NeuroImage 252 (2022) 119046

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Reliability and stability challenges in ABCD task fMRI data

James T. Kennedy[a,*], Michael P. Harms[a], Ozlem Korucuoglu[a], Serguei V. Astafiev[a], Deanna M. Barch[a], Wesley K. Thompson[b], James M. Bjork[c], Andrey P. Anokhin[a]

[a] Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, United States
[b] Division of Biostatistics and Department of Radiology, Population Neuroscience and Genetics Lab, University of California, San Diego, United States
[c] Department of Psychiatry, Virginia Commonwealth University, United States

A B S T R A C T

Trait stability of measures is an essential requirement for individual differences research. Functional MRI has been increasingly used in studies that rely on the assumption of trait stability, such as attempts to relate task related brain activation to individual differences in behavior and psychopathology. However, recent research using adult samples has questioned the trait stability of task-fMRI measures, as assessed by test-retest correlations. To date, little is known about trait stability of task fMRI in children. Here, we examined within-session reliability and long-term stability of individual differences in task-fMRI measures using fMRI measures of brain activation provided by the adolescent brain cognitive development (ABCD) Study Release v4.0 as an individual's average regional activity, using its tasks focused on reward processing, response inhibition, and working memory. We also evaluated the effects of factors potentially affecting reliability and stability. Reliability and stability (quantified as the ratio of non-scanner related stable variance to all variances) was poor in virtually all brain regions, with an average value of 0.088 and 0.072 for short term (within-session) reliability and long-term (between-session) stability, respectively, in regions of interest (ROIs) historically-recruited by the tasks. Only one reliability or stability value in ROIs exceeded the 'poor' cut-off of 0.4, and in fact rarely exceeded 0.2 (only 4.9%). Motion had a pronounced effect on estimated reliability/stability, with the lowest motion quartile of participants having a mean reliability/stability 2.5 times higher (albeit still 'poor') than the highest motion quartile. Poor reliability and stability of task-fMRI, particularly in children, diminishes potential utility of fMRI data due to a drastic reduction of effect sizes and, consequently, statistical power for the detection of brain-behavior associations. This essential issue urgently needs to be addressed through optimization of task design, scanning parameters, data acquisition protocols, preprocessing pipelines, and data denoising methods.

1. Introduction

Task-based functional magnetic resonance imaging (fMRI) has become a leading methodological approach in cognitive neuroscience. While initial application of fMRI focused on group-level effects such as average differences in regional brain activation between different stimuli, more recently fMRI has been increasingly applied to individual differences research such as across-subject correlation between task-related brain activation and other variables such as genetic markers, behavioral and cognitive performance, psychological traits, and psychopathology. Much of this research critically relies on the assumption that the magnitude of task-related regional activation is a stable trait-like measure, with individual differences between subjects prevailing over within-subject fluctuations between testing occasions, which is often quantified by test-retest reliability.[1]

However, recent studies have shown generally poor test-retest reliability of task-fMRI measures (Elliott et al., 2020; Herting et al., 2018; Noble et al., 2021). Importantly, reproducibility of group-averaged patterns of activation can still be high despite poor stability of intra-individual differences in the magnitude of activation (Chaarani et al., 2021; Herting et al., 2018), since averaging reduces error variance, as prescribed by basic statistical theory. In the most representative study to date, Elliott et al. (2020) performed a meta-analysis of 56 test-retest reliability studies using various sensory, motor, and cognitive tasks, finding an average reliability of 0.397. Task specific average reliability [limiting to studies that reported all reliabilities calculated, though most were region of interest (ROI) only and not whole brain] ranged from a low of −0.02 for an implicit memory encoding task (Brandt et al., 2013) to a high of 0.87 for a pain stimulation task (Taylor et al., 2009). All studies surveyed in Elliott et al. (2020) had sample sizes under 60 subjects, most

* Corresponding author: James T. Kennedy, Department of Psychiatry, Washington University School of Medicine, Campus Box 8134, 660 S. Euclid Ave, St. Louis, MO 63110.
E-mail address: jtkennedy@wustl.edu (J.T. Kennedy).

[1] It is important to distinguish between measures intended to capture a stable trait-like attribute versus measures that may be heavily influenced by state effects (such as attention, caffeine level, hydration, previous night sleep quality, current anxiety level, etc.). A measure could in principle have a high test-retest reliability if measured in a consistent and well-controlled subject state, yet empirically appear to have a low reliability because possible state influences are either not controlled, or the relevant state influences affecting the measurement are simply unknown. While it is highly valuable from a scientific perspective to study the effect of state on both within- and between-subject variance (and thus reliability), a measure that is only reliable under limited, state-specific conditions is by-definition not a stable "trait-like" measure.

https://doi.org/10.1016/j.neuroimage.2022.119046.
Received 9 October 2021; Received in revised form 25 February 2022; Accepted 28 February 2022
Available online 1 March 2022.

subjects were adults, and test-retest intervals were all under six months, with most under one month. Moderator analyses did not identify significant differences in reliabilities when comparing task type, event vs block design, scan duration, intertrial interval length, or clinical vs non-clinical sample, but did find lower reliabilities in subcortical relative to cortical brain regions. In a recent review, Noble et al. (2021) identified factors that tend to lead to higher test-retest reliability: shorter test-retest intervals, simple compared to complex tasks, brain regions with stronger activation, cortical regions rather than subcortical, and non-clinical populations. Recent studies in our lab examining the factors affecting test-retest reliability of fMRI measures from risk-taking and response inhibition tasks found that reliability increased with shorter interscan intervals, increasing scan duration, in ROIs relative to whole brain, and with lower subject movement, though the use of denoising via multirun spatial ICA (Glasser et al., 2018) plus FIX (Salimi-Khorshidi et al., 2014) ameliorated the negative impact of increased subject movement (Korucuoglu et al., 2021).

A major implication of poor reliability for research relying on individual differences is diminished measured effect sizes and statistical power for detecting associations with other variables, or diminished ability to detect changes over time in longitudinal or treatment studies (Elliott et al., 2020). Detecting small effects requires large samples, which is especially problematic for MRI research, given the high cost of assessments (Dick et al., 2021).

Most previous studies of test-retest stability of task-fMRI were conducted in adult samples, and evidence for temporal stability of individual differences in task-fMRI in children is scarce (Herting et al., 2018), despite the widespread use of task-fMRI in developmental research in pediatric samples. Stability of individual differences is particularly important for longitudinal studies that aim to establish prospective associations between developmental changes in task-related brain activations and behavior. As one of the primary goals of much of developmental psychiatric imaging research is to track how neurofunctional development is associated with future onset and course of mental disorders and substance use (Bjork et al., 2018; Feldstein Ewing et al., 2018; Giedd et al., 2008; Volkow et al., 2018), knowing what neurofunctional variables show stable individual differences is critical. Systematic age-related changes due to development do not necessarily preclude test-retest stability of individual differences, provided it is operationalized as rank-order stability, such as with measures of “consistency” or “relative” agreement rather than “absolute” agreement (Briesch et al., 2014; Shrout and Fleiss, 1979). However, individual variation in the *rate* of developmental changes will result in decreases in longitudinal test-retest stability because it would alter rank-ordering between individuals.

The Adolescent Brain Cognitive Development[SM] (ABCD) Study is an ongoing longitudinal project examining the neuropsychological development of ~12,000 individuals nine to ten years old at enrollment from 21 sites across the United States of America through adolescence (Casey et al., 2018). The ABCD Study[(R)] protocol included three fMRI tasks focused on neurocognitive constructs deemed essential for the understanding of adolescent development: response inhibition (Stop Signal Task; SST), reward anticipation and processing (Monetary Incentive Delay; MID), and working memory (nBack; Casey et al., 2018). However, reliability of brain activations elicited by these tasks in the ABCD data has not been established. The recent 4.0 release of ABCD data contains fMRI data for two longitudinal fMRI assessments conducted two years apart (baseline and the first follow-up), and each of these sessions has two approximately five-minute runs for each task. This enables test-retest reliability assessment at two time scales (within session and between sessions).

Our goal was to examine both within-session, between-run reliability (which is analogous to split-half internal consistency reliability in psychometrics; Heale and Twycross, 2015) and between-session longitudinal stability of regional brain activations elicited in the three ABCD fMRI tasks. This information is essential to evaluate potential utility of the task fMRI data for predictive and inter-individual association analyses, as well as to evaluate potential effects of different region- and subject-level factors on reliability such as relevance of the brain region to the targeted neurocognitive construct, the magnitude of activation, amount of in-scanner movement, and the effect of differences in pubertal development.

Due to the specifics of the ABCD Study, our approach differs somewhat compared to most fMRI reliability studies (surveyed in Elliott et al., 2020). While most studies use an intraclass correlation analysis approach, the multi-scanner, multi-site, family inclusive sample of the ABCD Study merited the use of a linear mixed-effect model (LME) based estimate of reliability and stability capable of controlling for these confounds. Moreover, the intervals between the scans being compared differ (same session or two years, vs. one day to six months) and the two-year span between currently available ABCD visits is occurring across a major period of brain development during adolescence. Our between-session analyses may thus be subject to developmental effects that could make a task appear less stable than reliability analyses using a short test-retest interval or a similarly-long interval between points in adulthood that would presumably be less impacted by developmental differences. For this reason, similar to Baranger et al. (2021), we avoid labeling between-session results as test-retest reliability and instead prefer the term longitudinal stability. (Note that the consistency-based LME measure used here allows for group level differences; stability does not decrease if everyone changes in the same direction and to the same extent.)

We hypothesized that both within-session reliability and longitudinal stability would be poor on average, given previous research for the MID, SST, and nBack (Blokland et al., 2017; Caceres et al., 2009; Fleissbach et al., 2010; S. Holiga et al., 2018; Korucuoglu et al., 2021; Plichta et al., 2012; Schlagenhauf et al., 2008; Zanto et al., 2014), with within-session reliabilities potentially negatively impacted by variable within-session change across individuals. We acknowledge upfront that stability values could be negatively impacted by the long retest interval (Noble et al., 2021) and developmental change between sessions (this was explored by including relevant pubertal variables from the ABCD Study in an expanded LME model examined in the supplement). Nonetheless, it is important to empirically establish the reliability and stability of the ABCD task fMRI data since the results have important implications for other studies using those data.

It is also important to investigate some of the factors that may influence reliability/stability, as a way to understand potential avenues for maximizing them. In that regard, we expected ROIs to have modestly higher values than regions with lesser task relevance and by extension less consistent incidence of activation in the literature. We expected within-session reliability to increase with age, as movement decreases with age in developmental samples (Engelhardt et al., 2017) and movement is a considerable source of additional variance in imaging research (Bright and Murphy, 2017; Diedrichsen and Shadmehr, 2005). Consistent with our previous findings in an adult sample (Korucuoglu et al., 2020, 2021), we expected more active regions to be modestly more reliable/stable. As developmental change typically occurs at different times and rates (Marceau et al., 2011) and the pubertal hormones associated with development are also related to functional activity in reward, emotional processing, and cognition processes targeted by the imaging tasks (Dai and Scherf, 2019), we expected regions that exhibit greater mean longitudinal change to also have lower longitudinal stability (i.e., a negative correlation of between-session change with between-session stability) as it seems likely (although not certain) that regions with greater mean longitudinal change will concurrently be more likely to have changes in relative ranking between individuals over that interval given the variance in onset and speed of change, and thus lower stabilities. Thus, the relationship between reliability/stability and activity and change are explored in the supplemental materials. An overview of research questions and main findings is summarized in Fig. 1.

Topic	Question	Findings	Results Sections	Tables	Figures
1. Reliability and Stability of ABCD task fMRI Data	How reliable and stable are ABCD task fMRI data, both for *a priori* ROIs and at the whole brain level?	Average reliability and stability are poor (<.1) for almost all regions, including ROIs.	3.1, 3.3	2, S2	2, 3
	How does data cleaning impact reliability and stability?	Data cleaning significantly increases reliability and stability, but generally by a negligible amount.	3.1, 3.3, S2.2	S3	S4
	How reliable/stable are low movement subjects relative to high movement subjects?	Low movement subjects had reliability and stability values three times higher than high movement subjects, though still poor overall.	3.1, 3.4	2, S4, S5	2, 4, 5, S8, S9
	Does controlling for pubertal variables change stability?	Controlling for pubertal development scale, DHEA levels, estradiol levels (females only), and testosterone increases stability, but only by an average of .011 across the whole brain.	S2.6	S13	
	How reliable/stable are within/between session change?	Within/between session change are unreliable/unstable (averages of .020 and .034 at the whole brain level, respectively).	S2.8	S15, S16	
	Does between session stability change based on the length of the intersession interval?	Stability increased with longer intersession intervals, but only by ~.005 across all regions in contrasts where a significant difference was observed.	S2.10	S17	
	Does stability change based on the amount of fMRI data?	Stability often stayed the same or even increased when stability was calculated using only the first run of data rather than the average of both runs.	S2.7	S14	
2. Differences in Reliability and Stability Based on Region or Contrast Type	Are reliability, stability, |activity|, and |change| greater in *a priori* ROIs relative to non-ROIs?	Reliability, stability, |activity|, and |change| are generally not significantly different between ROIs and non-ROIs.	3.2		S5, S6, S7
	Are any general brain regions more reliable, stable, or active?	Cortical regions are generally more reliable than subcortical. Occipital regions are more reliable and stable, but exhibit less change than the rest of the brain.	S2.3, S2.4	S10, S11	
	Do reliability and stability differ for condition vs condition and condition vs baseline contrasts?	Condition vs baseline contrasts are significantly more reliable and stable.	S2.5	S12	
	In what contrasts/datasets/regions were reliable activity observed?	The highest average reliability values were found using low movement subjects, but only ~10% of values were above .4, most were in condition vs baseline contrasts, and only 7 were in *a priori* ROIs.	S2.11		
3. Associations Between Reliability/Stability and Other Measures	How are reliability and stability associated with |activity|, |change|, and each other?	Reliable/stable regions were generally more active and exhibit greater inter-session change. Reliability and stability were highly correlated in most contrasts.	S2.1	S5, S6, S7	S4
	How does reliability at the follow-up session compare to reliability at the baseline session? What are the origins of these differences?	Reliability increased between sessions. Stable variance largely increased or stayed the same while residual variance decreased.	S2.1.5	S8, S9	S4

Fig. 1. Summary of findings and the relevant results sections, tables, and figures. Numbers under Results Sections, Tables, and Figures preceded by an S indicate they are in the Supplementary Materials. Regional results from each analysis can be found in Supplementary - Reliability and Stability Output, Supplementary - Full Output, and as parcellated scalar data in CIFTI format on BALSA. (https://balsa.wustl.edu/study/7qMqX).

2. Methods

2.1. Participants

The individuals and data used for our study come from the ABCD Study's "Curated Annual Release 4.0" (https://nda.nih.gov; DOI 10.15154/1,523,041). This data release includes two sessions worth of imaging data (structural, task fMRI, and resting state fMRI), with 10,814 individuals in the baseline session (having structural scans that passed ABCD's pre- and post-processing quality control), and approximately two thirds ($n = 7363$) having processed data available from their first follow-up visit (on average two years later). Task fMRI data was required to pass ABCD's quality control recommended inclusion flag,[2] leaving 7932 to 9353 individuals, depending on task, within the baseline session [mean (SD) age = 9.94 (0.63), 51% male across tasks] and 5979 to 6593 individuals at first follow-up [11.96 (0.65) years old, 53% male] (Table 1). Data for 15 participants were dropped as the scanner manufacturer associated with their data was inconsistent with the other participants from their site and we did not want to include possibly erroneous data in our random effects models. Participants were recruited primarily through school systems with the aim of reflecting American diversity in sex, urbanicity, race and ethnicity, and socioeconomic status (Garavan et al., 2018). Informed assent was gathered for ABCD participants and consent from their parents or guardians. All procedures were approved by the central ABCD Institutional Review Board (IRB) and/or the IRB for the local scanning site.

[2] Variables `imgincl_{mid,nback,sst}_include` of the `abcd_imgincl01` instrument. See "ABCD Release 4.0 release notes", available at DOI 10.15154/1523041

2.2. ABCD study: data, processing and task description

Each of the three fMRI tasks collected by ABCD consist of two approximately five-minute consecutive runs. The released task-activation data were processed through ABCD's "Data Analysis, Informatics and Resource Center" (DAIRC) image processing pipeline (Hagler et al., 2019), which includes motion correction and frame censoring by degree of movement, correction for susceptibility-induced distortions, functional-structural coregistration, activity normalization, and activity sampling onto the cortical surface, carried out using FreeSurfer (Fischl et al., 2002), FSL (Jenkinson et al., 2012), and AFNI (Cox, 1996). Imaging data quality and task performance were evaluated by ABCD's DAIRC as part of quality control. Based on their evaluation, at baseline, 21% of MID, 33% of nBack, and 30% of SST scans failed quality control; at follow-up those percentages were 16%, 21%, and 24%, respectively. A breakdown of the number of subjects who passed the ABCD's quality control measures is available in Table 1. Poor behavioral performance and insufficient fMRI frames (due to excess movement) appear to be the main causes of participant exclusion for both sessions. Additionally, a mismatch between the time stamps of the scans and their associated E-Prime behavioral files resulted in the cautionary exclusion of some participants. The ABCD Release 4.0 data provides estimated activation betas for each run and modeled contrast included in the task general linear model, for cortical parcels in the anatomically-defined Desikan-Killiany parcellation (68 parcels, Desikan et al., 2006) and a more granular gyral- and sulcal-specific Destrieux parcellation (148 parcels, Destrieux et al., 2010), as well as for thirty subcortical structures based on the FreeSurfer segmentations (Fischl et al., 2002). These approaches use the individual's own structural data to derive the boundaries of these different regions, rather than applying a generic common space labeled atlas. A more granular parcellation than the Destrieux parcellation is

Table 1
Participants with usable data by task, session, and number of participants passing QC criteria.

	Baseline Session			Follow-Up Session		
	MID	nBack	SST	MID	nBack	SST
Performance	11,388	9570	10,036	7663	7168	6916
fMRI	9625	9366	9490	6760	6643	6659
E-Prime	10,479	10,285	10,406	7154	7073	7115
Sample (Male)	9353 (4768)	7932 (4054)	8271 (4196)	6593 (3536)	6186 (3298)	5979 (3144)
Age (SD)	9.93 (0.63)	9.96 (0.63)	9.94 (0.63)	11.95 (0.65)	11.96 (0.65)	11.96 (0.65)

Performance: Participants with adequate behavioral performance (e.g., enough correct go trials on the SST); fMRI: Participants with usable fMRI data; E-Prime: Participants whose scanner and E-Prime time stamps matched; Sample (Male): Final usable sample size (# male) for each task at each session; Age (SD): Mean (standard deviation) age for the final sample.

not currently provided by ABCD, nor is data provided currently for a functionally-derived parcellation.

The ABCD fMRI task battery includes the Monetary Incentive Delay (MID), Stop-Signal (SST), and nBack tasks (Casey et al., 2018). The MID task is designed to elicit functional activity when people are anticipating and experiencing different magnitudes of reward and loss. The SST is designed to elicit response inhibition and error monitoring activity by asking participants to respond quickly to a "Go" cue, unless it is followed by a second "Stop" cue that prompts participants to cancel their response. The Emotional nBack is designed to elicit brain activations related to working memory, with a value-added probe of social information processing by showing participants blocks of images of places or emotional or neutral faces. The task requires participants to determine whether the current image matches a static target (0-back condition) or the image that occurred 2 images back (2-back condition). These tasks were implemented by ABCD because brain signatures of reward anticipation, response inhibition, error processing, and working memory change considerably during adolescence (Blakemore et al., 2010; Sheffield Morris et al., 2018) and have important implications for risk of substance use and psychopathology (Bjork et al., 2017; Giedd et al., 2008). For more information about these tasks, see the Supplemental Methods Section S1.1 and Casey et al. (2018).

2.3. Data analysis

Linear mixed-effects (LME) models were applied to beta values from the Destrieux parcellation (the most granular of the parcellations provided by the ABCD Release 4.0, allowing for better localized estimates of reliability, stability, activity, and change) and selected FreeSurfer subcortical structures (limited to those with gray matter, excluding ventricles and white matter, leaving 19 structures: left and right hemisphere accumbens, amygdala, caudate, cerebellum cortex, hippocampus, pallidum, putamen, thalamus, and ventral diencephalon, plus the brainstem, which contains both gray and white matter). Between-session stability analyses used beta values provided by ABCD which averaged activity from each run within a session, weighted by the number of usable frames (between session stability of specific runs is examined in Supplement Section S1.3.8). Reliability and stability were calculated from an LME model that included nested effects of scanner model [e.g., Siemens Prisma (Prisma Fit recoded as Prisma), GE Discovery MR750, Philips Achieva, and Philips Ingenia], site, family, and individual, comparing the stable variation associated with site, family, and individual to this stable variation plus the model residual; i.e.,

$$\text{Variance Ratio} = (Site + Family + IndividualVariance)/(Site + Family + Individual + ResidualVariance).$$

Consistent with the framework of Generalizability Theory (Briesch et al., 2014), the residual variance from the within-session reliability analyses was divided in half to make the reliability estimates (based on two five-minute runs) reflective of averaging across two runs. We excluded variance related to the scanner model from our calculations of reliability and stability as scanner-specific variance does not reflect individual differences in activity. This exclusion allows us to estimate reliability and stability as if they were derived from data all collected on the same scanner model. However, we decided to include site variance as part of the stable (numerator) variance as we cannot discount the possibility that there may be demographic differences between sites that are of interest. Namely, while some of the estimated site variance may be associated with differences in testing procedure and not reflect individual differences (e.g., research assistants at one site doing a better job of preparing participants to move less), site specific variance may also reflect valid community level differences (e.g., obesity rates differ by state and obesity is associated with neurobiological differences and increases in movement (K. Hodgson et al., 2017; Meng et al., 2020; Wang et al., 2020)). Thus, since the goal of the ABCD Study is to capture a representative sample of developing American children, we have kept site variance as a variance component in both the numerator and denominator terms of the variance ratio. Notably, the resulting reliability/stability values will always be higher than if site variance was excluded, which seems a reasonable 'positive' bias to accept given the general finding of poor reliability/stability (so that we do not unduly bias in the direction of overly pessimistic results). Per Cicchetti (1994), reliabilities below 0.4 are frequently considered poor, 0.4–0.59 as fair, 0.6–0.74 as good, and 0.75–1.0 as excellent.

LME models varied by analysis and were implemented using R version 4.1.0's (R Core Team, 2021) nlme package (Pinheiro et al., 2021). The within-session reliability analyses used the following LME model:

$$y \sim Age + Run,\ random = \sim 1|Scanner/Site/Family/Individual$$

while the stability analyses used this LME model:

$$y \sim \text{Age\_at\_Baseline} * \text{Time\_Between\_Sessions, random} = \sim 1|\text{Scanner/Site/Family/Individual}$$

Only intercept was allowed to vary as random slope LME models perform poorly when only two timepoints of data are available. Roughly 1% of LME models failed to converge; values from these analyses were omitted from summaries and statistics computed using the LME results. Reliability and stability values for all regions were also calculated using an intraclass correlation approach (ICC(3,2) for reliability and ICC(3,1) for stability, Shrout and Fleiss, 1979) and can be found in the Supplementary Output - ICC spreadsheet for the analyses covered in the main text. Reliability and stability for each model that converged can be found in the Supplementary Output - Reliability and Stability spreadsheet.

Analyses controlling for pubertal differences were performed to try to mitigate individual differences in development that would negatively impact stability. The pubertal measures included in these analyses were hormone levels for DHEA, estradiol, and testosterone and

the pubertal developmental score.[3] These pubertal variables were entered into the LME formula by including the value at baseline, the difference between follow-up and baseline values, and the interaction of the two (e.g., R syntax: `+ <measure>_at_Baseline * Difference_in_<measure> + ...`, so that the main effects of each and their interaction were modeled as three fixed effects). Only hormone values that passed ABCD's quality control were included in the analysis.[4] These models were run separately for each sex as estradiol was unavailable for males. All fixed effect variables were demeaned for the LME analyses. LME models of the reliability and stability of within and between session change, run specific stability, and stability for subgroups with high and low intersession intervals were also calculated – see Supplementary Methods Section S1.3 for details.

The variance component estimates from the LME model (scanner model, site, family, individual, and residual) are supplied in the Supplementary Output - Full spreadsheets, as well as the relative proportion of each to the total variance and total stable variance (including scanner). Those spreadsheets also include the value, standard error, degrees of freedom, t-statistic, p-value, and Cohen's D effect size for each fixed effect, the model loglikelihood, Akaike information criterion, Bayesian information criterion, and reliability/stability calculated with the scanner variance incorporated. These extensive tables are provided so that interested individuals can explore the quantitative model results in their entirety.

The initial ABCD quality controlled (QC) dataset was the basis for the creation of three additional datasets that were used to examine the effects of statistical approaches to data cleaning, namely outlier removal (QC+OR), motion regression (using the framewise displacement variable[5]) followed by outlier removal (QC+MV+OR), and rank normalization (QC+Rank). Results from these datasets were compared using paired-t tests. Group differences for datasets and other comparisons are sometimes expressed as Cohen's D effect sizes as the large number of values being compared (167 regions * 26 contrasts) may result in a weak effect appearing important due to it being highly statistically significant; reporting the actual effect size gives a sense of the strength of any difference. For more details, see the Sections S1.2 and S1.3.2 of the Supplement.

Within-session reliability, longitudinal stability, activity, and change statistics (activity/change methods described in Supplemental Section S1.3.1) for each contrast and dataset were converted into CIFTI 'pscalar' (parcellated scalar; Glasser et al., 2013) format for display and data dissemination purposes. Regional values from models that failed to converge are left blank. Data is available on BALSA at https://balsa.wustl.edu/study/7qMqX. Maps of significant activity and change were created for only the QC and outlier removed (QC+OR) datasets, as the movement regression (QC+MV+OR) and rank normalization (QC+Rank) approaches both mean center the data, rendering the computation of activity and change in those datasets moot. Region and contrast specific reliability, stability, activity, and change values are also provided as supplemental tables. The R code used to generate the datasets, reliability, stability, activity, change, and variance components are provided as supplements. All subsequent statistical analyses comparing reliability, stability, activity, and change were performed using SPSS v27 (IBM Corp, 2020).

2.4. Reliability, stability, activity, and change

2.4.1. Regions of interest

As our primary analysis we examined if the regions most consistently recruited by the cognitive demands of each specific task in previous research were more reliable, stable, significantly more active, or subject to greater within or between-session change. To this end, of the total 26 contrasts (10 MID, 9 nBack, 7 SST) included in the processing of the ABCD Release 4.0 data, we identified *a priori* ROIs for eight targeted contrasts by taking the coordinates of the reported cluster peak and subpeaks[6] from meta-analyses that report important regions for each process targeted by the task/contrast, converting to Montreal Neurological Institute (MNI) coordinates if necessary using the converter included with GingerALE version 3.0.2 (Eickhoff et al., 2011), and identifying the Destrieux parcel or subcortical structure in which this coordinate resides. These regions were not identified based on ABCD data but from previously published meta-analyses. Reliance on extant Destrieux parcels/FreeSurfer segmentations that overlap with literature-consensus activation maxima also avoids circularity compared to deriving ROIs from activation in the ABCD data itself. The same approach was used by Korucuoglu et al. (2021). This is not an ideal approach as the parcels/structures are originally generated based on an individual's specific anatomy and some variation in location within MNI space can be expected, but is reasonable given that the meta-analyses themselves report results in a common (MNI or Talairach) space. Moreover, reliability and stability are high when there are stable individual differences. We recognize that regions where individuals vary a great deal in their functional responses to a stimulus may nonetheless be very stable, but not significantly active at the group level, while conversely a stimulus may exhibit a strong group level response but be completely unreliable. As we are unaware of relevant meta-analyses focused specifically on identifying reliable/stable regions (without regard to group level activation), we defined our *a priori* ROIs from group level analyses instead.

The specific targeted contrasts and their associated meta-analyses are as follows: MID: anticipation of loss (large and small loss trials admixed) vs neutral, anticipation of reward (large and small reward trials admixed) vs neutral, and reward (positive) vs missed-reward (negative) notification in reward trials (henceforth referred to as reward feedback) from Oldham et al. (2017); nBack: 2- vs 0-back from Yaple et al. (2018), and emotional face vs neutral face and face vs place contrasts, both in Muller et al. (2018); SST: correct stop vs correct go from Swick et al. (2011) and incorrect stop vs correct go from Neta et al. (2015). There were a total of 57 unique regions (35 ignoring laterality) across the 8 contrasts. Between 7 and 20 ROIs were identified for each contrast, with 20 regions appearing in at least 2 contrasts. Supplemental Figures S1-S3 illustrate the location of the ROIs for each contrast and Supplemental Table 1 lists the ROIs by contrast.

2.4.2. ROI vs non-ROI comparison

The resulting ROIs can be considered to represent the regions that meta-analyses have established as among the most "task relevant" for the principal domains (i.e., contrasts) targeted by each task. To examine the impact of this "task relevance", for each of reliability, stability, activity, and change (both within and between-session), we directly compared the ROI results with the remaining regions ("non-ROIs", i.e., rest of the brain) for each of the 8 aforementioned contrasts, using an independent sample *t*-test. We used an FDR correction across the number of

[3] Hormone level variables were `hormone_scr_dhea_mean`, `hormone_scr_hse_mean` (estradiol), and `hormone_scr_ert_mean` (testosterone) from the file `abcd_hsss01` and the pubertal developmental score was the average of the first five values (4th and 5th sex specific) of the PDS scale from the `abcd_ppdms01` file.

[4] Exclusion variables were from the `abcd_hsss01` file and followed the naming pattern `hormone_scr_{dhea,hse,ert}_rep{1,2}_{ll,qns,nd}`(ll = "below lower level of sensitivity", qns = quantity not sufficient, and nd = none detected).

[5] `tfmri_{mid,nback,sst}_{all,run1,run2}_beta_mean.motion` using the harmonized "DEAP" variable name, as specified in the "21. abcd_4.0_mapping.csv" file in the ABCD Release 4.0 release notes, which also provides the mapping to the NDA instrument and corresponding NDA variable name in which the mean framewise displacement values can be located.

[6] A part of a large cluster with activity higher than its surrounding voxels that is not the highest point in the entire cluster.

contrasts, but the analyses for reliability, stability, activity, and change were each treated independently.

2.4.3. Whole brain

As there is no definitive consensus as to what regions should be considered ROIs, and since, to the best of our knowledge, meta-analyses to guide selection of ROIs were unavailable for 18 of the 26 available contrasts, unbiased, whole-brain analyses were also conducted for all contrasts. Of note, these whole brain analyses included four condition vs baseline contrasts, while the ROI analyses were restricted solely to condition vs condition contrasts.

2.5. Movement quartile comparison analyses

To examine the effects of in-scanner movement on reliability and stability, the QC dataset was first subdivided into four subgroups based on quartiles of mean framewise displacement, then framewise displacement was regressed from beta values (within quartile), and finally subject level outliers greater than 3 standard deviations from the mean were removed (within quartile recursively, until no new additional outliers were identified). Reliabilities and stabilities were computed separately for each movement quartile. Each quartile's regional values were compared against each other quartile for each reliability/stability measure using paired t-tests. Comparisons surviving an FDR correction for the number of contrasts are reported in the form of the average difference between quartiles. This was done for both ROIs from the 8 targeted contrasts and at the whole brain level across all contrasts. Further details can be found in the Supplemental Section S1.2.4.

2.6. Secondary analyses

Analyses of variables affecting reliability and stability are explored in the supplemental methods (Section S1). Unless otherwise noted, these analyses were based on LME model results from the QC+OR dataset and performed at the whole brain level for each contrast separately. These include the association between reliability/stability with other reliability measures (including the paired comparison of reliability at baseline and follow-up), the absolute value of activation, and the absolute value of within/between session change (Section S1.3.3), differences in reliability based on region (cortical vs subcortical, S1.3.4; occipital vs non-occipital; S1.3.5), comparison of results from condition vs condition and condition vs baseline contrasts (S1.3.6), differences in stability after accounting for pubertal variables (S1.3.7), the effect of the amount of data on stability (S1.3.8), comparison of the degree of change within and between sessions and the reliability/stability of that change (S1.3.9), differences in movement within and between sessions (S1.3.10), differences in stability based on differences in intersession interval (S1.3.11), and the relationship between an individual's absolute value of activity and standard error of the mean (S1.3.12).

3. Results

3.1. Reliability, stability, activity, and change in ROIs

Mean reliability and stability in ROIs, averaged across the targeted contrasts for all 3 tasks in the full "QC" dataset, was 0.076 (SD=0.060) for within-session reliability at baseline, 0.100 (0.068) for within-session reliability at follow-up, and 0.072 (0.066) for longitudinal stability (Fig. 2A). All ROI reliabilities and stabilities were poor (i.e., < 0.4) except for stability in the right inferior occipital in the face vs place contrast of the nBack. In the QC dataset, only 1.2% of ROI analyses had reliabilities or stabilities over 0.3 while only 4.9% had reliabilities or stabilities over 0.2. These poor reliability and stability values occurred despite the fact that the ROIs were indeed generally activated at the group level by their respective tasks (Fig. 2B) – 90 of 108 ROIs were statistically "active" (after FDR correction) at baseline and 87 were active at follow-up (one model, for the right amygdala in the face vs place contrast, did not converge). ROIs were also subject to statistically significant change in activation (Fig. 2C), but only in 67 ROIs within-session at baseline, 61 ROIs within-session at follow-up, and 37 ROIs between-session.

Data cleaning slightly increased mean reliabilities and stabilities (though average values remained poor) from 0.083 (0.066) for the QC dataset (mean (SD) across stability and both reliabilities) to 0.096 (0.071) for QC+OR, to 0.094 (0.071) for QC+MV+OR, and to 0.095 (0.070) for QC+Rank. While these increases in mean values were small, they occurred consistently, such that the increase was highly significant (all p values from paired t-tests comparing data cleaning types to QC dataset < 0.001, Cohen's D for paired comparisons of QC vs QC+OR dataset: −0.344; vs QC+MV+OR: −0.313; vs QC+Rank = −0.444; Fig. 2A).

A much bigger impact was observed by subsetting participants into different movement quartiles (using the QC dataset), where mean reliability and stability in ROIs was three times higher in the lowest movement group [1st quartile; average (SD) across reliability and stability of 0.166 (0.116)] compared to the highest movement group (4th quartile; 0.053 (0.046); paired comparison significance $p < .001$; 1st-4th Cohen's $D = 1.136$; Fig. 2D). Nonetheless, mean reliabilities and stabilities even of the lowest motion quartile remained well within the 'poor' range. Mean and standard deviations for ROIs by contrast and dataset can be found in Table 2.

The preceding analysis used mean reliabilities and stabilities across *a priori* defined ROIs as a way to broadly summarize our findings. However, the different tasks and contrasts are targeting different aspects of functional processing and it is natural to wonder if reliability and stability may be higher in particular contrasts. Thus, analyses were repeated at the contrast level. Mean reliability and stability values (across the ROIs for each contrast) was highest in the 2 vs 0-back contrast [mean (SD) for within-session reliability at baseline: 0.124 (0.065); for reliability at follow-up: 0.171 (0.091); longitudinal stability: 0.118 (0.073)] and was lowest in the emotion vs neutral face contrast [within-session reliability at baseline: 0.019 (0.033); for reliability at follow-up: 0.016 (0.023); longitudinal stability: 0.022 (0.016)]. Contrast specific comparisons of data cleaning approaches found statistically significant increases in reliability only in the baseline session; stability was significantly greater primarily in the QC+Rank vs QC comparisons (Table 2). Contrast specific comparisons of the 1st and 4th movement quartiles generally confirmed our finding of higher reliability and stability values in the lowest movement quartile for the individual contrasts, with a significant difference in 17 of 24 comparisons (Table 2).

3.2. Comparison of ROIs and non-ROIs

Reliability and stability values in the *a priori* ROIs were not statistically significantly higher than values in non-ROIs, regardless of data cleaning method (Fig. 2A), even though *post hoc* comparisons found ROIs were significantly more active (independent sample t-tests comparing the absolute value of the intercept of the within-session LME results of ROIs vs non-ROIs using the QC+OR dataset; baseline: Cohen's D 0.257, $p = .011$; follow-up: Cohen's D 0.246, $p < .015$, between session: Cohen's D 0.328, $p = .001$, Fig. 2B). ROIs were subject to greater within-session change at baseline and follow-up than non-ROIs, but only before data cleaning (QC Cohen's D baseline: 0.217, $p = .047$, follow-up: 0.222, $p = .047$, QC+OR baseline and follow-up $p = .161$; Fig. 2C).

Contrast specific analyses generally found no significant differences between ROIs and non-ROIs for reliability, stability, |Activity|, and |Change|, with a few exceptions. Stability values were statistically significantly greater in ROIs relative to non-ROIs only for the 2 vs 0-back contrast of the nBack (ROI stabilities 0.07 higher than non-ROIs), significantly weaker stability was observed in ROIs in the anticipation of

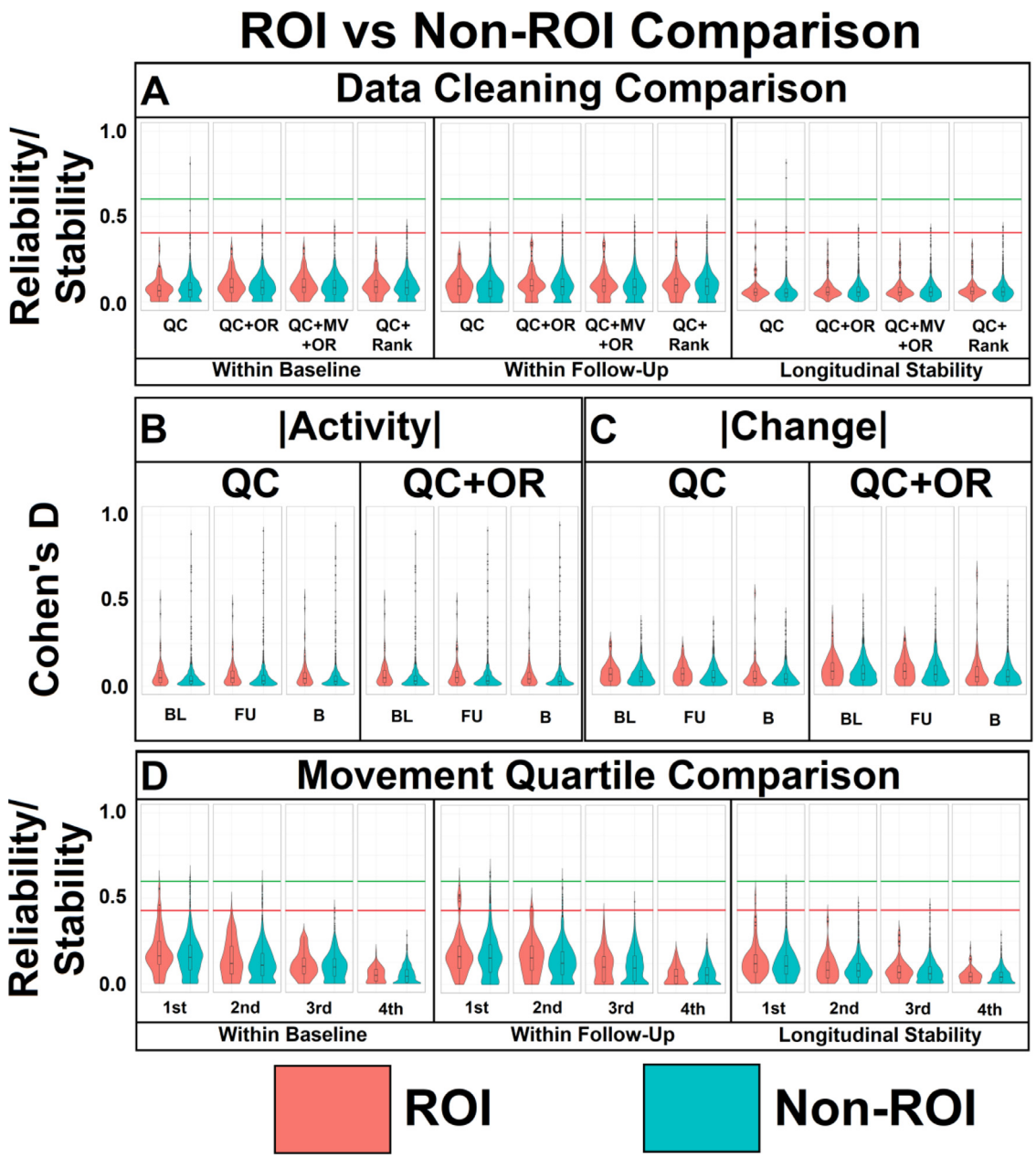

Fig. 2. Violin plots with embedded box plots showing the distribution of reliability/stability (A and D), absolute values of activity (B), and absolute values of within- and between-session change (C) for *a priori* ROIs (red) and non-ROIs (blue). Data was cleaned using different data-cleaning approaches (A) and also separated into movement quartiles (D) to assess the impact of those factors on reliability and stability. QC: all data that passed ABCD's quality control; QC+OR: QC dataset with outliers removed; QC+MV+OR: QC dataset with movement regressed and then outliers removed; QC+Rank: QC dataset with rank normalization. The movement quartiles analysis used the dataset with QC cleaning for the initial quartile separation and then had movement regressed out and outliers removed (separately for each quartile). Activity and change analyses are available only for the QC and QC+OR datasets as the movement regression and rank normalization processes demean the data, making meaningful between region comparisons impossible. For the embedded box plots, the horizontal dash indicates the median, with the box indicating the interquartile range (IQR, 25th to 75th percentile) and 'outliers' greater than 1.5 IQR from the median are shown with individual data points. ROI: Regions of Interest. Green line indicates the boundary for fair-good reliability/stability (0.6); red line indicates the boundary for poor-fair (0.4).

loss vs neutral contrast (ROI stabilities 0.02 lower than non-ROIs; Supplemental Figure 5 shows contrast specific violin plots). Differences in reliability between ROIs and non-ROIS were not observed at the contrast level. Greater absolute value of activity in ROIs relative to non-ROIs was observed in only the 2 vs 0-back at follow-up (|Activity| 0.06 higher in ROIs relative to non-ROIs; Supplemental Figure 6). Greater absolute value of change in ROIs relative to non-ROIs was found only in the SST between sessions (|Cohen's D| of intersession change 0.03 higher for correct stop vs correct go and 0.05 higher in incorrect stop vs correct Go in ROIs relative to non-ROIs; Supplemental Figure 7). Overall, separating ROIs into contrasts (some with as few as 7 ROIs) largely eliminated the significant effects of greater ROI relative to non-ROI |Activity|, while the increased specificity allowed us to identify ROI vs non-ROI differences in the 2 vs 0-back and anticipation of loss vs neutral contrasts.

3.3. *Whole brain analyses*

Since meta-analyses to guide ROI selection were not available for most (18 of 26) of the provided ABCD task contrasts, and since what

Table 2
Mean reliabilities and stabilities of the a priori ROIs by task, contrast, data cleaning approach, and movement quartile.

Task	Contrast	QC			QC+OR			QC+MV+OR			QC+Rank		
		W BL	W FU	Between	W BL	W FU	Between	W BL	W FU	Between	W BL	W FU	Between
MID	Anticipation loss vs neutral	0.06 (0.02)	0.10 (0.03)	0.04 (0.01)	0.07 (0.02)	0.09 (0.02)	0.04 (0.01)	0.06 (0.02)	0.09 (0.02)	0.04 (0.01)	0.07 (0.02)	0.10 (0.03)	0.04 (0.01)
MID	Anticipation reward vs neutral	0.08 (0.05)	0.10 (0.04)	0.07 (0.04)	**0.11 (0.05)**	0.12 (0.05)	0.07 (0.04)	**0.11 (0.05)**	0.11 (0.05)	0.07 (0.04)	**0.10 (0.05)**	0.11 (0.04)	**0.08 (0.04)**
MID	Reward positive vs negative feedback	0.08 (0.03)	0.15 (0.04)	0.05 (0.02)	**0.12 (0.02)**	0.12 (0.02)	0.05 (0.02)	**0.12 (0.02)**	0.12 (0.02)	0.05 (0.02)	**0.12 (0.02)**	0.14 (0.03)	0.05 (0.01)
nBack	2 back vs 0 back	0.12 (0.06)	0.17 (0.09)	0.12 (0.07)	**0.17 (0.08)**	0.23 (0.11)	**0.16 (0.07)**	**0.17 (0.08)**	0.23 (0.11)	**0.16 (0.07)**	**0.16 (0.08)**	0.23 (0.11)	**0.16 (0.08)**
nBack	Emotion vs neutral face	0.02 (0.03)	0.02 (0.02)	0.02 (0.02)	0.01 (0.01)	0.01 (0.02)	0.03 (0.02)	0.01 (0.01)	0.01 (0.02)	0.03 (0.02)	0.01 (0.01)	0.01 (0.02)	**0.03 (0.02)**
nBack	Face vs place	0.09 (0.10)	0.08 (0.08)	0.11 (0.12)	0.10 (0.10)	0.10 (0.11)	0.11 (0.10)	0.10 (0.10)	0.10 (0.11)	0.11 (0.10)	0.10 (0.10)	0.10 (0.10)	0.11 (0.10)
SST	Correct stop vs correct go	0.07 (0.04)	0.10 (0.05)	0.07 (0.02)	**0.12 (0.04)**	0.11 (0.04)	0.07 (0.02)	**0.12 (0.03)**	0.11 (0.04)	0.07 (0.02)	**0.10 (0.03)**	0.12 (0.04)	0.07 (0.02)
SST	Incorrect stop vs correct go	0.11 (0.02)	0.17 (0.04)	0.08 (0.02)	**0.15 (0.04)**	0.17 (0.03)	0.09 (0.02)	**0.15 (0.04)**	0.17 (0.03)	0.09 (0.02)	**0.14 (0.03)**	0.18 (0.03)	**0.09 (0.02)**

Task	Contrast	1st Movement Quartile			2nd Movement Quartile			3rd Movement Quartile			4th Movement Quartile		
		W BL	W FU	Between	W BL	W FU	Between	W BL	W FU	Between	W BL	W FU	Between
MID	Anticipation loss vs neutral	0.10 (0.04)	0.11 (0.06)	**0.08 (0.03)**	0.08 (0.06)	0.11 (0.06)	0.07 (0.05)	0.08 (0.04)	0.11 (0.04)	0.05 (0.04)	0.08 (0.03)	0.08 (0.05)	0.05 (0.04)
MID	Anticipation reward vs neutral	**0.18 (0.08)**	**0.21 (0.06)**	**0.14 (0.06)**	0.13 (0.09)	0.21 (0.06)	0.07 (0.05)	0.18 (0.05)	0.14 (0.07)	0.05 (0.03)	0.05 (0.03)	0.05 (0.04)	0.05 (0.03)
MID	Reward positive vs negative feedback	**0.16 (0.03)**	0.08 (0.04)	0.08 (0.04)	0.14 (0.05)	0.08 (0.04)	0.07 (0.04)	0.11 (0.04)	0.15 (0.06)	0.09 (0.05)	0.09 (0.03)	0.11 (0.06)	0.09 (0.05)
nBack	2 back vs 0 back	**0.34 (0.14)**	**0.35 (0.09)**	**0.25 (0.10)**	0.23 (0.15)	0.35 (0.09)	0.16 (0.11)	0.18 (0.11)	0.17 (0.10)	0.16 (0.09)	0.08 (0.04)	0.15 (0.05)	0.16 (0.09)
nBack	Emotion vs neutral face	0.02 (0.03)	0.02 (0.03)	0.04 (0.02)	0.01 (0.02)	0.02 (0.03)	0.04 (0.04)	0.04 (0.03)	0.01 (0.02)	0.06 (0.04)	0.01 (0.02)	0.02 (0.03)	0.06 (0.04)
nBack	Face vs place	**0.20 (0.16)**	**0.17 (0.14)**	**0.18 (0.16)**	0.15 (0.13)	0.17 (0.14)	0.13 (0.11)	0.10 (0.10)	0.07 (0.12)	0.11 (0.09)	0.08 (0.05)	0.04 (0.05)	0.11 (0.09)
SST	Correct stop vs correct go	**0.21 (0.06)**	**0.17 (0.06)**	**0.14 (0.05)**	0.22 (0.07)	0.17 (0.06)	0.10 (0.06)	0.12 (0.04)	0.10 (0.07)	0.08 (0.04)	0.02 (0.02)	0.03 (0.03)	0.08 (0.04)
SST	Incorrect stop vs correct go	**0.27 (0.08)**	**0.21 (0.06)**	**0.17 (0.04)**	0.20 (0.06)	0.21 (0.06)	0.12 (0.04)	0.07 (0.03)	0.20 (0.06)	0.03 (0.03)	0.06 (0.03)	0.08 (0.05)	0.03 (0.03)

Task, contrast, and data subset specific mean (standard deviation) reliability (within session) and between session stability values for the *a priori* ROI analyses. Results indicate reliability/stability are poor and increase only slightly with data cleaning and exclusion of high movement participants. Bold values under QC+OR, QC+MV+OR, or QC+Rank columns indicate values for that data cleaning approach were significantly greater than for the QC dataset. Bold values under the 1st Movement Quartile (lowest motion) columns indicate values were significantly greater than those for the 4th Movement Quartile (highest motion). Significance tested using paired *t*-test with FDR correction across 8 contrasts. W BL = Within Baseline session reliability; W FU = Within Follow-Up session reliability; Between = Between session stability.

qualifies as a "region of interest" is partly subjective, whole brain region-wise analyses were also performed for all contrasts. Using the QC dataset, across all regions and contrasts, mean (SD) within-session reliability was 0.098 (0.119) for baseline, 0.107 (0.105) for reliability at follow-up, and 0.072 (0.079) for longitudinal stability. Contrast and dataset specific mean (SD) reliability and stability values can be found in Supplemental Table 2. Figure 3 (top) shows the mean reliability and stability for each region for the QC dataset (across all available 26 contrasts). Occipital reliability and stability values tend to be higher than other brain regions, while subcortical and orbital frontal regions were lower than the rest of the brain Fig. 3. (bottom) shows histograms of reliability and stability values in ROIs, non-ROIs, condition vs baseline contrasts, and condition vs condition contrasts without identified ROIs (labeled "Other"). The histograms show that ROIs have similar distributions to non-ROIs and that the high end of the distribution is primarily regions from condition vs baseline contrasts. Complete data (and figures) for reliability and stability per region and for each contrast by data cleaning and reliability/stability type can be found on BALSA at https://balsa.wustl.edu/study/7qMqX.

The data cleaning comparison applied to the whole brain analysis found that removing outliers again slightly increased mean reliability and stability values – from a mean (SD) of 0.093 (0.101) for the QC dataset to 0.107 (0.113) for QC+OR ($p < .001$); see Supplemental section *S2.2* and Supplemental Table 3 for a comparison of values by data cleaning approaches. Scatterplots comparing region specific reliabilities/stabilities in the 2 vs 0-back contrast before and after outlier removal showed that values were greater in the QC+OR dataset relative to the QC dataset in most regions (132, 130, and 155 of 167 total regions for within-session reliability at baseline, within-session reliability at follow-up, and longitudinal stability, respectively; Supplemental Figure 4, panels A and B).

3.4. Movement quartile comparison analyses

A comparison of reliabilities and stabilities computed separately in each of the movement quartiles showed significantly higher values for both in the quartiles with less movement Fig. 2.D shows the average values for ROIs by quartile. For the whole brain, the average reliability increased from 0.066 for the 4th (highest) movement quartile to 0.184 for the 1st (lowest) movement quartile within the baseline session ($\Delta = 0.118$), from 0.079 to 0.183 within the follow-up session ($\Delta = 0.104$), and longitudinal stability increased from 0.053 to 0.130 ($\Delta = 0.077$). Increasing reliability and stability values with less movement was observed in 69 of 78 analyses (3 × 26 contrasts) when comparing 1st to 4th quartiles in paired t-tests, though a minority (2) were significantly less reliable or stable with less movement Fig. 4. shows similar results for the reliability values at follow-up across the whole brain, but separated into each of the 26 contrasts provided by ABCD (analogous violin plots for within-session reliability at baseline and between-session stability can be found as Supplemental Figures 8 and 9). Those results show that the nBack task had the contrasts with the highest reliability and stability values. Medial and lateral cortical maps of the reliability and stability values for the 2 vs 0-back contrast are shown for all movement quartiles in Fig. 5. This figure demonstrates decreasing reliability and stability values with increasing movement and greater reliability within-session at follow-up (when participants were older) relative to the baseline session (see also Figure S4, panel D) Table 2. provides the mean values for the 1st and 4th quartile by contrast for ROIs. A whole brain comparison of reliability and stability values and their components across quartiles for each of the 26 contrasts is provided in Supplemental Tables 4 and 5. Complete data for regional reliability and stability by movement quartile are available on BALSA at https://balsa.wustl.edu/study/7qMqX.

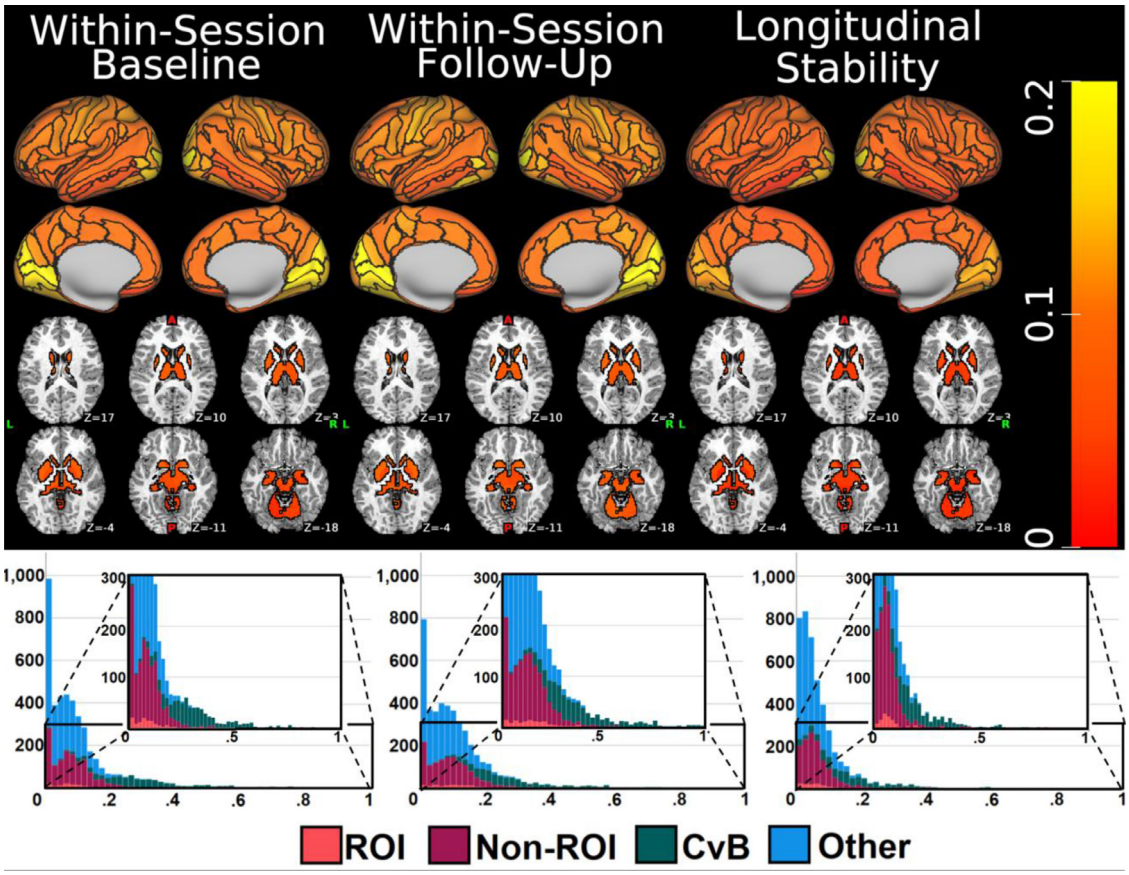

Fig. 3. Top: Task fMRI reliability and stability by region, averaged across all 26 contrasts released by the ABCD. Bottom: Histograms of reliabilities and stabilities across all contrasts and regions. Background histogram shows the full range of the distribution; the inset is zoomed in and thresholded at 300 to better display the distribution of values. Orange: Reliability/stability from *a priori* ROIs for the 8 condition vs condition contrasts for which meta-analyses to guide ROI identification were available; Red: Reliability/stability from non-ROIs for those same 8 contrasts; Green: Reliability/stability from condition vs baseline contrasts; Blue: Reliability/stability from the remaining (18) condition vs condition contrasts (for which meta-analyses to guide ROI identification were not available). ROI: Regions of Interest. CvsB: Condition vs baseline. All data based on the whole brain analysis using the QC dataset.

3.5. Other factors affecting reliability and stability

A number of additional analyses were conducted to better understand possible associations between reliability, stability, and other factors. A fuller overview of results can be found in the supplemental materials, but briefly, reliability and stability were positively correlated with each other, |activity|, and |change| (Supplementary Sections S2.1.1-S2.1.4, Tables S5, S6, and S7); within-session reliability increased from baseline to follow-up, predominantly due to a drop in residual variances (S2.1.5, Tables S8 and S9, Fig. 5 and Figure S4D); relative to subcortical regions, cortical regions were more reliable, stable, had greater |activity|, and had greater (predominantly between sessions) |change| (S2.3, Table S10); occipital regions were more reliable, stable, and active relative to non-occipital regions (S2.4 and Table S11); reliability and stability were significantly higher in condition vs baseline conditions relative to condition vs condition contrasts (S2.5, Table S12, Fig. 3); controlling for pubertal variables increased stability but only by an average of 0.011 across the whole brain (S2.6, Table S13); stability stayed the same or increased for most anticipatory MID contrasts and emotion vs neutral nBack contrasts when calculated with only one run (S2.7, Table S14); within-session change was very unreliable (average 0.037) and between-session change was unstable (0.020) across the whole brain, though correlated with reliability/stability of activity and the absolute values of activity and change (S2.8, Tables S15 and S16); movement increased within session and decreased between sessions (S2.9); and increased intersession interval was associated with higher stability (S2.10, Table S17).

4. Discussion

4.1. Poor overall within-session reliability and longitudinal stability

Our main finding was that within-session reliability and longitudinal stability of individual differences in task-related brain activation was consistently poor for the publicly released fMRI data from all three ABCD tasks. Data cleaning approaches like outlier removal, movement regression, and rank normalization led to a very small, albeit statistically significant, increase in reliability and stability (average change of less than 0.015). While the finding of poor within-session reliability and longitudinal stability in the ABCD task fMRI data is concerning, it did not come as a surprise, given the mounting evidence for generally lackluster reliability of task-fMRI in mostly adult samples (Elliott et al., 2020; Herting et al., 2018; Noble et al., 2021). However, the present estimates are far below the 0.397 average reliability of task-fMRI activation estimated in the meta-analysis by Elliott et al. (2020). The question then arises, what factors could contribute to this particularly disappointing outcome? Previous reliability studies largely involved adult participants who will likely move less in the scanner and used shorter retest intervals relative to the between session analyses (within-session analyses, which have no retest interval, would presumably be subject to habituation/automation/task-reorganization effects that would diminish over a few weeks; Spohrs et al., 2018). Although average reliability and stability values in the ABCD task fMRI data are poor overall, and thus subject to a "floor effect" with limited variability of values across tasks, contrasts, and brain regions, we have examined these and

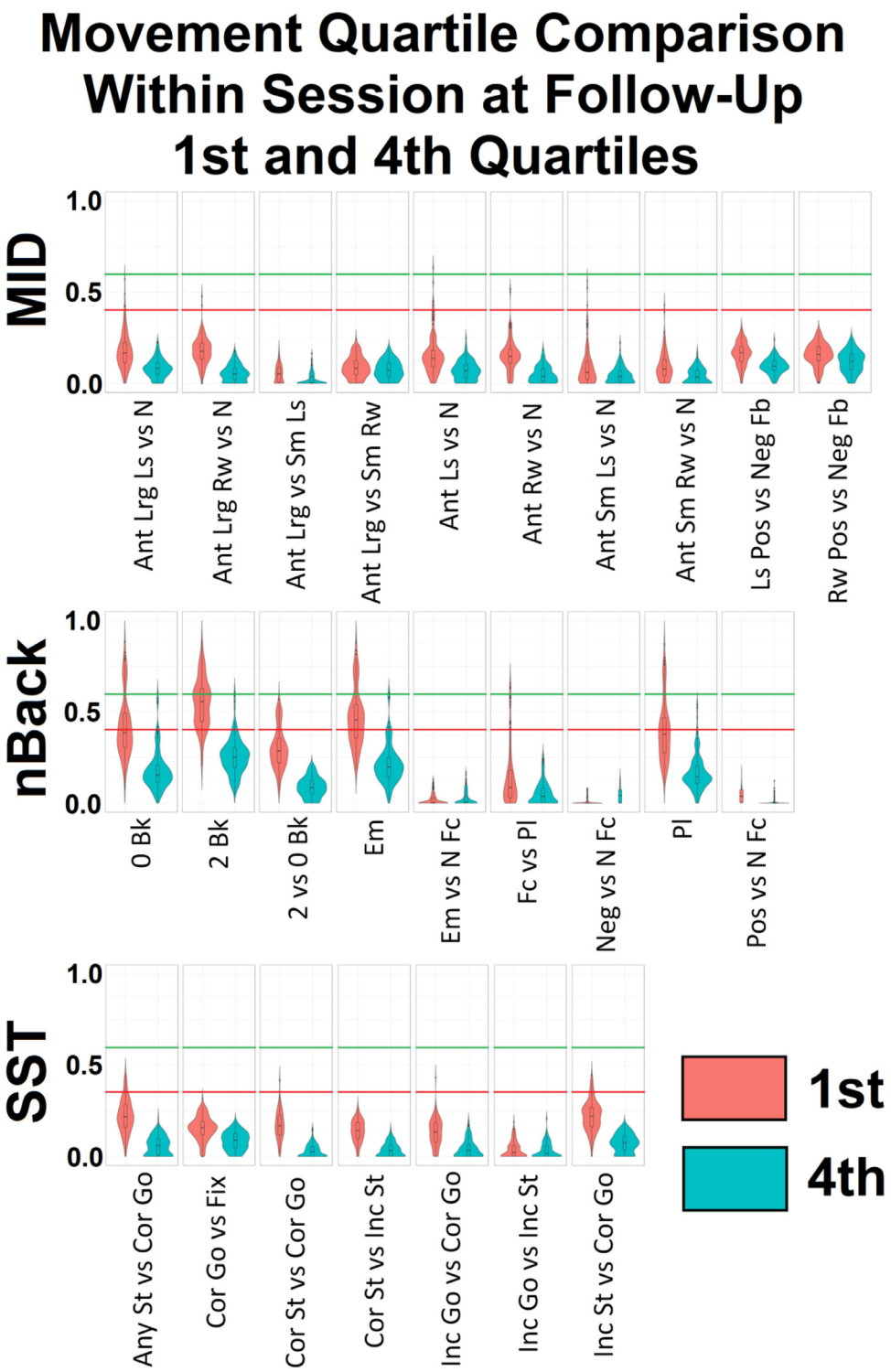

Fig. 4. Violin plots with embedded box plots showing the distribution of task and contrast specific reliability within-session at follow-up for the 1st and 4th movement quartiles, using all regions from the whole brain analysis. The movement quartiles analysis used the dataset with QC cleaning for the initial quartile separation and then had movement regressed out and outliers removed (separately for each quartile). Violin plots for within-session at baseline and between session are available as Supplemental Figures S8 and S9. Green line indicates the boundary for fair-good reliability (0.6); red line indicates the boundary for poor-fair (0.4). MID: Monetary incentive delay task, nBack: Emotional nBack task, SST: Stop signal task. Ant: Anticipation, Bk: Back, Cor: Correct, Em: Emotion, Fb: Feedback, Fc: Face, Fix: Fixation, Inc: Incorrect, Lrg: Large, Ls: Loss, N: Neutral, Neg: Negative, Pl: Place, Pos: Positive, Rw: Reward, Sm: Small, St: Stop.

other factors as potential determinants of reliability and stability. Some dataset/contrast combinations had values in the fair to excellent range, however these typically occurred in the condition vs baseline contrasts where activity is not specific to task relevant processing and in the low movement quartile datasets. Moreover, the highest reliabilities and stabilities were also in occipital lobe, raising the possibility that non-specific responses to actionable visual stimuli are what is most reliable and stable.

4.2. Factors affecting reliability and longitudinal stability

4.2.1. Task design and specific contrasts

Overall, reliability and stability were substantially higher for the working memory contrasts, although they were still in the poor range. These task differences may be related to the use of an adaptive procedure to equalize performance across subjects in the MID and SST tasks, which could also attenuate individual differences in task-related brain

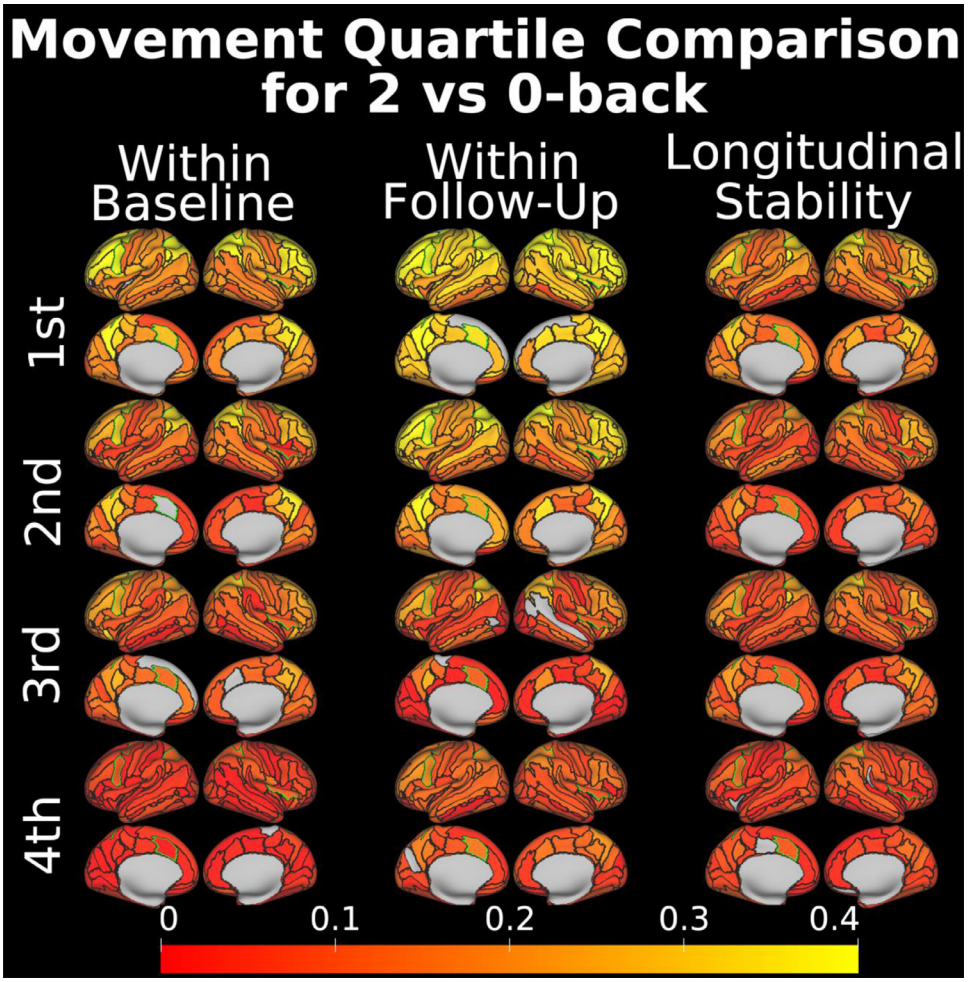

Fig. 5. Destrieux parcellation cortical reliability and stability for each movement quartile (1st = lowest, 4th = highest movement) for the nBack 2 vs 0-back contrast. LME models that did not converge are shown in gray. The *a priori* ROIs for this contrast are outlined in green.

activation, thereby reducing variance between individuals and consequently decreasing reliability/stability estimates.

Within tasks, there were differences in reliability and stability between specific contrasts, which was most evident for nBack task (because reliability for MID and SST was close to zero, there was too little variability to examine differences across contrasts within those tasks). Contrasts of an active condition vs a passive (e.g., fixation) baseline consistently showed higher reliabilities and stabilities than contrasts between two active conditions (e.g., greater reliability and stability of 2-back vs baseline compared with 2-back vs 0-back). This is consistent with psychometric and neurofunctional evidence (Baranger et al., 2021; Caruso, 2004; Infantolino et al., 2018) that contrast (difference) scores typically show lower reliability than their constituent measures because error variances of both constituents contribute to the error variance of the difference score and activity is highly correlated for different condition vs baseline contrasts (which represent the constituent measures for a direct condition vs condition contrast). For fMRI measures, this results in a trade-off between reliability or stability and validity of activation metrics. For example, an activation elicited by emotional faces relative to baseline shows higher reliabilities than activation of emotional faces relative to neutral faces (which is totally unreliable in the ABCD data). Similarly, Baranger et al. (2021) recently demonstrated using a number-guessing reward task that reward activation contrasted with baseline had greater reliability than reward contrasted directly with loss. However, contrasts with a passive baseline lack specificity because they may include nonspecific activation unrelated to the specific construct of interest (e.g., general sensory or motor related activation), resulting in poor discriminant validity. Thus, it is unclear whether the stable activation in the condition vs baseline working memory/face/emotion processing contrasts reflects functional activity related specifically to working memory/face/emotion processing.

One matter that cannot be addressed using the provided ABCD data but should be considered is if the task design and/or scanning parameters are ideal for capturing reliable and stable activity. Several of the studies analyzed in Elliott et al. (2020) meta-analysis of task reliability examined functional activity in the same domains as the tasks used in the ABCD data, with almost all finding substantially higher reliabilities than reported here, with most reporting reliability based on *a priori* ROIs (Blokland et al., 2016; Caceres et al., 2009; Cannon et al., 2017; Fliessbach et al., 2010; Fournier et al., 2014; Heckendorf et al., 2019; S. Holiga et al., 2018; Johnstone et al., 2005; Keren et al., 2018; Lois et al., 2018; Manoach et al., 2001; Nord et al., 2017; Plichta et al., 2012, 2014; Sauder et al., 2013, Schlagenhauf et al., 2007; van den Bulk et al., 2013; Wei et al., 2004; Zanto et al., 2014). While some of this can possibly be attributed to differences in demographics (e.g., age related movement differences) and scan length [highly variable, ranging from 4 min (S. Holiga et al., 2018) to nearly an hour (van den Bulk et al., 2013)], it is worth noting that there are alternate designs that may be more reliable (at least superficially in the absence of a direct comparison). For example, while the emotion processing studies examined by Elliott et al. (2021) generally had poor average reliabilities (Cannon et al., 2018; Fournier et al., 2014; S. Holiga et al., 2018; Lois et al., 2018; Nord et al., 2017; Plichta et al., 2012, 2014; Sauder et al., 2013; and van Den Bulk et al., 2013), these averages were closer to the 0.4 cutoff between 'poor' and 'fair' reliability than the 0.02 average we observed in ROIs. Due to time considerations, emotional

processing in the ABCD task fMRI is assessed as an implicit component of the nBack task, in which participants are asked to match if an image is the same as one shown at the beginning of a block (0-back) or two images earlier (2-back). But participants are not asked to examine or compare the emotions shown in the images themselves (Casey et al., 2018). In contrast, the emotional content of the images was explicitly assessed in most of the other studies (Cannon et al., 2018; S. Holiga et al., 2018; Lois et al., 2018; Nord et al., 2017; Plichta et al., 2012, 2014; Sauder et al., 2013; and van Den Bulk et al., 2013). It may be the case that emotion processing reliability is so poor in the ABCD task because emotional valence was not explicitly queried as part of the nBack task. Furthermore, scan parameters differ in some important aspects across the aforementioned earlier studies relative to ABCD, with voxel sizes typically greater than 3 mm, and repetition times (TRs) of 2 s or greater, since these studies did not use multiband acceleration. While the use of multiband acceleration in ABCD raises the possibility of some detrimental effects on reliability due to g-factor penalties and signal 'leakage' (Todd et al., 2016, 2017), the total acceleration used in the ABCD task fMRI scans is modest (multiband factor of 6, with no in-plane acceleration), and consistent with recommendations from other studies (Risk et al., 2018, 2021; Xu et al., 2013). Additionally, research from our lab (Korucuoglu et al., 2021) applying ABCD scan parameters to young adults found higher reliability estimates (∼.4) than with the children in the ABCD Study, undermining the hypothesis that scan parameters are responsible for these differences. An overview of the average reliabilities and scan parameters from studies examined in Elliott et al. (2020) meta-analysis that addressed the same domains as ABCD's fMRI tasks can be found in Supplemental Table S18.

4.2.2. Regions of interest

We hypothesized that *a priori* ROIs would be more reliable and stable than other ("non-ROI") brain regions. Task fMRI has historically been focused on the activity in particular regions engaged in particular cognitive processes, and it seemed reasonable that individual differences in the degree of activation under task performance would be more consistent in those regions than other regions that may be less constrained by the task and whose activity may fluctuate more (e.g., due to participant state). Indeed, this premise has been fundamental to the whole task fMRI endeavor. However, contrary to this expectation, our analyses show that, except for the 2 vs 0-back contrast of the nBack, *a priori* ROIs are not more reliable or stable than the rest of the brain. Furthermore, across tasks, higher reliability and stability values were observed largely in occipital regions that are generally of limited interest in the context of the neurocognitive constructs targeted by the tasks used in ABCD. Secondary analyses found that reliability and stability were also significantly correlated (across regions) with the absolute value of group-level (mean) activity in most contrasts. This was most prominent in the face vs place contrast of the nBack (correlations between 0.798–0.823), with most of these relationships having a correlation in the range of 0.4–0.5 (Table S7). This is inconsistent with our finding of greater activity in ROIs relative to non-ROIs but not accompanying greater reliability and stability in ROIs for most contrasts. A possible explanation is that the effect of activity on reliability/stability was not strong enough to manifest as greater values in ROIs relative to non-ROIs. It is worth noting that, unlike the other meta-analyses used to identify ROIs (Mueller et al., 2018; Neta et al., 2015; Oldham et al., 2017; and Swick et al., 2011), the meta-analysis used for the 2 vs 0-back was the only one based solely on children (Yaple et al., 2018). It may therefore be possible that the ROIs for the 2 vs 0-back contrast were more appropriate for the ABCD data, although this would require that activation shifts spatially in an appreciable manner with development. A different approach to identifying ROIs (e.g., data driven relative to based on published meta-analyses) may have given different results. While the lack of significant differences between ROIs and non-ROIs may make focusing on ROIs seem unwarranted, we believe it is important to recognize that having reliable and stable activity is more important in task specific regions. Additionally, without an ROI specific analysis, researchers could incorrectly assume that poor mean reliability and stability values, when averaged across the whole brain, were skewed downward by task non-relevant regions where one wouldn't necessarily expect activity to be consistent (rather, our secondary analyses found that activity in the occipital lobe was actually most reliable/stable, a region generally ignored in these specific tasks since visual activation is not the focus of the task). Notably, while we have used violin plots to illustrate distributions, all data is available as spatial maps within the BALSA database so researchers can look up the reliability/stability of a specific region for a specific contrast for a specific subset of data (i.e., cleaning method or movement quartile).

4.2.3. In-scanner movement

To examine the effect of movement on reliability and stability, participants were separated into quartiles based on movement and reliability and stability values were calculated separately for each quartile. The comparison of values between movement quartiles showed that the lowest quartile (the least moving participants) had an average whole brain reliability/stability of 0.143 while the highest movement quartile average was nearly half that value at 0.073. Although both quartile values are in the "poor" range, this significant difference indicates that efforts to mitigate the impact of movement (including frame censoring and motion parameter regression at the preprocessing stage, as well as excluding subjects with high movement at the point of defining the initial sample) did not fully control for the effect of movement on reliability and stability. Decreased values due to movement may be the result of either movement adding noise to estimated activity or a loss of data due to censoring frames with above threshold movement. Frame removal diminishes the amount of data available for analysis, which can reduce the precision of activation estimates and negatively affect reliability and stability values, resulting in a trade-off between data quality and quantity. Though this cannot be addressed using the released data, reprocessing ABCD data with different movement thresholds or removing an equal number of frames from low movement subjects (to match frame removal rates from high movement subjects) may better establish how movement affects reliability and stability. As amount of movement (mean frame-wise displacement) and number of censored frames are highly correlated ($r = 72$ in the MID task), we cannot say whether the loss of data or sub-threshold movement effects in the retained frames are responsible for the poorer reliability and stability values in high movement quartiles. More generally, our finding of a strong effect of our movement quartiles on reliability and stability values calls for approaches to reduce the impact of movement. While the large ABCD sample size means there is still sufficient power to identify effects with only a quarter of the sample, movement itself is frequently associated with measures of interest (e.g., fluid intelligence, externalizing behavior, adiposity; K. Hodgson et al., 2017; Lukoff et al., 2020; Siegel et al., 2017) and limiting analyses to only low movement samples may therefore lead to a biased sample with results that are not representative of the general population. Data driven noise removal (ICA-FIX, Salimi-Khorshidi et al., 2014) has been found to increase reliability in high movement adult participants, though by only 0.06–0.08 (Korucuoglu et al., 2021). However, given substantially more movement in children, ICA-FIX may potentially lead to larger reliability and stability gains in children, including ABCD data.

4.3. Implications of low reliability in the ABCD task fMRI data

The main (and certainly unwelcome) conclusion from the present analysis is that poor reliability and stability of child task fMRI activity in the MID, nBack, and SST tasks of the currently released ABCD data calls into question their suitability for many analyses focused on individual differences, as well as any analyses that rely on the assumption (explicit or implicit) that brain activation measures represent reliable and stable trait-like variables. Such studies include correlations between brain activations and individual differences in behavior or psychopathology (particularly, prospective longitudinal brain-behavior associations), within-

subject analyses of longitudinal changes, genetic associations, effects of individual differences in environmental exposures, and many other research designs. Reliability imposes the upper limit on the measurable correlation between variables (Nunnally et al., 1970; Vul et al., 2009), and traits with low reliability or stability cannot produce high correlations with other traits, even other highly reliable or stable ones. One particular positive of the ABCD Study is that its very large sample size affords enough statistical power to detect significant correlations even with low-reliability/stability traits. These correlations will predictably be very low, though that does in itself not preclude the ability to generate some predictive insights into biological mechanisms (Dick et al., 2021).

Since ABCD is an ongoing longitudinal study, a question arises whether there is a possibility that poor reliability and stability found in the present analysis is related to the participants' young age, and thus whether, in subsequent longitudinal waves, these values will improve. Some evidence supports this expectation. Our secondary analyses found that within-session reliability increased from the baseline to follow-up session for most contrasts, albeit by a small amount. Part of this increase is likely due to a reduction in detrimental movement-related effects, since movement decreased between sessions. However, the overall increase over two years was small, with the largest increase in average whole-brain contrast wide reliability being 0.04. Nevertheless, one can reasonably expect at least some small improvement of reliability/stability with age, at least until the propensity to move in the scanner stabilizes (around the mid-teenage years; Satterthwaite et al., 2012). Our recent study of test-retest reliability of the ABCD SST task in a sample of young adults showed fair and even good reliability for some contrasts/ROIs, though using a different preprocessing pipeline and parcellation (Korucuoglu et al., 2021).

Another parsimonious account of the lackluster reliability and stability values found here (as well as an account for the slight improvement with age from baseline to follow-up) is the possibility of inconsistent task engagement in children compared to adults. This has been evidenced not only by decreases in trial-to-trial reaction time variability from childhood to adulthood in signal detection tasks (Tamnes et al., 2012), but also evidenced in developmental pupillometry studies, where for example, task-demand-elicited noradrenergic activation (indexed by pupil dilation) waned during memory encoding in children, while remaining active in adults (Johnson et al., 2014). This effect was correlated with poorer recall in children. It stands to reason that as ABCD participants mature into more consistent task engagement, this will entail deeper and more consistent encoding of task information, that would lend itself to greater reliability and stability.

Researchers using ABCD task-fMRI data are strongly urged to select variables that show at least some trait stability and evaluate the upper boundary of expected correlations or effect sizes for other analyses. For example, attenuation of observed correlation between two variables can be easily estimated if reliabilities or stabilities of both variables are known with the formula $r_{ObservedA,ObservedB} = r_{A,B} * sqrt(Reliability_A * Reliability_B)$ where $r_{A,B}$ is the "true" correlation between two constructs (Nunnally, 1970); in the ABCD sample, longitudinal stability of non-imaging variables can be readily computed using data from subsequent assessment waves. However, reporting "reliability adjusted" correlations is generally inadvisable as the measurement errors responsible for low reliability or stability can be correlated between variables and applying the above formula can bias results, erroneously increasing or decreasing the estimated "true" correlations (Saccenti et al., 2020). For cognitive neuroscience research outside of ABCD, we suggest that establishing and reporting test-retest reliability and stability of task-fMRI phenotypes is imperative for planning studies and publishing results. In particular, computations of statistical power should account for imperfect reliability of task-fMRI data, because poor reliability leads to the reduction of the measured effect size and, consequently, increases the sample size needed (Baugh, 2002). *Post hoc* power analyses (calculated with the pwr.r.test R function; Champely, 2020) examining the sample sizes needed to find a significant (alpha = 0.05) correlation between a variable with a reliability of 0.8 and a true correlation of 0.3 with variables with reliabilities of 0.100, 0.110, and 0.076 (the average reliabilities within-session at baseline, at follow-up, and the average longitudinal stability for the *a priori* ROIs from the QC+OR dataset), found that sample sizes would need to be 1085, 988, and 1432 participants, respectively. While those numbers are far below the sample available for the ABCD Study, this greatly exceeds the average sample size for fMRI studies (Poldrack et al., 2017) and is consistent with research finding that large samples are needed to find a consistent correlation between imaging data and other variables (Marek et al., 2021).

The preponderance of small effects in imaging research that would necessarily result from poor reliability and stability is one of the reasons large, consortium-scale studies like the ABCD are needed (Dick et al., 2021). As the statistical approaches to increasing reliability/stability addressed here had small effects that did not increase reliabilities or stabilities out of the 'poor' range, developmental task fMRI researchers may need to plan studies around the limitation of poor task reliability. Since restricting analyses to low movement participants doubled reliability/stability relative to high movement, an even greater emphasis on accounting for movement, either through participant training or processing, may be warranted (although given the already profound emphasis on the movement confound, with no clear solution to date, progress on this front may be challenging). Neglecting the reliability and stability challenges in task fMRI research may result in further proliferation of small sample, underpowered studies and dissemination of spurious, false positive and non-replicable findings that undermine the credibility of cognitive neuroscience research relying on task-fMRI data.

Researchers planning for future studies may want to take notice of a recent review by Elliott et al. (2021) where they identify four strategies to obtaining reliable fMRI data: 1) obtaining longer runs of data (somewhat contradicted by our secondary analyses finding that more data can sometimes result in lower reliability, possibly due to factors such as changing arousal); 2) modeling trial by trial variance rather than just using average beta values (a concept explored by Chen et al. (2021) using Flanker data, who report that different modeling approaches can greatly increase reliability by removing error variance across trials, though this is not possible from the currently released ABCD data); 3) using a multi-echo fMRI scan acquisition to better identify blood-oxygen level-dependent signal from noise; 4) optimizing stimulus design, with them recommending using more naturalistic, ecologically valid stimuli rather than the abstract constructs on a black screen that are common in fMRI task research.

Our results do not necessarily mean that task fMRI activity is inherently unreliable or unstable. It remains unknown how much the reliability of task fMRI could be increased by acquiring more data per individual, although the resting-state literature suggests that the gains could be substantial (Birn et al., 2013; Gordon et al., 2017) if the challenges regarding learning and adaptation effects in task performance can be managed. Also, the task data released by the ABCD Study reflects only one approach to task fMRI processing and processing approaches can vary substantially, as do subsequent results, even when using the same data (Botvinik-Nezer et al., 2020). Identifying processing approaches that promote reliable and stable individual level data, rather than just maximizing the statistical significance of group-level activity, is vital to identifying reproducible individual differences in functional activity. Cohort (e.g., age) effects may also be a profound factor in the current results. For example, we have reported that SST task activity in young adults has fair to good reliability [using an intraclass correlation (ICC) approach] while using the same scanner, task design, and scan acquisition parameters as the ABCD Study, but processed using a different pipeline (though also with a shorter intersession interval of ~6 months; Korucuoglu et al., 2021). While an increase in reliability with age is expected (due to less motion), we cannot definitively say that this was the source of the higher reliability in that study, since there were processing

differences as well, including the use of the Human Connectome Project pipelines (Glasser et al., 2013) and parcellation using a more functionally relevant multi-modal parcellation (Glasser et al., 2016). Notably, ICA-FIX (Glasser et al., 2018; Salimi-Khorshidi et al., 2014) was able to increase reliability in subjects with high movement (Korucuoglu et al., 2021), albeit to a small extent (average whole brain increase in ICC of 0.06). Ultimately, alternate approaches to processing ABCD data and better accounting for noise and movement may result in more reliable data.

4.4. Limitations

These analyses are not without limitations. Reliability values will be influenced by differences in within-session change while stability will be influenced by between-session change/development that cannot be fully accounted for given the way the data was processed for public release (i.e., whole run beta values compared to a more granular analysis of possible temporal effects, such as block by block estimates of activation). Also, the number of sessions available (2) is currently a limitation, as more advanced statistical approaches to modeling and accounting for individual differences in change are not available with only two sessions of data (e.g., mixed effects modeling of nonlinear trajectories, Herting et al., 2018). Meta-analyses were only available for 8 contrasts, so *a priori* ROIs were not identified for the remaining 18 of 26 released contrasts. We investigated reliability in a univariate framework, and it is possible that more multivariate-oriented analyses will have higher reliability (Kragel et al., 2021), although this remains to be established. The inclusion of site variance in our stability estimates likely mixes stable variation due to demographic differences or sample ascertainment biases with stable variation due to undesired differences in data collection. (However, this biases the reliability/stability estimates in a strictly positive manner; excluding site variance would have resulted in even lower values.) As the ABCD Study initiated enrollment with a narrow age range, participant age and date are highly correlated across the longitudinal waves, thus age effects are potentially confounded with undesirable temporal changes (e.g., changes in scanner performance over time). Last, data was only available for structurally-based parcellations. However, functionally derived parcellations are frequently more granular, likely to be more relevant, and may be accompanied by increased reliability and stability.

5. Conclusions

Overall, reliability and stability of task-fMRI data in the ABCD sample was very poor. Movement decreases reliability and stability values, but even selecting only the lowest movement quartile for analysis didn't raise average reliability or stability out of the poor range. Reliability and stability values were only very minimally improved by the investigated data cleaning approaches. Reliability and stability were generally not better in *a priori* ROIs relative to the rest of the brain. Reliability/stability tended to be best in working memory related and condition vs baseline contrasts. Decreases in movement with age may somewhat increase reliability and stability in later ABCD assessment waves. For the amount of task fMRI data collected in the current study (~ 10 min per participant), using the ABCD imaging protocol and current ABCD analysis pipelines, the MID and SST tasks, and to a lesser extent the nBack task as well, may not be practical for other studies examining childhood development unless they can obtain sample sizes in the 1500+ range.

Data and code availability

All data used in these analyses can be obtained at https://dx.doi.org/10.15154/1523041.

All code is provided as a supplement.

Declaration of Competing Interest

None

Credit authorship contribution statement

James T. Kennedy: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Michael P. Harms:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Ozlem Korucuoglu:** Writing – original draft, Writing – review & editing. **Serguei V. Astafiev:** Writing – original draft, Writing – review & editing. **Deanna M. Barch:** Resources, Writing – original draft, Writing – review & editing. **Wesley K. Thompson:** Software, Formal analysis, Writing – original draft, Writing – review & editing. **James M. Bjork:** Writing – original draft, Writing – review & editing. **Andrey P. Anokhin:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development [SM] (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9–10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health (NIH) and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, U24DA041147. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. The ABCD data used in this report came from NIMH Data Archive DOI: 10.15154/1523041. Data analyses were partially supported by NIH grant R01HD083614.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119046.

References

Baranger, D.A.A., Lindenmuth, M., Nance, M., Guyer, A., Keenan, K., Hipwell, A.E., et al., 2021. The longitudinal stability of fMRI activation during reward processing in adolescents and young adults. Neuroimage 232, 117872.

Baugh, F., 2002. Correcting effect sizes for score reliability: a reminder that measurement and substantive issues are linked inextricably. Educ Psychol Meas 62 (2), 254–263.

Birn, R.M., Molloy, E.K., Patriot, R., Parker, T., Meier, T.B., Kirk, G.R., et al., 2013. The effect of scan length on the reliability of resting-state fMRI connectivity estimates. Neuroimage 83, 550–558.

Bjork, J.M., Straub, L.K., Provost, R.G., Neale, M.C, 2017. The ABCD study of neurodevelopment: identifying neurocircuit targets for prevention and treatment of adolescent substance abuse. Curr. Treat Options Psychiatry 4 (2), 196–209.

Blakemore, S.J., Burnett, S., Dahl, R.E., 2010. The role of puberty in the developing adolescent brain. Hum. Brain Mapp. 31 (6), 926–933.

Blokland, G.A.M., Wallace, A.K., Hansell, N.K., Thompson, P.M., Hickie, I.B., Montgomery, G.W., et al., 2017. Genome-wide association study of working memory brain activation. Int. J. Psychophysiol. 115, 98–111.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., et al., 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582 (7810), 84–88.
Brandt, D.J., Sommer, J., Krach, S., Bedenbender, J., Kircher, T., Paulus, F.M., Jansen, A., 2013. Test-retest reliability of fMRI brain activity during memory encoding. Front Psych. 4, 163.
Briesch, A.M., Swaminathan, H., Welsh, M., Chafouleas, S.M., 2014. Generalizability theory: a practical guide to study design, implementation, and interpretation. J. Sch. Psychol. 52 (1), 13–35.
Bright, M.G., Murphy, K., 2017. Cleaning up the fMRI time series: mitigating noise with advanced acquisition and correction strategies. Neuroimage 154, 1–3.
Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45 (3), 758–768.
Cannon, T.D., Cao, H., Mathalon, D.H., Gee, D.G.the NAPLS consortium, 2018. Reliability of an fMRI paradigm for emotional processing in a multisite longitudinal study: clarification and implications for statistical power. Hum. Brain Mapp. 39, 599–601.
Caruso, J.C., 2004. A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. Eur. J. Psychol. Assess 20, 166–171.
Casey, B.J., Cannonier, T., Conley, M.I., Cohen, A.O., Barch, D.M., Heitzeg, M.M., et al., 2018. The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. Dev. Cogn. Neurosci. 32, 43–54.
Chaarani, B., Hahn, S., Allgaier, N., Adise, S., Owens, M.M., Juliano, A.C., et al., 2021. Baseline brain function in the preadolescents of the ABCD Study. Nat. Neurosci. 24, 1176–1186.
Champely, S. (2020). pwr: basic Functions for Power analysis. R package version 1.3-0. [Computer software]. Retrieved From https://CRAN.R-project.org/package=pwr
Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6 (4), 284–290.
Chen, G., Pine, D.S., Brotmann, M.A., Smith, A.R., Cox, R.W., Haller, S.P., 2021. Trial and error: a hierarchical modeling approach to test-retest reliability. Neuroimage 245, 118547.
Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173.
Dai, J., Scherf, K.S., 2019. Puberty and functional brain development in humans: convergence in findings? Dev Cogn Neurosci 39, 100690.
Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980.
Destrieux, C., Fischl, B., Dale, A., Halgren, E., 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53 (1), 1–15.
Diedrichsen, J., Shadmehr, R., 2005. Detecting and adjusting for artifacts in fMRI time series data. Neuroimage 27 (3), 624–634.
Dick, A.S., Watts, A.L., Heeringa, S., Lopez, D.A., Bartsch, H., Fan, C.C., et al., 2021. Meaningful associations in the adolescent brain cognitive development study. Neuroimage 239, 118262.
Eickhoff, S.B., Bzdoc, D., Laird, A.R., Roski, C., Caspers, S., Zilles, K., et al., 2011. Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation. Neuroimage 57, 938–949.
Elliott, M.L., Knodt, A.R., Hariri, A.R., 2021. Striving toward translation: strategies for reliable fMRI measurement. Trends Cogn. Sci. 25 (9), 776–787.
Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., et al., 2020. What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. Psychol. Sci. 31 (7), 792–806.
Engelhardt, L.E., Roe, M.A., Juranek, J., DeMaster, D., Harden, K.P., Tucker-Drob, D.M., et al., 2017. Children's head motion during fMRI tasks is heritable and stable over time. Dev Cogn Neurosci 25, 58–68.
Feldstein Ewing, S.W., Bjork, J.M., Luciana, M, 2018. Implications of the ABCD study for developmental neuroscience. Dev Cogn Neurosci 32, 161–164.
Fischl, B., Salat, D.H., Busa, E., Albert, M., Dietrich, M., Haselgrove, C., et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.
Fliessbach, K., Rohe, T., Linder, N.S., Trautner, P., Elger, C.E., Weber, B., 2010. Retest reliability of reward-related BOLD signals. Neuroimage 50 (3), 1168–1176.
Fournier, J.C., Chase, H.W., Almeida, J., Phillips, M.L., 2014. Model specification and the reliability of fMRI results: implications for longitudinal neuroimaging studies in psychiatry. PLoS One 9 (8), e105169.
Garavan, H., Bartsch, H., Conway, K., Decastro, A., Goldstein, R.Z., Heeringa, S., et al., 2018. Recruiting the ABCD sample: design considerations and procedures. Dev Cogn Neurosci 32, 16–22.
Giedd, J.N., Keshavan, M., Paus, T., 2008. Why do many psychiatric disorders emerge during adolescence? Nat. Rev. Neurosci. 9 (12), 947–957.
Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., et al., 2013. The minimal preprocessing pipelines for the human connectome project. Neuroimage 80, 105–124.
Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., et al., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536 (7615), 171–178.
Glasser, M.F., Coalson, T.S., Bijsterbosch, J.D., Harrison, S.J., Harms, M.P., Anticevic, A., et al., 2018. Using temporal ICA to selectively remove global noise while preserving global signal in functional MRI data. Neuroimage 181, 692–717.
Gordon, E.M., Laumann, T.O., Gilmore, A.W., Newbold, D.J., Greene, D.J., Berg, J.J., et al., 2017. Precision functional mapping of individual human brains. Neuron 95 (4), 791–807.
Hagler Jr., D.J., Hatton, S.N., Cornejo, M.D., Makowski, C., Fair, D.A., Dick, A.S., et al., 2019. Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. Neuroimage 202, 116091.
Heale, R., Twycross, A., 2015. Validity and reliability in quantitative studies. Evid. Based Nurs. 18, 66–67.
Heckendorf, E., Bakermans-Kranenburg, M.J., van Ijzendoorn, M.H., Huffmeijer, R., 2019. Neural responses to children's faces: test–retest reliability of structural and functional MRI. Brain Behav 9, e01192.
Hodgson, K., Poldrack, R.A., Curran, J.E., Knowles, E.E., Mathias, S., Goring, H.H.H., et al., 2017a. Shared genetic factors influence head motion during MRI and body mass index. Cerebral Cortex 27 (12), 5539–5546.
Holiga, S., Sambataro, F., Luzy, C., Greig, G., Sarkar, N., Renken, R.J., Dukart, J., 2018a. Test-retest reliability of task-based and resting-state blood oxygen level dependence and cerebral blood flow measures. PLoS One 13 (11), e0206583.
Herting, M.M., Gautam, P., Chen, Z., Nezherm, A., Vetter, N.C., 2018. Test-retest reliability of longitudinal task-based fMRI: implications for developmental studies. Dev. Cogn. Neurosci. 33, 17–26.
IBM Corp, 2020. IBM SPSS Statistics for Windows, Version 27.0 [Computer Software]. IBM Corp, Armonk, NY.
Infantolino, Z.P., Luking, K.R., Sauder, C.L., Curtin, J.J., Hajcak, G., 2018. Robust is not necessarily reliable: from within-subjects fMRI contrasts to between-subjects comparisons. Neuroimage 173, 146–152.
Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. FSL. Neuroimage 62, 782–790.
Johnson, E.L., Miller Singley, A.T., Peckham, A.D., Johnson, S.L., Bunge, S.A., 2014. Task-evoked pupillometry provides a window into the development of short-term memory capacity. Front Psychol 5 (218). doi:10.3389/fpsyg.2014.00218.
Johnstone, T., Somerville, L.H., Alexander, A.L., Oakes, T.R., Davidson, R.J., Kalin, N.H., Whalen, P.J., 2005. Stability of amygdala BOLD response to fearful faces over multiple scan sessions. Neuroimage 25 (4), 1112–1123.
Keren, H., Chen, G., Benson, B., Ernst, M., Leibenluft, E., Fox, N.A., Stringaris, A., 2018. Is the encoding of reward prediction error reliable during development? Neuroimage 178, 266–276.
Korucuoglu, O., Harms, M.P., Astafiev, S.V., Kennedy, J.T., Golosheykin, S., Barch, D.M., et al., 2020. Test-retest reliability of fMRI-measured brain activity during decision making under risk. Neuroimage 214, 116759.
Korucuoglu, O., Harms, M.P., Astafiev, S.V., Golosheykin, S., Kennedy, J.T., Barch, D.M., et al., 2021. Test-retest reliability of neural correlates of response inhibition and error monitoring: an fMRI study of a stop-signal task. Front Neurosci. 15, 624911.
Kragel, P.A., Han, X., Kraynak, T.E., Gianaros, P.J., Wager, T.D., 2021. Functional MRI can be highly reliable, but it depends on what you measure: a commentary on elliott et al. (2020). Psychol. Sci. 32 (4), 622–626.
Lois, G., Kirsch, P., Sandner, M., Plichta, M.M., Wessa, M., 2018. Experimental and methodological factors affecting test-retest reliability of amygdala BOLD responses. Psychophysiology 55 (12), e13220.
Lukoff, J., Bashford-Largo, J., Zhang, R., Elowsky, J., Carollo, E., Debbertin, M., et al., 2020. Association of different types of externalizing conditions with head motion (HM) during fMRI acquisition. Biol. Psychiatry 87 (9), S365.
Manoach, D.S., Halpern, E.F., Kramer, T.S., Chang, Y., Goff, D.C., Rauch, S.L., Gollub, R.L., 2001. Test-retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. Am. J. Psychiatry 158 (6), 955–958.
Marceau, K., Ram, N., Houts, R.M., Grimm, K.J., Susman, E.J., 2011. Individual differences in boys' and girls' timing and tempo of puberty: modeling development with nonlinear growth models. Dev. Psychol. 47 (5), 1389–1409.
Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S. et al. (2021). Towards reproducible brain-wide association studies. Manuscript submitted for publication.
Meng, X., Huang, D., Ao, H., Wang, X., Gao, X., 2020. Food cue recruits increased reward processing and decreased inhibitory control processing in the obese/overweight: an activation likelihood estimation meta-analysis of fMRI studies. Obes. Res. Clin. Pract. 14 (2), 127–135.
Muller, V.L., Hohner, Y., Eickhoff, S.B., 2018. Influence of task instructions and stimuli on the neural network of face processing: an ALE meta-analysis. Cortex 103, 240–255.
Neta, M., Miezin, F.M., Nelson, S.M., Dubis, J.W., Dosenbach, N.U.F., Schlagger, B.L., et al., 2015. Spatial and temporal characteristics of error-related activity in the human brain. J. Neurosci. 35 (1), 253–266.
Noble, S., Scheinost, D., Constable, R.T., 2021. A guide to the measurement and interpretation of fMRI test-retest reliability. Curr Opin Behav Sci 40, 27–32.
Nord, C.L., Gray, A., Charpentier, C.J., Robinson, O.J., Roiser, J.P., 2017. Unreliability of putative fMRI biomarkers during emotional face processing. Neuroimage 156, 119–127.
Nunnally Jr., J.C., 1970. Introduction to Psychological Measurement. McGraw-Hill.
Oldham, S., Murawski, C., Fornito, A., Youssef, G., Yucel, M., Lorenzetti, V., 2017. The anticipation and outcome phases of reward and loss processing: a neuroimaging meta–analysis of the monetary incentive delay task. Hum. Brain Mapp. 39, 3398–3418.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R. Core Team (2021). nlme: linear and nonlinear mixed effects models. R package version 3.1-152, https://CRAN.R-project.org/package=nlme
Plichta, M.M., Schwarz, A.J., Grimm, O., Morgen, K., Mier, D., Haddad, L., et al., 2012. Test–retest reliability of evoked BOLD signals from a cognitive–emotive fMRI test battery. Neuroimage 60, 1746–1758.
Plichta, M.M., Grimm, O., Morgen, K., Mier, D., Sauer, C., Haddad, L., et al., 2014. Amygdala habituation: a reliable fMRI phenotype. Neuroimage 103, 383–390.
Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafo, M.R., et al., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. 18, 115–126.
R Core Team, 2021. R: A language and Environment For Statistical Computing [Computer

Software]. R Foundation for Statistical Computing, Vienna, Austria Retrieved from https://www.R-project.org/ .
Risk, B.B., Kociuba, M.C., Rowe, D.B, 2018. Impacts of simultaneous multislice acquisition on sensitivity and specificity in fMRI. Neuroimage 172, 538–553.
Risk, B.B., Murden, R.J., Wu, J., Nebel, M.B., Venkatarman, A., Zhang, Z., et al., 2021. Which multiband factor should you choose for your resting-state fMRI study? Neuroimage 234, 117965.
Saccenti, E., Hendriks, M.H.W.b., Smilde, A.K., 2020. Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models. Sci. Rep. 10, 438.
Salimi-Khorshidi, G., Douaud, G., Beckmann, C.F., Glasser, M.F., Griffanti, L., Smith, S.M., 2014. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. Neuroimage 90, 449–468.
Satterthwaite, T.D., Wolf, D.H., Loughead, J., Ruparel, K., Elliott, M.A., Hakonarson, H., et al., 2012. Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth. Neuroimage 60 (1), 623–632.
Sauder, C.L., Hajcak, G., Angstadt, M., Phan, K.L., 2013. Test-retest reliability of amygdala response to emotional faces. Psychophysiology 50 (11), 1147–1156.
Schlagenhauf, F., Juckel, G., Koslowski, M., Kahnt, T., Knutson, B., Dembler, T., et al., 2008. Reward system activation in schizophrenic patients switched from typical neuroleptics to olanzapine. Psychopharmacology (Berl.) 196 (4), 673–684.
Sheffield Morris, A., Squeglia, L.M., Jacobus, J., Silk, J.S, 2018. Adolescent brain development: implications for understanding risk and resilience processes through neuroimaging research. J. Res. Adolesc. 28 (1), 4–9.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull., 86 (2), 420–428.
Siegal, J.S., Mitra, A., Laumann, T.O., Seitzman, B.A., Raichle, M., Corbetta, M., et al., 2017. Data quality influences observed links between functional connectivity and behavior. Cerebral. Cortex 27 (9), 4492–4502.
Spohrs, J., Bosch, J.E., Dommes, L., Beschoner, P., Stingl, J.C., Geiser, F., et al., 2018. Repeated fMRI in measuring the activation of the amygdala without habituation when viewing faces displaying negative emotions. PloSOne 13 (6), e0198244.
Swick, D., Ashley, V., Turken, U., 2011. Are the neural correlates of stopping and not going identical? Quantitative meta-analysis of two response inhibition tasks. Neuroimage 56 (3), 1655–1665.
Tamnes, C.K., Fjell, A.M., Westlye, L.T., Ostby, Y., Walhovd, K.B., 2012. Becoming consistent: developmental reductions in intraindividual variability in reaction time are related to white matter integrity. J. Neurosci. 32 (3), 972–982.
Taylor, K.S., Davis, K.D., 2009. Stability of tactile- and pain-related fMRI brain activations: an examination of threshold-dependent and threshold-independent methods. Hum. Brain Mapp. 30 (7), 1947–1962.
Todd, N., Moeller, S., Auerbach, E.J., Yacoub, E., Flandin, G., Weiskopf, N., 2016. Evaluation of 2D multiband EPI imaging for high-resolution, whole-brain, task-based fMRI studies at 3T: sensitivity and slice leakage artifacts. Neuroimage 124 (Part A), 34–42.
Todd, N., Josephs, O., Zeidman, P., Flandin, G., Moeller, S., Weiskopf, N., 2017. Functional sensitivity of 2D simultaneous multi-slice echo-planar imaging: effects of acceleration on g-factor and physiological noise. Front Neurosci 11 (158). doi:10.3389/fnins.2017.00158.
van den Bulk, B.G., Koolschijn, P.C.M.P., Meens, P.H.F., van Lang, N.D.J., van der Wee, N.J.A., Rombouts, S.A.R.B., Crone, E.A., 2013. How stable is activation in the amygdala and prefrontal cortex in adolescence? A study of emotional face processing across three measurements. Dev Cogn Neurosci 4, 65–76.
Volkow, N.D., Koob, G.F., Croyle, R.T., Bianchi, D.W., Gordon, J.A., Koroshetz, W.J., et al., 2018. The conception of the ABCD study: from substance use to a broad NIH collaboration. Dev Cogn Neurosci 32, 4–7.
Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect Psychol Sci 4 (3), 274–290.
Wang, Y., Beydoun, M.A., Min, J., Xue, H., Kaminsky, L.A., Cheskin, L.J., 2020. Has the prevalence of overweight, obesity and central obesity leveled off in the United States? Trends, patterns, disparities, and future projections for the obesity epidemic. Int. J. Epidemiol. 49 (3), 820–823.
Wei, X., Yoo, S.-S., Dickey, C.C., Zou, K.H., Guttmann, C.R.G., Panych, L.P., 2004. Functional MRI of auditory verbal working memory: long-term reproducibility analysis. Neuroimage 21 (3), 1000–1008.
Xu, J., Moeller, S., Auerback, E.J., Strupp, J., Smith, S.M., Feinberg, D.A., et al., 2013. Evaluation of slice accelerations using multiband echo planar imaging at 3 T. Neuroimage 83, 991–1001.
Yaple, Z., Arsalidou, M., 2018. N-back Working Memory Task: meta-analysis of Normative fMRI Studies With Children. Child Dev. 89 (6), 2010–2022.
Zanto, T.P., Pa, J., Gazzaley, A., 2014. Reliability measures of functional magnetic resonance imaging in a longitudinal evaluation of mild cognitive impairment. Neuroimage 84, 443–452.