability; and (5) the median ICC estimate varied across sample size
854 but the variability decreased with increased sample size. Excluding some differences, the results
855 are relatively consistent across the three samples, runs and sessions, providing a comprehensive
856 overview of how analytic decisions at the GLM impact reliability of estimated BOLD in
857 commonly used versions of the MID task.

858 The findings from these multiverse analyses confirm previous reports that ICC estimates
859 are relatively low in univariate task-fMRI and in the current state are inadequate measures for
860 use in individual differences research (Elliott et al., 2020; Kennedy et al., 2022). Consistent with
861 Elliott et al (2020), reliability estimates in the sub-threshold (or non-target mask) are lower than
862 the supra-threshold of the MID task (target mask). The range of median ICCs varied across
863 analytic decisions. Using commonly employed cut-offs (Cicchetti & Sparrow, 1981; Elliott et al.,
864 2020; Noble et al., 2019), ICC estimates for *Large Gain* versus *Neutral* contrast are in the 'Poor'

865 range and the *Large Gain* versus *Implicit Baseline* contrast ranged between ‘Poor’ and ‘Fair’
866 across the three samples. Test-retest reliability for the *Large Gain* (*Small Gain*) versus *Implicit*
867 *Baseline* contrast are modulated by Model Parameterization, whereby the Cue Model had a
868 meaningfully higher reliability than the Anticipation Model. However, this may come at the cost
869 of validity, which is discussed below. Nevertheless, based on voxelwise distributions from the
870 top performing model (Model: Cue Model, Contrast: *Small Gain* versus *Implicit Baseline*,
871 Motion Correction: None, Smoothing: 8.4 mm kernel), visual and motor regions had the highest
872 ICCs, in the ‘Fair’ to ‘Good’ range. *Post hoc* analyses of the bilateral NAc illustrate that, on
873 average, ICC estimates in this region of interest are in the ‘Poor’ range. Notably, ICCs in this
874 *post hoc* region were not meaningfully impacted by Model Parameterization but were impacted
875 by Contrast and Motion correction, suggesting that test-retest reliability may be uniquely
876 impacted by analytic strategy depending on the voxels under consideration. These findings
877 illustrate that the test-retest reliability of the MID task is relatively low, even in the most
878 common ROI such as the Left and Right NAc. While Kennedy et al. (2022, p. 13) speculated that
879 low reliabilities in the ABCD sample may be attributed to the participants’ young age, our results
880 demonstrate that median ICC estimates are *higher* in older than younger samples but reliability
881 estimates in the MID task remain consistently low across early adolescents and late
882 adolescents/young adults. To understand how analytic strategies differentially impact ICCs in
883 different brain regions, we encourage future researchers to use the publicly available estimated
884 maps to probe this question further.

885 Consistent with Fröhner et al. (2019), the group-level maps are not always representative
886 of the individual-level maps across analytic decisions. On average, the Spearman *rho*, Jaccard
887 coefficients and median ICC estimates are higher for the between-session than between-run
888 estimates. Consistently, Spearman *rho* estimates are meaningfully higher for supra-threshold
889 group maps than supra-threshold median ICC estimates derived from individual maps. This
890 suggests that across each of the three samples, the MID task is relatively effective at eliciting a
891 group-level activation map; however, the individual estimates are lower and more variable. In
892 the context of the MID task, the between-run and between-session effects may be the result of
893 within-session effects *decreasing* across runs (Demidenko, Mumford, et al., 2024). Notably, the
894 higher between-session than between-run reliabilities is inconsistent with values reported in
895 previous work (Fröhner et al., 2019), this is likely the result of those between-run estimates being

896 based on randomly split-half (within runs) which are inflated as a result of dependencies in the
897 model estimates within runs (Mumford et al., 2014). Nevertheless, the results here emphasize
898 that group-level maps and group similarity are not a good indicator of individual-level
899 reliabilities. This is unsurprising, considering that the MID task design was optimized to elicit
900 activity in anatomical regions at a group-level and for averaged time-courses within an
901 anatomical region (Knutson et al., 2003).

902 A major question of these analyses was: Are there decisions that *consistently* result in
903 higher individual- (continuous) and/or group-level reliability estimates (continuous/binary)? The
904 results across the analytic choices illustrate that reliability estimates are impacted most by
905 contrast, model parameterization and smoothing decisions. Across the three samples, for
906 between-run and between-session estimates, the contrast type had the largest influence of
907 individual and group reliability estimates. Consistent with previous reports (Baranger et al.,
908 2021; Han et al., 2022; Kennedy et al., 2022; Vetter et al., 2015, 2017), the contrast *Large Gain*
909 (and *Small Gain*) versus *Implicit Baseline* had meaningfully higher estimated ICC, Jaccard and
910 Spearman *rho* similarity estimates than the *Large Gain* versus *Neutral* contrast. The estimated
911 ICC and Spearman *rho* coefficients for contrasts are modulated by the model parameterization,
912 whereby the conditions including the *Implicit Baseline* are highest for the Cue Model
913 parameterization. Conversely, ICC and similarity estimates are relatively stable across the three
914 model parameterizations when comparisons are against the *Neutral* condition. Whether using
915 contrasts or percent signal changes, estimates of BOLD activity suffer from decreases in
916 reliability due to difference scores (Hedge et al., 2018). Where gains are observed from the less
917 reliable *Large Gain* versus *Neutral* to the more reliable *Large Gain* versus *Implicit Baseline*
918 contrast, it comes at the cost of interpretability and face validity that is expected in the estimated
919 BOLD activity. Finally, higher FWHM smoothing kernels positively impacted between-run and
920 between-session median ICC estimates and Spearman *rho* similarity estimates whereas motion
921 correction strategies had a smaller but negative impact on these estimates (i.e., more stringent
922 motion correction reduced reliability estimates). Decisions to smooth in the MID task are
923 especially important given that larger smoothing kernels have been reported to spatially bias
924 reward-related activity in the MID task (Sacchet & Knutson, 2013). In general, variability in
925 reliability estimates decreased with large sample sizes.

926 Improvements in estimated reliability as a function of contrast selection may come at the
927 cost of interpretability. For example, in the context of the *Large Gain* versus *Neutral* contrast,
928 despite differences in the estimated efficiencies the ICC estimates are relatively stable across the
929 model parameterizations in each of the three samples and the activation patterns are interpretable
930 at the group-level. In the context of the *Large Gain* versus *Implicit Baseline* contrast, there are
931 meaningful differences in the ICC estimates across model parameterizations, whereby the Cue
932 and Fixation models demonstrate a substantial improvement over the Anticipation model
933 parameterization, but the group-level activity patterns are less interpretable. As a researcher
934 looking for BOLD estimates that are consistent from run-to-run or session-to-session for
935 individual participants, the *Implicit Baseline* suggests a considerable and valuable improvement
936 on the reliability of estimated values. However, the difference of means for the *Implicit Baseline*
937 is complicated by the intercept in the GLM at the first level. For example, in the Cue Model
938 parameterization, the intercept takes on the average for the unmodeled phase of the task which
939 includes the fixation cross (between cue and probe phase) and the probe response phase. In this
940 instance, isolating the difference of [Cue *Large Gain*] - [Fixation + Probe phase] to a specific
941 cognitive function becomes especially challenging (Poldrack & Yarkoni, 2016; Price & Friston,
942 1997). It is well recognized that different definitions of “baseline”, whether rest, passive or task-
943 related, in task-fMRI will result in different activation patterns (Newman et al., 2001). The use of
944 “neutral” or “fixation” is a cause for caution as it impacts interpretability in various fMRI task
945 designs (Balodis & Potenza, 2015; Filkowski & Haas, 2017). Here, we illustrated how contrasts
946 with the unmodeled phases of a task (*Implicit Baseline*) may improve reliability estimates but
947 may be heavily biased by the activity patterns throughout the task and diminish the validity of
948 the measure. It is reasonable to suspect that subtle modeling deviations between similar and
949 different task designs would further complicate comparisons between studies when using an
950 *Implicit Baseline* condition.

951 In the context of test-retest reliability of estimated BOLD activity, it is important to
952 consider alternative methods to improve reliability, estimation procedures and considerations of
953 what a ‘reliable’ BOLD estimate implies. In general, the evidence here illustrates that the test-
954 retest reliability for the modified version of the MID task is consistently low using the intraclass
955 correlation (ICC[3,1]), even at its maximum. The analytic decisions at the GLM modeling phase
956 demonstrated improvements in reliability from between-run to between-session. Higher between-

957 session reliability may be related to decreasing activity from early to later runs (Demidenko,
958 Mumford, et al., 2024) or based on the sessions being an average of two runs/increased trials
959 (Han et al., 2022; Ooi et al., 2024). In the current analyses, we focused on univariate maps and
960 the parametric, voxelwise ICC estimation procedures (ICC[3,1]). Parametric and non-parametric
961 multivariate methods are reported to improve reliability estimates over univariate estimates using
962 multi-dimensional BOLD data (Gell et al., 2023; Noble et al., 2021). For example, I2C2 is a
963 parametric method that pools variance across images to estimate a global estimate of reliability
964 using a comparable ratio as ICC (Shou et al., 2013) and the discriminability statistic is a non-
965 parametric statistic that is a global index of reliability testing whether the between-subject
966 distance between voxels is greater than the within-subject voxels (Bridgeford et al., 2021). Each
967 of these metrics uniquely summarizes the within- and between-subject variability of the
968 estimated BOLD data and so a consensus and definition of reliability in task-fMRI remains a
969 challenge (Bennett & Miller, 2010). In our analyses we used the ICC as it estimated the
970 reliability for each voxel in an easy-to-interpret coefficient that is useful in common brain-
971 behavior studies. Cut-offs from the self-report literature (Cicchetti & Sparrow, 1981) are often
972 leveraged in fMRI research (Elliott et al., 2020; Noble et al., 2019); however, these cut-offs
973 should depend on the optimal level of precision necessary for the question and reasonable for the
974 methods (Bennett & Miller, 2010; Lance et al., 2006). Some recommendations have been made
975 to use bias-corrections in developmental samples to adjust for suboptimal levels of reliability
976 (Herting et al., 2017), but these corrections should be used cautiously as they do not account for
977 the underlying problems of the measure or the complexities in the data that prevent accurate
978 measurement of the latent process (Nunnally, 1978).

979 Study Considerations

980 The analytic decisions in the current analyses focused primarily on a subset of decisions
981 at the First Level GLM model and its impact on estimates and supra/sub-threshold masks. As a
982 result, other decisions were not considered that may arise at the preprocessing (Li et al., 2021),
983 assumed hemodynamic response function (Kao et al., 2013; Lindquist et al., 2009), cardiac and
984 respiratory correction (Allen et al., 2022; Birn et al., 2006), and the effects of different methods
985 of signal distortion correction (Montez et al., 2023). Furthermore, we focused on voxelwise
986 estimates of reliability which are typically noisier than *a priori* anatomical regions. It is unclear

987 how much interpretation would change if ICC estimates were compared across variable
988 parcellations. Nevertheless, we shared all aggregate maps for the three samples and the
989 preprocessed data for the MLS/AHRB samples to facilitate reanalysis.

990 The results provide a comprehensive overview of individual and group reliability
991 estimates for the modified version of the MID task, but it is challenging to infer how reflective
992 these results are of alternate MID designs and different reward tasks. Based on prior reports of
993 low test-retest reliabilities in task fMR, if a sufficient sample size is used, we suspect that results
994 may be comparable to other MID and reward task designs. Future research should consider how
995 reliability estimates change as a function of modeling decisions in different task paradigms.

996 Conclusion

997 With the increasing interest in test-retest reliability in task fMRI and methods for
998 improving reliability estimates of BOLD, the current study evaluated which decisions at the
999 GLM model improved group and individual reliability estimates of reliability. In general, the
1000 findings illustrate that the MID task group activation maps are more reliable than individual
1001 maps across testing occasions and independent samples. Across group and individual models,
1002 between-session estimates are consistently higher than between-run estimates of reliability.
1003 Furthermore, estimates of reliability were more variable at the median fMRI sample size and
1004 stabilized with N . While individual estimates of reliability are low (ICC[3,1]), contrasts and
1005 model parameterization meaningfully improved test-retest reliability. However, the improvement
1006 in reliability came at the cost of interpretability and may be region specific in the current version
1007 of the MID task. This underscores the importance of evaluating reliability in larger samples sizes
1008 and ensuring improved estimates reflect the neural processes of interest. While Model
1009 Parameterization and Contrast selection had the largest impact on voxelwise ICCs, further work
1010 is needed to expand on these findings by evaluating alternative brain regions and analytic
1011 decisions that may result in improved test-retest reliability that may be meaningful in individual
1012 differences research.

1013 **Data & Code Availability Statement**

1014 *Adolescent Brain Cognitive Development* (ABCD) data: The ABCD BIDS data, MRIQC v23.1.0
1015 and fMRIPrep v23.1.4 derivatives can be accessed through the ABCD-BIDS Community
1016 Collection (ABCC) with an established Data Use Agreement (see <https://abcdstudy.org/>). The
1017 data used in these analyses will be available at a future release onto the National Institute of
1018 Mental Health Data Archive. The complete set of group-level and ICC maps are publicly
1019 available on Neurovault for ABCD (6180 images;
1020 <https://identifiers.org/neurovault.collection:17171>).

1021 *Michigan Longitudinal Study* (MLS) and *Adolescent Health Risk Behavior* (AHRB) data: The
1022 BIDS inputs, fMRIPrep v23.1.4 and MRIQC v23.1.0 derivatives are available on OpenNeuro.org
1023 (MLS: <https://doi.org/10.18112/openneuro.ds005027.v1.0.1> AHRB:
1024 <https://doi.org/10.18112/openneuro.ds005012.v1.0.1>). The complete set of group-level and ICC
1025 maps are publicly available on Neurovault for MLS (2400 images;
1026 <https://identifiers.org/neurovault.collection:16606>) and AHRB (2400 images;
1027 <https://identifiers.org/neurovault.collection:16605>)

1028 *R and Python code*: The *.html* and *.rmd* file containing the code to be run on extracted estimates
1029 from reliability maps are available on Github with the associated output files containing the
1030 estimates across the models and samples. Likewise, all of the code for first level, fixed effect,
1031 group and ICC models are available online at
1032 https://github.com/demidenm/Multiverse_Reliability.

1033 **Acknowledgements**

1034 MID is funded by the Ruth L. Kirschstein Postdoctoral Individual National Research Service
1035 Award through the National Institute on Drug Abuse (F32 DA055334-01A1). RAP is supported
1036 by the National Institute of Mental Health (R01MH117772 and R01MH130898). Thanks to Dr.
1037 Daniel Keating for agreeing to share the Adolescent Health Risk Behavior (AHRB;
1038 R01HD075806) study data and to Dr. Mary Heitzeg for agreeing to share the Michigan
1039 Longitudinal Study (MLS; R01HD075806) data for this project. The authors would also like to
1040 thank the research participants and staff involved in data collection of the Adolescent Brain
1041 Cognitive Development (ABCD) Study data. The ABCD Study is a multisite, longitudinal study
1042 designed to recruit more than 10,000 children ages 9 and 10 and follow them over 10 years into
1043 early adulthood. The ABCD Study is supported by the National Institutes of Health (NIH) and
1044 additional federal partners under award numbers U01DA041048, U01DA050989,
1045 U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987,
1046 U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134,
1047 U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120,
1048 U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, and
1049 U24DA041147. The list of supporters is available at <https://abcdstudy.org/federal-partners.html>.
1050 The list of participating sites and study investigators is available at [https://abcdstudy.org/study-](https://abcdstudy.org/study-sites/)
1051 [sites/](https://abcdstudy.org/study-sites/).

1052 Thanks to members of the Cognitive Development and Neuroimaging Lab (CDNI) at the
1053 University of Minnesota, specifically Eric Feczko, rae McCollum and Audrey Houghton, for
1054 assisting and providing access to the ABCD-BIDS Community Collection (ABCC) data. The
1055 analyses here are based on data available from CDNI as of February 2024. The ABCC data
1056 repository grows and changes over time (<https://collection3165.readthedocs.io/>). Thanks to
1057 Krisanne Litinas at the University of Michigan for providing expert advice and scripts to convert
1058 the AHRB data into BIDS format. Thanks to Mary Soules and Ryan Klaus (with assistance from
1059 Krisanne Litinas) at the University of Michigan for working to convert the MLS data to BIDS
1060 format.

1061 **Author's Contribution**

1062 MID obtained data sharing agreements. MID conceptualized the study with critical input from
1063 RAP. MID defined the methodology with critical input from RAP and JAM. MID curated the

1064 analytic code and performed the formal analysis and interpretation with input from RAP and
1065 JAM. MID wrote the original draft and curated the visualizations. RAP and JAM reviewed,
1066 edited, and provided critical feedback on the draft and all revisions.

1067 **Conflicts of Interest**

1068 The authors declare that they have no conflicts of interest.

1069

References

- 1070
- 1071 Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A.,
1072 Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-
1073 learn. *Frontiers in Neuroinformatics*, 8.
1074 <https://www.frontiersin.org/articles/10.3389/fninf.2014.00014>
- 1075 Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. A. (2019). Raincloud plots:
1076 A multi-platform tool for robust data visualization. *Wellcome Open Research*, 4, 63.
1077 <https://doi.org/10.12688/wellcomeopenres.15191.1>
- 1078 Allen, M., Varga, S., & Heck, D. H. (2022). Respiratory rhythms of the predictive mind.
1079 *Psychological Review*, No Pagination Specified-No Pagination Specified.
1080 <https://doi.org/10.1037/rev0000391>
- 1081 Balodis, I. M., & Potenza, M. N. (2015). Anticipatory reward processing in addicted populations:
1082 A focus on the monetary incentive delay task. *Biological Psychiatry*, 77(5), 434–444.
1083 <https://doi.org/10.1016/j.biopsych.2014.08.020>
- 1084 Baranger, D. A. A., Lindenmuth, M., Nance, M., Guyer, A. E., Keenan, K., Hipwell, A. E.,
1085 Shaw, D. S., & Forbes, E. E. (2021). The longitudinal stability of fMRI activation during
1086 reward processing in adolescents and young adults. *NeuroImage*, 232, 117872.
1087 <https://doi.org/10.1016/j.neuroimage.2021.117872>
- 1088 Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability.
1089 *Psychological Reports*, 19(1), 3–11. <https://doi.org/10.2466/pr0.1966.19.1.3>
- 1090 Bates, D., Maechler, M., Bolker, B., cre, Walker, S., Christensen, R. H. B., Singmann, H., Dai,
1091 B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., & Krivitsky, P. N. (2020).

1092 *lme4: Linear mixed-effects models using “Eigen” and S4* (1.1-26) [Computer software].
1093 <https://CRAN.R-project.org/package=lme4>

1094 Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic
1095 resonance imaging? *Annals of the New York Academy of Sciences*, *1191*(1), 133–155.
1096 <https://doi.org/10.1111/j.1749-6632.2010.05446.x>

1097 Bennett, C. M., & Miller, M. B. (2013). fMRI reliability: Influences of task and experimental
1098 design. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(4), 690–702.
1099 <https://doi.org/10.3758/s13415-013-0195-1>

1100 Birn, R. M., Diamond, J. B., Smith, M. A., & Bandettini, P. A. (2006). Separating respiratory-
1101 variation-related fluctuations from neuronal-activity-related fluctuations in fMRI.
1102 *NeuroImage*, *31*(4), 1536–1548. <https://doi.org/10.1016/j.neuroimage.2006.02.048>

1103 Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with
1104 desired precision. *Statistics in Medicine*, *21*(9), 1331–1335.
1105 <https://doi.org/10.1002/sim.1108>

1106 Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M.,
1107 Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M.,
1108 Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., ...
1109 Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by
1110 many teams. *Nature*, *582*(7810), 84–88. <https://doi.org/10.1038/s41586-020-2314-9>

1111 Bowring, A., Nichols, T. E., & Maumet, C. (2022). Isolating the sources of pipeline-variability in
1112 group-level task-fMRI results. *Human Brain Mapping*, *43*(3), 1112–1128.
1113 <https://doi.org/10.1002/hbm.25713>

1114 Bridgeford, E. W., Wang, S., Wang, Z., Xu, T., Craddock, C., Dey, J., Kiar, G., Gray-Roncal,
1115 W., Colantuoni, C., Douville, C., Noble, S., Priebe, C. E., Caffo, B., Milham, M., Zuo,
1116 X.-N., Reproducibility, C. for R. and, & Vogelstein, J. T. (2021). Eliminating accidental
1117 deviations to minimize generalization error and maximize replicability: Applications in
1118 connectomics and genomics. *PLOS Computational Biology*, *17*(9), e1009279.
1119 <https://doi.org/10.1371/journal.pcbi.1009279>

1120 Caballero-Gaudes, C., & Reynolds, R. C. (2017). Methods for cleaning the BOLD fMRI signal.
1121 *NeuroImage*, *154*, 128–149. <https://doi.org/10.1016/j.neuroimage.2016.12.018>

1122 Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C. R., & Mehta, M. A. (2009). Measuring
1123 fMRI reliability with the intra-class correlation coefficient. *NeuroImage*, *45*(3), 758–768.
1124 <https://doi.org/10.1016/j.neuroimage.2008.12.035>

1125 Carp, J. (2012). On the plurality of (methodological) worlds: estimating the analytic flexibility of
1126 fMRI experiments. *Frontiers in Neuroscience*, *6*.
1127 <https://doi.org/10.3389/fnins.2012.00149>

1128 Chen, G., Taylor, P. A., Haller, S. P., Kircanski, K., Stoddard, J., Pine, D. S., Leibenluft, E.,
1129 Brotman, M. A., & Cox, R. W. (2017). Intraclass correlation: Improved modeling
1130 approaches and applications for neuroimaging. *Human Brain Mapping*, *39*(3), 1187–
1131 1206. <https://doi.org/10.1002/hbm.23909>

1132 Churchill, N. W., Spring, R., Afshin-Pour, B., Dong, F., & Strother, S. C. (2015). An automated,
1133 adaptive framework for optimizing preprocessing pipelines in task-based functional MRI.
1134 *PLOS ONE*, *10*(7), e0131520. <https://doi.org/10.1371/journal.pone.0131520>

- 1135 Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater
1136 reliability of specific items: Applications to assessment of adaptive behavior. *American*
1137 *Journal of Mental Deficiency*, 86, 127–137.
- 1138 Cohen, M. S., & DuBois, R. M. (1999). Stability, repeatability, and the expression of signal
1139 magnitude in functional magnetic resonance imaging. *Journal of Magnetic Resonance*
1140 *Imaging: JMRI*, 10(1), 33–40. [https://doi.org/10.1002/\(sici\)1522-](https://doi.org/10.1002/(sici)1522-2586(199907)10:1<33::aid-jmri5>3.0.co;2-n)
1141 [2586\(199907\)10:1<33::aid-jmri5>3.0.co;2-n](https://doi.org/10.1002/(sici)1522-2586(199907)10:1<33::aid-jmri5>3.0.co;2-n)
- 1142 Demidenko, M. I., Huntley, E. D., & Keating, D. P. (2024). *Adolescent Health Risk Behavior*
1143 *Study*. (ds005012; 1.0.1) [dataset]. OpenNeuro.
1144 <https://doi.org/www.doi.org/10.18112/openneuro.ds005012.v1.0.1>
- 1145 Demidenko, M. I., Klaus, R., Soules, M., & Heitzeg, M. M. (2024). *Michigan Longitudinal*
1146 *Study*. (ds005027; 1.0.1) [dataset]. OpenNeuro.
1147 <https://doi.org/www.doi.org/10.18112/openneuro.ds005027.v1.0.1>
- 1148 Demidenko, M. I., Mumford, J. A., Ram, N., & Poldrack, R. A. (2024). A multi-sample
1149 evaluation of the measurement structure and function of the modified monetary incentive
1150 delay task in adolescents. *Developmental Cognitive Neuroscience*, 65, 101337.
1151 <https://doi.org/10.1016/j.dcn.2023.101337>
- 1152 Demidenko, M., Mumford, J. & Poldrack, R. (2024). *PyReliMRI: An open-source python tool for*
1153 *estimates of reliability in MRI data* (2.1.0) [Computer software].
1154 <https://zenodo.org/record/8387971>
- 1155 Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI.
1156 *Trends in Cognitive Sciences*, 20(6), 425–443. <https://doi.org/10.1016/j.tics.2016.03.014>

1157 Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L.,
1158 Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What is the test-retest reliability of
1159 common task-functional MRI measures? New empirical evidence and a meta-analysis.
1160 *Psychological Science*, *31*(7), 792–806. <https://doi.org/10.1177/0956797620916786>

1161 Esteban, O., Baratz, Z., Markiewicz, C. J., MacNicol, E., Provins, C., & Hagen, M. P. (2023).
1162 *MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites*
1163 [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.8034748>

1164 Esteban, O., Markiewicz, C. J., Burns, C., Goncalves, M., Jarecka, D., Ziegler, E., Berleant, S.,
1165 Ellis, D. G., Pinsard, B., Madison, C., Waskom, M., Notter, M. P., Clark, D., Manhães-
1166 Savio, A., Clark, D., Jordan, K., Dayan, M., Halchenko, Y. O., Loney, F., ... Ghosh, S.
1167 (2022). *nipy/nipype: 1.8.3* [Computer software]. Zenodo.
1168 <https://doi.org/10.5281/zenodo.6834519>

1169 Esteban, O., Markiewicz, C. J., Goncalves, M., Provins, C., Kent, J. D., DuPre, E., Salo, T.,
1170 Ciric, R., Pinsard, B., Blair, R. W., Poldrack, R. A., & Gorgolewski, K. J. (2022).
1171 *fMRIPrep: A robust preprocessing pipeline for functional MRI* [Computer software].
1172 Zenodo. <https://doi.org/10.5281/zenodo.7117719>

1173 Feczko, E., Conan, G., Marek, S., Tervo-Clemmens, B., Cordova, M., Doyle, O., Earl, E.,
1174 Perrone, A., Sturgeon, D., Klein, R., Harman, G., Kilamovich, D., Hermosillo, R.,
1175 Miranda-Dominguez, O., Adebimpe, A., Bertolero, M., Cieslak, M., Covitz, S.,
1176 Hendrickson, T., ... Fair, D. A. (2021). *Adolescent Brain Cognitive Development*
1177 *(ABCD) community MRI collection and utilities* (p. 2021.07.09.451638). bioRxiv.
1178 <https://doi.org/10.1101/2021.07.09.451638>

1179 Filkowski, M. M., & Haas, B. W. (2017). Rethinking the use of neutral faces as a baseline in
1180 fMRI neuroimaging studies of Axis-I psychiatric disorders. *Journal of Neuroimaging*,
1181 27(3), 281–291. <https://doi.org/10.1111/jon.12403>

1182 Fisher, R. A. (1934). Statistical methods for research workers. In F. A. E. Crew & D. W. Cutler
1183 (Eds.), *Statistical methods for research workers* (5th ed., rev). Oliver and Boyd.

1184 Fröhner, J. H., Teckentrup, V., Smolka, M. N., & Kroemer, N. B. (2019). Addressing the
1185 reliability fallacy in fMRI: Similar group effects may arise from unreliable individual
1186 effects. *NeuroImage*, 195, 174–189. <https://doi.org/10.1016/j.neuroimage.2019.03.053>

1187 Gell, M., Eickhoff, S. B., Omidvarnia, A., Küppers, V., Patil, K. R., Satterthwaite, T. D., Müller,
1188 V. I., & Langner, R. (2023). *The burden of reliability: How measurement noise limits*
1189 *brain-behaviour predictions* (p. 2023.02.09.527898). bioRxiv.
1190 <https://doi.org/10.1101/2023.02.09.527898>

1191 Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis—"A"
1192 garden of forking paths"—Explains why many statistically significant comparisons don't
1193 hold up. *American Scientist*, 102(6), 460–466.

1194 Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not
1195 itself statistically significant. *The American Statistician*, 60(4), 328–331.
1196 <https://doi.org/10.1198/000313006X152649>

1197 Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility
1198 mean? *Science Translational Medicine*, 8(341), 341ps12-341ps12.
1199 <https://doi.org/10.1126/scitranslmed.aaf5027>

1200 Gorgolewski, K. J., Burns, C., Madison, C., Clark, D., Halchenko, Y., Waskom, M., & Ghosh, S.
1201 (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing

1202 framework in Python. *Frontiers in Neuroinformatics*, 5.
1203 <https://www.frontiersin.org/articles/10.3389/fninf.2011.00013>

1204 Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., Whittle, I., & Pernet, C. (2013). Single subject
1205 fMRI test–retest reliability metrics and confounding factors. *NeuroImage*, 69, 231–243.
1206 <https://doi.org/10.1016/j.neuroimage.2012.10.085>

1207 Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., Sochat,
1208 V. V., Nichols, T. E., Poldrack, R. A., Poline, J.-B., Yarkoni, T., & Margulies, D. S.
1209 (2015). NeuroVault.org: A web-based repository for collecting and sharing unthresholded
1210 statistical maps of the human brain. *Frontiers in Neuroinformatics*, 9.
1211 <https://www.frontiersin.org/articles/10.3389/fninf.2015.00008>

1212 Grady, C. L., Rieck, J. R., Nichol, D., Rodrigue, K. M., & Kennedy, K. M. (2020). Influence of
1213 sample size and analytic approach on stability and interpretation of brain-behavior
1214 correlations in task-related fMRI data. *Human Brain Mapping*.
1215 <https://doi.org/10.1002/hbm.25217>

1216 Han, X., Ashar, Y. K., Kragel, P., Petre, B., Schelkun, V., Atlas, L. Y., Chang, L. J., Jepma, M.,
1217 Koban, L., Losin, E. A. R., Roy, M., Woo, C.-W., & Wager, T. D. (2022). Effect sizes
1218 and test-retest reliability of the fMRI-based neurologic pain signature. *NeuroImage*, 247,
1219 118844. <https://doi.org/10.1016/j.neuroimage.2021.118844>

1220 Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks
1221 do not produce reliable individual differences. *Behavior Research Methods*, 50(3), 1166–
1222 1186. <https://doi.org/10.3758/s13428-017-0935-1>

1223 Herting, M. M., Gautam, P., Chen, Z., Mezher, A., & Vetter, N. C. (2017). Test-retest reliability
1224 of longitudinal task-based fMRI: Implications for developmental studies. *Developmental*
1225 *Cognitive Neuroscience*, 33, 17–26. <https://doi.org/10.1016/j.dcn.2017.07.001>

1226 Kao, M.-H., Majumdar, D., Mandal, A., & Stufken, J. (2013). Maximin and maximin-efficient
1227 event-related fMRI designs under a nonlinear model. *The Annals of Applied Statistics*,
1228 7(4), 1940–1959. <https://doi.org/10.1214/13-AOAS658>

1229 Kennedy, J. T., Harms, M. P., Korucuoglu, O., Astafiev, S. V., Barch, D. M., Thompson, W. K.,
1230 Bjork, J. M., & Anokhin, A. P. (2022). Reliability and stability challenges in ABCD task
1231 fMRI data. *NeuroImage*, 252, 119046. <https://doi.org/10.1016/j.neuroimage.2022.119046>

1232 Knutson, B., Fong, G. W., Bennett, S. M., Adams, C. M., & Hommer, D. (2003). A region of
1233 mesial prefrontal cortex tracks monetarily rewarding outcomes: Characterization with
1234 rapid event-related fMRI. *NeuroImage*, 18(2), 263–272. [https://doi.org/10.1016/S1053-](https://doi.org/10.1016/S1053-8119(02)00057-5)
1235 [8119\(02\)00057-5](https://doi.org/10.1016/S1053-8119(02)00057-5)

1236 Knutson, B., Westdorp, A., Kaiser, E., & Hommer, D. (2000). FMRI visualization of brain
1237 activity during a monetary incentive delay task. *NeuroImage*, 12(1), 20–27.
1238 <https://doi.org/10.1006/nimg.2000.0593>

1239 Kragel, P. A., Han, X., Kraynak, T. E., Gianaros, P. J., & Wager, T. D. (2021). Functional MRI
1240 can be highly reliable, but it depends on what you measure: A Commentary on Elliott et
1241 al. (2020). *Psychological Science*, 0956797621989730.
1242 <https://doi.org/10.1177/0956797621989730>

1243 Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported
1244 cutoff criteria: What did they really say? *Organizational Research Methods*, 9(2), 202–
1245 220. <https://doi.org/10.1177/1094428105284919>

1246 Lenth, R. V., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl,
1247 H., & Singmann, H. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares*
1248 *Means* (1.8.4-1) [Computer software]. <https://CRAN.R-project.org/package=emmeans>

1249 Li, X., Ai, L., Giavasis, S., Jin, H., Feczko, E., Xu, T., Clucas, J., Franco, A., Heinsfeld, A. S.,
1250 Adebimpe, A., Vogelstein, J. T., Yan, C.-G., Esteban, O., Poldrack, R. A., Craddock, C.,
1251 Fair, D., Satterthwaite, T., Kiar, G., & Milham, M. P. (2021). *Moving beyond processing*
1252 *and analysis-related variation in neuroscience* (p. 2021.12.01.470790).
1253 <https://doi.org/10.1101/2021.12.01.470790>

1254 Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation—A discussion
1255 and demonstration of basic features. *PloS One*, *14*(7), e0219854.
1256 <https://doi.org/10.1371/journal.pone.0219854>

1257 Lindquist, M. A., Meng Loh, J., Atlas, L. Y., & Wager, T. D. (2009). Modeling the
1258 hemodynamic response function in fMRI: Efficiency, bias and mis-modeling.
1259 *NeuroImage*, *45*(1, Supplement 1), S187–S198.
1260 <https://doi.org/10.1016/j.neuroimage.2008.10.065>

1261 Liu, J., Tang, W., Chen, G., Lu, Y., Feng, C., & Tu, X. M. (2016). Correlation and agreement:
1262 Overview and clarification of competing concepts and measures. *Shanghai Archives of*
1263 *Psychiatry*, *28*(2), 115–120. <https://doi.org/10.11919/j.issn.1002-0829.216045>

1264 Liu, S., Abdellaoui, A., Verweij, K. J. H., & van Wingen, G. A. (2023). Replicable brain–
1265 phenotype associations require large-scale neuroimaging data. *Nature Human Behaviour*,
1266 1–13. <https://doi.org/10.1038/s41562-023-01642-5>

1267 Maitra, R. (2010). A re-defined and generalized percent-overlap-of-activation measure for
1268 studies of fMRI reproducibility and its use in identifying outlier activation maps.
1269 *NeuroImage*, 50(1), 124–135. <https://doi.org/10.1016/j.neuroimage.2009.11.070>

1270 Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S.,
1271 Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala,
1272 S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J.,
1273 Cordova, M., Doyle, O., ... Dosenbach, N. U. F. (2022). Reproducible brain-wide
1274 association studies require thousands of individuals. *Nature*, 1–7.
1275 <https://doi.org/10.1038/s41586-022-04492-9>

1276 Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E.,
1277 Hardcastle, N., Wexler, J., Esteban, O., Goncavles, M., Jwa, A., & Poldrack, R. (2021).
1278 The OpenNeuro resource for sharing of neuroscience data. *eLife*, 10, e71774.
1279 <https://doi.org/10.7554/eLife.71774>

1280 Martz, M. E., Trucco, E. M., Cope, L. M., Hardee, J. E., Jester, J. M., Zucker, R. A., & Heitzeg,
1281 M. M. (2016). Association of marijuana use with blunted nucleus accumbens response to
1282 reward anticipation. *JAMA Psychiatry*, 73(8), 838–844.
1283 <https://doi.org/10.1001/jamapsychiatry.2016.1161>

1284 Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error
1285 and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315.
1286 <https://doi.org/10.1016/j.jml.2017.01.001>

1287 McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation
1288 coefficients. *Psychological Methods*, 1, 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>

1289 Montez, D. F., Van, A. N., Miller, R. L., Seider, N. A., Marek, S., Zheng, A., Newbold, D. J.,
1290 Scheidter, K., Feczko, E., Perrone, A. J., Miranda-Dominguez, O., Earl, E. A., Kay, B. P.,
1291 Jha, A. K., Sotiras, A., Laumann, T. O., Greene, D. J., Gordon, E. M., Tisdall, M. D., ...
1292 Dosenbach, N. U. F. (2023). Using synthetic MR images for distortion correction.
1293 *Developmental Cognitive Neuroscience, 60*, 101234.
1294 <https://doi.org/10.1016/j.dcn.2023.101234>

1295 Mumford, J. A., Davis, T., & Poldrack, R. A. (2014). The impact of study design on pattern
1296 estimation for single-trial multivariate pattern analysis. *NeuroImage, 103*, 130–138.
1297 <https://doi.org/10.1016/j.neuroimage.2014.09.026>

1298 Newman, S. D., Twieg, D. B., & Carpenter, P. A. (2001). Baseline conditions and subtractive
1299 logic in neuroimaging. *Human Brain Mapping, 14*(4), 228–235.
1300 <https://doi.org/10.1002/hbm.1055>

1301 Nikolaidis, A., Chen, A. A., He, X., Shinohara, R., Vogelstein, J., Milham, M., & Shou, H.
1302 (2022). *Suboptimal phenotypic reliability impedes reproducible human neuroscience* (p.
1303 2022.07.22.501193). bioRxiv. <https://doi.org/10.1101/2022.07.22.501193>

1304 Noble, S., Scheinost, D., & Constable, R. T. (2019). A decade of test-retest reliability of
1305 functional connectivity: A systematic review and meta-analysis. *NeuroImage, 203*,
1306 116157. <https://doi.org/10.1016/j.neuroimage.2019.116157>

1307 Noble, S., Scheinost, D., & Constable, R. T. (2021). A guide to the measurement and
1308 interpretation of fMRI test-retest reliability. *Current Opinion in Behavioral Sciences, 40*,
1309 27–32. <https://doi.org/10.1016/j.cobeha.2020.12.012>

1310 Nunnally, J. C. (1978). An Overview of Psychological Measurement. In B. B. Wolman (Ed.),
1311 *Clinical diagnosis of mental disorders: A handbook* (pp. 97–146). Springer US.
1312 https://doi.org/10.1007/978-1-4684-2490-4_4

1313 Ooi, L. Q. R., Orban, C., Nichols, T. E., Zhang, S., Tan, T. W. K., Kong, R., Marek, S.,
1314 Dosenbach, N. U. F., Laumann, T., Gordon, E. M., Zhou, J. H., Bzdok, D., Eickhoff, S.
1315 B., Holmes, A. J., & Yeo, B. T. T. (2024). *MRI economics: Balancing sample size and*
1316 *scan duration in brain wide association studies* (p. 2024.02.16.580448). bioRxiv.
1317 <https://doi.org/10.1101/2024.02.16.580448>

1318 Plichta, M. M., Schwarz, A. J., Grimm, O., Morgen, K., Mier, D., Haddad, L., Gerdes, A. B. M.,
1319 Sauer, C., Tost, H., Esslinger, C., Colman, P., Wilson, F., Kirsch, P., & Meyer-
1320 Lindenberg, A. (2012). Test–retest reliability of evoked BOLD signals from a cognitive–
1321 emotive fMRI test battery. *NeuroImage*, *60*(3), 1746–1758.
1322 <https://doi.org/10.1016/j.neuroimage.2012.01.129>

1323 Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R.,
1324 Nichols, T. E., Poline, J.-B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon:
1325 Towards transparent and reproducible neuroimaging research. *Nature Reviews*
1326 *Neuroscience*, *18*(2), Article 2. <https://doi.org/10.1038/nrn.2016.167>

1327 Poldrack, R. A., & Yarkoni, T. (2016). From brain maps to cognitive ontologies: informatics and
1328 the search for mental structure. *Annual Review of Psychology*, *67*(1), 587–612.
1329 <https://doi.org/10.1146/annurev-psych-122414-033729>

1330 Price, C. J., & Friston, K. J. (1997). Cognitive conjunction: A new approach to brain activation
1331 experiments. *NeuroImage*, *5*(4 Pt 1), 261–270. <https://doi.org/10.1006/nimg.1997.0269>

- 1332 Rombouts, S. A., Barkhof, F., Hoogenraad, F. G., Sprenger, M., & Scheltens, P. (1998). Within-
1333 subject reproducibility of visual activation patterns with functional magnetic resonance
1334 imaging using multislice echo planar imaging. *Magnetic Resonance Imaging, 16*(2), 105–
1335 113. [https://doi.org/10.1016/s0730-725x\(97\)00253-1](https://doi.org/10.1016/s0730-725x(97)00253-1)
- 1336 Sacchet, M. D., & Knutson, B. (2013). Spatial smoothing systematically biases the localization
1337 of reward-related brain activity. *NeuroImage, 66*, 270–277.
1338 <https://doi.org/10.1016/j.neuroimage.2012.10.056>
- 1339 Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal*
1340 *of Research in Personality, 47*(5), 609–612. <https://doi.org/10.1016/j.jrp.2013.05.009>
- 1341 Shou, H., Eloyan, A., Lee, S., Zipunnikov, V., Crainiceanu, A. N., Nebel, M. B., Caffo, B.,
1342 Lindquist, M. A., & Crainiceanu, C. M. (2013). Quantifying the reliability of image
1343 replication studies: The image intraclass correlation coefficient (I2C2). *Cognitive,*
1344 *Affective, & Behavioral Neuroscience, 13*(4), 714–724. [https://doi.org/10.3758/s13415-](https://doi.org/10.3758/s13415-013-0196-0)
1345 [013-0196-0](https://doi.org/10.3758/s13415-013-0196-0)
- 1346 Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability.
1347 *Psychological Bulletin, 86*(2), 420–428. <https://doi.org/10.1037//0033-2909.86.2.420>
- 1348 Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed
1349 flexibility in data collection and analysis allows presenting anything as significant.
1350 *Psychological Science, 22*(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- 1351 Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature*
1352 *Human Behaviour, 1*–7. <https://doi.org/10.1038/s41562-020-0912-z>
- 1353 Soares, J. M., Magalhães, R., Moreira, P. S., Sousa, A., Ganz, E., Sampaio, A., Alves, V.,
1354 Marques, P., & Sousa, N. (2016). A Hitchhiker’s Guide to Functional Magnetic

1355 Resonance Imaging. *Frontiers in Neuroscience*, *10*, 515.
1356 <https://doi.org/10.3389/fnins.2016.00515>

1357 Spearman, C. (1904). The proof and measurement of association between two things. *The*
1358 *American Journal of Psychology*, *15*(1), 72–101. <https://doi.org/10.2307/1412159>

1359 Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency
1360 through a multiverse analysis: *Perspectives on Psychological Science*.
1361 <https://doi.org/10.1177/1745691616658637>

1362 Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power
1363 in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3),
1364 e2000797. <https://doi.org/10.1371/journal.pbio.2000797>

1365 Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the
1366 replicability of task-based fMRI studies. *Communications Biology*, *1*(1), Article 1.
1367 <https://doi.org/10.1038/s42003-018-0073-z>

1368 Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software*, *3*(31), 1026.
1369 <https://doi.org/10.21105/joss.01026>

1370 Vetter, N. C., Pilhatsch, M., Weigelt, S., Ripke, S., & Smolka, M. N. (2015). Mid-adolescent
1371 neurocognitive development of ignoring and attending emotional stimuli. *Developmental*
1372 *Cognitive Neuroscience*, *14*, 23–31. <https://doi.org/10.1016/j.dcn.2015.05.001>

1373 Vetter, N. C., Steding, J., Jurk, S., Ripke, S., Mennigen, E., & Smolka, M. N. (2017). Reliability
1374 in adolescent fMRI within two years – a comparison of three tasks. *Scientific Reports*,
1375 *7*(1), Article 1. <https://doi.org/10.1038/s41598-017-02334-7>

1376 Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D.,
1377 Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M.,

1378 Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E.,
1379 ... van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing
1380 in Python. *Nature Methods*, 17(3), Article 3. <https://doi.org/10.1038/s41592-019-0686-2>

1381 Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J.,
1382 Pérez-Stable, E. J., Riley, W. T., Bloch, M. H., Conway, K., Deeds, B. G., Dowling, G.
1383 J., Grant, S., Howlett, K. D., Matochik, J. A., Morgan, G. D., Murray, M. M., Noronha,
1384 A., Spong, C. Y., ... Weiss, S. R. B. (2018). The conception of the ABCD study: From
1385 substance use to a broad NIH collaboration. *Developmental Cognitive Neuroscience*, 32,
1386 4–7. <https://doi.org/10.1016/j.dcn.2017.10.002>

1387 Wilson, R. P., Colizzi, M., Bossong, M. G., Allen, P., Kempton, M., Abe, N., Barros-
1388 Loscertales, A. R., Bayer, J., Beck, A., Bjork, J., Boecker, R., Bustamante, J. C., Choi, J.
1389 S., Delmonte, S., Dillon, D., Figuee, M., Garavan, H., Hagele, C., Hermans, E. J., ...
1390 MTAC. (2018). The neural substrate of reward anticipation in health: A meta-analysis of
1391 fMRI findings in the monetary incentive delay task. *Neuropsychology Review*, 28(4),
1392 496–506. <https://doi.org/10.1007/s11065-018-9385-5>

1393 Witt, S. T., Laird, A. R., & Meyerand, M. E. (2008). Functional neuroimaging correlates of
1394 finger-tapping task variations: An ALE meta-analysis. *NeuroImage*, 42(1), 343–356.
1395 <https://doi.org/10.1016/j.neuroimage.2008.04.025>

1396 Zucker, R. A., Fitzgerald, H. E., Refior, S. K., Puttler, L. I., Pallas, D. M., & Ellis, D. A. (2000).
1397 The clinical and social ecology of childhood for children of alcoholics: Description of a
1398 study and implications for a differentiated: Description of a study and implications for a
1399 differentiated social policy. In *Children of Addiction*. Routledge.

1400 Zuo, X.-N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., Breitner, J. C. S.,
1401 Buckner, R. L., Calhoun, V. D., Castellanos, F. X., Chen, A., Chen, B., Chen, J., Chen,
1402 X., Colcombe, S. J., Courtney, W., Craddock, R. C., Di Martino, A., Dong, H.-M., ...
1403 Milham, M. P. (2014). An open science resource for establishing reliability and
1404 reproducibility in functional connectomics. *Scientific Data*, 1(1), Article 1.
1405 <https://doi.org/10.1038/sdata.2014.49>
1406

Supplemental Materials

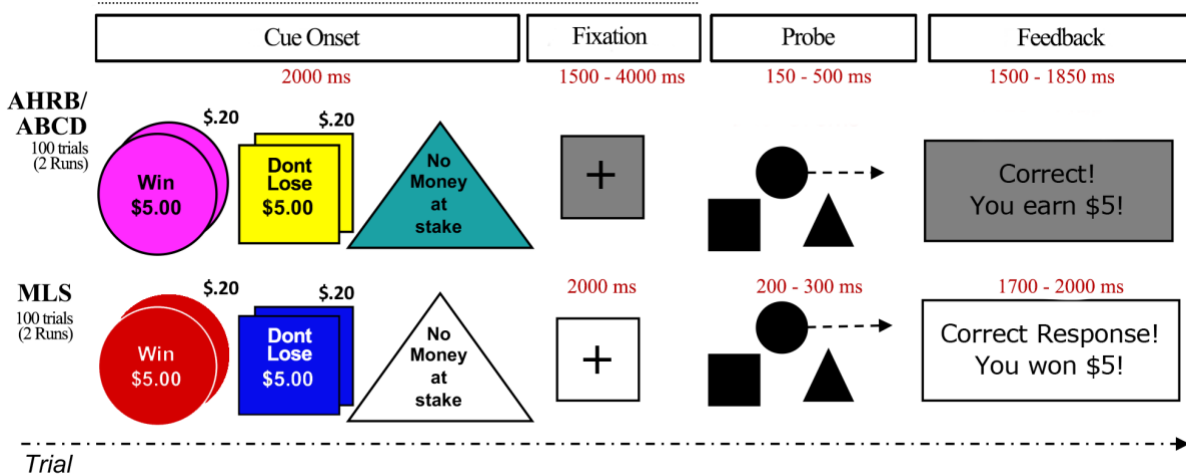
Section 1 – Analytic Decisions, FMRI Task, Data & Preprocessing

1.1 Description of analytic decisions

The effect of smoothing is evaluated by selecting smoothing kernels that range from 1.5x - 3.5x the voxel size (in half point increments). This range is used in place of specific sized smoothing kernels (e.g., 4 mm) because the MLS and ABCD/AHRB differ in their voxel size, 4mm & 2.4mm, respectively. To avoid inflating the smoothing kernel in the MLS dataset, we scale the magnitude (e.g., voxel 4 mm x 2) by a magnitude of .60 (e.g., 2.4 mm/4 mm voxel size) for MLS data. The approximate smoothness between the two datasets is evaluated using Nipype's (Gorgolewski et al., 2018) interface of FSL's *SmoothEstimate()* applied to the model residuals to ensure the resulting smoothing in the BOLD data is comparable between the ABCD/AHRB and MLS samples. A range of liberal (e.g., no motion correction) to conservative strategies (e.g., censoring high motion volumes, excluding high motion subjects, and regressing estimated motion, their derivatives and eight anatomically derived noise components) are used to reduce the effects of motion and other artifacts that are historically acknowledged to increase variance in signal (Tomarken, 1995). Finally, over the years there have been several different modeling techniques for the MID task. For example, the cue phase (Demidenko et al., 2021; Srirangarajan et al., 2021) or fixation phase (Bjork et al., 2004; Sacchet & Knutson, 2013) may be modeled as the 'anticipation'. Below, **Figure S2**, suggests that these modeling decisions impact the efficiency of the design which may alter the variance structure across contrasts with lower and higher BOLD activity.

For demonstration purposes, the MID task events data from the AHRB study are used to generate the regressors for efficiency using the *neuRosim* package (Welvaert et al., 2011). Events information from 101 subjects (**for this demonstration**, some do not have the necessary outcome events which prevent the use of data in this case) is used for BOLD time series with a TR 800 ms and 407 volumes. The design of the task in the AHRB sample (as well as MLS/ABCD) is presented in **Figure S1**. The models that are calculated include different 'anticipation' model versions observed in the literature over the years (also included the 10-feedback variation duration regressors [hit/miss for each of the five cue types]):

- Cue Model: Cue onset + Cue Duration (2sec)
- Ant Model: Cue onset + (Cue Duration [2sec] + Fixation Duration [variable, 1.5-4sec])
- Fix Model: Fixation onset + Fixation Duration (variable, 1.5-4sec)



34 *Trial*
35 **Figure S1.** Task schematic for the AHRB, ABCD and MLS studies.

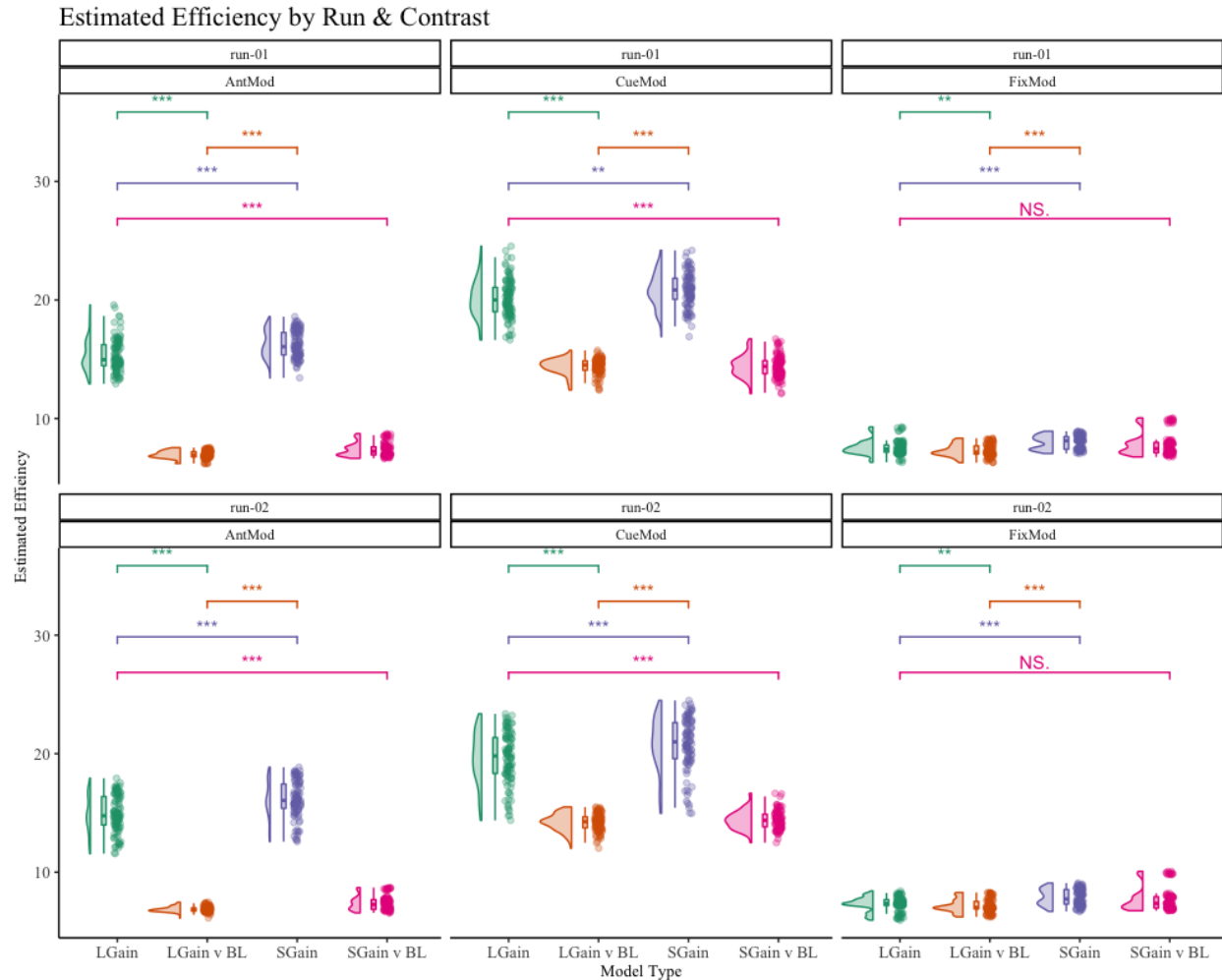
36 Schematic of the MID task design for the AHRB, ABCD and MLS samples. Both studies acquired 100 trials across
37 two runs. Each task trial starts with a Cue indicating the trial time (Win [\$5 or \$0.20]; Lose [\$5 or \$0.20]; or
38 Neutral). The cue lasts for 2000 ms. Following the cue is the Fixation cross. In the AHRB/ABCD samples, Fixation
39 duration is variable (1500-4000 ms) but constant in MLS (2000 ms). The probe duration is a variable duration in all
40 three samples. It is dependent on the participants performance. The probe window increases/decreases as the
41 participants probe hit rate increases/decreases below a target of ~60%. The feedback phase of both the three studies
42 is a variable duration and is adjusted based on the probe phase.

43 For the regressor estimates generated based on the provided behavioral data, efficiency
44 can be calculated across model types. **Figure S2** displays the distribution and difference in
45 estimated efficiency between the three model types across runs and the four contrasts for the
46 Stage 1 Registered report (**NOTE:** in Stage 2 we learned of an error in *neuRosim* that impacted
47 the interpretation of ‘most efficient’ model. See results in **Section 2.2 & Figure S7**). These data
48 suggest that across both runs the least efficient model is the Fixation Model (FixMod) and the
49 most efficient model is the Cue Model (CueMod). While there is more similarity between the
50 Anticipation Model (AntMod) and the Cue Model (CueMod), the latter in this is marginally
51 better comparing vectors (via t-tests) as implemented in R using `ggsignif::geom_signif`
52 (Ahlmann-Eltze & Patil, 2021). Efficiency is impacted by the modeled trial duration, number of
53 trials, collinearity and other factors. The efficiency of a model’s design matrix only reflects part

54 of the first level model's variance, which is the product of the inverse of the efficiency and the
55 residual variance. The most efficient design matrix may not fit the data well, increasing the
56 residual variance and the overall variance of the estimated contrast. For example, consider
57 CueMod and AntMod for the LGain v BL contrast. CueMod has higher efficiency due to lower
58 overlap between the anticipation regressor (only modeled during Cue Onset + Cue Duration) and
59 the Feedback regressor, but if the anticipation-based brain activation continues throughout the
60 fixation period, CueMod will not capture this variability as well as AntMod. Whether CueMod
61 outperforms AntMod for this contrast depends on whether the increased efficiency of CueMod is
62 overshadowed by an increase in residual variance due to poor model fit.

63 The impact of model efficiency on reliability will be considered in parallel with how the
64 residual variance estimate also varies. These modeling decisions may have an underlying impact
65 on the underlying contrasts, as is shown in the figure below representing models across each run
66 and contrast type. However, the impact on reliability estimates remains to be empirically tested
67 across these different modeling approaches but one may hypothesize that the least efficient
68 model (FixMod) and contrast (Small Gain v Neutral & Small Gain v Implicit Baseline) would
69 have a lower reliability than the other models and contrasts.

- 70 ● LGain: Large Gain > Neut
- 71 ● SGain: Small Gain > Neut
- 72 ● LGain v BL: Large Gain > Implicit Baseline
- 73 ● SGain v BL: Small Gain > Implicit Baseline



74

75 *Figure S2: Modeling Efficiency Across Model, Run and Four MID Contrasts.*

76 Comparing the model efficiencies between the four contrast types across the three model types. The Models are
 77 plotted for each run (run 01 and run 02) separately. LGain: Large Gain > Neut; SGain: Small Gain > Neut;
 78 LGain v BL: Large Gain > Implicit Baseline; SGain v BL: Small Gain > Implicit Baseline; CueMod: Cue onset +
 79 Cue Duration; AntMod: Cue onset + (Cue Duration + Fixation Duration; FixMod: Fixation onset + Fixation
 80 Duration. **Deprecated result:** We identified an error in neuRosim with how convolution is estimated. This does
 81 not impact other efficiency estimates as Nilearn is used in Stage 2 analyses.

82 1.2 Monetary Incentive Delay task description

83 As described elsewhere (Bjork, 2020; Demidenko et al., 2021; Knutson & Greer, 2008),
 84 the monetary incentive delay (MID) task measures reward anticipation. Apart from some minor
 85 differences, the MID task across the ABCD, AHRB and MLS samples are nearly identical. For
 86 example, during the MID task each trial starts with a cue type and consists of three phases:
 87 anticipation, probe and outcome (that is, feedback). The task regressors include different cue

88 (five) and feedback types (ten), totaling 15-task regressors that are included in the GLM. **Table**
 89 **S1**, below, summarizes the trials, runs, cue types, timing and targeted accuracy information for
 90 the MID task across the three samples.

91 *Table S1.* Monetary Incentive Delay Task Details Across AHRB, ABCD and MLS samples.

Sample	Trials	Runs	Cue Types (Trials)	Cue Duration (ms)	Fixation Duration (ms)	Probe Duration (ms)	Feedback Duration (ms)	Target Accuracy
AHRB	50	2	Win \$5.00 (10), Win \$0.20 (10), Neutral (10), Don't Lose \$5.00 (10), Don't Lose \$0.20 (10)	2000	1500 - 4000	150 - 500	1500 - 1850	60%
ABCD	50	2	Win \$5.00 (10), Win \$0.20 (10), Neutral (10), Don't Lose \$5.00 (10), Don't Lose \$0.20 (10)	2000	1500 - 4000	150 - 500	1500 - 1850	60%
MLS	50	2	Win \$5.00 (10), Win \$0.20 (10), Neutral (10), Don't Lose \$5.00 (10), Don't Lose \$0.20 (10)	2000	2000	300 - 500	1700 - 2000	60%

92

93 1.3 FMRI Acquisition details

94 *Table S3.* Acquisition parameters for structural and functional data across *four* samples.

	Scanner	Scan	TR (ms)	TE (ms)	Flip Angle	FOV (cm)	Voxel (mm)	Matrix
AHRB	GE MR750	Structural	7	2.9	8	25.6	1	256x256
ABCD	GE MR750	Structural	2500	2	8	25.6	1	256x256
	Philips	Structural	6.31	2.9	8	25.6	1	256x256
	Siemens	Structural	2500	2.88	8	25.6	1	256x256
MLS	GE Signa	Structural	12	5.2	15	19.5	1.2	256x256

AHRB	GE MR750	BOLD*	800	30	52	21.6	2.4	90x90
ABCD	GE MR750	BOLD*	800	30	52	21.6	2.4	90x90
	Philips	BOLD*	800	30	52	21.6	2.4	90x90
	Siemens	BOLD*	800	30	52	21.6	2.4	90x90
MLS	GE Signa	BOLD	2000	30	90	20	4	64x64

95 *BOLD runs are multiband 6 factor acquisition & Fieldmaps were collected. TR: Time Repetition; TE = Echo time;
 96 FOV: Field of view. ABCD & AHRB data are isotropic voxels (2.4 x2.4 x 2.4) and MLS data are anisotropic (3.125
 97 x 3.125 x 4)
 98

99 1.4. Preprocessing MRI & fMRI Data

100 Preprocessing of anatomical data. T1-weighted images are corrected for intensity non-
 101 uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs
 102 2.3.3 (RRID:SCR_004757; Avants et al., 2008) and used as T1w-reference throughout the
 103 fMRIPrep workflow. The T1w-reference is then skull-stripped with a Nipype implementation of
 104 the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as the target template.
 105 Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter
 106 (GM) is performed on the brain-extracted T1w using fast (FSL 6.0.5.1:57b01774,
 107 RRID:SCR_002823; Zhang et al., 2001). Brain surfaces are reconstructed using recon-all
 108 (FreeSurfer 7.2.0, RRID:SCR_001847; Dale et al., 1999), and the brain mask estimated
 109 previously is refined with a custom variation of the method to reconcile ANTs-derived and
 110 FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle
 111 (RRID:SCR_002438; Klein et al., 2017). Volume-based spatial normalization to one standard
 112 space (MNI152NLin2009cAsym) is performed through nonlinear registration with
 113 antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the
 114 T1w template. The following template are selected for spatial normalization: ICBM 152
 115 Nonlinear Asymmetrical template version 2009c (RRID:SCR_008796; TemplateFlow ID:
 116 MNI152NLin2009cAsym; Fonov et al., 2009)

117 Preprocessing of functional data. For each of the 2 BOLD functional runs, the following
 118 preprocessing steps are performed. First, a reference volume and its skull-stripped version are
 119 generated using a custom methodology of fMRIPrep. The estimated fieldmap was then aligned

120 with rigid-registration to the target EPI (echo-planar imaging) reference run. The field
121 coefficients were mapped on to the reference EPI using the transform. The BOLD reference was
122 then co-registered to the T1w reference using bregister (FreeSurfer) which implements
123 boundary-based registration (Greve & Fischl, 2009). Co-registration was configured with six
124 degrees of freedom. The BOLD time-series were resampled into standard space, generating a
125 preprocessed BOLD run in MNI152NLin2009cAsym space. Head-motion parameters with
126 respect to the BOLD reference (transformation matrices, and six corresponding rotation and
127 translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL
128 6.0.5.1:57b01774; Jenkinson et al., 2002). The estimated fieldmap is then aligned with rigid-
129 registration to the target EPI. Framewise displacement (FD) is calculated based on the
130 preprocessed BOLD. Principal components are estimated after high-pass filtering the
131 preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for anatomical
132 (aCompCor). For the aCompCor decomposition, the k components with the largest singular
133 values are retained, such that the retained components' time series are sufficient to explain 50
134 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining
135 components are dropped from consideration. The confounded time series derived from head
136 motion estimates were expanded with the inclusion of temporal derivatives and quadratic terms
137 for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.9 mm FD or 1.5
138 standardized DVARS were annotated as motion outliers.

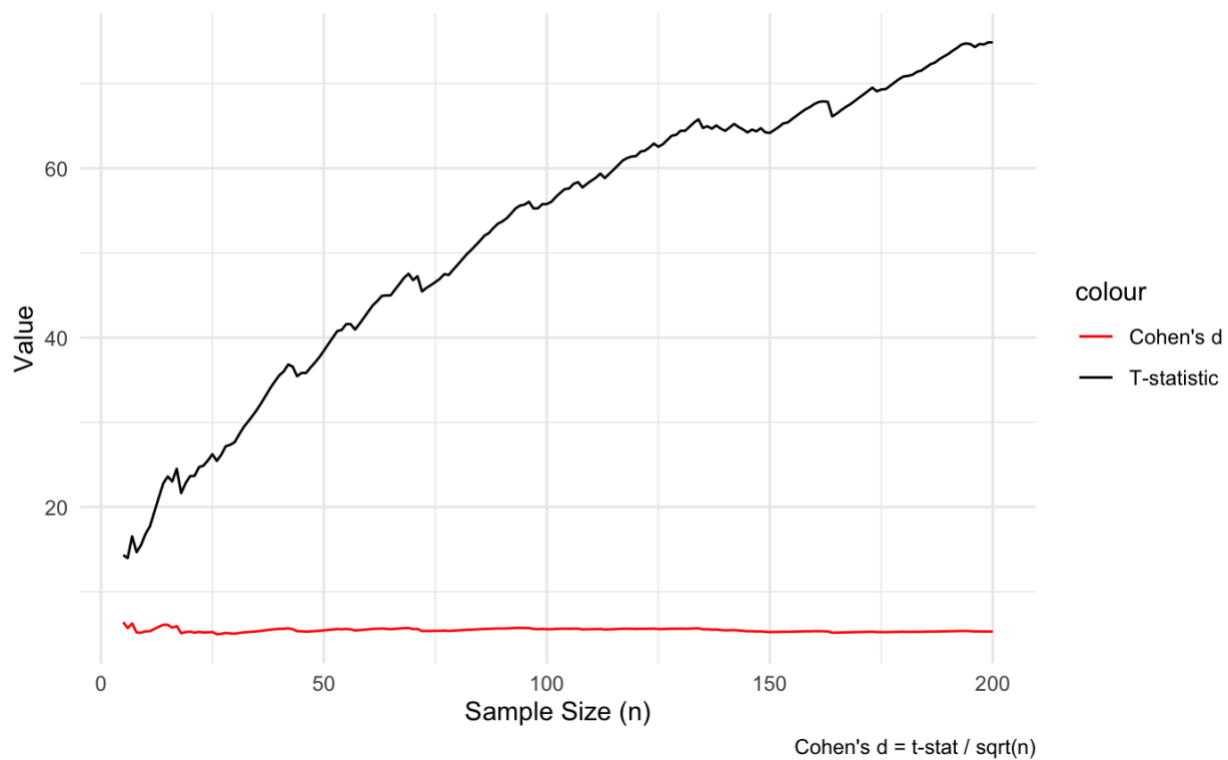
139 Section 2 – Results

140 The analytic code to recreate figures and estimates are available in the python notebooks
141 and R markdown files shared in within the Stage 2 github repository. Specifically, the html
142 reports include expanded information from the between-run and between-session HLM,

143 emmeans, Specification Curves and other plots within the R html reports and may be
144 recreated/reanalyzed using the share output files within the github Stage 2 repository.

145 2.1 Analytic modifications

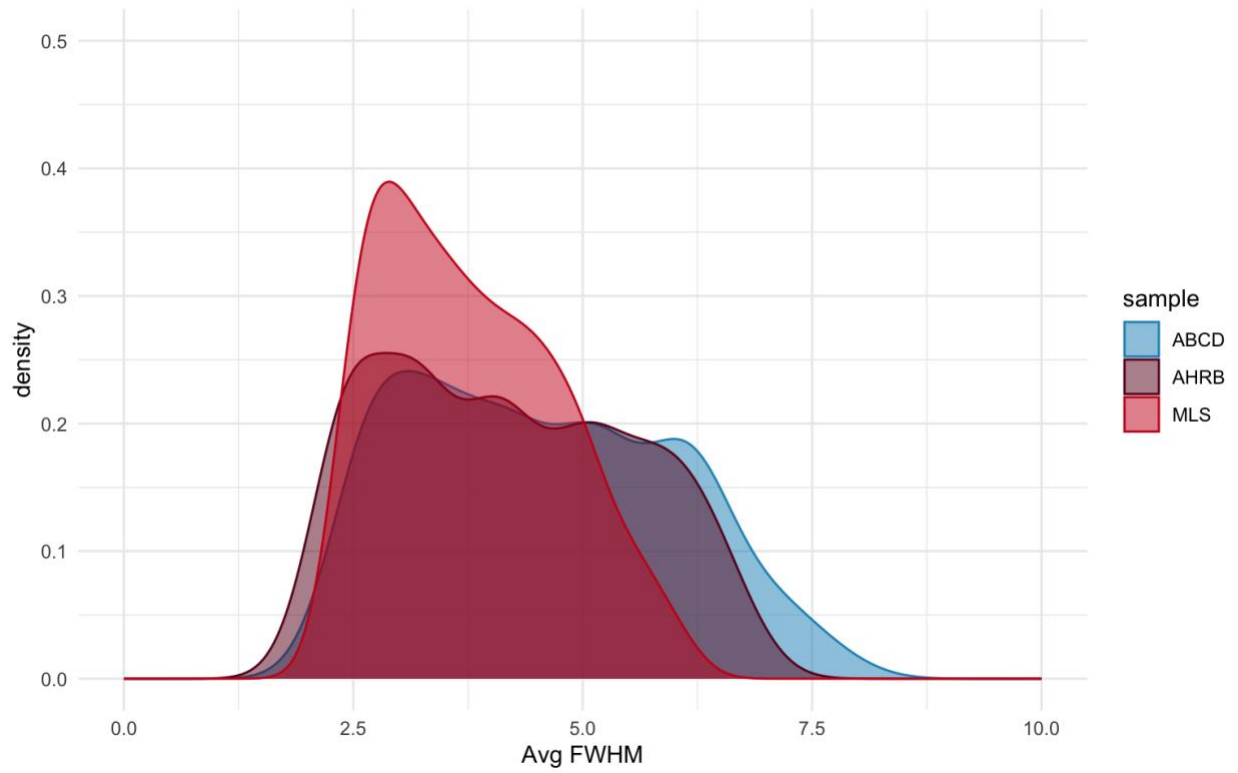
146 For Aim 1b, instead of thresholding images by $p < .001$ (or t -stat 3.2) we converted the
147 group t -stat to Cohen's d 3D effect size maps using the formula: $\frac{t\text{-statistic}}{\sqrt{N}}$. This is to avoid
148 differences in N s between some models because of failures during preprocessing (e.g., $N = 15$ in
149 ABCD failed aCompCor WM/GM/CSF masks).



150
151 *Figure S3:* Change in t -statistic and Cohen's d across $N = 0$ to $N = 200$ in a randomly simulated
152 data with $\mu = 5$ and $\sigma = 1$. The population mean for t -test is assumed to be zero.
153

154 We ran the model permutations on the ABCD/AHRB (2.4mm data) and MLS (4mm) data
155 with a weighted .50 FWHM smoothing parameter, we estimated the smoothness of the *group*
156 *residual variance* maps for the data. Since the model permutations differed in several decisions,
157 the smoothness is estimated across the 240 pipelines spanning four contrasts, four motion
158 options, three model parameterizations and five smoothness parameters. The estimated *average*
159 smoothness ($\text{Resel}^{[1/3]}$) for the ABCD 4.5 (SD = 1.4), AHRB 4.2 (SD = 1.3) and MLS 3.8 (SD

160 = 1.0). The distribution of estimated smoothness across group-level maps for the ABCD, AHRB
161 and MLS data are reported in **Figure S4**.



162
163 *Figure S4:* Estimates of smoothing of group level residual 3D volumes across 240 permutations
164 for the Michigan Longitudinal (MLS), Adolescent Health Risk Behavior (AHRB) and
165 Adolescent Brain Cognitive Development (ABCD) imaging data.
166

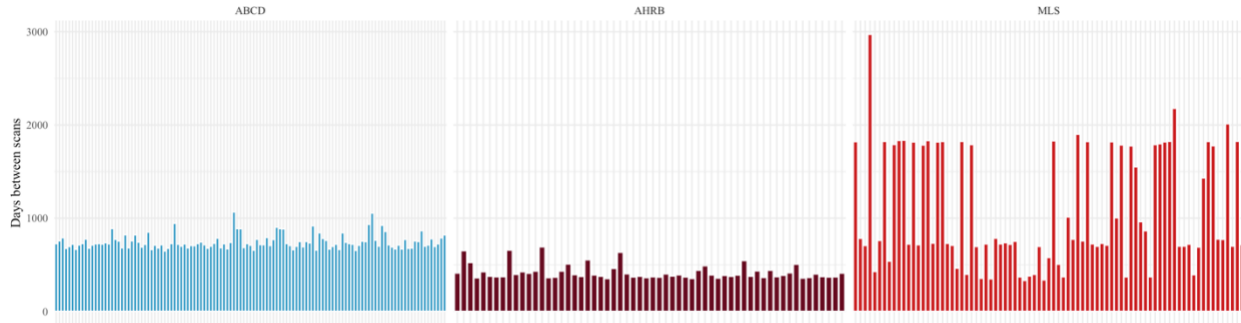
167 **2.2 Descriptive Results**

168 *Demographics Across Samples:* The demographic information is reported in **Table S4** and the
169 days between sessions are visually represented in **Figure S5**.

170 *Table S4.* Age, Sex, Race/Ethnicity from Session 1 and Days Between Sessions Across ABCD,
171 AHRB and MLS

	ABCD	AHRB	MLS
	(N=119)	(N=60)	(N=81)
	<i>Mean (SD)</i>		
Age	9.8 (0.6)	19.3 (1.3)	20.7 (2.3)
Days Btwn Session	747 (79.1)	419 (80.1)	1090 (624)
Sex	<i>N (%)</i>		
Female	58 (48.7%)	35 (58.3%)	31 (38.3%)
Male	61 (51.3%)	25 (41.7%)	50 (61.7%)
Race/Ethnicity			
Asian	4 (3.4%)	0 (0%)	0 (0%)
Black	14 (11.8%)	10 (16.7%)	2 (2.5%)
Hispanic	8 (6.7%)	3 (5.0%)	5 (6.2%)
Other	15 (12.6%)	5 (8.3%)	1 (1.2%)
White	78 (65.5%)	42 (70.0%)	73 (90.1%)

172 Note: *MLS* participants reported on “caucasian”, “African American”, “Native American”,
173 “Asian American”, “Filipino or Pacific Islander”, “Bi-Racial” and “Hispanic-caucasian race”,
174 and *AHRB* “White Non-Hispanic”, “Black Non-Hispanic”, “Hispanic/Latinx”, and “american
175 Indian/Alaska/Native Hawaiian”, “Other” for simplicity refactor to match ABCD
176 “Race/Ethnicity” variable in *acspsw03*
177



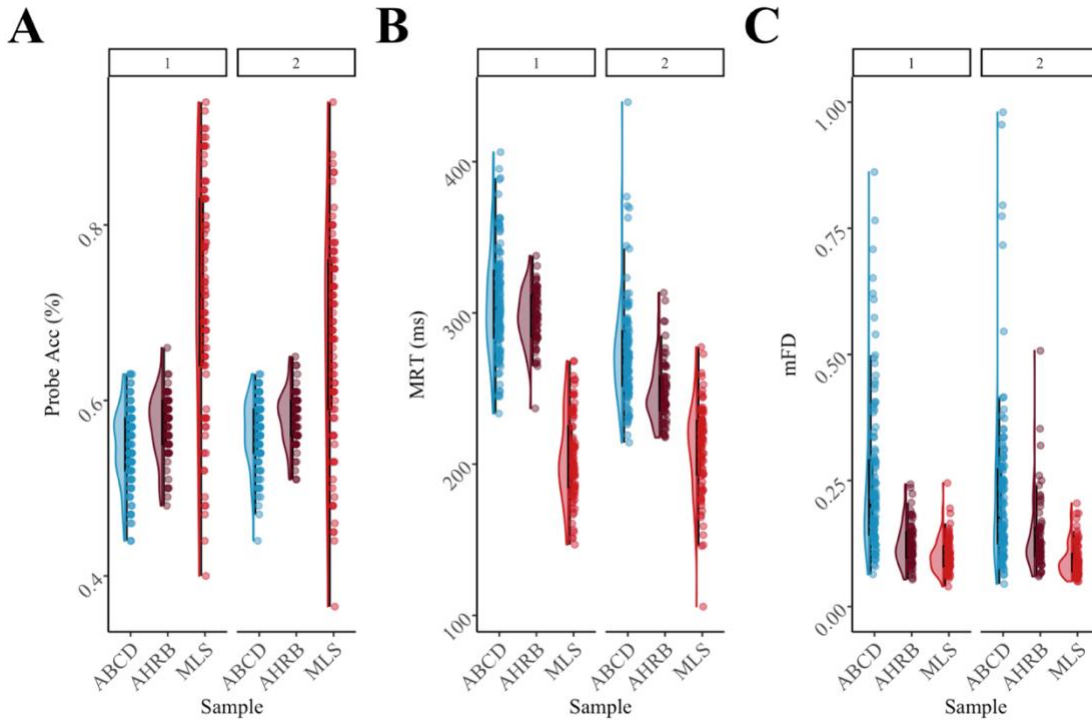
178
179 *Figure S5.* The number of days between sessions for subjects across ABCD, AHRB and MLS
180 samples.

181 *Task Behavior Across Samples:* The Mean and Standard Deviation for the run average Mean
182 Framewise Displacement, Average Probe Response Times and Average Probe Accuracies are
183 reported in **Table S5** and **Figure S6**.

184 *Table S5:* The run average Mean FD, Average Probe Accuracy and Mean RT across samples and
185 sessions.

Sample	Session	Mean	SD	Min	Max
<i>Mean Framewise Displacement</i>					
ABCD	1	0.25	0.15	0.06	0.86
AHRB	1	0.12	0.04	0.05	0.24
MLS	1	0.10	0.03	0.04	0.25
ABCD	2	0.25	0.23	0.05	1.29
AHRB	2	0.14	0.08	0.06	0.51
MLS	2	0.09	0.03	0.05	0.21
<i>Average Probe Accuracy (%)</i>					
ABCD	1	0.55	0.04	0.44	0.63
AHRB	1	0.57	0.04	0.48	0.66
MLS	1	0.72	0.13	0.40	0.94
ABCD	2	0.56	0.04	0.44	0.63
AHRB	2	0.58	0.03	0.51	0.65
MLS	2	0.67	0.12	0.37	0.94
<i>Average probe MRT (ms)</i>					
ABCD	1	306.8	34.5	233.5	406.2
AHRB	1	297.1	18.5	236.8	337.8
MLS	1	204.3	28.9	146.8	268.2

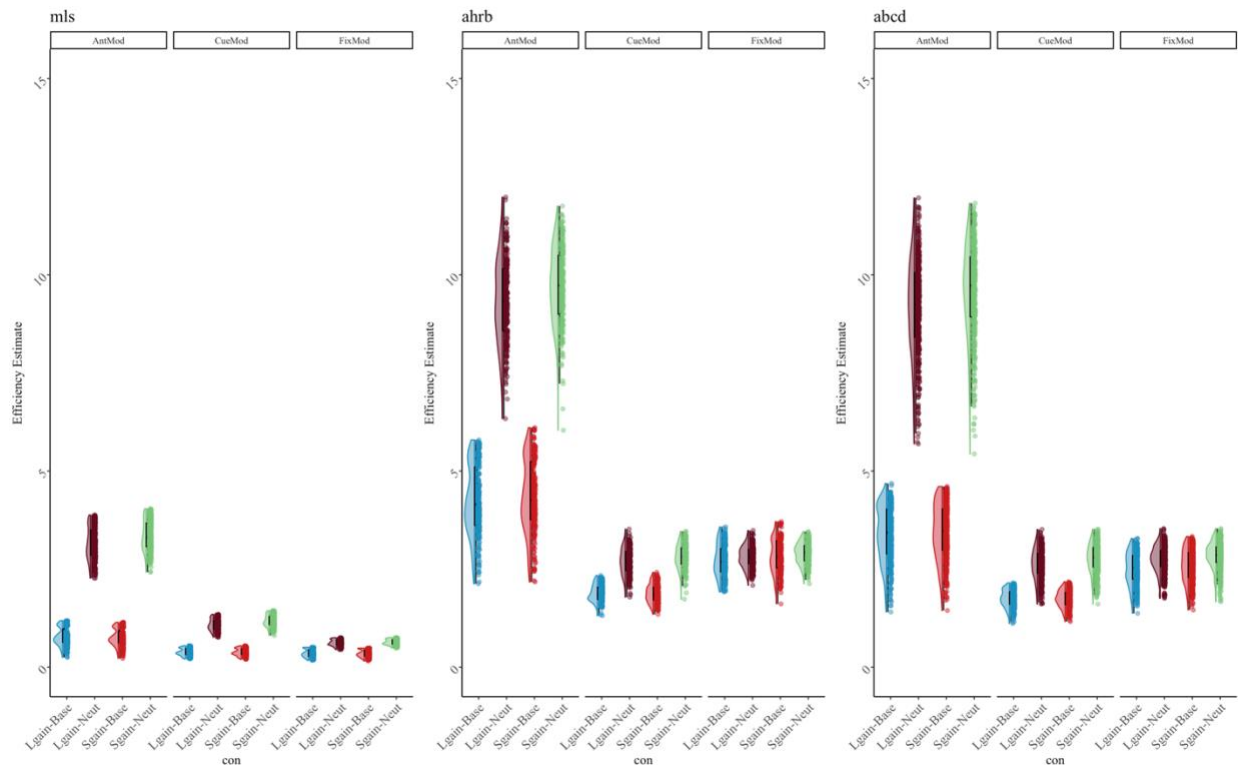
ABCD	2	274.3	34.1	214.3	439.3
AHRB	2	248.5	21.5	217.6	313.3
MLS	2	210.1	30.0	105.8	277.4



186 *Figure S6.* Distribution of (A) Mean Framewise Displacement, (B) Mean Probe RTs (ms) and
 187 (C) Mean Probe Accuracy (%) across Sessions and ABCD, AHRB and MLS samples.

188 *Task Efficiency Across Samples*: The model efficiency was calculated as the inverse proportion
 189 of variance based on the design matrix. The design matrix varied only as a function of
 190 parameterization and motion regressors for the four contrasts. The formula used is:

191 $Efficiency = \frac{1}{c(X'X)^{-1}c'}$. As is observed from **Figure S7**, contrary to the above/incorrect
 192 *neuRosim Figure S2*, the most efficient design (compared within a category) is the Anticipation
 193 Model ('AntMod'). Furthermore, consistent with our hypothesis, the most efficient contrast
 194 within a model is the *Large Gain* versus *Neutral* contrast.



195
 196 **Figure S7.** Distribution of estimated model efficiencies from design matrices for Model
 197 Parameterization and Contrast type across ABCD, AHRB and MLS samples.

199 *Between-run and Between-session similarity estimates*: Overall, the between-session ICC,
 200 Jaccard and Spearman Similarity estimates were higher than the Session 1 between-run estimates
 201 (**Table S5**).

202 **Table S5.** Session 1 Between-run and Between-session Median, Mean, Standard Deviation (SD),
 203 Minimum and Maximum of median Intraclass Correlation Coefficient (ICC) and Jaccard and
 204 Spearman Similarity and from 240 analytic models across ABCD, AHRB and MLS Samples.

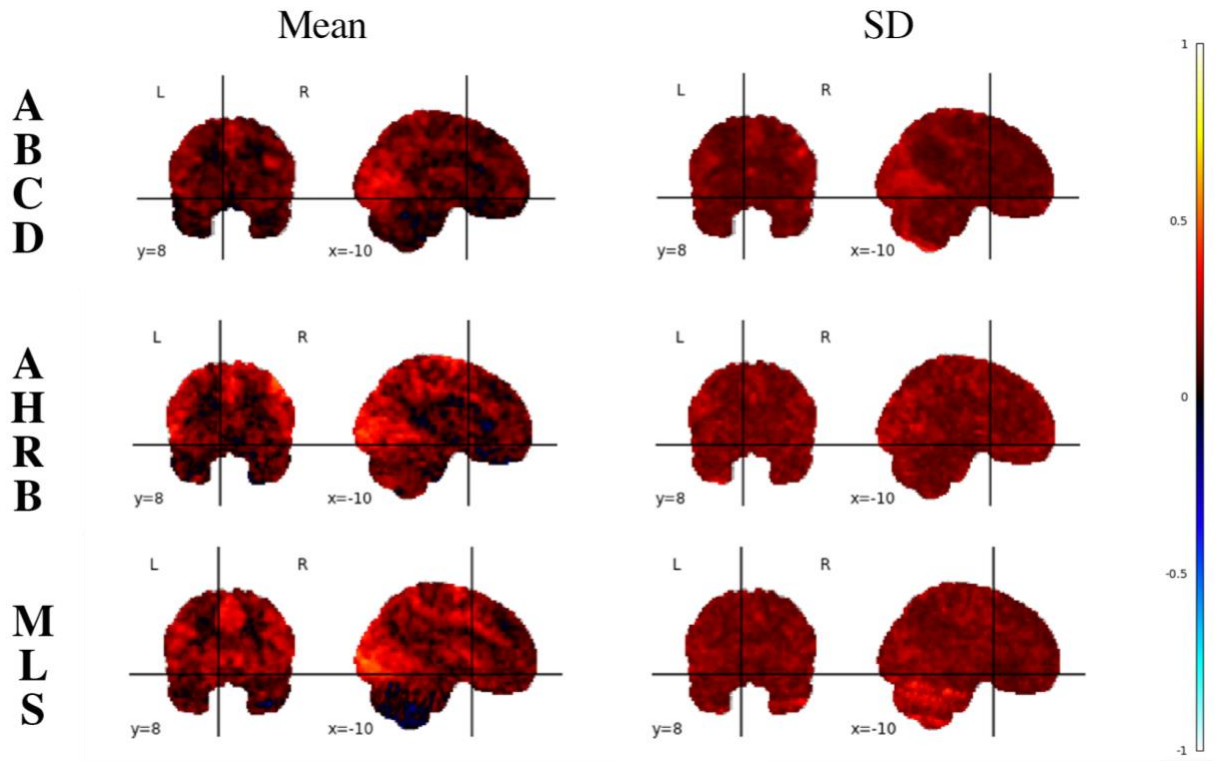
study	estimate	median	mean	sd	min	max
<i>Session 1: Between-runs</i>						
ABCD	ICC*	.11	.15	.12	-.07	.43
AHRB	ICC*	.18	.20	.13	.00	.52
MLS	ICC*	.18	.21	.13	.04	.55
ABCD	Jaccard	.09	.11	.09	.01	.45
AHRB	Jaccard	.18	.21	.15	.01	.64
MLS	Jaccard	.34	.34	.11	.15	.60
ABCD	Spearman*	.68	.68	.14	.35	.89
AHRB	Spearman*	.73	.68	.22	.22	.96
MLS	Spearman*	.84	.80	.12	.47	.95
<i>Between-sessions</i>						
ABCD	ICC*	.15	.16	.07	.03	.34
AHRB	ICC*	.21	.23	.13	.04	.53
MLS	ICC*	.21	.22	.10	.06	.47
ABCD	Jaccard	.25	.26	.13	.02	.61
AHRB	Jaccard	.30	.32	.19	.04	.73
MLS	Jaccard	.42	.43	.12	.20	.74
ABCD	Spearman*	.80	.76	.13	.40	.94
AHRB	Spearman*	.82	.74	.21	.32	.97
MLS	Spearman*	.87	.85	.09	.59	.97

206 *Supra-threshold mask

207 **2.3 Aim 1 results**

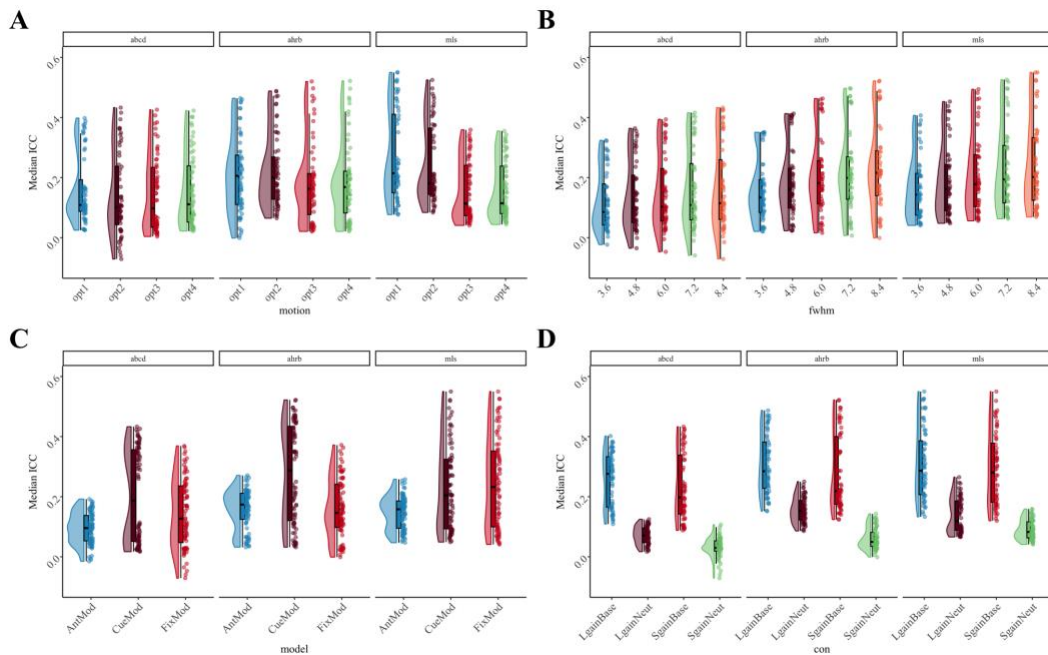
208 *A. Between-Run Individual Reliability:*

209 The average and standard deviation across model permutations for each sample are
 210 reported in **Figure S8**. The distribution of median ICC estimates across [four] analytic options is
 211 reported in **Figure S9**. The complete supra-threshold specification curve for between-run median
 212 ICCs are reported in **Figure S9** and the sub-threshold in **Figure S11**.



213

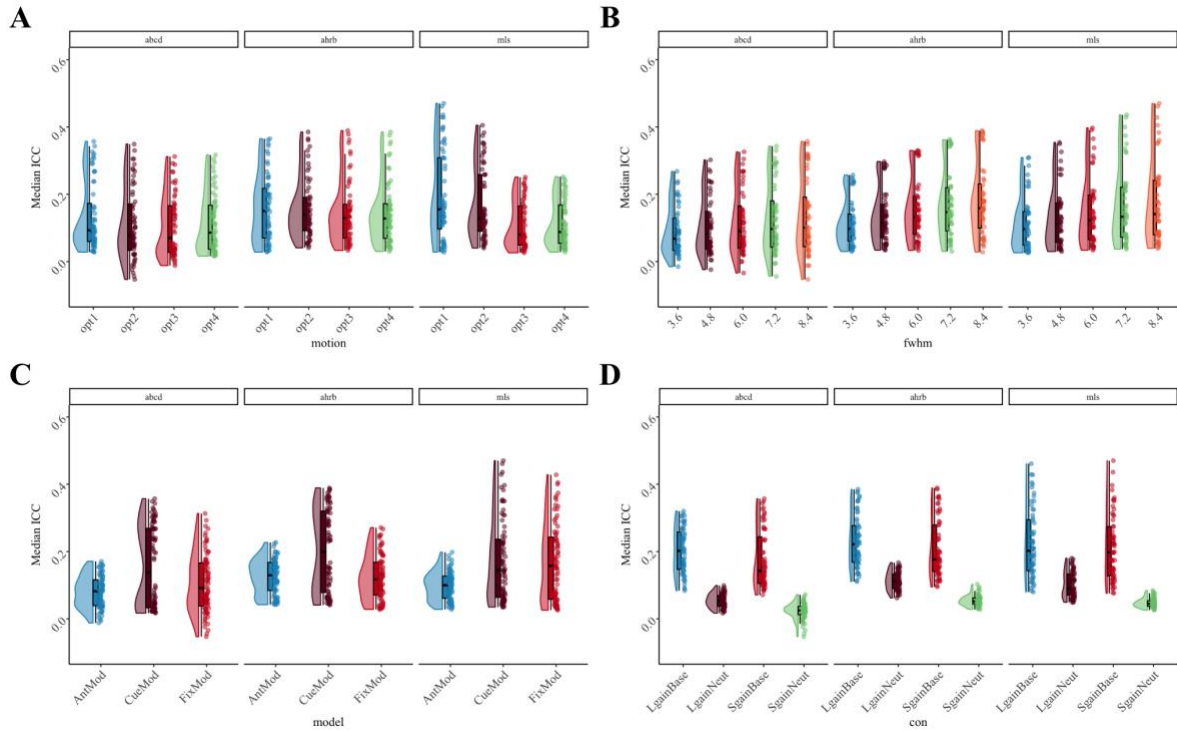
214 **Figure S8: Mean and SD of ICC estimates across 240 permutations for the Adolescent Brain**
 215 **Cognitive Development (ABCD), Adolescent Health Risk Behavior (AHRB) and Michigan**
 216 **Longitudinal (MLS) 3D volumes.**



217

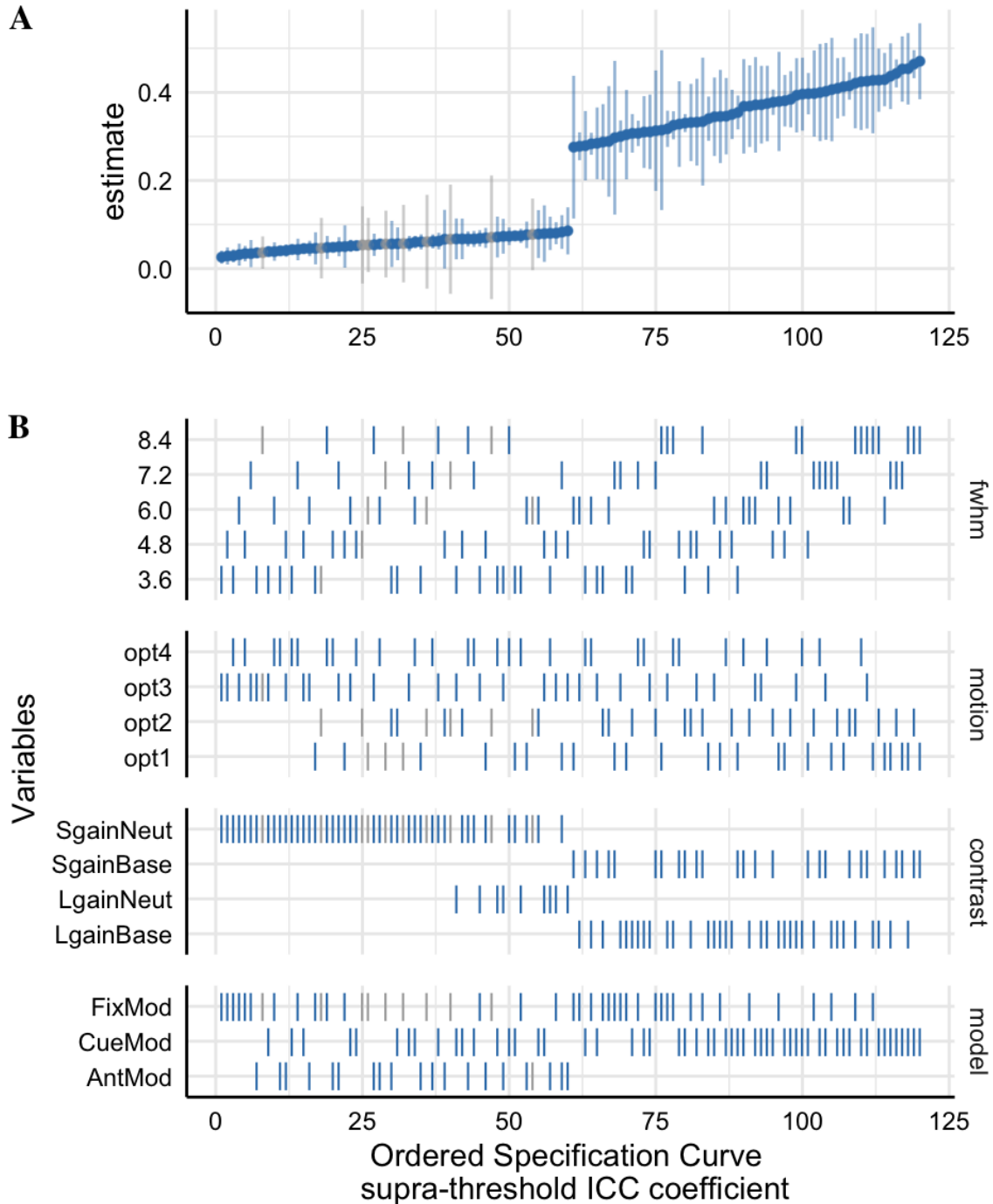
218 **Figure S9. Supra-threshold Median ICC Session 1 between-run reliability estimates for (A)**
 219 **Motion, (B) FWHM, (C) Model Parameterization and (D) Contrasts analytic options across the**
 220 **ABCD, AHRB and MLS samples. Expanded version of in-text Figure 2.**

221



222
223
224

Figure S10. Sub-threshold Median ICC Session 1 between-run reliability estimates for Contrast (con) and Model Parameterization analytic options across the ABCD, AHRB and MLS samples.



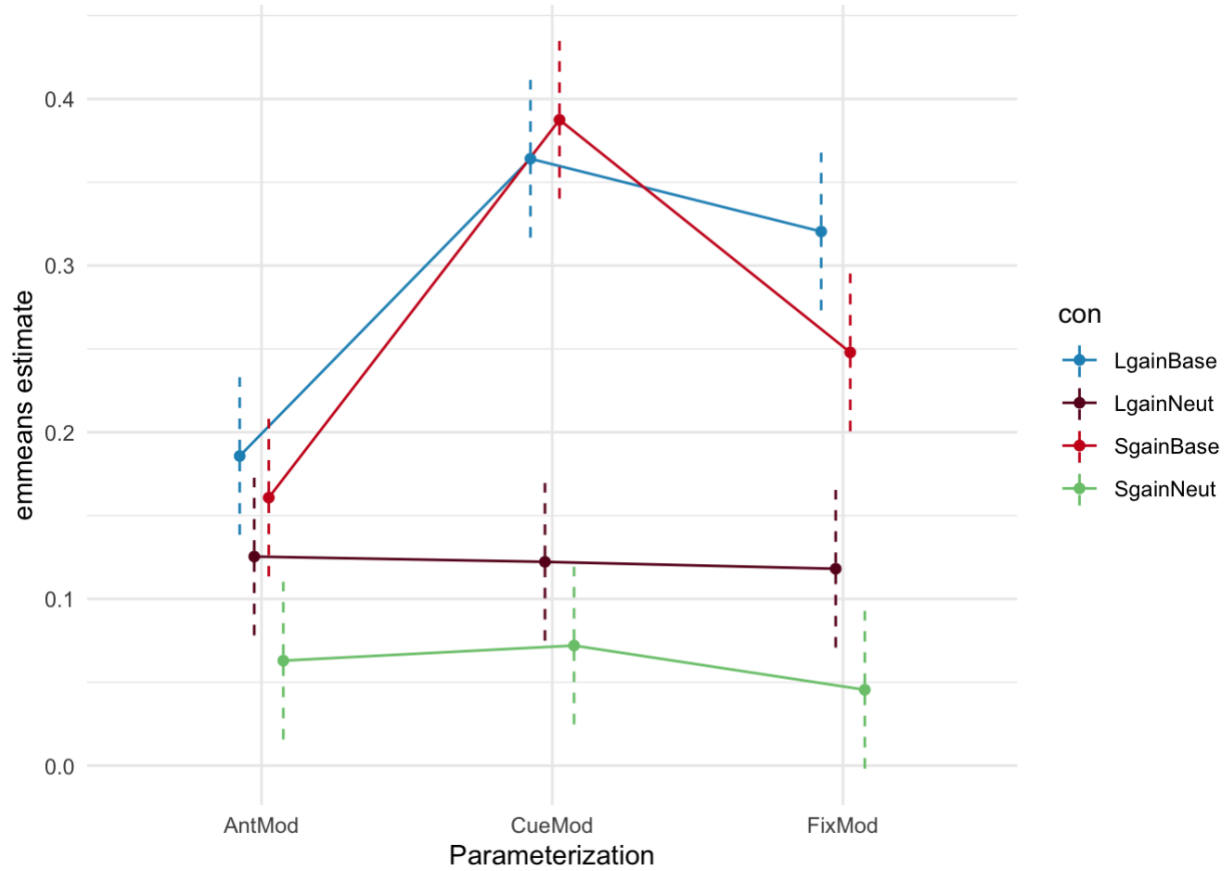
225
226 **Figure S11:** The 25th and 75th percentile supra-threshold Specification Curve of the Session 1
227 Between-run Median ICC estimates across 240 pipeline permutations for the ABCD, AHRB and
228 MLS samples. Full length of estimates reported in **Figure 4**.
229 A. The distribution of the point estimate (average) and distribution (error bars) across the three samples. B. The
230 model options (four) associated with each estimate.
231

232
233
234
235

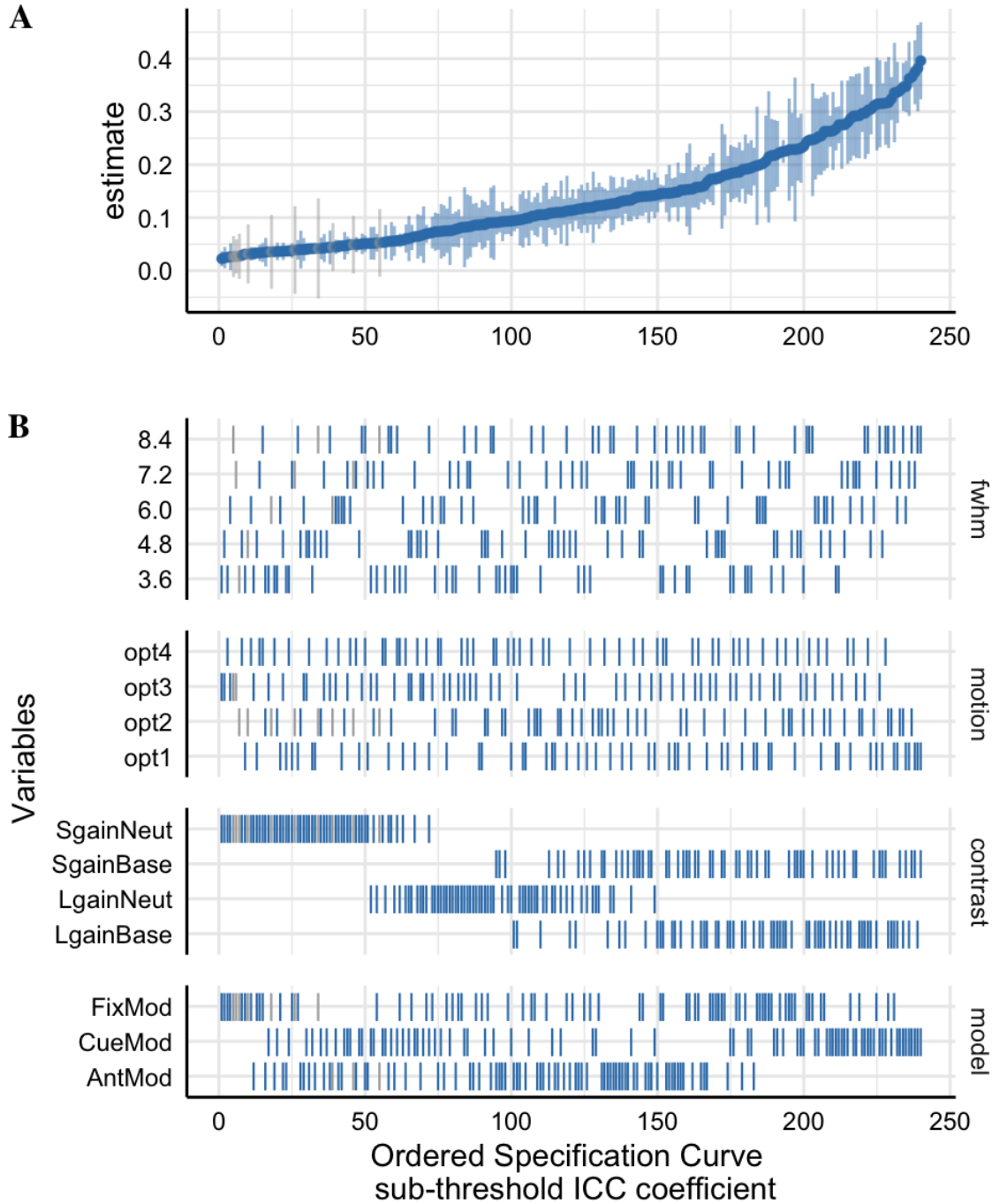
Table S6: Tukey's HSB Estimate Means Differences for between-run Supra-threshold ICC Model Parameters in-text Table 3.

Contrast	Est	SE	Low.CI	Up.CI	<i>p</i>
fwhm3.6 - fwhm4.8	-.02	.01	-.04	.00	.023
fwhm3.6 - fwhm6.0	-.04	.01	-.06	-.02	.000
fwhm3.6 - fwhm7.2	-.06	.01	-.08	-.04	.000
fwhm3.6 - fwhm8.4	-.07	.01	-.09	-.05	.000
fwhm4.8 - fwhm6.0	-.02	.01	-.04	.00	.098
fwhm4.8 - fwhm7.2	-.03	.01	-.05	-.01	.000
fwhm4.8 - fwhm8.4	-.04	.01	-.06	-.02	.000
fwhm6.0 - fwhm7.2	-.01	.01	-.04	.01	.299
fwhm6.0 - fwhm8.4	-.03	.01	-.05	-.01	.006
fwhm7.2 - fwhm8.4	-.01	.01	-.03	.01	.575
LgainBase - LgainNeut	.17	.01	.15	.19	.000
LgainBase - SgainBase	.02	.01	.01	.04	.001
LgainBase - SgainNeut	.23	.01	.21	.25	.000
LgainNeut - SgainBase	-.14	.01	-.16	-.13	.000
LgainNeut - SgainNeut	.06	.01	.04	.08	.000
SgainBase - SgainNeut	.21	.01	.19	.22	.000
opt1 - opt2	.01	.01	-.01	.03	.283
opt1 - opt3	.05	.01	.03	.07	.000
opt1 - opt4	.05	.01	.03	.06	.000
opt2 - opt3	.04	.01	.02	.06	.000
opt2 - opt4	.03	.01	.02	.05	.000
opt3 - opt4	.00	.01	-.02	.01	.940
AntMod - CueMod	-.10	.01	-.12	-.09	.000
AntMod - FixMod	-.05	.01	-.06	-.04	.000
CueMod - FixMod	.05	.01	.04	.07	.000

236



237
238 *Figure S12: Median ICC estimate: Interaction plot of emmeans fitted model of Contrast-by-*
239 *Model parameterization for Session 1 Between-run supra-threshold estimates using emmip().*
240 *Point estimate is a linear median ICC estimate from emmeans function. Dashed bars are*
241 *estimated confidence intervals by emmeans.*



242
243
244
245
246
247
248
249

Figure S13: The sub-threshold Specification Curve of the Median Intraclass Correlation Coefficient (ICC[3,1]) estimates across 240 pipeline permutations for the ABCD, AHRB and MLS estimate.

A. The distribution of the point estimate (average) across the three studies and distribution across the three samples.
B. The model options (four) associated with each estimate.

250 **Table S7: Hierarchical Linear Model: (A) Linear associations between the analytic decisions and**
 251 **the *Session 1 Between-run* median Intraclass Correlation Coefficient (ICC[3,1]), Between-**
 252 **subject (BS) and Within-subject variance (WS) from **sub-threshold mask** and (B) the impact of**
 253 **the analytic category on the marginal R².**
 254

A. HLM Estimates for Sub-threshold Mask									
<i>Predictors</i>	Median ICC			Median BS			Median WS		
	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>
(Intercept)	.17	.15 – .20	<.00 1	.31	.21 – .41	<.00 1	1.34	1.05 – 1.64	<.00 1
Reference [3.6]									
fwhm [4.8]	.02	.01 – .03	.001 <.00 1	-.02	-.06 – .01	.18 <.00 1	-.35	-.42 – -.28	<.00 1
fwhm [6.0]	.03	.02 – .05	.001 <.00 1	-.04	-.08 – -.01	.02 <.00 1	-.55	-.62 – -.48	<.00 1
fwhm [7.2]	.05	.04 – .06	.001 <.00 1	-.06	-.09 – -.02	.002 <.00 1	-.67	-.74 – -.60	<.00 1
fwhm [8.4]	.06	.05 – .07	.001 <.00 1	-.07	-.10 – -.03	.001 <.00 1	-.75	-.82 – -.68	<.00 1
Reference [opt1]									
motion [opt2]	-.02	-.03 – -.01	.003 <.00 1	-.07	-.10 – -.04	.001 <.00 1	-.14	-.21 – -.08	<.00 1
motion [opt3]	-.04	-.05 – -.03	.001 <.00 1	-.14	-.17 – -.11	.001 <.00 1	-.29	-.35 – -.23	<.00 1
motion [opt4]	-.04	-.05 – -.03	.001 <.00 1	-.14	-.17 – -.11	.001 <.00 1	-.30	-.36 – -.24	<.00 1
Reference [AntMod]									
model [CueMod]	.08	.07 – .08	.001 <.00 1	.18	.15 – .20	.001 <.00 1	.34	.29 – .40	<.00 1
model [FixMod]	.03	.02 – .04	.001 <.00 1	.13	.10 – .15	.001 <.00 1	.38	.33 – .44	<.00 1
Reference [LgainBase]									
con [LgainNeut]	-.13	-.14 – -.12	.001 <.00 1	-.25	-.28 – -.22	.001 <.00 1	-.46	-.52 – -.40	<.00 1
con [SgainBase]	-.02	-.03 – -.01	.001 <.00 1	-.03	-.06 – .01	.12 <.00 1	.01	-.06 – .07	.84 <.00 1
con [SgainNeut]	-.18	-.19 – -.17	.001 <.00 1	-.27	-.31 – -.24	.001 <.00 1	-.49	-.55 – -.43	<.00 1

B. Analytic Category Model Impact

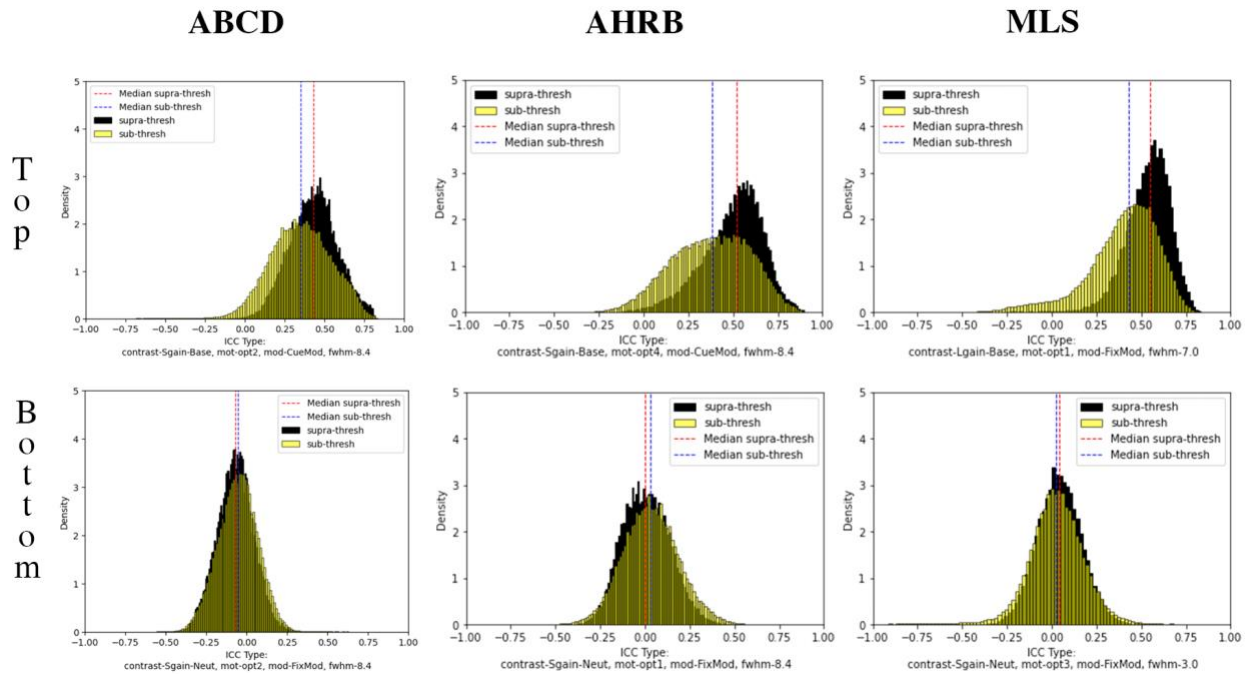
Comparison												
	χ^2	Orig R2	New R2	ΔR^2	χ^2	Orig R2	New R2	ΔR^2	χ^2	Orig R2	New R2	ΔR^2
[Full] vs [New - fwhm]	123	.73	.69	.04	16	.45	.44	.01	428	.53	.31	.22
[Full] vs [New - motion]	84	.73	.71	.02	94	.45	.39	.06	115	.53	.49	.04
[Full] vs [New - model]	252	.73	.63	.10	147	.45	.36	.09	209	.53	.44	.09
[Full] vs [New - con]	867	.73	.17	.56	362	.45	.17	.28	360	.53	.36	.17

256
257
258

Table S8: Tukey's HSB Estimate Means Differences for Sub-threshold Between-run ICC Model Parameters in Table S6.

Contrast	Est	SE	Low.CI	Up.CI	<i>p</i>
fwhm3.6 - fwhm4.8	-.02	.01	-.03	.00	.013
fwhm3.6 - fwhm6.0	-.03	.01	-.05	-.02	.000
fwhm3.6 - fwhm7.2	-.05	.01	-.06	-.03	.000
fwhm3.6 - fwhm8.4	-.06	.01	-.07	-.04	.000
fwhm4.8 - fwhm6.0	-.02	.01	-.03	.00	.044
fwhm4.8 - fwhm7.2	-.03	.01	-.04	-.01	.000
fwhm4.8 - fwhm8.4	-.04	.01	-.06	-.02	.000
fwhm6.0 - fwhm7.2	-.01	.01	-.03	.00	.134
fwhm6.0 - fwhm8.4	-.02	.01	-.04	-.01	.000
fwhm7.2 - fwhm8.4	-.01	.01	-.03	.00	.317
LgainBase - LgainNeut	.13	.01	.12	.14	.000
LgainBase - SgainBase	.02	.01	.01	.03	.000
LgainBase - SgainNeut	.18	.01	.16	.19	.000
LgainNeut - SgainBase	-.11	.01	-.12	-.09	.000
LgainNeut - SgainNeut	.05	.01	.04	.06	.000
SgainBase - SgainNeut	.16	.01	.14	.17	.000
opt1 - opt2	.02	.01	.00	.03	.018
opt1 - opt3	.04	.01	.03	.05	.000
opt1 - opt4	.04	.01	.02	.05	.000
opt2 - opt3	.03	.01	.01	.04	.000
opt2 - opt4	.02	.01	.01	.04	.000
opt3 - opt4	.00	.01	-.02	.01	.913
AntMod - CueMod	-.08	.00	-.09	-.07	.000
AntMod - FixMod	-.03	.00	-.04	-.02	.000
CueMod - FixMod	.04	.00	.03	.06	.000

259



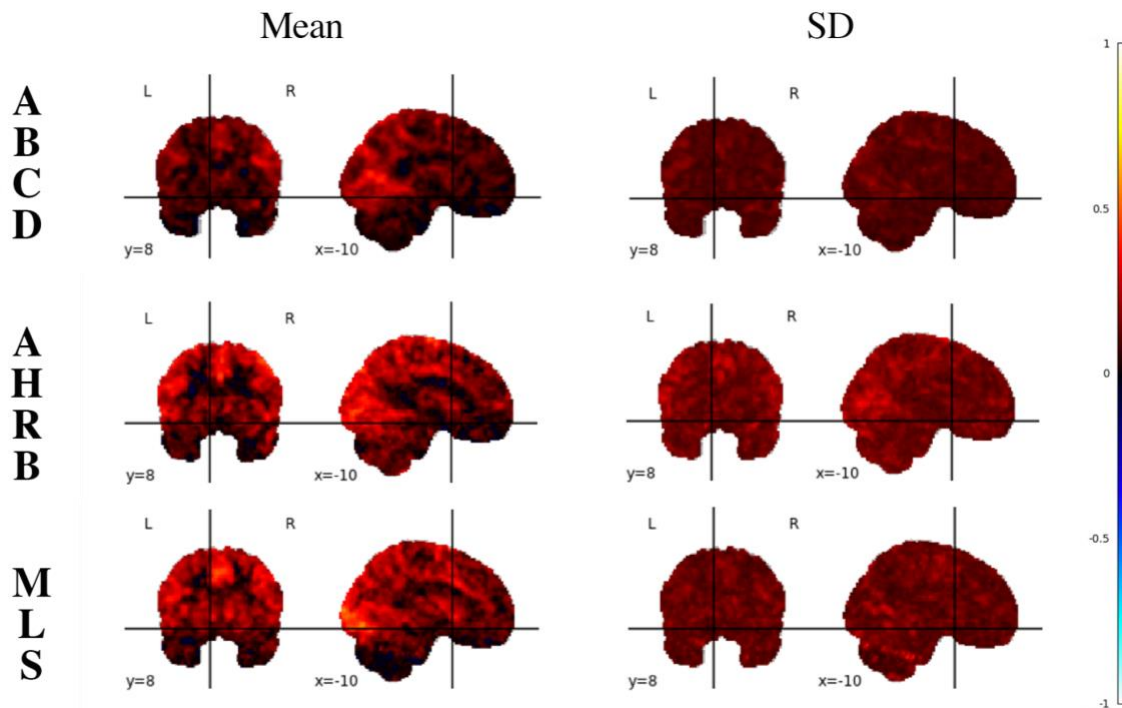
260
261
262
263
264

Figure S14: Voxelwise Distribution of ICCs for Supra- and Sub-threshold mask for highest (Top) and Lowest (Bottom) estimates from in-text Figure 3 and Figure S13 Across ABCD, AHRB and MLS samples.

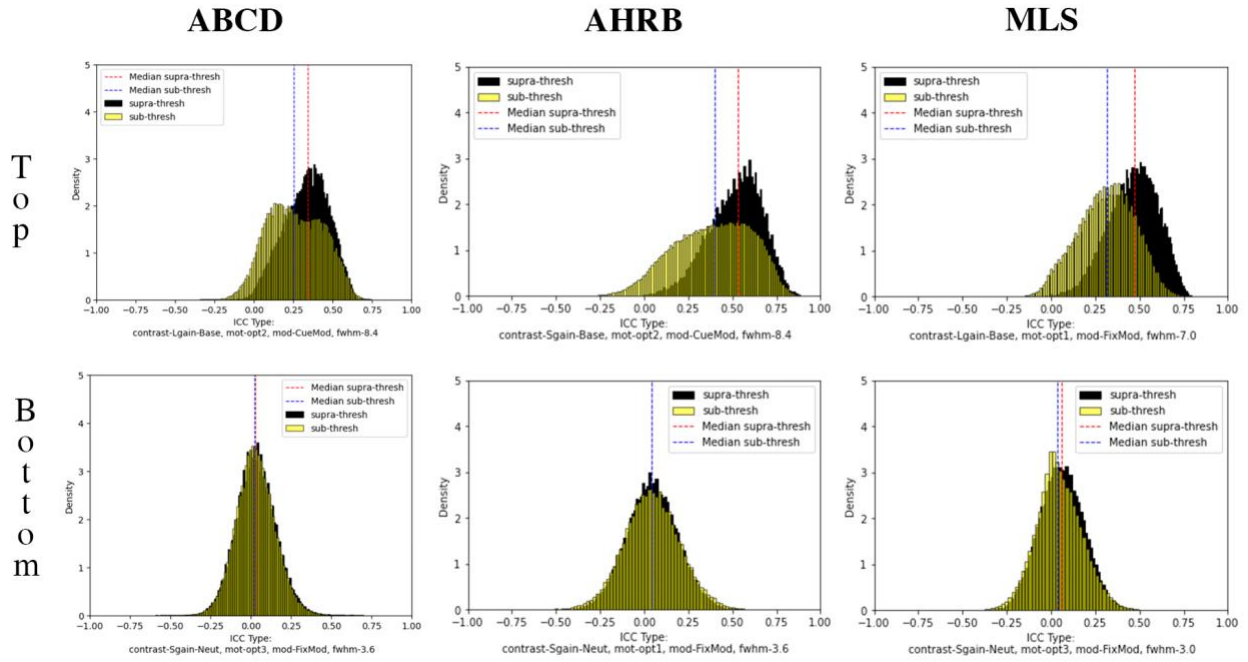
265

266 *A. Between-Session Individual Reliability:*

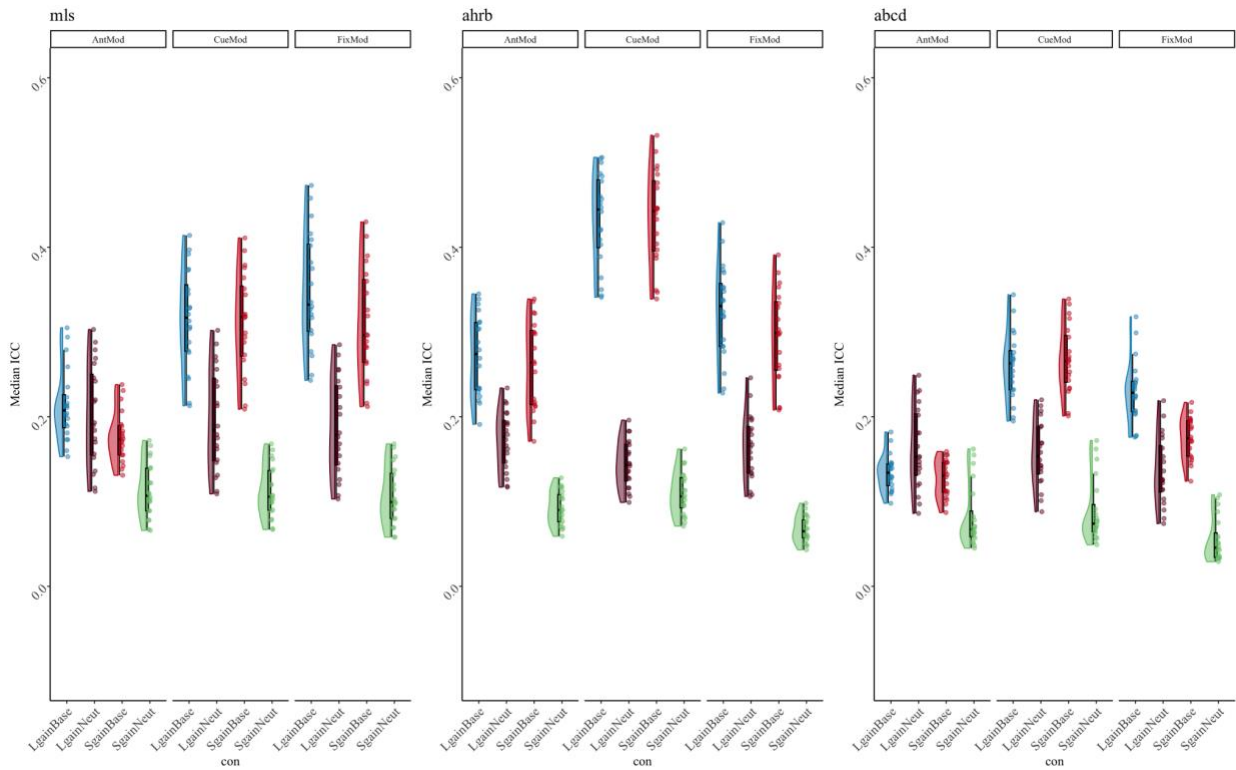
267 The mean and standard deviation (**Figure S15**) of the 3D volumes across the 240 analytic
 268 decisions illustrate a consistent pattern, whereby the highest nose is within CSF and high noise
 269 regions across the three samples. Consistent with the Session 1 between-run median ICC
 270 estimates, variability in the median ICC estimate across 240 pipelines and three samples is best
 271 explained by contrast (marginal ΔR^2 : .51) and model parameterization (marginal ΔR^2 : .07), see
 272 **Table S9**. Compared to the between-run, the FWHM had a higher impact on the between-session
 273 model fit (marginal ΔR^2 : .06) but motion remained negligible (marginal ΔR^2 : .02). Like the
 274 between-run estimates, the *Implicit Baseline* is the main contributor to the model
 275 parameterization differences (**Figure S17**).
 276



277 **Figure S15:** Mean and SD of ICC estimates across 240 permutations for the Adolescent Brain
 278 Cognitive Development (ABCD), Adolescent Health Risk Behavior (AHRB) and Michigan
 279 Longitudinal (MLS) 3D volumes.
 280
 281



282
283 **Figure S16.** Voxelwise Distribution of ICCs for Supra- and Sub-threshold mask for highest
284 (Top) and Lowest (Bottom) estimates from in-text Figure 3 and Figure S13 Across ABCD,
285 AHRB and MLS samples.



286
287 **Figure S17.** Supra-threshold Median ICC between-session reliability estimates for Contrast (con)
288 and Model Parameterization analytic options across the ABCD, AHRB and MLS samples.
289

290
291
292
293
294

Table S9. Hierarchical Linear Model: (A) Linear associations between the analytic decisions and the *Between Session* median Intraclass Correlation Coefficient (ICC[3,1]), *Between-subject (BS)* and *Within-subject variance (WS)* from **supra-threshold mask and (B) the impact of the analytic category on the marginal R².**

A. HLM Estimates for Supra-threshold Mask									
	Median ICC			Median BS			Median WS		
<i>Predictors</i>	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>
(Intercept)	.22	.18 – .26	<.001	.15	.11 – .20	<.001	.49	.39 – .60	<.001
Reference [3.6]									
fwhm [4.8]	.03	.01 – .04	<.001	-.01	-.03 – .00	.11	-.11	-.14 – -.09	<.001
fwhm [6.0]	.05	.03 – .06	<.001	-.02	-.04 – -.01	.01	-.18	-.20 – -.15	<.001
fwhm [7.2]	.06	.05 – .07	<.001	-.03	-.05 – -.01	<.001	-.22	-.24 – -.19	<.001
fwhm [8.4]	.07	.06 – .09	<.001	-.04	-.05 – -.02	<.001	-.25	-.27 – -.22	<.001
Reference [opt1]									
motion [opt2]	.00	-.01 – .01	.50	-.02	-.03 – -.00	.01	-.05	-.08 – -.03	<.001
motion [opt3]	-.03	-.04 – -.02	<.001	-.06	-.07 – -.04	<.001	-.12	-.14 – -.09	<.001
motion [opt4]	-.03	-.04 – -.02	<.001	-.06	-.07 – -.04	<.001	-.12	-.14 – -.10	<.001
Reference [AntMod]									
model [CueMod]	.07	.06 – .08	<.001	.08	.07 – .10	<.001	.17	.15 – .19	<.001
model [FixMod]	.03	.02 – .04	<.001	.07	.06 – .08	<.001	.16	.14 – .18	<.001
Reference [LgainBase]									
con [LgainNeut]	-.11	-.12 – -.10	<.001	-.12	-.13 – -.11	<.001	-.23	-.25 – -.20	<.001
con [SgainBase]	-.02	-.03 – -.01	.00	-.02	-.03 – -.00	.03	-.01	-.03 – .02	.52
con [SgainNeut]	-.19	-.20 – -.18	<.001	-.14	-.15 – -.12	<.001	-.24	-.26 – -.22	<.001

B. Analytic Category Model Impact												
Comparison	χ^2	Orig R2	New R2	ΔR^2	χ^2	Orig R2	New R2	ΔR^2	χ^2	Orig R2	New R2	ΔR^2
[Full] vs [New - fwhm]	159	.66	.60	.06	25	.49	.48	.01	336	.59	.43	.16
[Full] vs [New - motion]	65	.66	.64	.02	94	.49	.44	.05	126	.59	.54	.05
[Full] vs [New - model]	174	.66	.59	.07	185	.49	.38	.11	275	.59	.47	.12

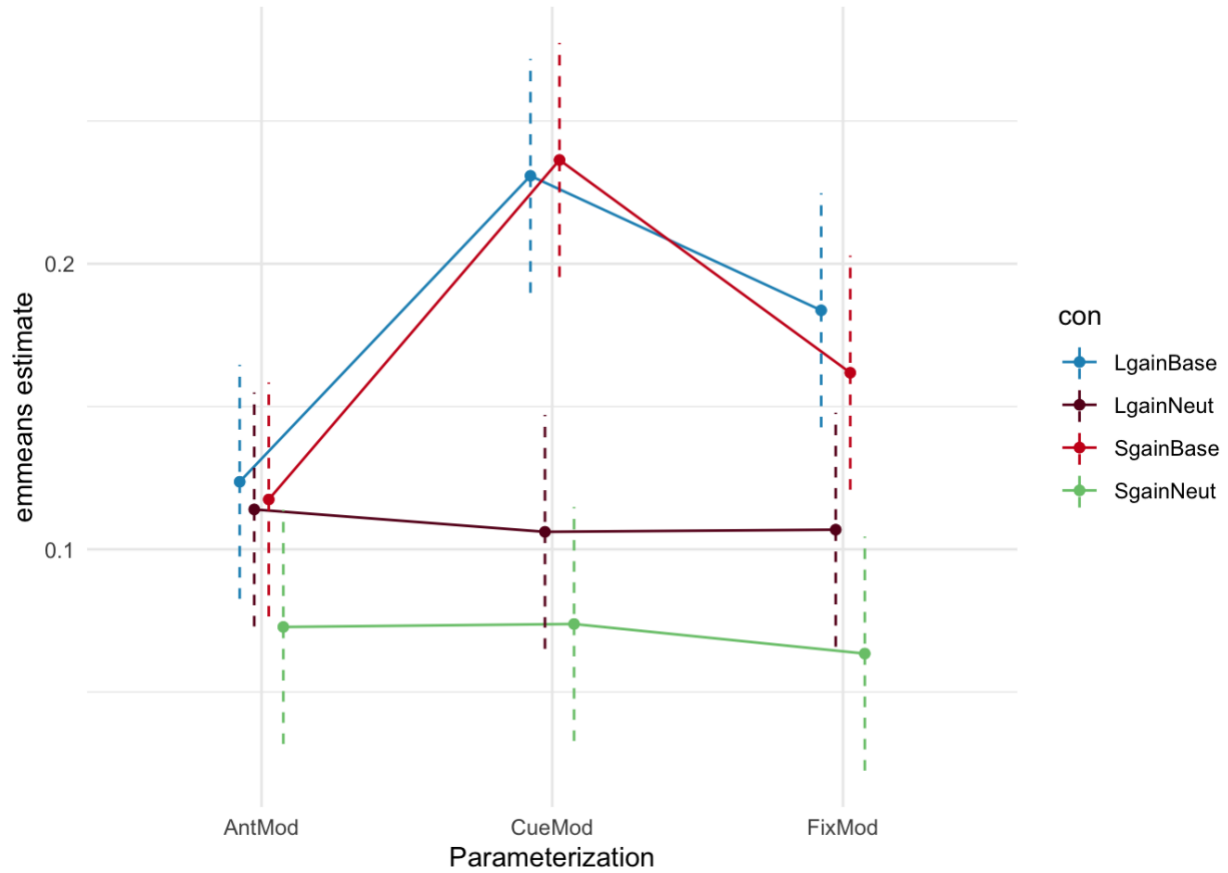
[Full] vs [New - con]	800	.66	.15	.51	421	.49	.18	.31	507	.59	.32	.27
--------------------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

295
296
297

298 **Table S10: Tukey's HSB Estimate Means Differences for Supra-threshold Between-session ICC**
 299 **Model Parameters in Table S9.**
 300

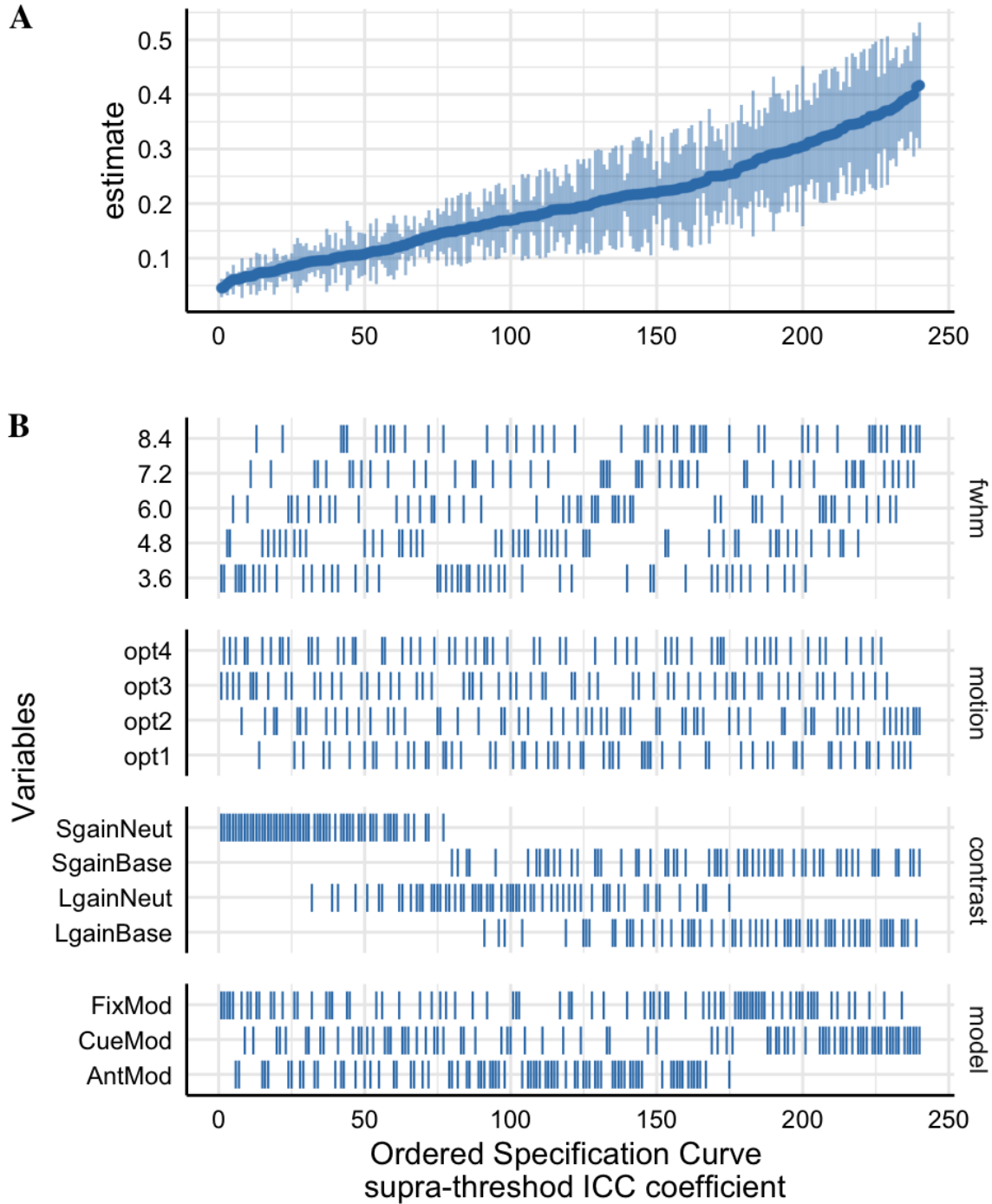
Contrast	Est	SE	Low.CI	Up.CI	<i>p</i>
fwhm3.6 - fwhm4.8	-.03	.01	-.04	-.01	.001
fwhm3.6 - fwhm6.0	-.05	.01	-.06	-.03	.000
fwhm3.6 - fwhm7.2	-.06	.01	-.08	-.05	.000
fwhm3.6 - fwhm8.4	-.07	.01	-.09	-.06	.000
fwhm4.8 - fwhm6.0	-.02	.01	-.04	.00	.009
fwhm4.8 - fwhm7.2	-.04	.01	-.05	-.02	.000
fwhm4.8 - fwhm8.4	-.05	.01	-.07	-.03	.000
fwhm6.0 - fwhm7.2	-.02	.01	-.03	.00	.089
fwhm6.0 - fwhm8.4	-.03	.01	-.04	-.01	.000
fwhm7.2 - fwhm8.4	-.01	.01	-.03	.01	.342
LgainBase - LgainNeut	.11	.01	.10	.13	.000
LgainBase - SgainBase	.02	.01	.00	.03	.005
LgainBase - SgainNeut	.19	.01	.18	.20	.000
LgainNeut - SgainBase	-.09	.01	-.11	-.08	.000
LgainNeut - SgainNeut	.08	.01	.06	.09	.000
SgainBase - SgainNeut	.17	.01	.16	.19	.000
opt1 - opt2	.00	.01	-.02	.01	.906
opt1 - opt3	.03	.01	.01	.04	.000
opt1 - opt4	.03	.01	.02	.05	.000
opt2 - opt3	.03	.01	.02	.05	.000
opt2 - opt4	.04	.01	.02	.05	.000
opt3 - opt4	.00	.01	-.01	.02	.922
AntMod - CueMod	-.07	.00	-.08	-.06	.000
AntMod - FixMod	-.03	.00	-.04	-.02	.000
CueMod - FixMod	.04	.00	.02	.05	.000

301



302
303
304
305
306

Figure S18: Interaction plot of *emmeans* fitted model of Contrast-by-Model parameterization for Between-session supra-threshold median ICC estimates using *emmip()*. Point estimate is a linear estimate from *emmeans* function. Dashed bars are estimated confidence intervals by *emmeans*.



307
308 **Figure S19:** The supra-threshold Specification Curve of the Between-Session Median ICC
309 estimates across 240 pipeline permutations for the ABCD, AHRB and MLS estimate.
310 A. The distribution of the point estimate (average) across the three studies and distribution across the three samples.
311 B. The model options (four) associated with each estimate.

312
313
314
315
316

Table S11. Hierarchical Linear Model: (A) Linear associations between the analytic decisions and the *Between Session* median Intraclass Correlation Coefficient (ICC[3,1]), Between-subject (BS) and Within-subject variance (WS) from **sub-threshold mask** and (B) the impact of the analytic category on the marginal R².

A. HLM Estimates for Sub-threshold Mask												
	Median ICC			Median BS			Median WS					
<i>Predictors</i>	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>			
(Intercept)	.13	.10 – .16	<.001	.14	.10 – .19	<.001	.84	.67 – 1.02	<.001			
Reference [3.6]												
fwhm [4.8]	.02	.01 – .03	<.001	-.01	-.02 – .01	.24	-.19	-.23 – -.15	<.001			
fwhm [6.0]	.04	.03 – .05	<.001	-.02	-.03 – -.00	.04	-.30	-.34 – -.26	<.001			
fwhm [7.2]	.05	.04 – .06	<.001	-.02	-.04 – -.01	.01	-.37	-.41 – -.33	<.001			
fwhm [8.4]	.07	.06 – .08	<.001	-.03	-.04 – -.01	.00	-.41	-.46 – -.37	<.001			
Reference [opt1]												
motion [opt2]	.00	-.01 – .01	.87	-.02	-.04 – -.01	.00	-.11	-.15 – -.07	<.001			
motion [opt3]	-.03	-.03 – -.02	<.001	-.07	-.08 – -.05	<.001	-.22	-.26 – -.18	<.001			
motion [opt4]	-.03	-.03 – -.02	<.001	-.07	-.08 – -.05	<.001	-.22	-.26 – -.18	<.001			
Reference [AntMod]												
model [CueMod]	.05	.05 – .06	<.001	.09	.08 – .11	<.001	.26	.22 – .29	<.001			
model [FixMod]	.02	.01 – .03	<.001	.06	.05 – .07	<.001	.25	.21 – .28	<.001			
Reference [LgainBase]												
con [LgainNeut]	-.07	-.08 – -.06	<.001	-.12	-.13 – -.10	<.001	-.37	-.41 – -.33	<.001			
con [SgainBase]	-.01	-.02 – .00	.07	-.01	-.02 – .00	.11	-.02	-.06 – .02	.41			
con [SgainNeut]	-.11	-.12 – -.10	<.001	-.13	-.14 – -.12	<.001	-.39	-.43 – -.35	<.001			
B. Analytic Category Model Impact												
Comparison	χ^2	Orig R2	New R2	ΔR^2	χ^2	Orig R2	New R2	ΔR^2	χ^2	Orig R2	New R2	ΔR^2
[Full] vs [New - fwhm]	225	.62	.51	.11	14	.51	.50	.01	343	.58	.42	.16
[Full] vs [New - motion]	79	.62	.59	.03	122	.51	.44	.07	153	.58	.52	.06
[Full] vs [New - model]	205	.62	.52	.10	216	.51	.38	.13	236	.58	.48	.10

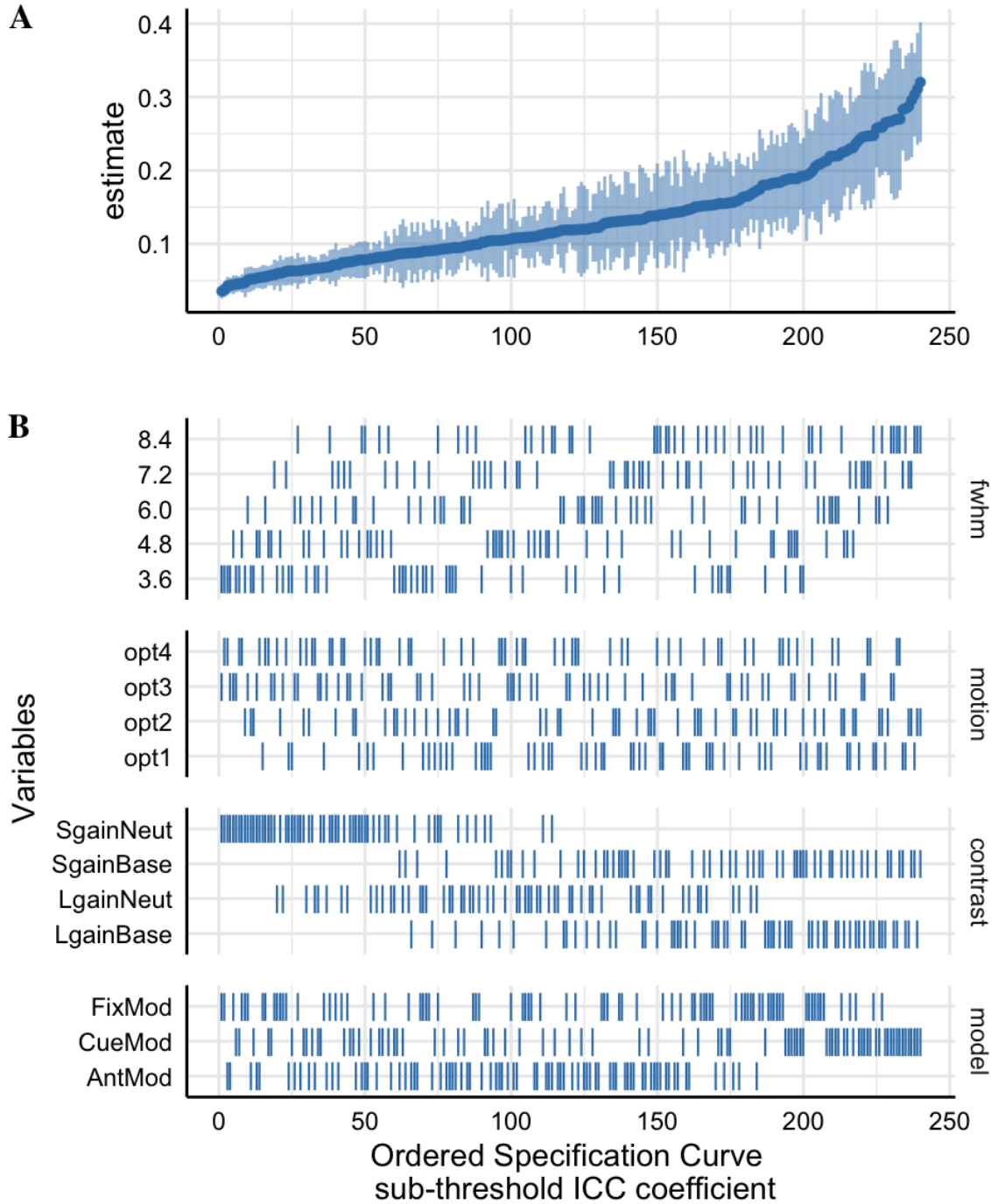
[Full] vs [New - con]	609	.62	.24	.38	424	.51	.21	.30	484	.58	.33	.25
--------------------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

317

318 **Table S12: Tukey's HSB Estimate Mean Differences for Sub-threshold Between-session ICC**
 319 **Model Parameters in Table S11.**

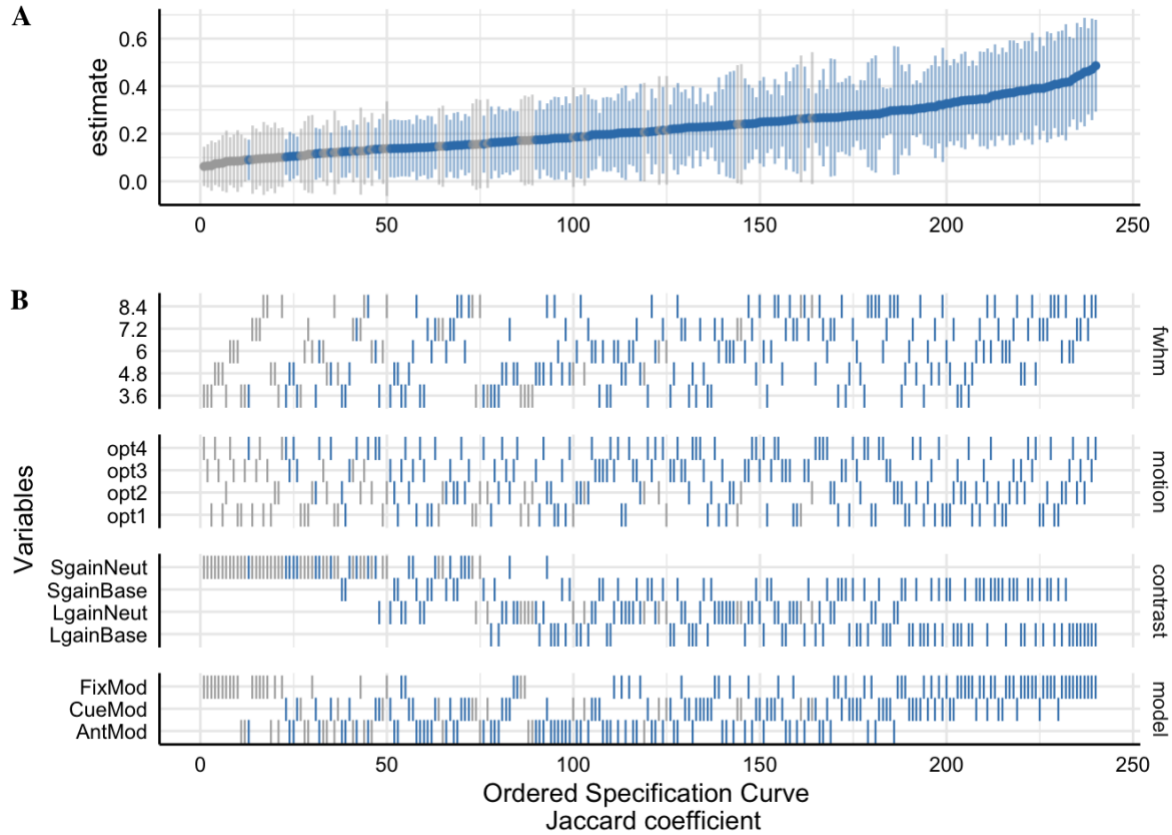
Contrast	Est	SE	Low.CI	Up.CI	<i>p</i>
fwhm3.6 - fwhm4.8	-.02	.00	-.03	-.01	.000
fwhm3.6 - fwhm6.0	-.04	.00	-.05	-.03	.000
fwhm3.6 - fwhm7.2	-.05	.00	-.07	-.04	.000
fwhm3.6 - fwhm8.4	-.07	.00	-.08	-.05	.000
fwhm4.8 - fwhm6.0	-.02	.00	-.03	-.01	.001
fwhm4.8 - fwhm7.2	-.03	.00	-.05	-.02	.000
fwhm4.8 - fwhm8.4	-.05	.00	-.06	-.03	.000
fwhm6.0 - fwhm7.2	-.02	.00	-.03	.00	.008
fwhm6.0 - fwhm8.4	-.03	.00	-.04	-.02	.000
fwhm7.2 - fwhm8.4	-.01	.00	-.03	.00	.042
LgainBase - LgainNeut	.07	.00	.06	.08	.000
LgainBase - SgainBase	.01	.00	.00	.02	.274
LgainBase - SgainNeut	.11	.00	.10	.12	.000
LgainNeut - SgainBase	-.06	.00	-.07	-.05	.000
LgainNeut - SgainNeut	.04	.00	.03	.05	.000
SgainBase - SgainNeut	.10	.00	.09	.11	.000
opt1 - opt2	.00	.00	-.01	.01	.999
opt1 - opt3	.03	.00	.02	.04	.000
opt1 - opt4	.03	.00	.02	.04	.000
opt2 - opt3	.03	.00	.02	.04	.000
opt2 - opt4	.03	.00	.02	.04	.000
opt3 - opt4	.00	.00	-.01	.01	1.000
AntMod - CueMod	-.05	.00	-.06	-.05	.000
AntMod - FixMod	-.02	.00	-.03	-.01	.000
CueMod - FixMod	.03	.00	.02	.04	.000

320



321
322 **Figure S20:** The sub-threshold Specification Curve of the Between-Session Median ICC
323 estimates across 240 pipeline permutations for the ABCD, AHRB and MLS estimate.
324 A. The distribution of the point estimate (average) across the three studies and distribution across the three samples.
325 B. The model options (four) associated with each estimate.

326 *B. Between-Run Group Reliability:*



327

328 **Figure S21: The Specification Curve of the Session 1 Between-run Jaccard Similarity estimates**
329 **across 240 pipeline permutations for the ABCD, AHRB and MLS samples.**

330 A. The distribution of the point estimate (average) across the three studies and distribution across the three samples.

331 B. The model options (four) associated with each estimate.

332

333 **Table S13: Tukey's HSB Estimate Means Differences for (A) Jaccard and (B) Spearman Model**
334 **Parameters in-text Table 4.**

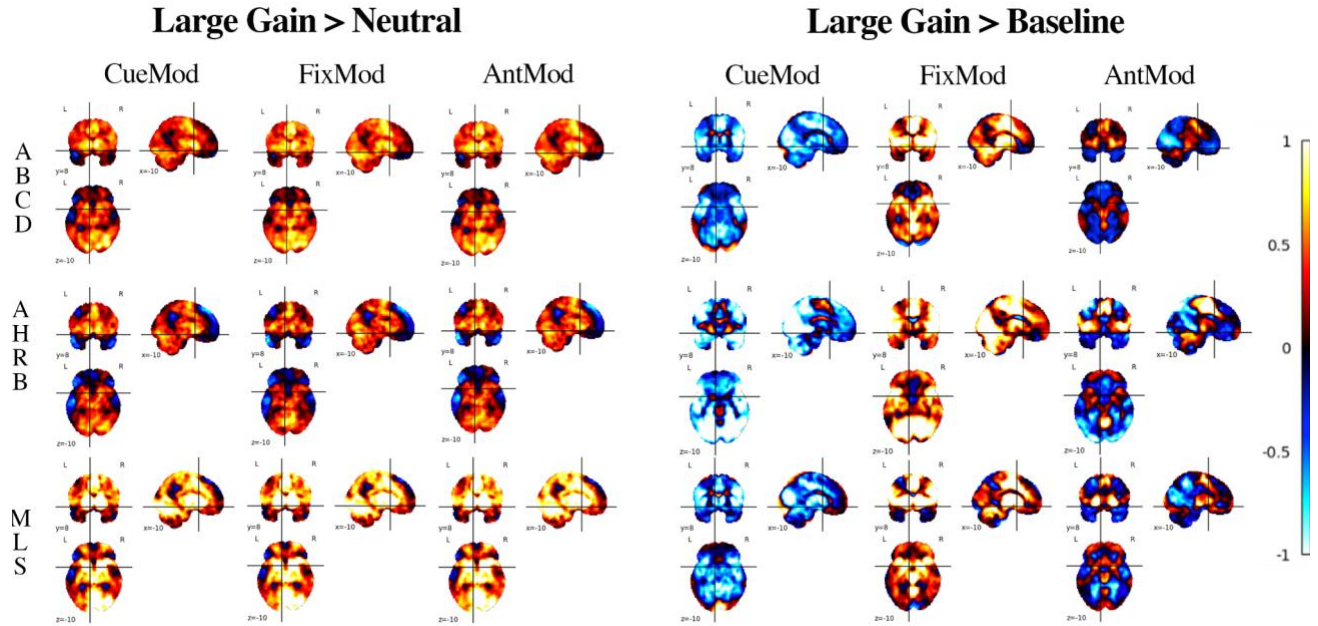
Contrast	Est	SE	Low.CI	Up.CI	<i>p</i>
A. Jaccard Similarity					
fwhm3.6 - fwhm4.8	-.03	.01	-.06	.00	.037
fwhm3.6 - fwhm6	-.05	.01	-.08	-.02	.000
fwhm3.6 - fwhm7.2	-.07	.01	-.10	-.04	.000
fwhm3.6 - fwhm8.4	-.08	.01	-.11	-.06	.000
fwhm4.8 - fwhm6	-.02	.01	-.05	.01	.171
fwhm4.8 - fwhm7.2	-.04	.01	-.07	-.01	.001
fwhm4.8 - fwhm8.4	-.05	.01	-.08	-.03	.000
fwhm6 - fwhm7.2	-.02	.01	-.05	.01	.448
fwhm6 - fwhm8.4	-.03	.01	-.06	.00	.031
fwhm7.2 - fwhm8.4	-.01	.01	-.04	.02	.737
LgainBase - LgainNeut	.09	.01	.06	.11	.000
LgainBase - SgainBase	.03	.01	.01	.05	.008
LgainBase - SgainNeut	.18	.01	.16	.21	.000
LgainNeut - SgainBase	-.05	.01	-.08	-.03	.000
LgainNeut - SgainNeut	.10	.01	.07	.12	.000
SgainBase - SgainNeut	.15	.01	.13	.18	.000
opt1 - opt2	-.01	.01	-.04	.01	.437
opt1 - opt3	.00	.01	-.02	.03	.998
opt1 - opt4	.00	.01	-.03	.02	.979
opt2 - opt3	.02	.01	-.01	.04	.332
opt2 - opt4	.01	.01	-.01	.03	.687
opt3 - opt4	-.01	.01	-.03	.02	.938
AntMod - CueMod	-.05	.01	-.07	-.03	.000
AntMod - FixMod	-.08	.01	-.10	-.07	.000
CueMod - FixMod	-.03	.01	-.05	-.01	.000
B. Spearman Supra-threshold Similarity					
fwhm3.6 - fwhm4.8	-.05	.01	-.07	-.03	.000
fwhm3.6 - fwhm6	-.09	.01	-.11	-.07	.000
fwhm3.6 - fwhm7.2	-.11	.01	-.14	-.09	.000
fwhm3.6 - fwhm8.4	-.13	.01	-.16	-.11	.000
fwhm4.8 - fwhm6	-.04	.01	-.06	-.02	.000
fwhm4.8 - fwhm7.2	-.06	.01	-.09	-.04	.000

fwhm4.8 - fwhm8.4	-.08	.01	-.11	-.06	.000
fwhm6 - fwhm7.2	-.03	.01	-.05	.00	.008
fwhm6 - fwhm8.4	-.05	.01	-.07	-.02	.000
fwhm7.2 - fwhm8.4	-.02	.01	-.04	.00	.107
LgainBase - LgainNeut	.20	.01	.18	.22	.000
LgainBase - SgainBase	.01	.01	-.01	.03	.531
LgainBase - SgainNeut	.34	.01	.32	.36	.000
LgainNeut - SgainBase	-.19	.01	-.21	-.17	.000
LgainNeut - SgainNeut	.14	.01	.12	.16	.000
SgainBase - SgainNeut	.33	.01	.31	.35	.000
opt1 - opt2	-.01	.01	-.03	.00	.217
opt1 - opt3	-.01	.01	-.03	.01	.578
opt1 - opt4	-.01	.01	-.03	.01	.305
opt2 - opt3	.00	.01	-.01	.02	.915
opt2 - opt4	.00	.01	-.02	.02	.998
opt3 - opt4	.00	.01	-.02	.02	.967
AntMod - CueMod	-.02	.01	-.04	-.01	.001
AntMod - FixMod	-.01	.01	-.02	.01	.384
CueMod - FixMod	.01	.01	.00	.03	.054

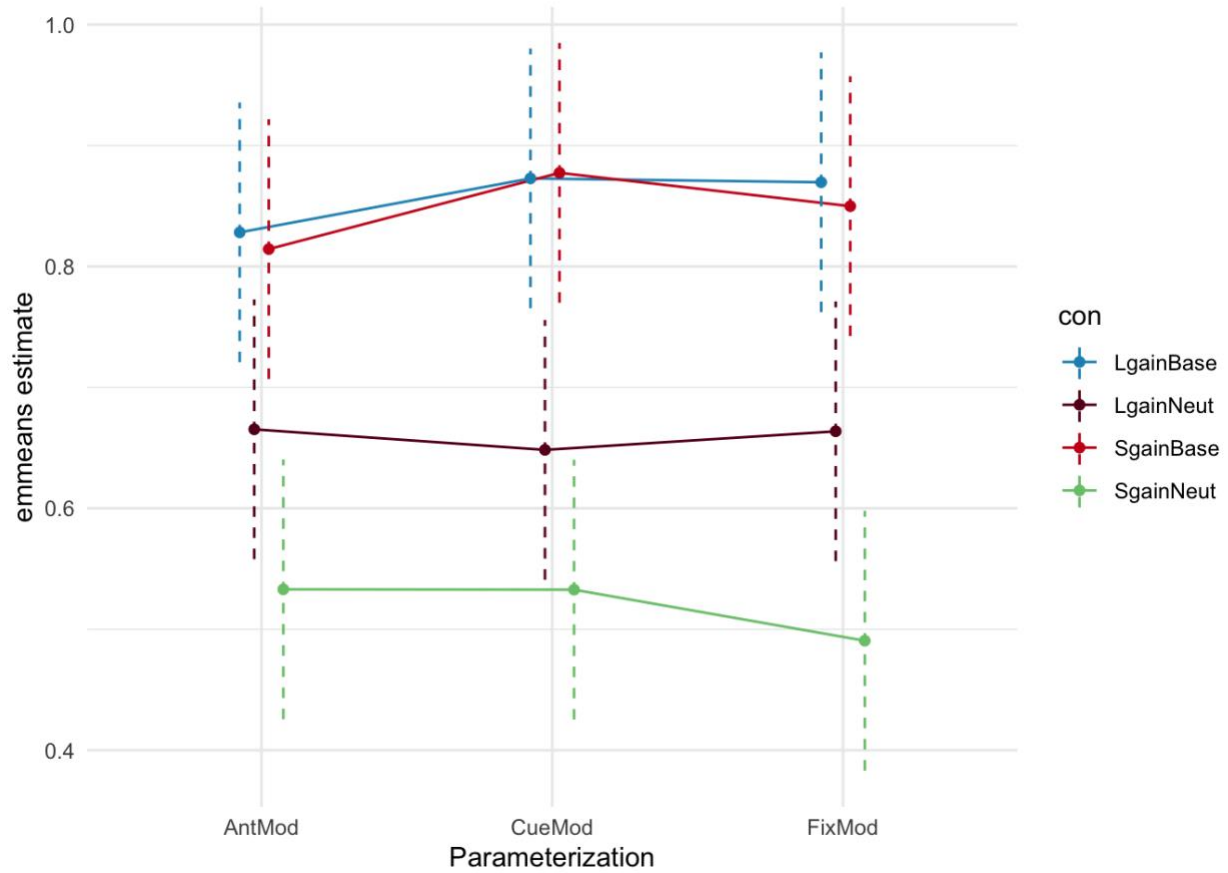
335

336

337



338
339 **Figure S22: Comparing Lgain-Neut & Lgain-Base contrasts for Session 1 run average group activity**
340 **for Cue, Fixation and Anticipation Parameterization for Motion opt2 and FWHM 8.4 (MLS 7.0)**
341 **across ABCD, AHRB and MLS samples.**
342 Note: For quick access on NeuroVault, example image search: “*_type-session_contrast-Lgain-Base_mask-mni152_mot-*
343 *opt2_mod-CueMod_fwhm-8.4_stat-cohensd.nii.gz*”



344
345 *Figure S23: Spearman rho: Interaction plot of emmeans fitted model of Contrast-by-Model*
346 *parameterization for Between-run supra-threshold Spearman Similarity estimates using emmip().*
347 *Point estimate is a linear spearman rho estimate from emmeans function. Dashed bars are*
348 *estimated confidence intervals by emmeans.*

349 *B. Between-Session Group Reliability:*

350 **Table S14. Hierarchical Linear Model: (A) Linear associations between the analytic decisions**
 351 **and the Jaccard and Spearman supra-threshold mask between-session similarity and (B) the**
 352 **impact of the analytic category on the marginal R².**

A. HLM Group-map Estimates								
<i>Predictors</i>	Jaccard			Spearman				
	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>		
(Intercept)	.29	.20 – .38	<.001	.82	.76 – .87	<.001		
Reference [3.6]								
fwhm [4.8]	.04	.02 – .06	<.001	.04	.03 – .06	<.001		
fwhm [6.0]	.07	.05 – .10	<.001	.07	.05 – .08	<.001		
fwhm [7.2]	.10	.08 – .12	<.001	.09	.07 – .10	<.001		
fwhm [8.4]	.12	.10 – .14	<.001	.10	.08 – .12	<.001		
Reference [opt1]								
motion [opt2]	.04	.02 – .06	<.001	.03	.02 – .04	<.001		
motion [opt3]	.03	.01 – .05	.00	.05	.03 – .06	<.001		
motion [opt4]	.04	.02 – .06	<.001	.05	.04 – .06	<.001		
Reference [AntMod]								
model [CueMod]	.00	-.01 – .02	.64	-.01	-.02 – .00	.12		
model [FixMod]	.10	.08 – .12	<.001	-.01	-.02 – .01	.31		
Reference [LgainBase]								
con [LgainNeut]	-.06	-.08 – -.04	<.001	-.15	-.16 – -.14	<.001		
con [SgainBase]	-.04	-.06 – -.02	<.001	-.01	-.03 – -.00	.05		
con [SgainNeut]	-.24	-.26 – -.22	<.001	-.32	-.34 – -.31	<.001		
B. Analytic Category Model Impact								
Comparison	χ^2	Orig R2	New R2	ΔR^2	χ^2	Orig R2	New R2	ΔR^2
[Full] vs [New - fwhm]	124	.47	.40	.07	184	.74	.69	.05
[Full] vs [New - motion]	22	.47	.45	.02	61	.74	.73	.01
[Full] vs [New - model]	149	.47	.39	.08	3	.74	.74	.00
[Full] vs [New - con]	468	.47	.15	.32	1141	.74	.07	.67

353
354
355

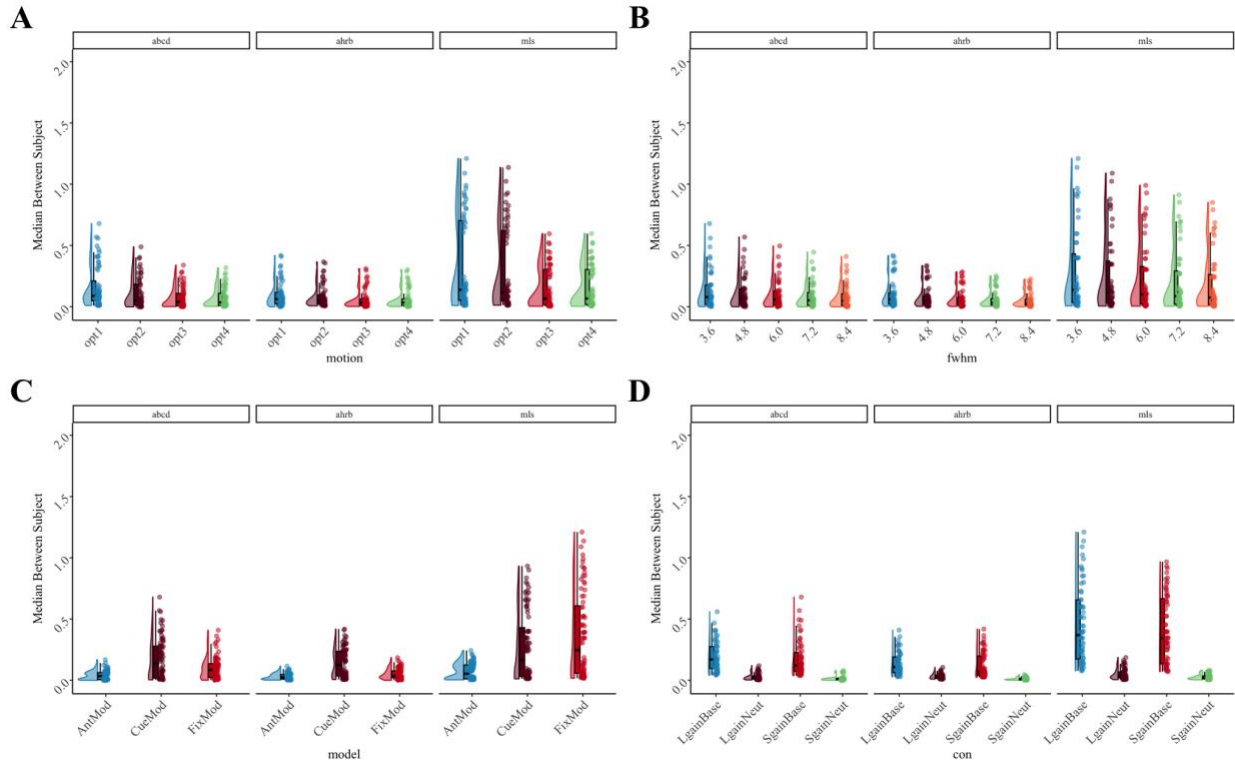
356 **Table S15: Tukey's HSB Estimate Means Differences for (A) Jaccard and (B) Spearman Model**
357 **Parameters in-text Table S14.**

Contrast	Est	SE	Low.CI	Up.CI	<i>p</i>
A. Jaccard Similarity					
fwhm3.6 - fwhm4.8	-.04	.01	-.07	-.01	.003
fwhm3.6 - fwhm6	-.07	.01	-.11	-.04	.000
fwhm3.6 - fwhm7.2	-.10	.01	-.13	-.07	.000
fwhm3.6 - fwhm8.4	-.12	.01	-.15	-.09	.000
fwhm4.8 - fwhm6	-.03	.01	-.06	.00	.040
fwhm4.8 - fwhm7.2	-.06	.01	-.09	-.03	.000
fwhm4.8 - fwhm8.4	-.08	.01	-.11	-.04	.000
fwhm6 - fwhm7.2	-.02	.01	-.06	.01	.209
fwhm6 - fwhm8.4	-.04	.01	-.08	-.01	.002
fwhm7.2 - fwhm8.4	-.02	.01	-.05	.01	.455
LgainBase - LgainNeut	.06	.01	.03	.08	.000
LgainBase - SgainBase	.04	.01	.01	.06	.001
LgainBase - SgainNeut	.24	.01	.21	.27	.000
LgainNeut - SgainBase	-.02	.01	-.04	.01	.338
LgainNeut - SgainNeut	.18	.01	.16	.21	.000
SgainBase - SgainNeut	.20	.01	.18	.23	.000
opt1 - opt2	-.04	.01	-.07	-.02	.000
opt1 - opt3	-.03	.01	-.06	.00	.013
opt1 - opt4	-.04	.01	-.07	-.01	.001
opt2 - opt3	.01	.01	-.01	.04	.654
opt2 - opt4	.00	.01	-.02	.03	.976
opt3 - opt4	-.01	.01	-.03	.02	.880
AntMod - CueMod	.00	.01	-.03	.02	.886
AntMod - FixMod	-.10	.01	-.12	-.08	.000
CueMod - FixMod	-.10	.01	-.12	-.08	.000
B. Spearman Supra-threshold Similarity					
fwhm3.6 - fwhm4.8	-.04	.01	-.06	-.02	.000
fwhm3.6 - fwhm6	-.07	.01	-.09	-.05	.000
fwhm3.6 - fwhm7.2	-.09	.01	-.11	-.07	.000
fwhm3.6 - fwhm8.4	-.10	.01	-.12	-.08	.000
fwhm4.8 - fwhm6	-.03	.01	-.05	-.01	.004
fwhm4.8 - fwhm7.2	-.05	.01	-.07	-.02	.000
fwhm4.8 - fwhm8.4	-.06	.01	-.08	-.04	.000

fwhm6 - fwhm7.2	-.02	.01	-.04	.00	.119
fwhm6 - fwhm8.4	-.03	.01	-.05	-.01	.001
fwhm7.2 - fwhm8.4	-.01	.01	-.03	.01	.463
LgainBase - LgainNeut	.15	.01	.13	.17	.000
LgainBase - SgainBase	.01	.01	.00	.03	.196
LgainBase - SgainNeut	.32	.01	.31	.34	.000
LgainNeut - SgainBase	-.14	.01	-.16	-.12	.000
LgainNeut - SgainNeut	.17	.01	.15	.19	.000
SgainBase - SgainNeut	.31	.01	.29	.33	.000
opt1 - opt2	-.03	.01	-.05	-.01	.000
opt1 - opt3	-.05	.01	-.06	-.03	.000
opt1 - opt4	-.05	.01	-.07	-.03	.000
opt2 - opt3	-.02	.01	-.03	.00	.106
opt2 - opt4	-.02	.01	-.04	.00	.024
opt3 - opt4	.00	.01	-.02	.01	.943
AntMod - CueMod	.01	.01	.00	.02	.265
AntMod - FixMod	.01	.01	-.01	.02	.568
CueMod - FixMod	.00	.01	-.02	.01	.850

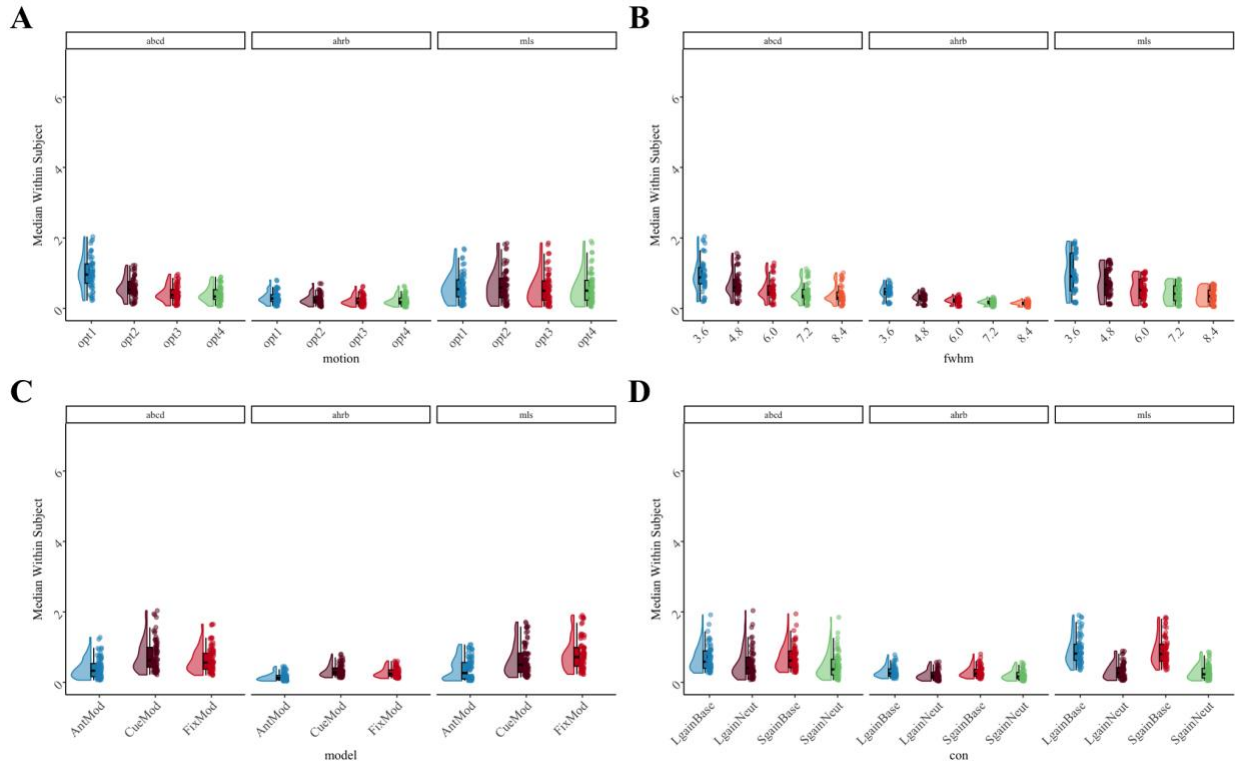
359 **2.4 Aim 2 results**

360 *Between-Run Reliability:*



361

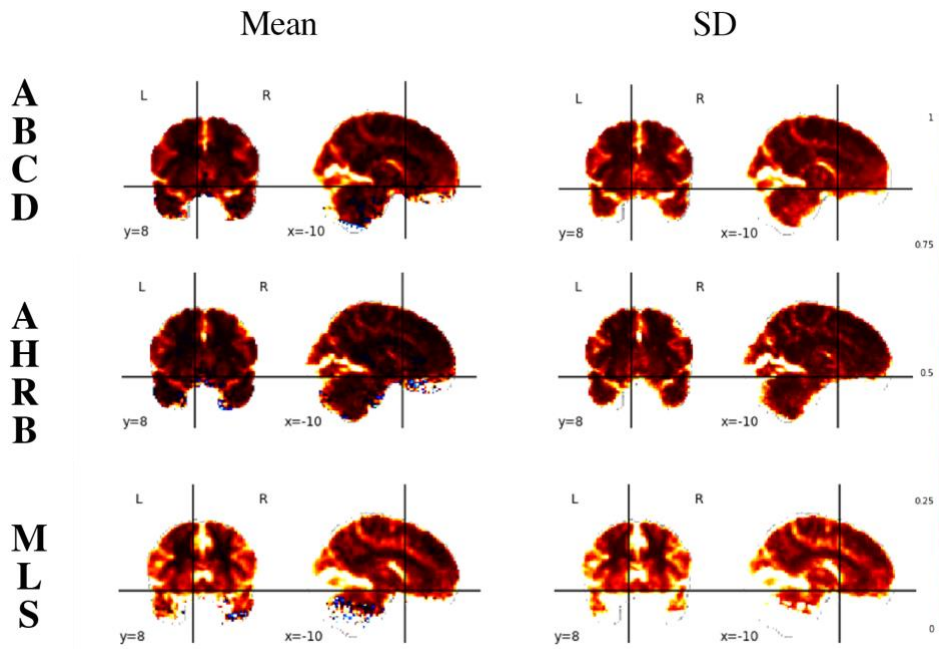
362 **Figure S24. Session 1 Between-run: Supra-threshold Median **Between-subject variance****
 363 **estimates across (A) Motion, (B) FWHM, (C) Model Parameterization and (D) Contrast analytic**
 364 **options for between-run reliability across the ABCD, AHRB and MLS samples.**
 365



366
367
368
369
370

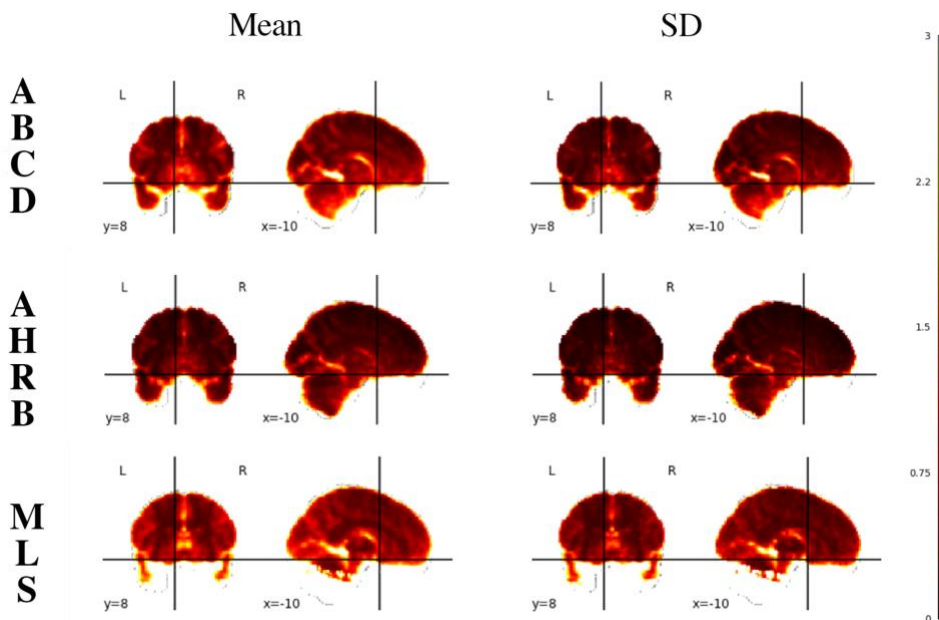
Figure S25. Session 1 Between-run: Supra-threshold Median Within-subject variance estimates across (A) Motion, (B) FWHM, (C) Model Parameterization and (D) Contrast analytic options for between-run reliability across the ABCD, AHRB and MLS samples.

371



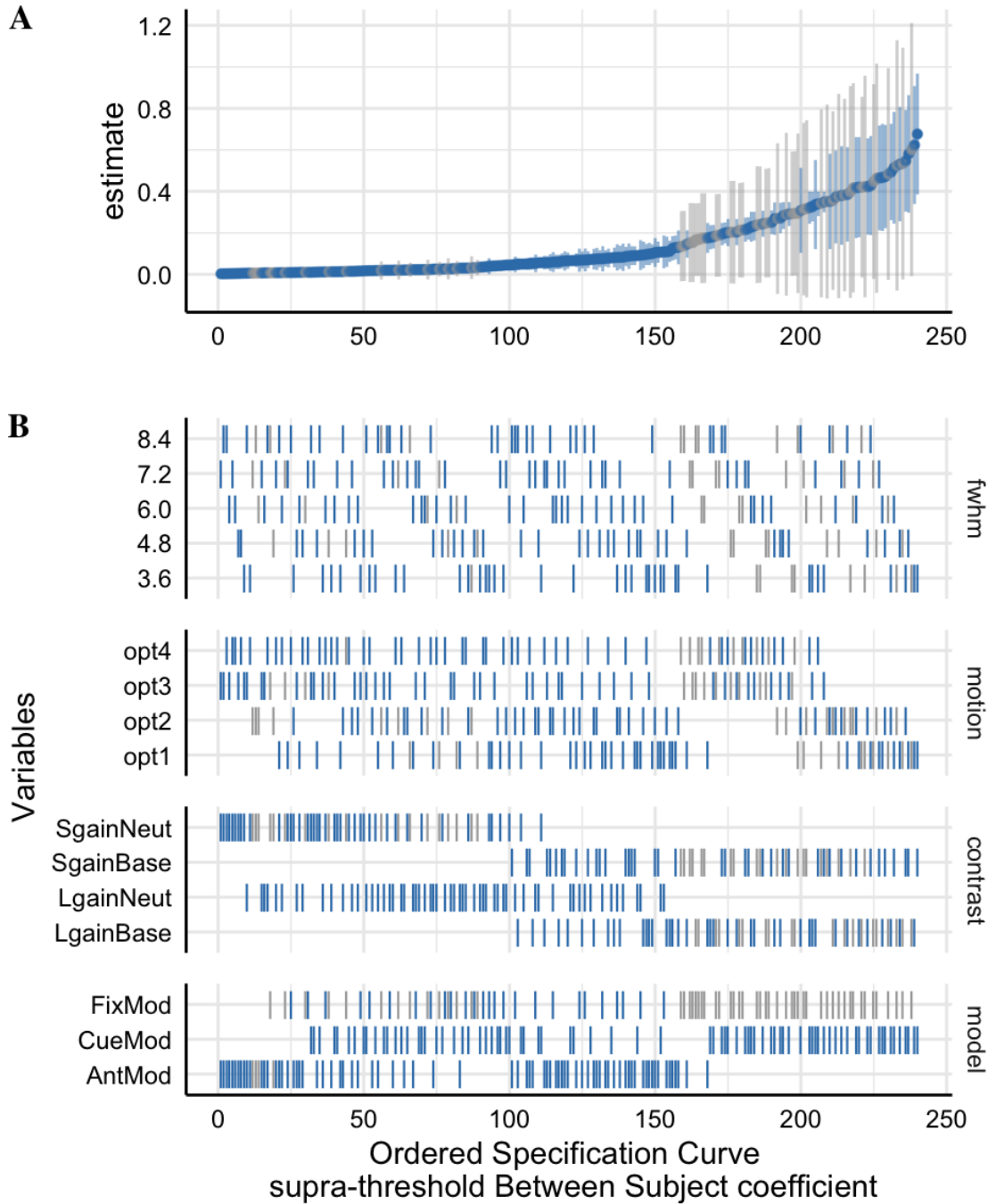
372
373
374
375
376

Figure S26: Mean and SD of Between-subject variance (σ_r^2) estimates across 240 permutations for the Adolescent Brain Cognitive Development (ABCD), Adolescent Health Risk Behavior (AHRB) and Michigan Longitudinal (MLS) 3D volumes.



377
378
379
380
381

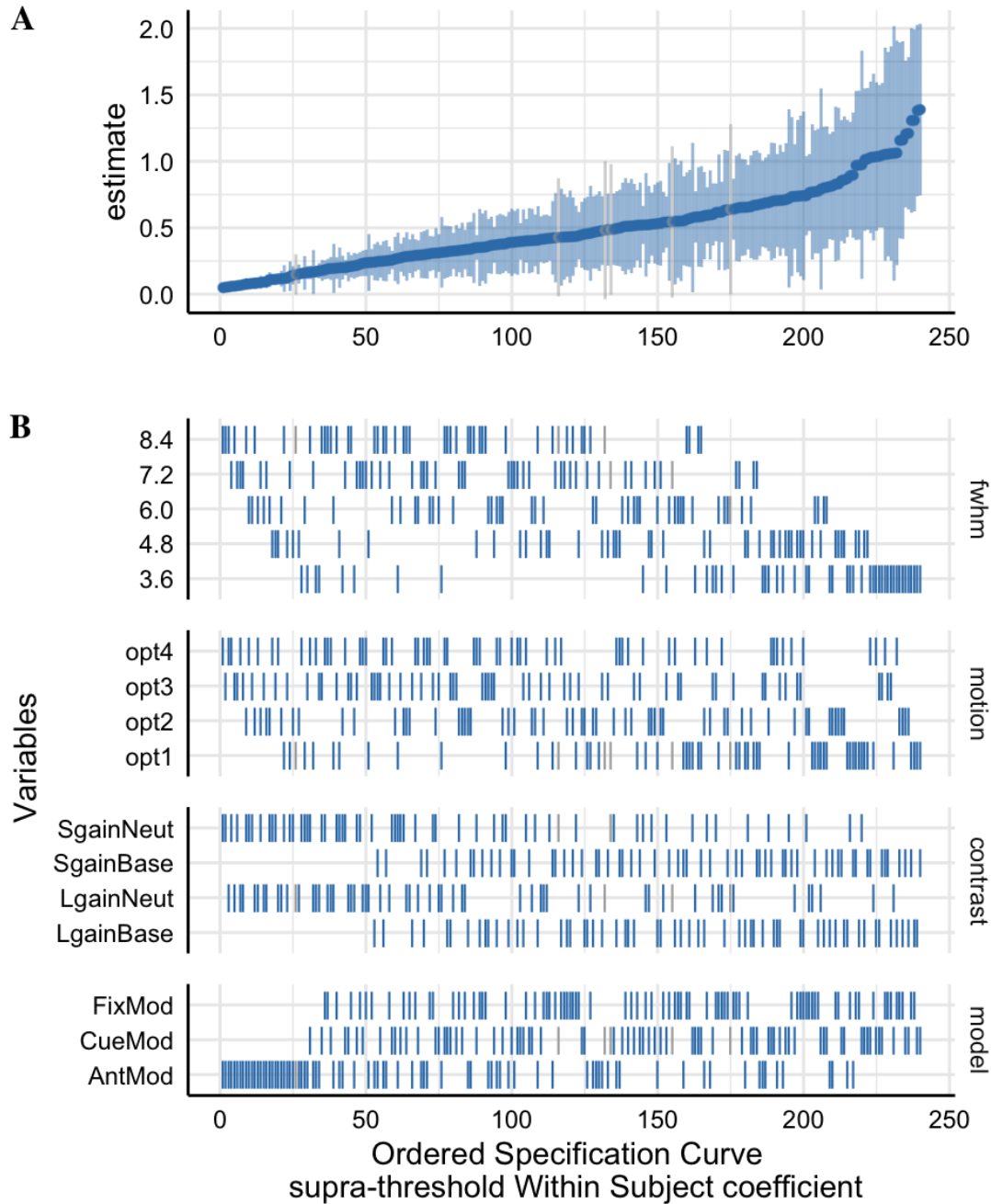
Figure S27: Mean and SD of Within-subject variance estimates (σ_v^2) across 240 permutations for the Adolescent Brain Cognitive Development (ABCD), Adolescent Health Risk Behavior (AHRB) and Michigan Longitudinal (MLS) 3D volumes.



382
383
384
385
386
387
388

Figure S28: Session 1 Between-run: The supra-threshold Specification Curve of the Median Between-subject variance (σ_r^2) estimates across 240 pipeline permutations for the ABCD, AHRB and MLS estimate.

A. The distribution of the point estimate (average) across the three studies and distribution across the three samples.
B. The model options (four) associated with each estimate.



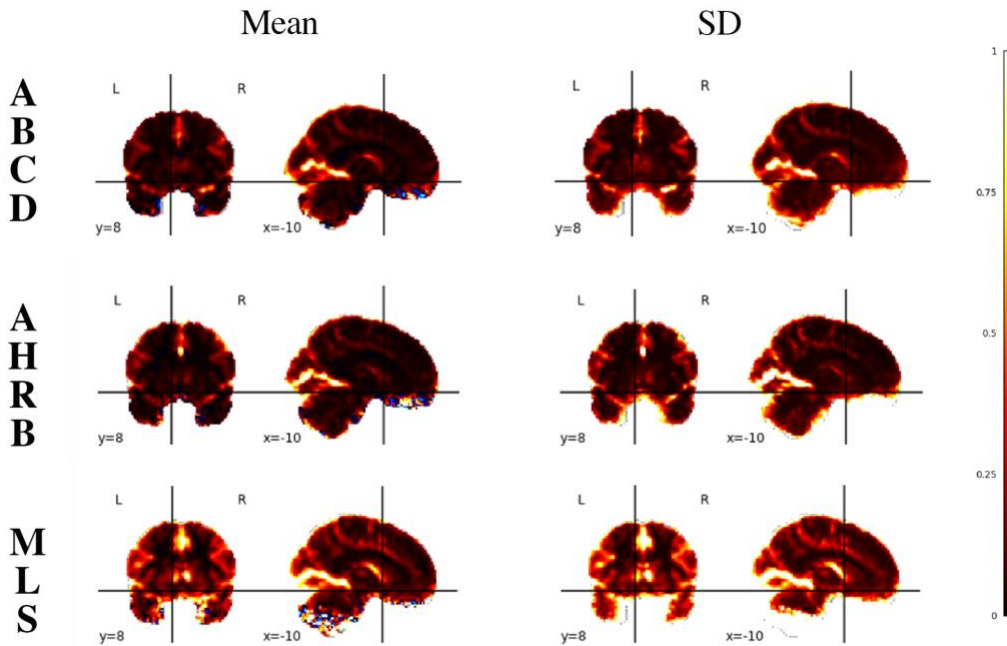
389
390 **Figure S29: Session 1 Between-run: The supra-threshold Specification Curve of the Median**
391 **Within-subject variance (σ_v^2) estimates across 240 pipeline permutations for the ABCD, AHRB**
392 **and MLS estimate.**

393 A. The distribution of the point estimate (average) across the three studies and distribution across the three samples.

394 B. The model options (four) associated with each estimate.

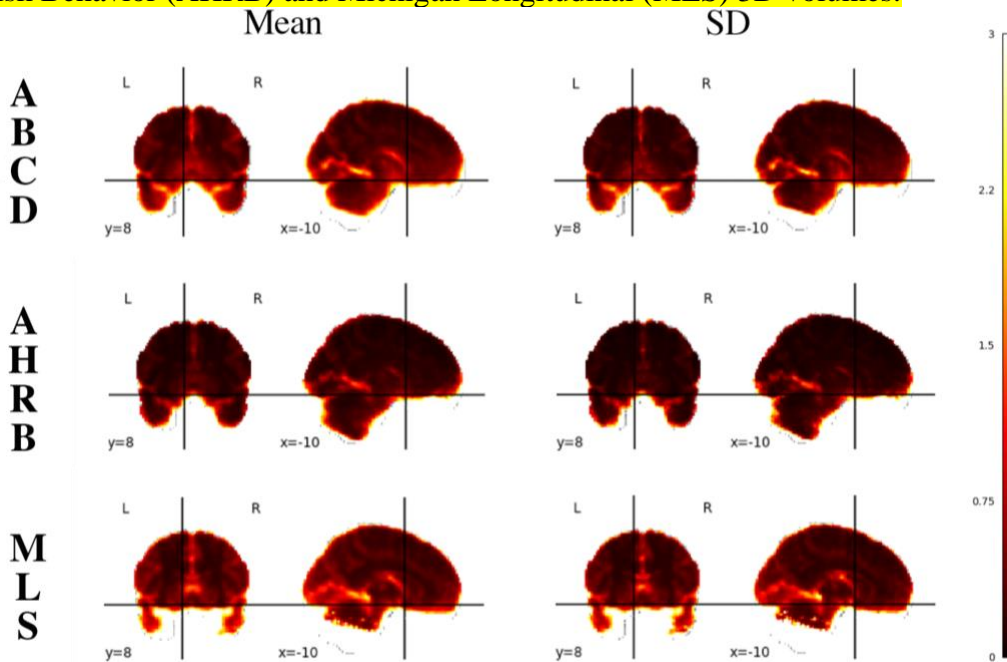
395

396 *Between-Session Reliability:*



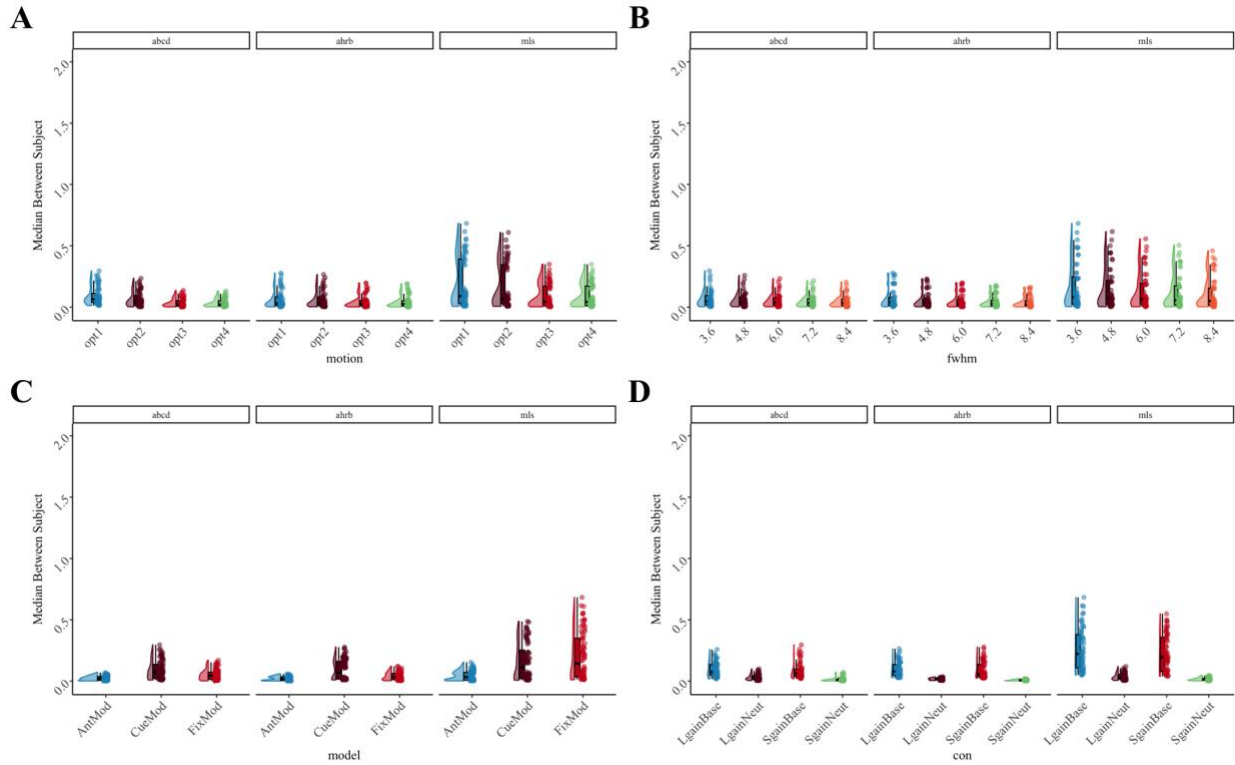
397

398 *Figure S30: Between-session Mean and SD of Between-subject variance (σ_r^2) estimates across*
 399 *240 permutations for the Adolescent Brain Cognitive Development (ABCD), Adolescent Health*
 400 *Risk Behavior (AHRB) and Michigan Longitudinal (MLS) 3D volumes.*



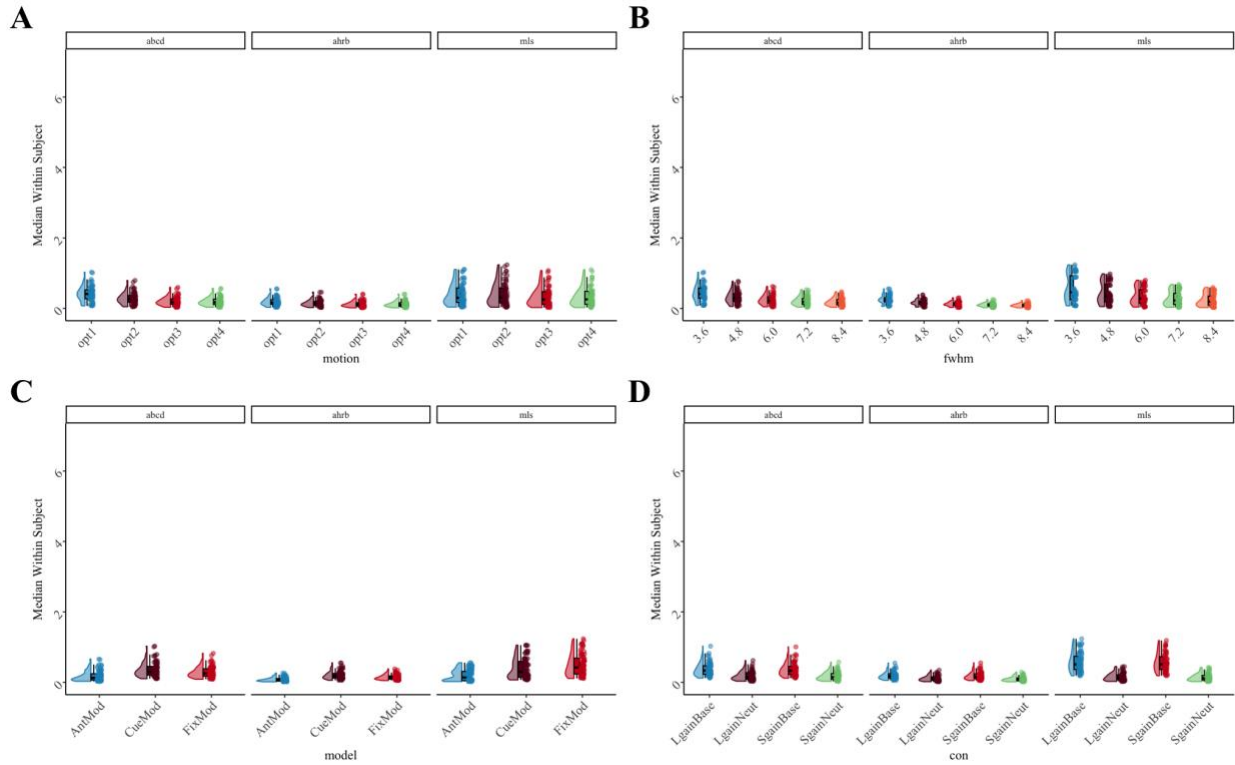
401

402 *Figure S31: Between-session Mean and SD of Within-subject variance (σ_v^2) estimates across*
 403 *240 permutations for the Adolescent Brain Cognitive Development (ABCD), Adolescent Health*
 404 *Risk Behavior (AHRB) and Michigan Longitudinal (MLS) 3D volumes.*



405
406
407
408
409

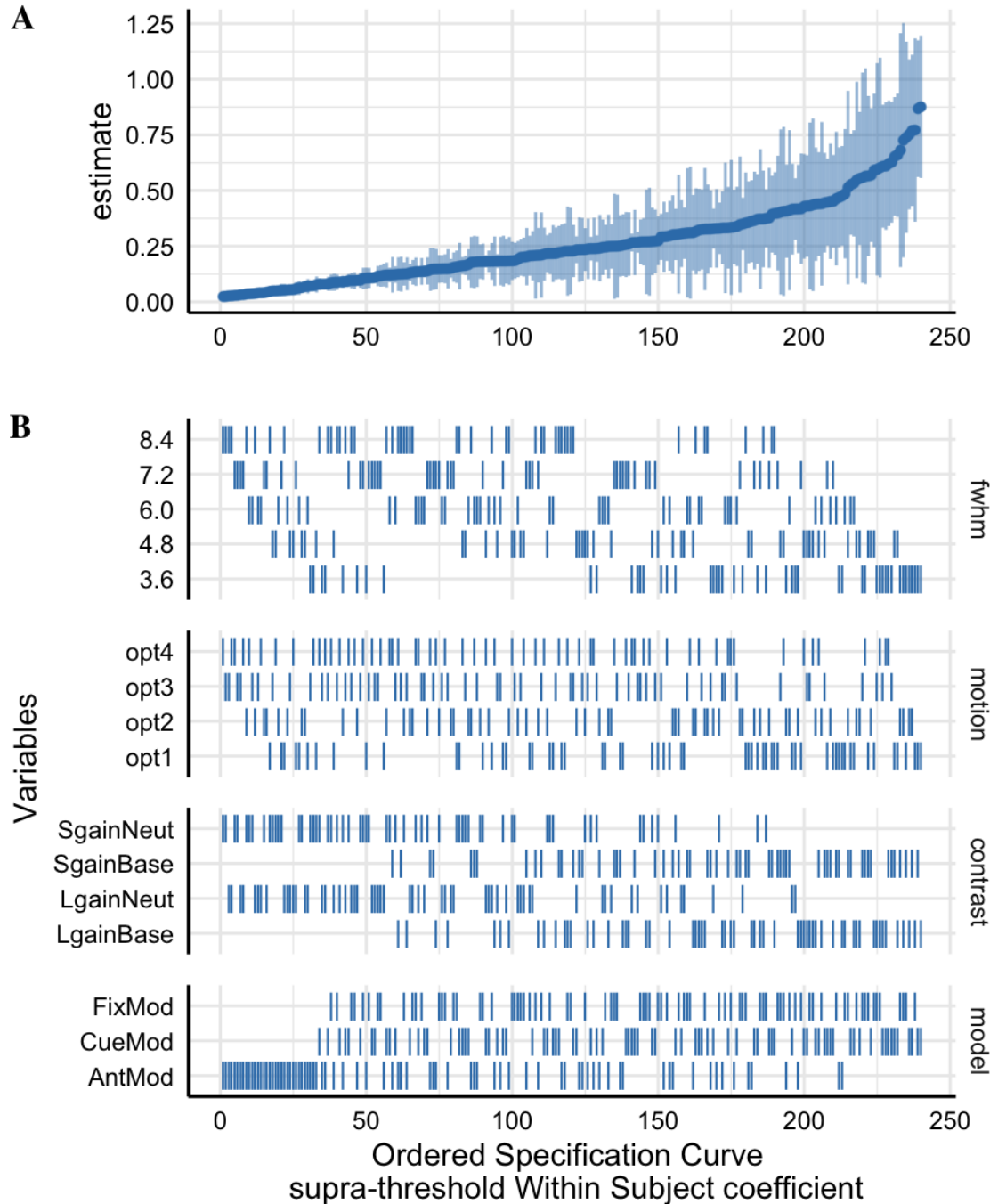
Figure S32. Between-session: Supra-threshold Median Between-subject variance (σ_r^2) estimates across (A) Motion, (B) FWHM, (C) Model Parameterization and (D) Contrast analytic options for between-run reliability across the ABCD, AHRB and MLS samples.



410
411
412
413

Figure S33. Between-session: Supra-threshold Median Within-subject variance (σ_v^2) estimates across (A) Motion, (B) FWHM, (C) Model Parameterization and (D) Contrast analytic options for between-run reliability across the ABCD, AHRB and MLS samples.

414
415 **Figure S34: Between-session: The supra-threshold Specification Curve of the Median Between-**
416 **subject variance (σ_r^2) estimates across 240 pipeline permutations for the ABCD, AHRB and**
417 **MLS estimate.**
418 A. The distribution of the point estimate (average) across the three studies and distribution across the three samples.
419 B. The model options (four) associated with each estimate.



420

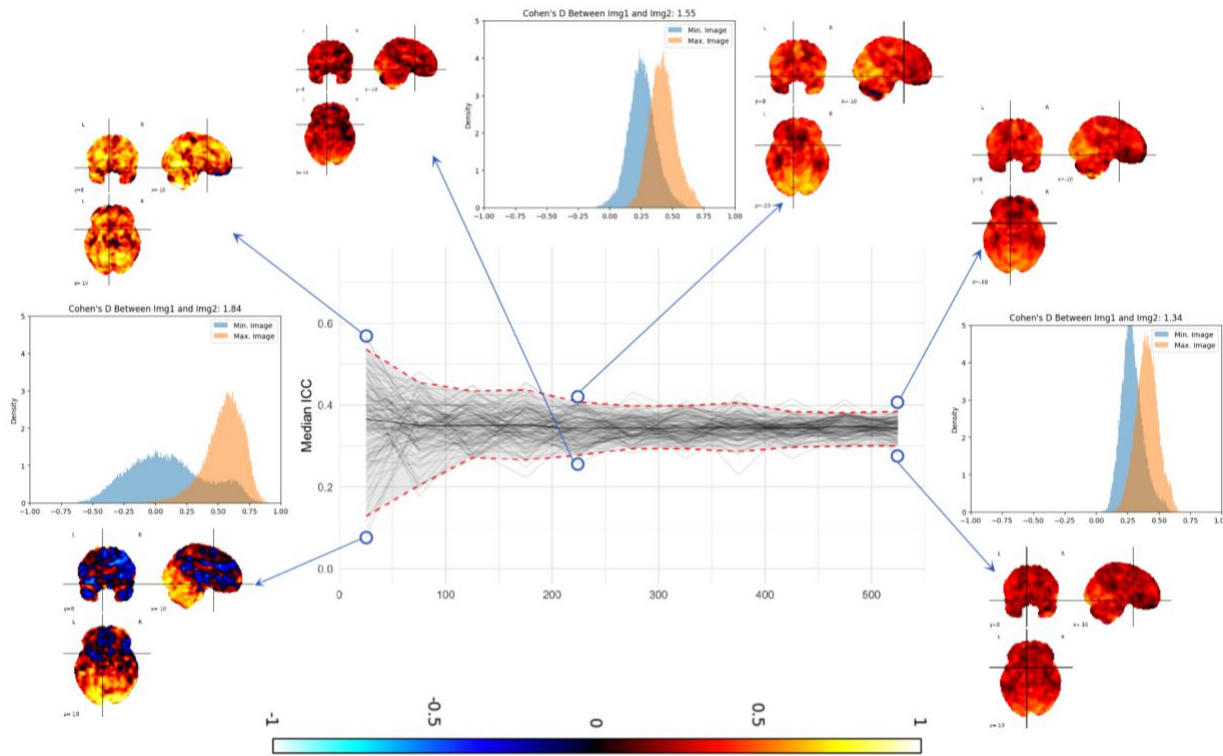
421 **Figure S35: Between-session: The supra-threshold Specification Curve of the Median Within-**
 422 **subject variance (σ_v^2) estimates across 240 pipeline permutations for the ABCD, AHRB and**
 423 **MLS estimate.**

424 A. The distribution of the point estimate (average) across the three studies and distribution across the three samples.

425 B. The model options (four) associated with each estimate.

426 **2.5 Aim 3 results**

427 *Between-Run Stability Effect Size:*



428

429 *Figure S36: Changes in the Median ICC (Supra-threshold mask) estimate in the ABCD sample*
 430 *from N 25 to 525 with 100 bootstraps at each N for Top Model in Figure 2: Small Gain versus*
 431 *Baseline Contrast, Cue Model, Motion option 1 and FWHM 8.4. The associated 3D volumes are*
 432 *plotted for the maximum and minimum median ICC value at N 25, 225 and 525 (circled) and*
 433 *associated voxelwise distribution of maps and Cohen's d between maps are provided.*

434 *Note: Upper and Lower dashed red lines: +/- 95% Confidence Intervals for the median estimates; black solid line is*
 435 *the average of the median estimates; light gray lines are individual subsamples, N 25 to N 525, for each bootstrap.*

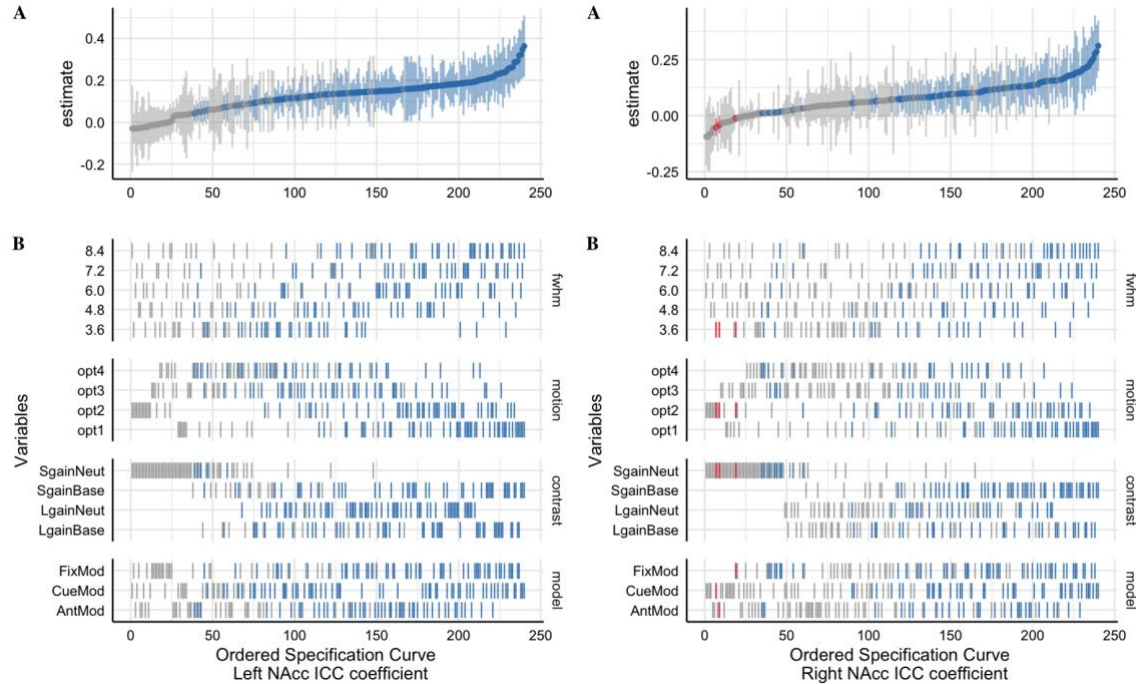
436 **2.6 Post Hoc Analyses**437 *Modeling impacts on Left/Right NAcc:*438 Effect of analytic decisions on ICC estimate for Left and Right Nucleus
439 Accumbens440 For the MID task, researchers are often interested in the activation of the bilateral nucleus
441 accumbens (NAc). The strength of the median ICC estimate from 3D volumes is that it is
442 agnostic to small, anatomical biases and captures the central tendency of ICC estimates across
443 the brain. However, a weakness is that it lacks specificity that is often of interest to brain-
444 behavior researchers. A *post hoc* analysis of the Left and Right NAc was performed using the
445 NAc region of interest from the Harvard-Oxford subcortical atlas (procedure described in
446 Demidenko et al., 2023) for the Session 1 between-run data.447 The specification curve and the HLM results are reported for the Left and Right NAc in
448 supplemental **Figure S37** and **Table S16**, respectively. The average ICC estimate across the 240
449 pipelines varied across the three samples for the *Left NAc* (ABCD = .09 [Min: -0.06, Max: .32];
450 AHRB = .11 [Min: -.23, Max: .46]; MLS = .17 [Min: .03, Max: .44]) and *Right NAc* (ABCD =
451 .08 [Min: -0.04, Max: .32]; AHRB = .03 [Min: -.25, Max: .42]; MLS = .11 [Min: -.07, Max:
452 .40]). In general, model parameterization had a near zero impact on the ICC estimates for the
453 Left (ΔR^2 : .00) and Right NAc (ΔR^2 : .01). The analytic decision that explained the largest
454 amount of variance in the ICC estimates is contrast selection for the Left (ΔR^2 : .27) and Right
455 NAc (ΔR^2 : .24). For example, the change from the contrast of *Large Gain* versus *Implicit*
456 *Baseline* to *Large Gain* versus *Neutral* results in a $b = .01$ decrease in the ICC estimate for the
457 Left NAc and $b = -.02$ decrease for the Right NAc. The largest effect on the ICC estimates is the
458 change from the contrast of *Large Gain* versus *Implicit Baseline* to *Small Gain* versus *Neutral*
459 which results in a $b = .13$ decrease for the Left NAc and $b = .10$ decrease for the Right NAc
460 estimate. Consistent with the Aim 1a results, for Left NAc and Right NAc, the highest average
461 ICC estimate across the three studies is for the *Small Gain* versus *Implicit Baseline* contrast for
462 the Cue Model with no motion correction and 8.4mm FWHM.

463

464 **Table S16: Hierarchical Linear Model: (A) Linear associations between the analytic decisions**
 465 **and the ICC estimate for Left and Right NAc and (B) the impact of the analytic category on the**
 466 **marginal R².**
 467

A. HLM Nucleuss Accumbens (NAc) Estimates								
<i>Predictors</i>	Left Nac			Right Nac				
	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>		
(Intercept)	.16	.11 – .20	<.001	.11	.07 – .14	<.001		
Reference [3.6]								
fwhm [4.8]	.02	.00 – .04	.02	.01	-.01 – .03	.23		
fwhm [6.0]	.04	.02 – .06	<.001	.02	.01 – .04	.01		
fwhm [7.2]	.05	.03 – .07	<.001	.04	.02 – .05	<.001		
fwhm [8.4]	.06	.04 – .08	<.001	.05	.04 – .07	<.001		
Reference [opt1]								
motion [opt2]	-.05	-.06 – -.03	<.001	-.04	-.06 – -.02	<.001		
motion [opt3]	-.06	-.08 – -.05	<.001	-.06	-.08 – -.05	<.001		
motion [opt4]	-.07	-.09 – -.06	<.001	-.06	-.08 – -.05	<.001		
Reference [AntMod]								
model [CueMod]	.02	.00 – .03	.01	.01	-.01 – .02	.27		
model [FixMod]	.01	-.00 – .03	.05	.03	.01 – .04	<.001		
Reference [LgainBase]								
con [LgainNeut]	-.01	-.02 – .01	.28	-.02	-.04 – -.01	.01		
con [SgainBase]	.00	-.02 – .02	.98	.03	.01 – .04	<.001		
con [SgainNeut]	-.13	-.14 – -.11	<.001	-.10	-.12 – -.09	<.001		
B. Analytic Category Model Impact								
Comparison	χ^2	Orig R2	New R2	ΔR^2	χ^2	Orig R2	New R2	ΔR^2
[Full] vs [New - fwhm]	57	.38	.34	.04	48	.36	.33	.03
[Full] vs [New - motion]	91	.38	.31	.07	83	.36	.30	.06
[Full] vs [New - model]	7	.38	.38	.00	16	.36	.35	.01
[Full] vs [New - con]	305	.38	.11	.27	260	.36	.12	.24

468
469



470
471 **Figure S37: The Specification Curve of the ICC estimates for left and right NAcc across 240**
472 **pipeline permutations for the ABCD, AHRB and MLS samples.**
473 **A. The distribution of the point estimate (average) across the three studies and distribution across the three samples.**
474 **B. The model options (four) associated with each estimate.**

475 **Group-level Cohen’s *d* association with estimated ICC**

476 Given the potential association between estimated ICCs and group-level activations
477 magnitudes, the correlation between run and session maps was evaluated for the supra-threshold
478 mask using Spearman *rho*. Across the 240 pipeline permutations, the *rho* coefficient between
479 Session 1 group-level Cohen’s *d* maps and Session 1 between-run ICC maps are low on average
480 but vary widely for *Run 1* (ABCD = -.05 [Min: -.43; Max: .22]; AHRB = .09 [Min: -.41; Max:
481 .50]; MLS = .08 [Min: -.35, Max: .43) and *Run 2* (ABCD = -.04 [Min: -.47; Max: .26]; AHRB =
482 .10 [Min: -.40; Max: .51]; MLS = .08 [Min: -.38, Max: .46). This pattern is consistent for the
483 session-level estimates, whereby the associations between the session group-level maps and the
484 between-session ICC maps are low on average but vary widely for *Session 1* (ABCD = .01 [Min:
485 -.40; Max: .29]; AHRB = .11 [Min: -.45; Max: .53]; MLS = .12 [Min: -.28, Max: .43) and
486 *Session 2* (ABCD = -.01 [Min: -.46; Max: .30]; AHRB = .12 [Min: -.43; Max: .53]; MLS = .11
487 [Min: -.31, Max: .39]).
488

References

- 489
490 Ahlmann-Eltze, C., & Patil, I. (2021). *ggsignif: R Package for Displaying Significance Brackets*
491 *for “ggplot2.”* <https://doi.org/10.31234/osf.io/7awm6>
- 492 Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic
493 image registration with cross-correlation: Evaluating automated labeling of elderly and
494 neurodegenerative brain. *Medical Image Analysis, 12*(1), 26–41.
495 <https://doi.org/10.1016/j.media.2007.06.004>
- 496 Bjork, J. M. (2020). The Ups and Downs of Relating Nondrug Reward Activation to Substance
497 Use Risk in Adolescents. *Current Addiction Reports*. [https://doi.org/10.1007/s40429-020-](https://doi.org/10.1007/s40429-020-00327-7)
498 [00327-7](https://doi.org/10.1007/s40429-020-00327-7)
- 499 Bjork, J. M., Knutson, B., Fong, G. W., Caggiano, D. M., Bennett, S. M., & Hommer, D. W.
500 (2004). Incentive-elicited brain activation in adolescents: Similarities and differences
501 from young adults. *The Journal of Neuroscience: The Official Journal of the Society for*
502 *Neuroscience, 24*(8), 1793–1802. <https://doi.org/10.1523/JNEUROSCI.4862-03.2004>
- 503 Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis: I.
504 Segmentation and Surface Reconstruction. *NeuroImage, 9*(2), 179–194.
505 <https://doi.org/10.1006/nimg.1998.0395>
- 506 Demidenko, M. I., Weigard, A. S., Ganesan, K., Jang, H., Jahn, A., Huntley, E. D., & Keating,
507 D. P. (2021). Interactions between methodological and interindividual variability: How
508 Monetary Incentive Delay (MID) task contrast maps vary and impact associations with
509 behavior. *Brain and Behavior, 11*(5), e02093. <https://doi.org/10.1002/brb3.2093>

- 510 Fonov, V., Evans, A., McKinstry, R., Almlı, C., & Collins, D. (2009). Unbiased nonlinear
511 average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102.
512 [https://doi.org/10.1016/S1053-8119\(09\)70884-5](https://doi.org/10.1016/S1053-8119(09)70884-5)
- 513 Gorgolewski, K. J., Esteban, O., Markiewicz, C. J., Ziegler, E., Ellis, D. G., Jarecka, D., Notter,
514 M. P., Johnson, H., Burns, C., Manhães-Savio, A., Hamalainen, C., Yvernault, B., Salo,
515 T., Goncalves, M., Jordan, K., Waskom, M., Wong, J., Modat, M., Loney, F., ... Ghosh,
516 S. (2018). *nipy/nipype: 1.1.5* [Computer software]. Zenodo.
517 <https://doi.org/10.5281/zenodo.1480713>
- 518 Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-
519 based registration. *NeuroImage*, *48*(1), 63–72.
520 <https://doi.org/10.1016/j.neuroimage.2009.06.060>
- 521 Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the
522 robust and accurate linear registration and motion correction of brain images.
523 *NeuroImage*, *17*(2), 825–841.
- 524 Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter,
525 M., Neto, E. C., & Keshavan, A. (2017). Mindboggling morphometry of human brains.
526 *PLOS Computational Biology*, *13*(2), e1005350.
527 <https://doi.org/10.1371/journal.pcbi.1005350>
- 528 Knutson, B., & Greer, S. (2008). Anticipatory affect: Neural correlates and consequences for
529 choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*,
530 *363*(1511), 3771–3786. <https://doi.org/10.1098/rstb.2008.0155>

- 531 Sacchet, M. D., & Knutson, B. (2013). Spatial smoothing systematically biases the localization
532 of reward-related brain activity. *NeuroImage*, *66*, 270–277.
533 <https://doi.org/10.1016/j.neuroimage.2012.10.056>
- 534 Srirangarajan, T., Mortazavi, L., Bortolini, T., Moll, J., & Knutson, B. (2021). Multi-band fMRI
535 compromises detection of mesolimbic reward responses. *NeuroImage*, *244*, 118617.
536 <https://doi.org/10.1016/j.neuroimage.2021.118617>
- 537 Tomarken, A. J. (1995). A psychometric perspective on psychophysiological measures.
538 *Psychological Assessment*, *7*, 387–395. <https://doi.org/10.1037/1040-3590.7.3.387>
- 539 Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C.
540 (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*,
541 *29*(6), 1310–1320. <https://doi.org/10.1109/TMI.2010.2046908>
- 542 Welvaert, M., Durnez, J., Moerkerke, B., Berdoolaege, G., & Rosseel, Y. (2011). neuRosim: An
543 R Package for Generating fMRI Data. *Journal of Statistical Software*, *44*, 1–18.
544 <https://doi.org/10.18637/jss.v044.i10>
- 545 Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden
546 Markov random field model and the expectation-maximization algorithm. *IEEE*
547 *Transactions on Medical Imaging*, *20*(1), 45–57. <https://doi.org/10.1109/42.906424>
- 548