# Longitudinal stability of grey matter measures varies across brain regions, imaging metrics, and testing sites in the ABCD study

Parsons[1], S., Brandmaier[2,3,4], A. M., Lindenberger[2,4], U., & Kievit[1,5], R.

[1] Donders Institute for Brain, Cognition and Behavior, Radboud University Medical Center, Nijmegen, Netherlands

[2] Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany

[3] Department of Psychology, MSB Medical School Berlin, Berlin, Germany

[4] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany

[5]Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge



Author note:



We would like to thank Léa Michel for sharing code to map ABCD grey matter measures to the ggseg package for visualisation.

## Abstract

Magnetic resonance imaging (MRI) is a vital tool for the study of brain structure and function. It is increasingly being used in individual differences research to examine brain-behaviour associations. Prior work has demonstrated low test-retest stability of functional MRI measures, highlighting the need for more work to characterise the longitudinal stability (test-retest reliability across long timespans) of MRI measures across brain regions and imaging metrics, particularly in adolescence. In this study, we examined the longitudinal stability of grey matter measures (cortical thickness, surface area, and volume) across brain regions, and testing sites in the Adolescent Brain Cognitive Development (ABCD) study release v4.0. We used Intra-Class Effect Decomposition (ICED) to estimate between-subjects variance and error variance, and assess the relative contribution of each across brain regions and testing sites on longitudinal stability. Our results highlight meaningful heterogeneity in longitudinal stability across brain regions, structural measures (cortical thickness in particular), and ABCD testing sites. We found that differences in longitudinal stability across brain regions was largely driven by between-subjects variance, whereas differences in longitudinal stability across testing sites was largely driven by differences in error variance. We argue that investigations such as this are essential to capture patterns of longitudinal stability heterogeneity that would otherwise go undiagnosed. Such improved understanding allows the field to more accurately interpret results, compare effect sizes, and plan more powerful studies.

## Introduction

Brain imaging techniques, including Magnetic Resonance Imaging (MRI), are indispensable for studying brain function and structure and its role in supporting cognitive development across the lifespan. In recent years, MRI has been increasingly used to examine individual differences, suggesting that, for instance, individuals with (regional) differences in cortical morphology or structural connectivity also demonstrate differences in phenotypes such as cognitive performance (Kievit et al., 2014; Magistro et al., 2015; Muetzel et al., 2015; Schnack et al., 2015). Moreover, the crucial role of (differences in) change and maturation in brain structure across the lifespan have prompted longitudinal investigations collecting multiple brain scans from individuals across the lifespan (e.g. Casey et al., 2018; Healthy Brain Study Consortium et al., 2021; von Rhein et al., 2015; Walhovd et al., 2018). Addressing individual differences questions, whether cross-sectionally or longitudinally, rests on the assumption that brain imaging measures are reliable. In other words, the inferences we can draw from such longitudinal datapoints depend on the extent to which they capture stable between-subjects differences with little contamination by within-subject fluctuations or measurement error.

More commonly than not, we do not know how reliable our measures are (Flake et al., 2017; Gawronski et al., 2011; Hussey & Hughes, 2018; Parsons et al., 2019). This basic psychometric concern does not only relate to questionnaires, but also cognitive measurements (Parsons et al., 2019) and neuroimaging metrics (Anand et al., 2022; Brandmaier, Wenger, et al., 2018; Noble et al., 2017; Wenger et al., 2021; Zuo et al., 2019). Low reliability translates to low statistical power and related challenges, including a decreased likelihood that a significant finding reflects a true effect (Button et al., 2013), is in the correct direction (Type 2 “Sign” error; Gelman & Carlin, 2014), and an inherent overestimation of the true effect size (Type M “Magnitude” error; Gelman & Carlin, 2014). In other words, if reliability is not assessed then it is impossible to gauge its impact on our results and therefore the confidence we should have them. Failing to assess reliability can become a greater, more complex, problem when we wish to compare effect sizes from different regions, measures, or studies (for examples, see Cooper et al., 2017). For example, a study may conclude that there is no difference in brain atrophy between an experimental medicine group and a control group, when in fact the clinical benefits are attenuated or hidden because of low reliability. Similarly, within studies, marked differences in reliability between brain regions could lead researchers to make incorrect conclusions about the similarity of brain-behaviour associations across these regions. As such, we propose that mapping reliability across brain regions and measures provides vital information about reliability heterogeneity. Further, exploring reliability heterogeneity may allow us to uncover sources of unreliability, and account for this in our study designs to improve precision, statistical power, and efficiency (Brandmaier et al., 2015; Brandmaier, Wenger, et al., 2018; Noble et al., 2017; Zuo et al., 2019).

**Reliability in brain imaging**

Various tools exist to examine reliability. Readers may commonly see Cronbach's alpha reported to index the internal-consistency reliability of a questionnaire, though alternatives like MacDonalds Omega are likely more suitable (McNeish, 2018). Readers may also commonly see a Pearson correlation to index the test-retest reliability of a measure. Similarly, an improved approach to examining test-retest reliability would be to use Intraclass Correlation Coefficients (Koo & Li, 2016). Broadly, the ICC quantifies the proportion of variance attributed to between-subjects variance compared to all sources of variance (including not only error, but also within-participant variance including between-sessions variance). Due to these strengths, ICCs are becoming more commonly reported in brain imaging (Noble et al., 2021). Various extensions and generalisations of the ICC exist which focus on distinct aspects such as the reliability of a single measurement, or the average of more than one, and whether one wishes to capture absolute agreement or consistency across repeated measures (for a complete introduction, see Koo & Li, 2016).

Although empirical investigations into the (un)reliability of (f)MRI are somewhat limited, existence evidence strongly suggest this reliability is considerably worse than assumed. For instance, analyses of the ABCD data showed very poor within-session reliability and two-year stability, of task-based fMRI measures, with estimates (proportion of non-scanner related between-subjects variance to all sources of variance) rarely exceeding 0.2 (Kennedy et al., 2022). One review of reported test-retest estimates (ICC) also found fMRI measures to have low reliability (mean ICC = 0.44; Bennett & Miller, 2010) and concluded that studies are needed that examine the factors that influence reliability. In Bennett and Miller, study test-retest intervals varied from less than one hour to 59 weeks, and the authors highlight a trend for lower reliability (stability) in studies with test-retest intervals longer than three months, relative to studies with intervals less than one hour. A recent meta-analysis including 90 experiments using common fMRI tasks found ICC to be around 0.4 (Elliott et al., 2020). Test-retest intervals varied from one day to 1,008 days, however, unlike Bennett & Miller, the authors found no moderating effect of test-retest interval on the meta-analytic ICC estimate. The authors identified various design factors, including scanner, subject, task, and study factors, which may help improve test-retest reliability of fMRI measures in studies of development. It is likely these recommendations are also applicable to structural MRI. Two related considerations are the size of the contribution to reliability and how difficult it is to modify (e.g. adapting study design, increasing the number of scans, etc). For example, Karch and colleagues found increased time between scans and scanning at inconsistent times of day (within and between participants) predicted reliability of several brain volume estimates (Karch et al., 2019). Maintaining the same scanning time for a participant should be a relatively easy way to boost reliability by a small increment. In contrast, additional scanning sessions quickly increase the time and cost of a study.

There is some evidence that structural measures (e.g. cortical thickness) are more reliable than functional measures (Elliott et al., 2020; Han et al., 2006). However, these studies also highlight variation in reliability across brain measures and brain regions. As

discussed above, if left undiagnosed this reliability heterogeneity can have impactful downstream consequences on the inferences we can draw from brain imaging research.

**Reliability and stability**

Consider two brain scans collected from the same individual. If, hypothetically, we observe no differences between brain images, we can infer our measure is perfectly reliable. If the second scan was obtained immediately following the first, we can assume that any differences between scans are due to some measurement error introduced during the scans or image processing, and that greater differences between these scans indicate lower reliability. However, as the time between scans increases, the difference between successive scans will reflect a combination of (un)reliability as well as true differences, or changes, in brain structure. For example, time of day (Karch et al., 2019) and hydration levels (e.g. Trefler et al., 2016) may induce differences between scans. When scans are taken months or years apart (Casey et al., 2018; Kennedy et al., 2022), it is highly likely that developmental processes have occurred: brain structure changes over time, and the rate of change depends on the region, imaging modality and lifespan stage (Bethlehem et al., 2022). Moreover, impactful events that occur in between scans such as learning new skills (e.g. Wenger et al., 2021), or adverse events such as brain injury (e.g. Lindberg et al., 2019), will lead to lasting differences. As such, differences in brain images over years necessarily reflect a combination of measurement reliability and longitudinal stability.

Traditional models used to estimate reliability focus on measurement properties and thus implicitly assume stability, i.e. no systematic changes or individual differences in change over time (Nesselroade, 1991). When we use these models to estimate reliability over long durations individual differences in change will appear as error in our model. To address this challenge, prior work tracing back to Cronbach and Furby (Cronbach & Furby, 1970; also see Hertzog & Nesselroade, 2003) has denoted reliability estimates from these models as *stability* (Brandmaier, Wenger, et al., 2018; Deary et al., 2013; Kennedy et al., 2022). The difference in interpretation relies on the tenability of the assumption that true change in the underlying system is negligible for the purposes of our repeated measurements or not. To reflect this inherent ambiguity, we follow previous work and use the term **longitudinal stability** to describe what is captured by our estimates. At the same time, we emphasise that our estimates capture a mixture of *both reliability and stability* due to expected individual differences in changes in brain structure over the lifespan. With appropriate study designs it will be possible to disentangle these distinct sources of variance, but the vast majority of longitudinal designs do not (yet) allow for this - An issue we consider further in the discussion. With the emergence of developmental or lifespan studies with long inter-scan intervals (several years in some studies), it is crucial that methodological work allows us to characterise the distinct sources of longitudinal stability across developmental time.

**Generating detailed maps of test-retest stability**

In this study we make use of the Adolescent Brain Cognitive Development longitudinal study imaging data (ABCD; Casey et al., 2018; Compton et al., 2019; https://abcdstudy.org/) to map longitudinal stability of structural brain imaging measures. The ABCD study is a

collaboration across 21 research sites across the United States including a representative sample of over 11,000 children aged 9-10, with plans to follow-up participants into young adulthood. For our purposes, the data include two brain imaging sessions at baseline and two-year follow-up. Relative to prior investigations of structural and functional MRI longitudinal stability, ABCD also offers a considerably larger sample size. For example, the estimates reported by Elliot et al. (2020) from the large-scale Human Connectome Project (Van Essen et al., 2013) and Dunedin study (Poulton et al., 2015), included only 45 and 20 participants with repeated measures, respectively. Further, with the ABCD data we had a decent test-retest sample size for each site (minimum site n = 336), allowing us to isolate these sources of (un)reliability giving us confidence in the precision of our multigroup analyses across testing sites.

In addition, we note that the opportunities to examine brain-behaviour associations using the ABCD data are vast (Feldstein Ewing et al., 2018). Hundreds of studies using the ABCD data have already been published. Given this, there is increasing importance to generate maps of longitudinal stability specifically for this cohort, to inform data users about potential undiagnosed heterogeneity in longitudinal stability. We had two main questions. First, what is the longitudinal stability of grey matter measures in the ABCD study, and do they differ across brain regions, structural metrics, and testing sites? Second, are these differences in longitudinal stability driven more by individual differences or measurement error?

## Methods

### ABCD data

We used imaging data from the Adolescent Brain Cognitive Development study (Casey et al., 2018), data release 4.0 (http://dx.doi.org/10.15154/1523041; see supplemental materials for full acknowledgement). Full design information about the ABCD study have been described previously, including: recruitment and sampling procedures (Compton et al., 2019), imaging protocol (Casey et al., 2018), details of image processing (Hagler et al., 2019), guides for researchers using this data (Saragosa-Harris et al., 2022) and an open access data from an adult equivalent of ABCD with an accelerated design (Rapuano et al., 2022).

**MRI imaging.** The raw imaging data were processed using FreeSurfer, version 5.3.0 (Laboratory for Computational Neuroimaging) by the ABCD Data Acquisition and Integration Core with a standardised ABCD pipeline (Hagler et al., 2019). Participant's images were excluded if severe imaging artifacts were detected in manual quality control checks. The Desikan-Killiany-Tourville atlas (Desikan et al., 2006) was used to parcellate images into 34 regions per hemisphere. We extracted the three derived cortical measures: cortical thickness, surface area, and volume – calculated in FreeSurfer as product of cortical thickness and surface area, though more accurate methods exist (Winkler et al., 2018) and are implemented in more recent versions of FreeSurfer (6.0.0).

**Participants.** We included data from 7,269 participants (3,354 female, 3,915 male), for whom there were two available structural MRI scans. We removed 12 participants that were labelled as belonging to a 22$^{nd}$ site, as we were unable to determine if they belonged to

any of the actual 21 ABCD testing sites. Time from baseline to follow-up scan was on average 24.5 months (SD = 2.33) apart. Mean participants age at baseline was 9 years 11 months (range 9 years 1 month to 11 years 1 month), and at the 2 year follow up was 11 years 11 months (range 10 years 5 months to 13 years 10 months).

**ICED model**

We used a two-timepoint ICED model implemented in the SEM framework (Brandmaier, Wenger, et al., 2018). Figure 1 (Left) presents a path diagram depicting the unique contribution of each source of variance. The two observed measurements are presented as rectangles, while the latent variables are presented as circles representing the sources of variance. Between-subject variance ($\sigma^2_B$) captures variance attributable to individual differences between participants. Error variance ($\sigma^2_E$) captures the remaining variance that cannot be attributed to between-subjects differences, e.g. within-subject fluctuations (hence sometimes being called *residual variance*). Single-headed arrows represent fixed regression loadings (set to 1), and double-headed arrows indicate the variance of the latent variables. The variance estimates for the two error latent variables (E1 and E2) were constrained to be equal.

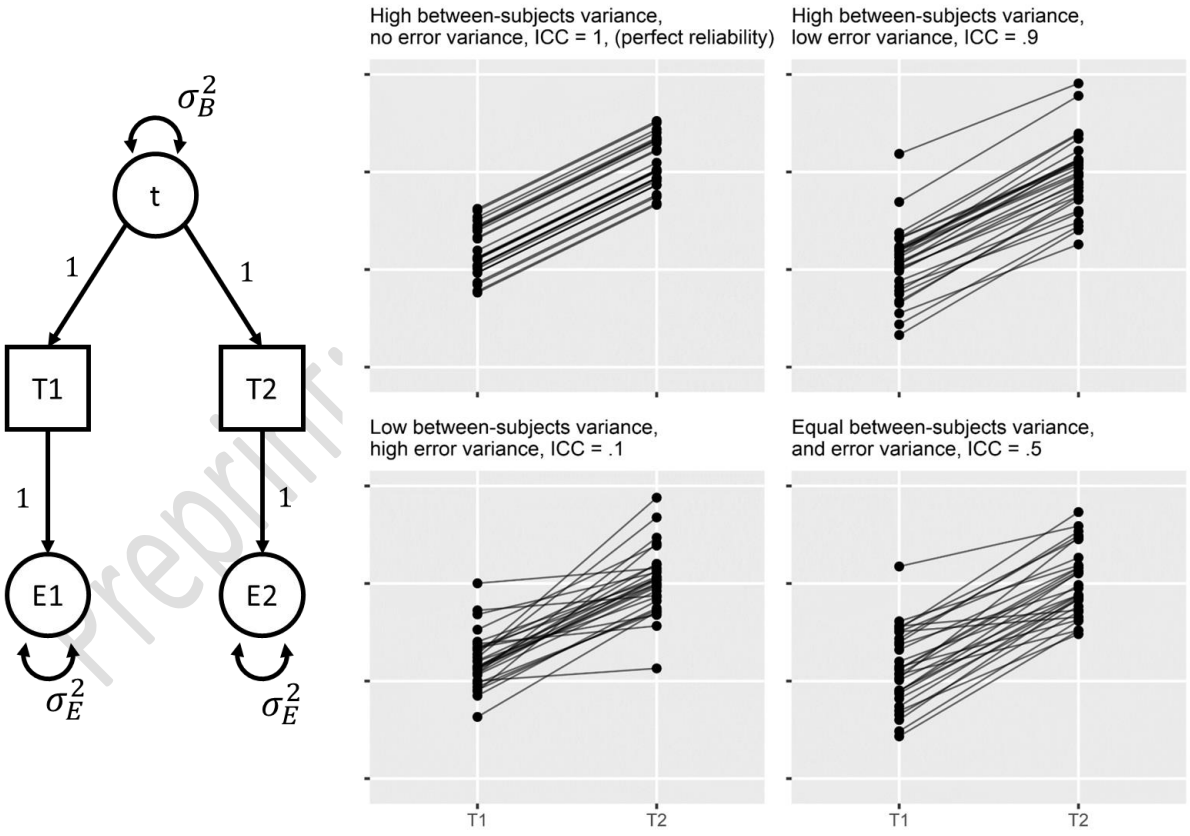


*Figure 1.* Left: Path diagram of the two timepoint ICED model used to estimate Between-subjects ($\sigma_B^2$) and Error variance ($\sigma_E^2$) components. Right: Four plots visualising hypothetical differing levels of between-subjects and error-variance, with equal total variance, to depict the relationship between test-retest reliability and rank-ordering individuals.

Using this model, longitudinal stability is estimated as Intraclass Correlation Coefficients (ICCs), which are the most common measures use test-retest reliability and longitudinal stability in neuroimaging research (Noble et al., 2021). ICC captures the reliability of an individual assessment. ICC is calculated using the between-subject variance ($\sigma_B^2$) and error variance ($\sigma_E^2$) estimates as the proportion of between-subjects variance to total observed variance (Formula 1). Higher ICCs result from the between-subjects differences (i.e. individual differences) outweighing other sources of variance, in this case "error" (other sources of variance would be added to the denominator). Figure 1 (Right) presents four sets of simulated data to demonstrate the relationship between these two sources of variance and ICC estimates. One practical important take-home message from this figure is that test-retest reliability reflects how well we are able to rank-order participants and therefore how consistent this rank ordering is over time. In the first scenario (top left), if there were no measurement error, we would observe an ICC of 1, "perfect" reliability.[1] Note that the rank ordering of participants remains the same over time. With near-perfect reliability (top right) there are some disruptions to the rank ordering, but overall we are very able to distinguish between individuals. With very low reliability (bottom left) there is very little consistency in the ordering of participants over time and we have little information with which we can distinguish between individuals. Between these two, when we have equal parts between-subjects differences and error variances (bottom right) there is some consistency, but half of our signal (that we aim to use as a measure of individual differences) is unrelated to the construct we wish to measure. Note that across each simulation the total variance and the average difference over time is the same – we highlight the latter to reinforce that we are interested in between-subjects differences instead of differences in the mean over time (which may be the use of "stability" that some readers are more familiar with). To calculate ICC from our ICED model we extract the between-subject variance ($\sigma_B^2$) and error variance estimates ($\sigma_E^2$), and use this formula:

$$ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_E^2} \quad (1)$$

Often we are interested in the reliability of the underlying construct, rather than individual indicators or observed measures. The estimate therefore takes into account the number of measurements – increasing the number of measures typically increases the reliability of the overall measure (e.g. Cronbach's alpha or ICC). In the case of ICED models, the measurement structure is also incorporated into the model (e.g. capturing repeated measures nested within days; Brandmaier, Wenger, et al., 2018). To capture the construct-level reliability using this approach, we compute the *effective error* that would emerge as the residual error if we were to directly measure the construct. Effective error is derived from power-equivalence theory (Brandmaier, von Oertzen, et al., 2018; Oertzen, 2010) and is a function of the combination of all sources of error (i.e. non between-subjects differences). Effective error can be calculated by generating a power equivalent model using the algorithm

[1] Note: unless reliability is explicitly specified in the model, most statistical tools assume perfect reliability.

provided by von Oertzen (2010) or calculating a numerical estimate following the equations in Brandmaier et al. 2018, (supplementary material 3). This provides a flexible framework to calculate effective error for any complex study design. We can then calculate the construct-level reliability as ICC2 (Bliese, 2000), as follows:

$$ICC2 = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_{EFF}^2} \quad (2)$$

Where the between-subjects variance ($\sigma_B^2$) remains the same as in the ICC calculation. All other sources of variance are incorporated into the effective error term ($\sigma_{EFF}^2$). In our two timepoint models, consisting of between-subjects and error variance, $\sigma_{EFF}^2$ is calculated as $\frac{\sigma_E^2}{N}$ where $N$ is the number of repeated measures (this follows from the multiple-indicator theorem; (Oertzen, 2010). From this, we can see that ICC2 will always show higher reliability relative to ICC, scaled by the number of independent measurements assuming measurements are independent and have identical variance. For example, our previous example of ICC = 0.5 corresponds to an ICC2 of 0.66 with two independent measurement occasions and an ICC2 of 0.75 with three independent measurement occasions. This reflects the improvement of reliability of an average score over repeated measurements by adding extra measurements. This is directly comparable to how one might improve the reliability of a questionnaire measure by increasing the number of items, for instance with reliability metrics like Cronbach's alpha (Cronbach, 1951).

We used the R package *ICED* (Parsons et al., 2022; https://github.com/sdparsons/ICED) to run these analyses, which acts as a wrapper around the *lavaan* package (Rosseel, 2012). Note that the Maximum Likelihood estimator assumes multivariate normality.

Additionally, ICED benefits from the powerful toolkit SEM offers that allow flexible modelling accommodating complex, nested study designs, including latent variables modelled by multiple indicators (e.g. left and right hemispheres as examined by Anand et al., 2022), (in)equality constraints, multigroup modelling and model comparison techniques which allow for symmetric quantification of evidence for multiple competing models (Rodgers, 2010). We make extensive use of these in this study to capture distinct sources of variance and longitudinal stability.

**Data analyses**

To address our first question (what is the longitudinal stability of grey matter measures in the ABCD study, and do they differ across brain regions, structural metrics, and testing sites?) we ran a series of ICED models (Brandmaier, Wenger, et al., 2018; for other applied studies, see Anand et al., 2022; Wenger et al., 2021). We estimated between-subject and error variances for three grey matter measures (cortical thickness, surface area, and volume) across regions of interest. We present test-retest ICCs to provide a "map" of test-retest stability across structural measures and brain regions. To address our second question (are these differences in longitudinal stability driven more by individual differences or measurement error?) we used a multigroup SEM and a series of model comparisons. We

compared the relative influence of between-subjects variance and error variance across testing sites.

Given the challenges often associated with estimating such models, we implemented an approach that balances model optimisation and generalisability (proposed by Srivastava, 2018 and others). Specifically, we initially estimated the ICED modify the model on a randomly selected subset of all the data (495 participants), to make any necessary modifications to the model needed for estimation, prior to estimating the model on the full dataset (minus the initial exploratory subset). This ensures our final model estimation is more likely to converge and yield reliable estimation whilst being less likely to be overfit to the idiosyncrasies of a specific subset of the data. Based on this test-set, we multiplied surface area and grey matter volume by an arbitrary constant (0.001) to ensure comparable variances across the three structural metrics.

**Code and data availability**

The code used for these analyses can be found in the OSF (https://osf.io/rxmn2/) and github (https://github.com/sdparsons/Longitudinal_Stability_ABCD_Grey_Matter) repositories for this project. Data from ABCD may be obtained via application from the NIMH Data Archive (https://nda.nih.gov/abcd/). The ABCD data used in this report came from release 4.0 (http://dx.doi.org/10.15154/1523041; accessed on 21st February 2022).

Readers may be interested in applying these methods to their own data or in reproducing our analyses. To make these analyses accessible and to make the *ICED* package easy to use, we simulated data for each structural measure based on the ICED variance estimates (separately for each testing site and matching the sample size at each site) and provided these in the supplemental materials.

## Results

### 1. Stability estimates

To estimate the longitudinal stability of grey matter measures, we fit our ICED model to each region across each structural measure. From each model we extracted ICC and ICC2 estimates. Figures 2 and 3 visualise the ICC and ICC2 estimates, respectively, across measures and Desikan-Killiany Cortical Atlas (Desikan et al., 2006) regions of interest using the R package *ggseg* (Mowinckel & Vidal-Piñeiro, 2019).

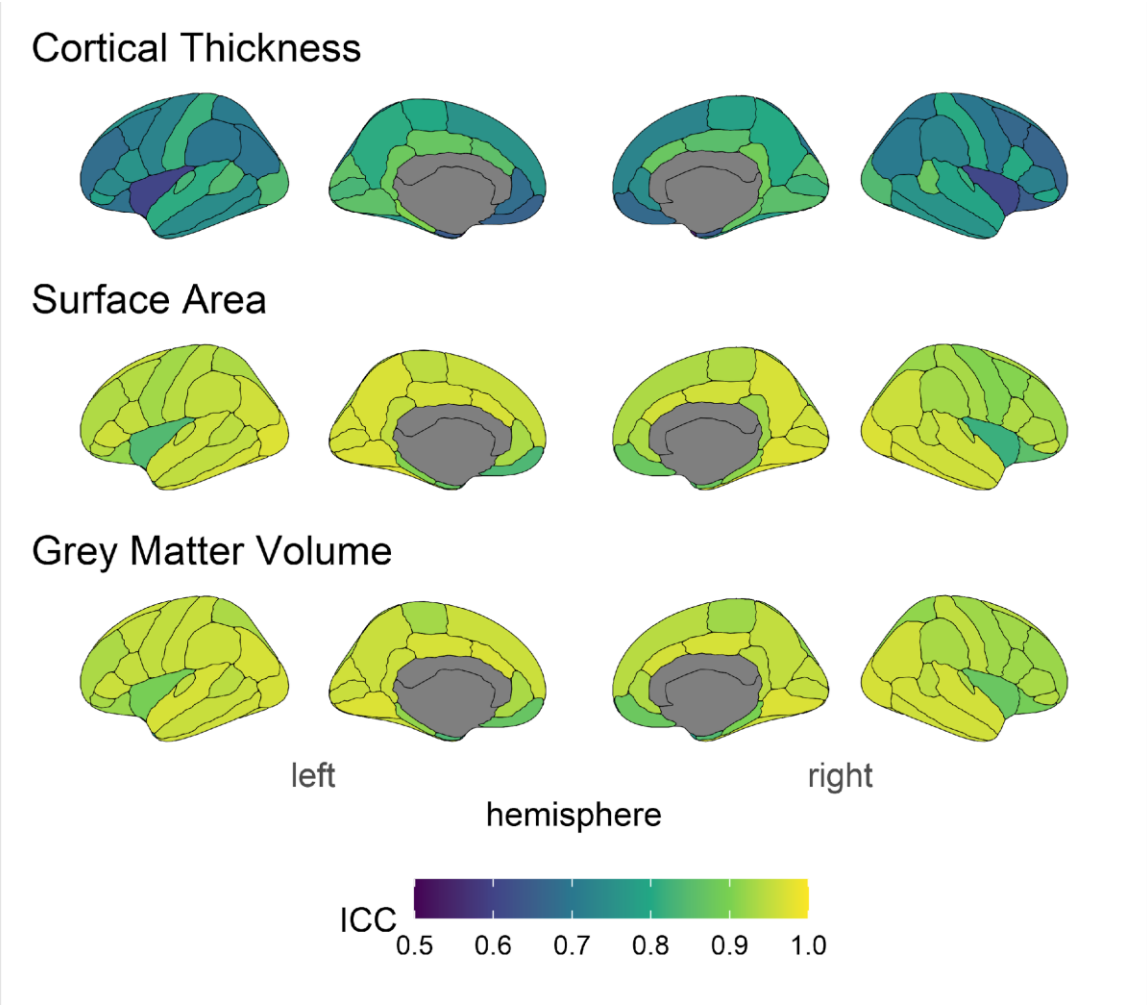


*Figure 2.* ICC estimates across structural measures and ROIs. Lighter colours indicate higher stability.

ICC estimates the longitudinal stability of an individual indicator or measurement – essentially, how reliable do we expect a single measure to be? Mean ICCs for each measure were: cortical thickness .76 (range = .54 - .90), surface area .93, (range = .82 - .97), and volume (mean = .93, range = .76 - .97). Comparing measures, cortical thickness showed an overall poorer pattern of longitudinal stability. While all estimates for surface area and volume above commonly used cut offs for "good"[2] longitudinal stability (>0.75), for cortical thickness 46% of regions had stability lower than 0.75[2]. This relatively low longitudinal stability of this measure means that true patterns or associations will likely be attenuated and/or rendered non-significant purely because of lower stability estimates. We discuss these practical implications in more detail below.

[2] There have been several recommendations for standards to judge test-retest reliability (here, longitudinal stability) estimates. For example, a common historical rule of thumb is < 0.4 is poor, 0.4 – 0.59 is good, 0.6 – 0.74 is good, and 0.75 – 1 is excellent (Cicchetti & Sparrow, 1981; Fleiss, 1986). Others have proposed stricter rules of thumb of < 0.5 is poor, 0.5 – 0.75 is moderate, 0.75 – 0.9 is good, and 0.9 – 1 is excellent (Koo & Li, 2016). In this paper we avoid adopting any specific threshold to describe our reliability estimates. We aim to avoid dichotomous thinking about whether estimates are 'good' or 'bad' in favour of considering the influence of relative differences in reliability across different measures, regions, and sites.

The ICC2 provides an estimate of longitudinal stability at the level of the construct. ICC2 estimates (Figure 3) show the same pattern of ICCs across ROIs, albeit higher values. The mean ICC2s for each measure were: Cortical Thickness = 0.86 (range = 0.70 - 0.95), surface area = 0.96 (range = 0.90 - 0.99), and volume = 0.96 (range = 0.86 - 0.99).

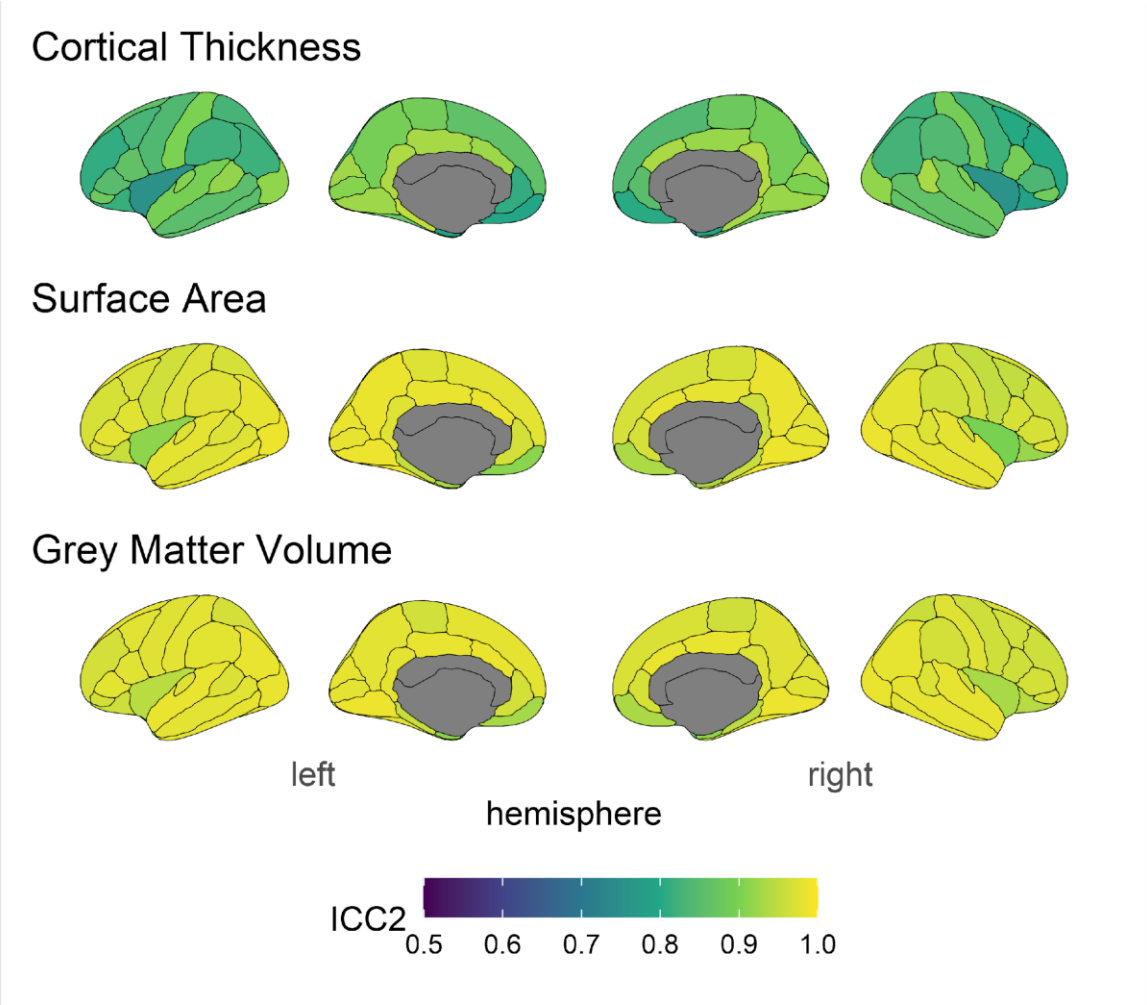


*Figure 3.* ICC2 estimates across structural measures and ROIs. Lighter colours indicate higher stability.



## 2. Examining sources of longitudinal (in)stability

To probe potential variability in stability across additional factors, we re-ran the ICED model across each of the 21 sites, again separately for each ROI. For brevity, and because Cortical Thickness showed the largest heterogeneity in ICC across brain regions, we present results from Cortical Thickness only (analysis output and figures for surface area and grey matter volume can be found in the supplemental materials). We then decomposed these longitudinal stability estimates into the between-subjects and error variance components. This allowed us to quantify the relative contributions of both variance components across brain regions and testing sites.

**Region differences**. To explore the sources of differences in stability estimates across brain regions we compared the relative size of between-subjects and error variances across each brain region. Figure 4 plots the between-subject (left panel) and error variance estimates

(right panel) for each region of interest, with each point representing a different testing site. As expected from a visual inspection of Figure 4, on average, the variance of the between-subjects variance estimates was 2.8 times larger than the error variance estimates. This suggests that differences in stability estimates across regions is likely driven more by differences in the between-subject variance than site differences in measurement error.

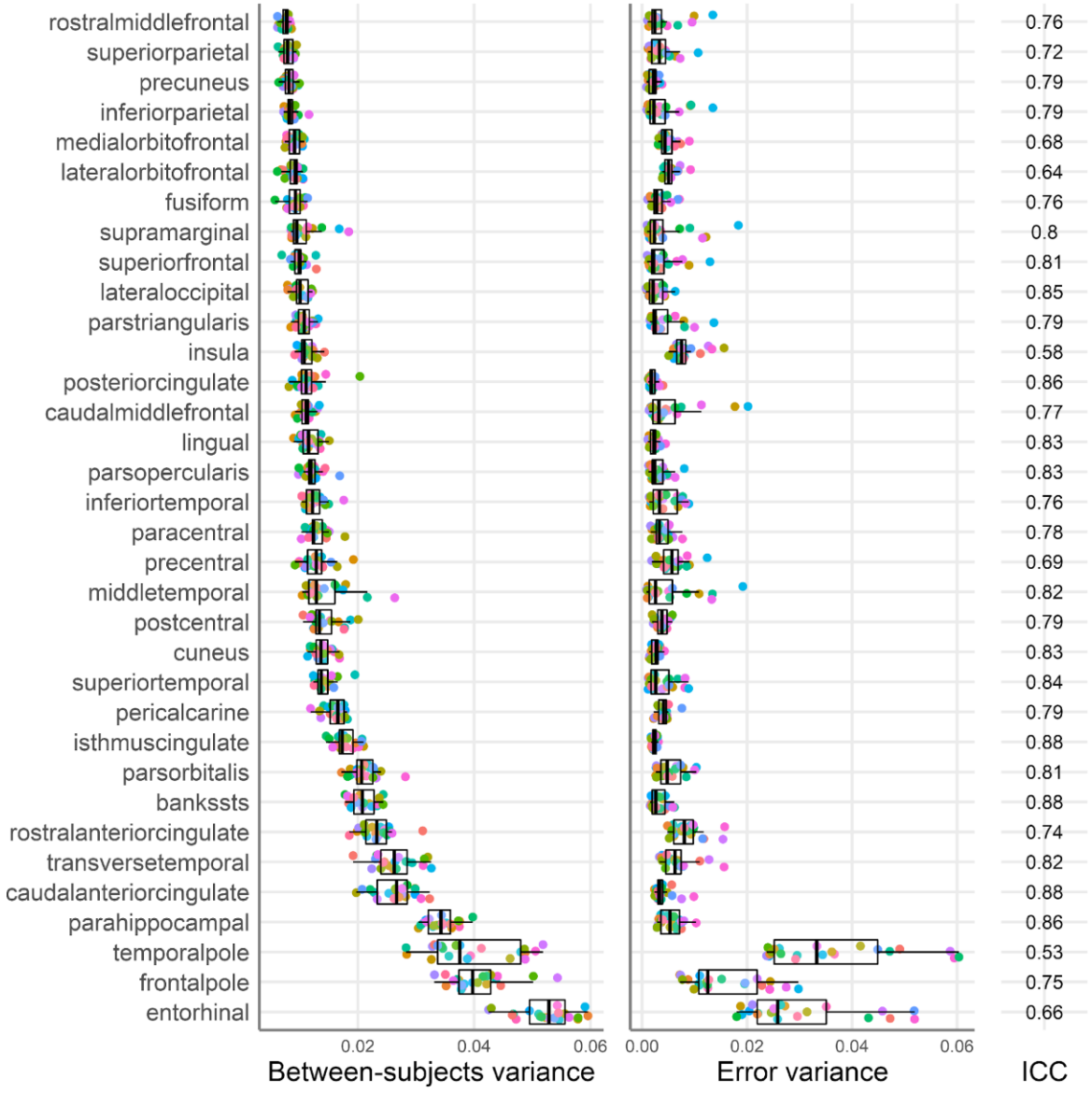


*Figure 4.* Between-subjects (left panel), error variance estimates (middle panel), and median ICC (right panel) for each region of interest (y-axis). Regions are ordered by the median between-subjects variance. For clarity we present only the right hemisphere regions. Each point represents a different testing site, and the colour mapping is the same as in Figure 5. The boxplots present the median and the 25$^{th}$ and 75$^{th}$ percentiles, the whiskers extend at a maximum to 1.5 times the interquartile range from the box.

**Site differences.** Figure 5 plots the latent between-subject and error variances across ROIs separately for each site. In contrast to Figure 4, the distributions of between-subject variance estimates are largely overlapping across sites. In contrast, the distributions of error variance differ markedly across sites in both the median estimate and the interquartile ranges. To help quantify the difference in contributions from between-subject and error variance, we extracted the median variance estimates for each region and calculated the variance of these estimates to compare the spread of between-subject variance and error variance. Across sites, there was 11.5 times more variance in the median error variance estimate than the median between-subject variance estimate. This suggests that differences in stability across testing sites is driven mainly by differences in error across sites, rather than genuine differences between people in each location. We later discuss potential causes of these differences in error.

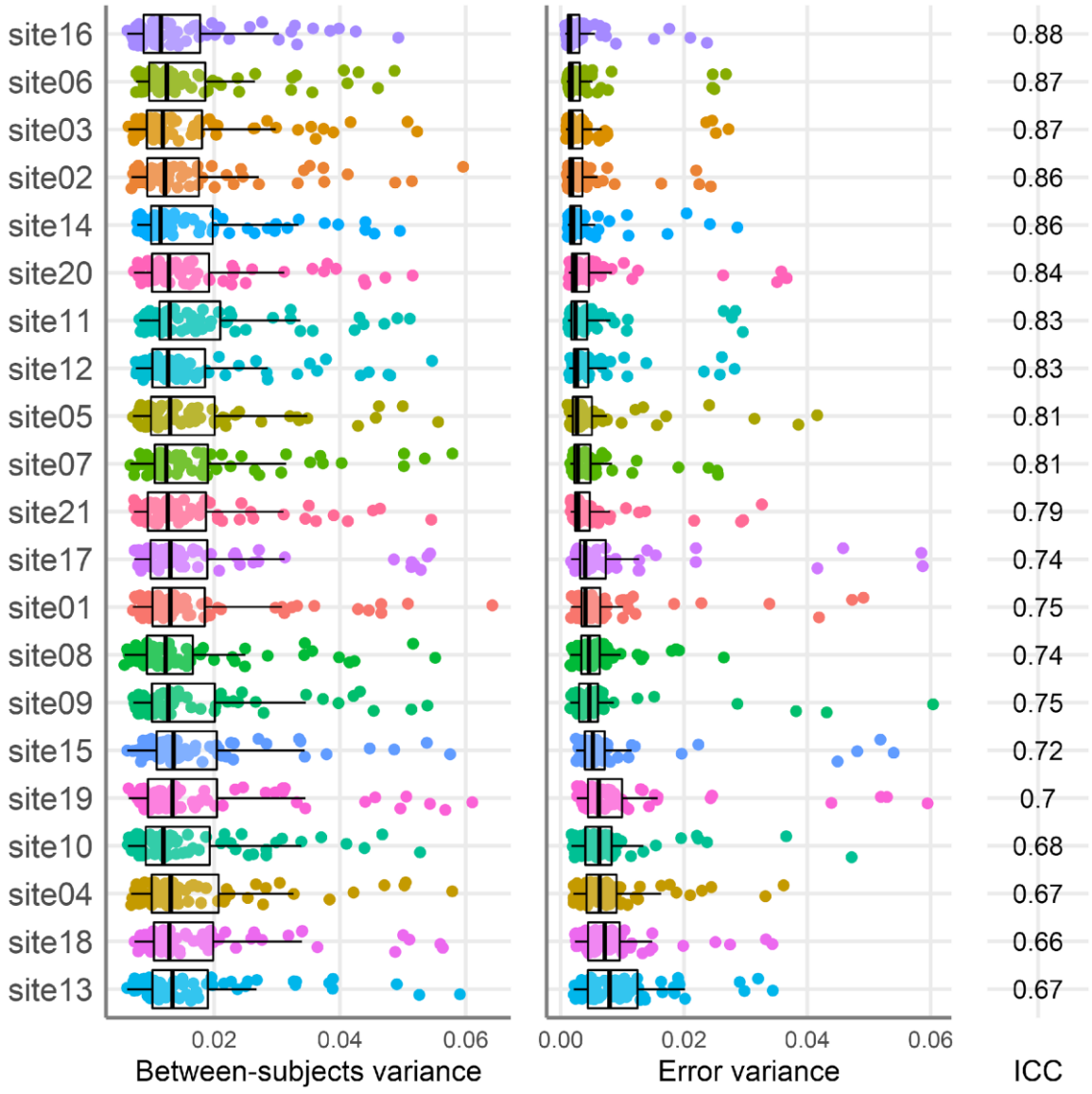

*Figure 5.* Between-subjects (left panel), error variance (middle panel) estimates, separately per testing site. Sites are ordered by the median error variance. Each point represents a different ROI, and the site colour maps to Figure 4. Cortical thickness only. The boxplots present the median and the 25$^{th}$ and 75$^{th}$ percentiles, the whiskers extend at a maximum to 1.5 times the interquartile range from the box.

**Rank order stability of ICC estimates.** To assist interpretation, we also calculated rank order stability of ICC, between-subjects variance, and error variance estimates. We did this separately for region differences and site differences, allowing us to capture the extent to which the same region, or the same site, is (un)reliable. Table 1 reports ICC (2,1) and ICC (3,1) estimates (Koo & Li, 2016). ICC(3,1) indexes *consistency agreement* and can be conceptualised as degree to which scores can be equated to each other, with some systemic error. ICC(2,1) is a more conservative index of *absolute agreement* across measures that additionally penalises for any systemic error. To illustrate, consider two repeated measures for which participants score the exact same number (Time1 = Time2). Here we have perfect longitudinal stability, both ICC(2,1) and ICC(3,1) equal 1. Now, consider instead that due to some practice effects all participants score 2 points higher in the second measure (Time1 = Time2 - 2). Here, ICC(3,1) = 1, indicating perfect longitudinal stability, while our ICC(2,1) will be lower as a result. These estimates give an indication of whether the stability estimates for brain regions are consistent across testing sites, and whether the same testing sites are consistently more or less reliable across brain regions.

*Table 1.*

ICC2,1 and ICC3,1 estimates separately comparing the rank-order stability of site and brain region estimates of ICC, between-subjects variance, and error variances.

| | **By region: How stable are regional estimates across sites?** | | **By site: How stable are site estimates across regions?** | |
|---|---|---|---|---|
| | ICC2 | ICC3 | ICC2 | ICC3 |
| ICC | 0.45 | 0.64 | 0.30 | 0.54 |
| Between-subjects variance | 0.94 | 0.95 | < 0.01 | 0.07 |
| Error variance | 0.76 | 0.82 | 0.07 | 0.30 |

The rank-order stability of brain region estimates (ICC, between-subjects variance, and error variance) suggests that across testing sites the same brain regions tend to have higher, or lower, longitudinal stability. Supporting our previous analyses, it is particularly clear that different brain regions typically have differing levels of between subjects variance. In contrast, the rank order stability of site estimates is considerably lower (particularly for the variance estimates), suggesting that we cannot discern that particular testing sites show higher or lower longitudinal stability across brain regions.

**Multigroup models for site differences.** To more formally assess potential cross-site variation in stability across measures and ROIs we performed a series of four multigroup ICED models, such that each site is represented by a different group. Specifically, the four models were (1) a *constrained* model, in which all groups were constrained to have equal between-

subject and error variances; (2) a *between-subject varying* model in which the between-subjects variance parameter was free to vary across groups (while we set an equality constraint on the error variance parameter across groups); (3) an *error varying* model in which the error variance parameter was free to vary between groups (while we set an equality constraint on the between-subject variance parameter across groups); and (4) an *unconstrained* model in which both variance components were allowed to vary between groups. Including comparisons with the between-subject and error varying models allows us to make some inferences about the sources of differences in stability across sites – i.e. whether stability differences across sites are due to different levels of between-subjects differences or measurement error. To compare model fit, we extracted the Comparative Fit Index (CFI; Bentler, 1990) for each model and computed the difference in CFI (ΔCFI) for 5 model comparisons: (A) constrained - between-subjects varying, (B) between-subjects varying - unconstrained, (C) constrained – error varying, (D) error varying - unconstrained, (E) constrained - unconstrained. Figure 6 presents the models and model comparisons visually. Greater ΔCFI values indicate larger improvements in model fit for the less-constrained model. ΔCFI values greater than 0.01 (Cheung & Rensvold, 2002) and a more conservative 0.02 (Meade et al., 2008) have been proposed as thresholds to determine differences in fit.[3]

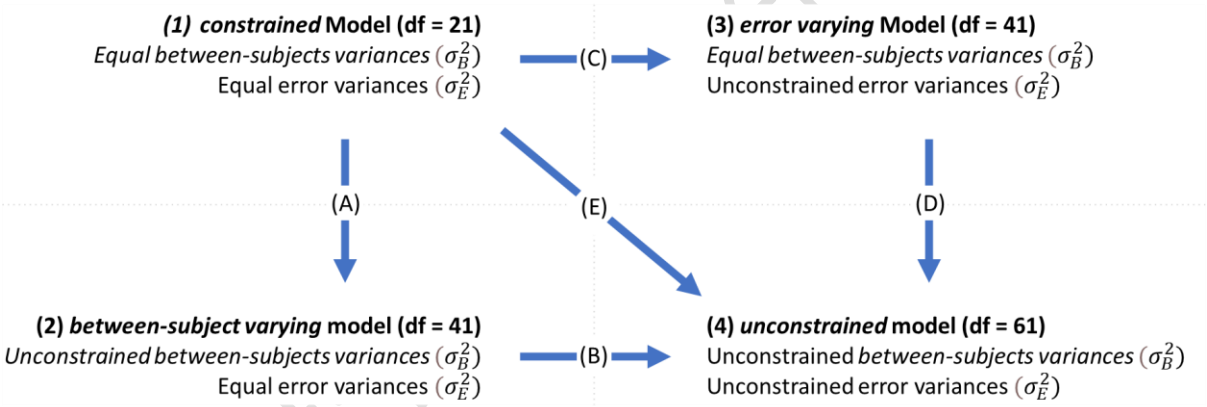


*Figure 6.* Representation of multigroup ICED models (numbers 1-4), and model comparisons (arrows A-E). The model descriptions refer to whether the between-subjects variance ($\sigma_B^2$) and error variance ($\sigma_E^2$) parameters were allowed to vary across sites (unconstrained) or were set to be equal across sites (equal). The arrows represent the model comparisons (ΔCFI) in the direction towards the less constrained model.

Figure 7 presents ΔCFI values for each model comparison across each brain region. Higher values indicate that the more complex model (with more free parameters) better fit the data even when penalizing for the additional complexity. Allowing the error variance to vary across sites (comparisons B, C, and E) meaningfully improved model fit in almost all cases (ΔCFI greater than 0.02 in over 97% brain regions). This suggests that testing sites are characterised by differing levels of measurement error. In contrast, allowing between-subjects variance to vary across sites (comparisons A and D) typically led to negligible or

[3] We also present the AIC and BIC model comparisons in the supplemental materials.

negative (1.5% of brain regions in comparison A and 13.2% of brain regions in comparison D) improvements in model fit, thus favouring the more parsimonious model, suggesting between subject variance did not differ systematically between sites. Allowing between-subjects variance to vary across sites improved the fit (ΔCFI greater than 0.02) in 19% of brain regions compared to the fully constrained model (comparison A) and in 0 regions compared to the error varying model (comparison D). This suggests that the between-subjects variance components are highly similar across testing sites and allowing between-subjects variances to vary across sites does not improve model fit over allowing error variances to vary across sites.

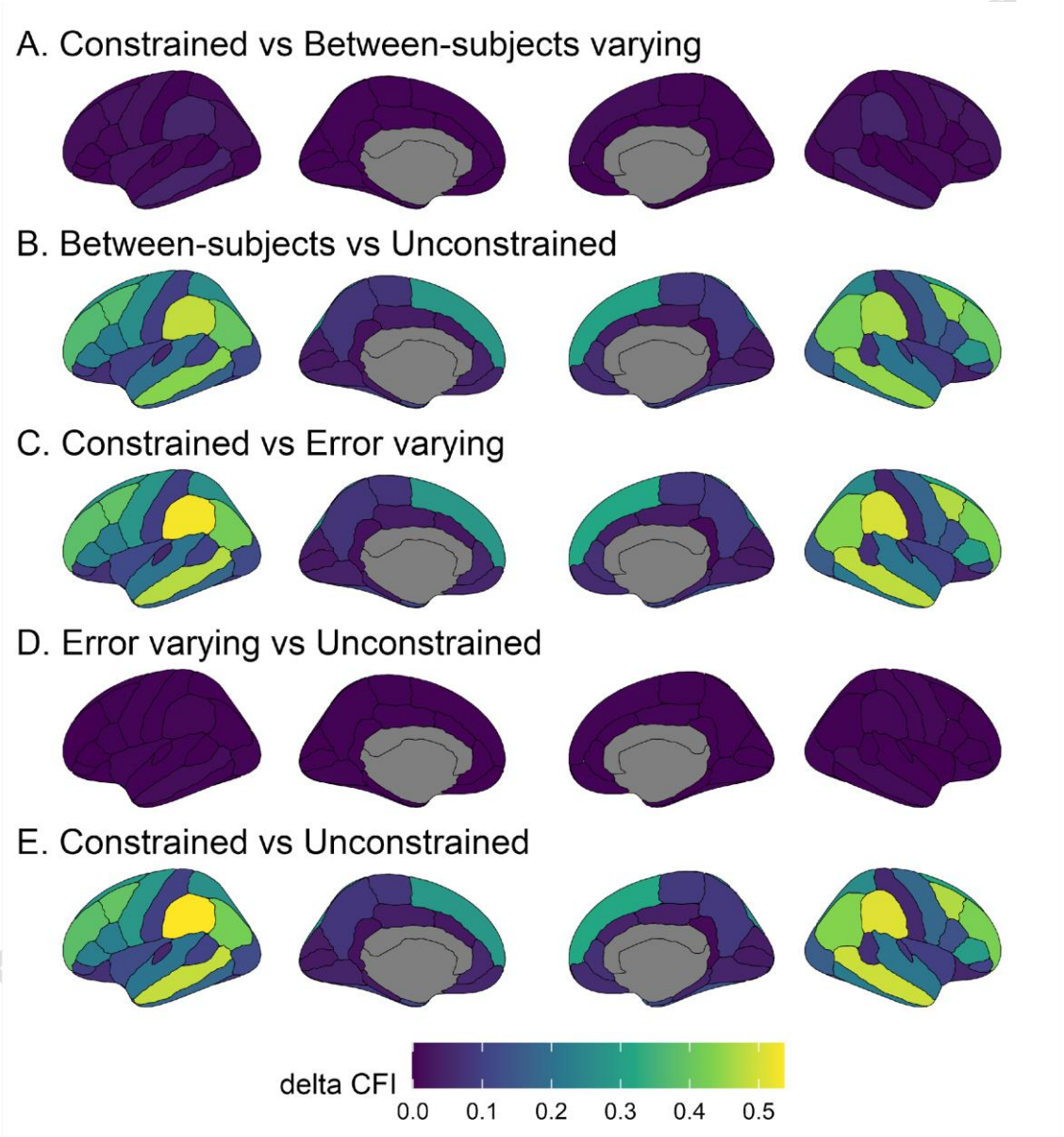


*Figure 7.* ΔCFI for each model comparison (panels A-E) across regions. Higher values (lighter and more yellow coloured) indicate improved model fit with more free parameters. In comparisons A and D, between-subject variance is allowed to vary compared to the preceding model. In comparisons B and C, error variance is allowed to vary compared to the preceding model. In panel E both between-subject and error variances are allowed to vary compared to the fully constrained model.

**Follow-up multigroup analyses by scanner manufacturer.** We expanded these analyses to explore the influence of MRI scanner on between-subjects variance and error variance. We ran the series of multigroup models and model comparisons described above, treating MRI scanner manufacturer (Siemens, 13 sites; Philips Medical Systems, 3 sites; and GE Medical Systems, 5 sites) as the grouping variable. From these analyses, we generated Figures 4, 5, and 7 for each metric (cortical thickness, surface area, and volume) and provide these in the supplement. For Cortical thickness; scanners from Siemens, Philips Medical Systems, and GE Medical Systems had average ICCs across brain regions of .83, .72, and .69, respectively. The multigroup model comparisons also showed a near identical pattern of results (Figure7_CT_scanners in the supplement) as those presented above treating testing site as the grouping variable (Figure 7).

We then ran three series of multigroup models by site (as described in the previous section), separately for the sites with each scanner manufacturer (supplemental figures: Figure7_CT_Siemens, Figure7_CT_Philips, and Figure7_CT_GE). For each scanner manufacturer, allowing between-subjects variance to vary between sites did not generally improve model fit – matching the general pattern of results. However, the patterns of results for allowing error variance to vary between sites differed markedly across brain regions within each scanner manufacturer. Together, these patterns of results suggest that there are both scanner and site level influences on the amount of error variance in our measures of grey matter, and that these influences differ across brain regions.

### 3. Practical implications

Above, we quantified the longitudinal stability of three grey matter measures. We can use these estimates to answer pragmatic questions about study design choices, including; how many repeated brain scans do we need to achieve high longitudinal stability? And, what influence are differences in longitudinal stability across brain regions likely to have on the attenuation of our results?

**How many repeated measures do we need to achieve high longitudinal stability?** We answered this question assuming that the stability estimates are proxies for reliability estimates. To put these estimates into context we performed a brief decision-study (Shavelson & Webb, 1991; Vispoel et al., 2018; Webb et al., 2006), using the Cortical Thickness estimates. We estimated the number of repeated measures needed to achieve an ICC2 longitudinal stability of greater than 0.9 - “excellent” longitudinal stability, following Koo and Li’s standards (2016). We can reformulate the ICC2 formula for this purpose.

$$ICC2 = \frac{\sigma_B^2}{\sigma_B^2 + \frac{\sigma_E^2}{N}} \quad (3)$$

$$N > \frac{ICC2 \cdot \sigma_E^2}{(1 - ICC2) \cdot \sigma_B^2} \quad (4)$$

Then, for ICC2 > .9

$$N > \frac{9 \cdot \sigma_E^2}{\sigma_B^2} \quad (5)$$

As visualised in Figure 8, our estimates suggest that most (48 of 68, or 70.5%) regions would require 3 or more timepoints to achieve an ICC2 longitudinal stability greater than .9. Further, 45.6% regions would require 4 or more timepoints. Performing poorest were the left and right temporal pole regions – both would require eight repeated scans to achieve high longitudinal stability. Given there are relatively few longitudinal brain imaging studies (Kievit & Simpson-Kent, 2020), and most of these contain only two timepoints, these results suggest that we will be unlikely to achieve sufficient longitudinal stability in some brain regions. Substantively, our findings therefore suggest that the absence of findings in these regions in similarly designed studies may therefore reflect low power (caused by suboptimal longitudinal stability) rather than a true absence of effects or differences between individuals or groups.

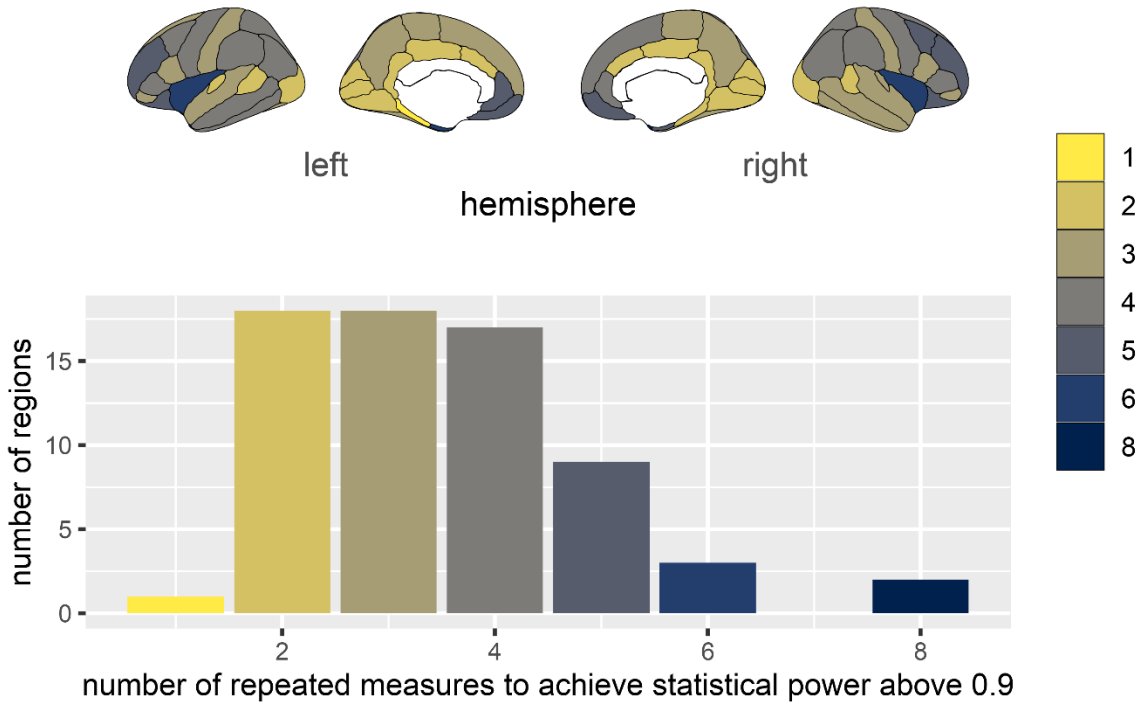


*Figure 8.* Number of timepoints required to achieve an ICC2 longitudinal stability estimate of .9 or greater for Cortical Thickness ROIs (assuming no individual differences in change in cortical thickness over two years).

**How attenuated are our estimates likely to be?** A related practical implication is that our standardised effect sizes will be more attenuated for regions with lower longitudinal stability. To demonstrate this, we extracted estimates from regions with the highest (parahippocampal gyrus, left hemisphere ICC = .9) and lowest (temporal pole, right hemisphere ICC = .54) Cortical Thickness longitudinal stability estimates. For example, assuming a 'true' correlation between a hypothetical measure and each ROI is .3, and the hypothetical measure has a longitudinal stability of .9. Using Spearman's attenuation correction formula (Spearman, 1904), we expect

the parahippocampal gyrus correlation to be attenuated to .27 and the temporal pole to be attenuated to .21. We can use these attenuated effect size estimates to compare expected statistical power for a straightforward correlation analysis. Given the attenuation we would require almost 70% more participants (105 vs 175) to detect the more severely attenuated correlation with 80% statistical power with a 5% alpha.

## Discussion

In this study, we used a series of ICED models (Brandmaier, Wenger, et al., 2018) to generate brain maps of (two-year) Longitudinal stability to provide a nuanced overview of the stability of grey matter across imaging measures and brain regions in the ABCD study imaging data (Casey et al., 2018). Our first analyses demonstrated heterogeneity in longitudinal stability estimates of longitudinal stability across brain regions. Further, of the grey matter structural measures (thickness, surface area, and volume) "one of these is not like the other"[4]. Specifically, cortical thickness showed a lower average longitudinal stability, and a wider range of longitudinal stability estimates, across brain regions. In contrast, surface area and grey matter volume showed near identical patterns of high longitudinal stability.

We extended our analyses to examine the relative contributions of between-subjects variance and error variance on differences in patterns of stability across brain region and ABCD's 21 testing sites. Stability estimates were heterogeneous across regions, and this appeared to be driven by differences in between-subject variance, suggesting these differences are due more to actual between-subjects differences. This observation is encouraging insofar as we can be more certain that observing individual differences across brain regions are likely a result of those individual differences, instead of differences in the amount of error captured in each region (perhaps with exceptions of the temporal pole, frontal pole, and entorhinal cortex).

In contrast, we found that differences in longitudinal stability across testing sites was largely driven by differences in error variance. It is not yet clear why some sites contribute more error than others. The ABCD consortium has gone to great lengths to ensure consistency in scanning parameters, data processing, quality control, and data harmonisation (Casey et al., 2018; Hagler et al., 2019). In our follow-up analyses we found the average contribution of error variance differed between scanner manufacturers, and an overall similar pattern of results in the multigroup analyses. We also saw differing patterns of site-related differences in error variances when analysing sites with each scanner type separately. Given the study design, with sites using a single scanner type, we are unable to fully disentangle the contributions of site-related and scanner-related influences on error. In addition to site-related differences (e.g. MRI scanner and image acquisition), site-related sampling differences, including demographics like age, related to each site may also be impactful. It is also plausible that sites were differentially affected by recruitment, retesting, and COVID-19 related delays. We expect that time between scans moderates the longitudinal stability and stability of the measures (as discussed in the introduction). At the site level, different patterns

[4] Sesame Street, Episode 1056 (1977)

of time lags between scans may capture differing levels of individual differences in change over time – which in these models would lead to higher estimated error.

**Practical Implications**

Our results have several implications. First, we should expect associations between cortical thickness and a phenotypic variable to be more attenuated on average than associations between surface area or volume and the same phenotypic variable. We demonstrated that for cortical thickness, three or more repeated measures would be needed for most brain regions to achieve high longitudinal stability to ensure true associations are not overly attenuated. We also highlighted that differences in longitudinal stability between regions can lead to requiring as many as 70% more participants to achieve the same level of statistical power. Of course, the relationship is nuanced, depending on the particular region of interest, the 'true' association of interest, and other characteristics of the model and sample. For instance, in very underpowered studies we are just as likely to see attenuation as over-estimation of our effects, also known as Type M (magnitude) errors (Gelman & Carlin, 2014). This effectively increases the chances of false-positive effects observed in small sample studies, or studies with too few repeated measure studies, further exacerbated in the case of significance threshold driven publication bias (Loken & Gelman, 2017).

Second, our results highlight the challenge inherent to comparing relative contributions of brain regions and structures without assessing the measurement properties across measures and regions. Estimating and reporting longitudinal stability and stability as standard practice (c.f. Parsons et al., 2019) affords us the opportunity to correct our estimates (e.g. Cooper et al., 2017; Schmidt & Hunter, 1996), or use approaches that integrate longitudinal stability into the model (e.g. for cognitive measures see; (Haines et al., 2020; Rouder & Haaf, 2018)). Both would facilitate comparisons across elements where we know longitudinal stability and stability likely differ (region, measure, sample, etc).

We stress that the implications of these analyses stretch further than grey matter measures in the ABCD data. Although the precise longitudinal stability estimates of these metrics will likely vary in other samples as a function of the nature of the study design, participant demographics, scanner specification and other aspects, we believe several of our high level findings are likely to generalize. First and foremost, reliability and longitudinal stability are likely to vary across brain regions, measures, and samples. For example, MRI and fMRI show distinct patterns of longitudinal stability (Elliott et al., 2020), shorter term reliability has also been shown to differ across channels in functional near infrared spectroscopy (Blasi et al., 2014) and EEG components (McEvoy et al., 2000). Beyond brain measures and regions, it has been demonstrated that different fMRI data processing pipelines can lead to marked variation in results, even using the same data (Li et al., 2021). Similarly, in behavioural data even basic data cleaning decisions can lead to large variation in reliability and longitudinal stability (Parsons, 2022).

In sum, we may find different patterns of longitudinal stability and longitudinal stability across: imaging modalities (e.g. EEG, NIRS), analyses pipelines, brain regions and

parcellations, populations and studies, as well as over the lifespan. We argue that reliability (and longitudinal stability) vary across a number of factors, and the unrevealed variation in reliability poses a danger to our inferences. Much more work is needed to ensure we understand the psychometric properties of our tools, and the heterogeneity of these properties across modalities. In future studies, ICED models (Brandmaier, Wenger, et al., 2018) could be expanded to directly examine predictors of error, such as time between scans and demographic characteristics (Bauer, 2017). By systematically accounting for these between-site and between-subject features, we can further improve longitudinal stability and stability estimates, while investigating how researchers could minimise these sources of error in future study designs.

**Limitations and opportunities for future research**

The central limitation of this paper is the reliance on two-timepoint data. Currently ABCD (Casey et al., 2018) has collected and released access to two timepoints of imaging data (with an average of two years between scans). As such, we did not examine sources of variance that could be possible in more complex testing schemes with three or more timepoints (e.g. (Anand et al., 2022; Brandmaier, Wenger, et al., 2018; Wenger et al., 2021) Prior work has examined the within-session reliability of fMRI measures within ABCD (Kennedy et al., 2022). However to our knowledge ABCD did not collect similar within-session repeated structural measures – we therefore focused on longitudinal stability. Future investigations would benefit from including repeated measures within session to enable the teasing apart of reliability and longitudinal stability and allow us to investigate predictors of both. In this paper we chose instead to capitalise on the multi-site nature of ABCD to examine sources of variance across brain regions and testing sites.

As we highlight in the introduction, individual differences in rate of change in brain structure over time will reduce our stability estimates. With two timepoints, we cannot uniquely identify these individual differences in change. As such, while high stability suggests we can adequately rank order participants over this time period, it does not suggest that participants brain structure remained stable across that time period (e.g. if all participants cortical thickness increased by 1mm, the stability estimates would be identical here). On the other hand, low stability indicates that we are unable to adequately rank order participants in this time course. This could result from population level instability; it could suggest that the rate of change between participants differs substantially. However, these estimates alone do not give us information about the other sources of within-subject variance. Lifespan charts of brain development (Bethlehem et al., 2022) highlight periods of rapid change and stability in brain structure, as well as periods characterised by greater between-subject variance. Moving forward, developmental neuroscience needs models that capture the reliability of change, alongside a sufficient number of repeated measures (longitudinal and ideally within session). We suggest two ways this might be achieved with extensions of the ICED modelling approach.

First, to model change in two timepoint data many studies calculate change scores (or annualised change scores to account for differential timings between scans). Difference

scores can be modelled equivalently within the SEM framework as latent difference scores (for a tutorial, see Kievit et al., 2018). It is possible to extract reliability estimates for these change scores, with some adaptations to the ICED approach (the difference score model is a special case of a two-timepoint latent growth curve model). It is worth noting that the literature on the reliability of difference scores indicates we should expect generally lower reliability than individual measures (e.g. Lord, 1956; Thomas & Zumbo, 2012; Zimmerman & Williams, 1998). Unfortunately, in standard latent difference score models the error variance is not uniquely identifies; instead they only capture the variance of the intercept and the change, which are both confounded with error in a single-indicator model. In effect, the model specification assumes perfect reliability of the change score if the intercept and change are to be interpreted as "pure" constructs. Given an estimate of reliability, or multiple indicators at each timepoint (e.g. multiple scans per session), the reliability of the change score can be estimated. Further work in this direction would enable mapping the reliability of change, given only two timepoints, as we have done in this paper across measures and brain regions.

Second, with three or more timepoints (e.g. when further waves of ABCD data are released) the ICED models can be expanded into latent growth curve models (Brandmaier, von Oertzen, et al., 2018; Brandmaier, Wenger, et al., 2018). This powerful and flexible extension allows for the simultaneous modelling of the intercept and slope reliability, termed 'Effective Curve Reliability'. This approach would provide key insights for investigations of individual differences in trajectories of change in existing and future data. Psychometrically, this would allow us to expand the grey matter structure reliability maps presented in this paper into reliability maps of change trajectories allowing us to gauge how well we can detect individual differences, their antecedents, correlates, and consequences. Effective curve reliability is a valuable tool for planning future studies for desired levels of precision, expected reliability, and statistical power, given variance estimates from studies such as ours and the planned longitudinal sampling. These considerations become especially important in clinical applications, such as drug trials intending to decelerate atrophy in MS or dementia, given the time and expense required to conduct longitudinal neuroscience.

**Summary**

In this study, we mapped the (two-year) test-retest stability of grey matter measures across brain regions using the first two timepoints from the ABCD study (Casey et al., 2018). This study complements previous examinations of the reliability and longitudinal stability of fMRI measures (Kennedy et al., 2022; Taylor et al., 2020). It also adds to existing research on the test-retest reliability and longitudinal stability of structural MRI measures (Elliott et al., 2020; Han et al., 2006), focusing on a longer timescale. Previous studies have used relatively short inter-scan intervals, e.g. 2 weeks: we moved beyond prior investigations and examined longitudinal stability in a very large sample with 21 testing sites across a longer developmental period (2 years). We found patterns of stability to differ across structural measures, brain regions, and testing sites. Decomposing these estimates allowed us to highlight that differences in stability across brain regions appears to be largely due to genuine between-

subjects differences. In contrast, differences in stability across testing sites was driven by variations in error, hinting at important cross-site differences causing increases in measurement error. Heterogeneity in reliability or longitudinal stability is not a problem in itself, but it does highlight the importance of examining the reliability of our measurements, and further investigating the sources of this (un)reliability, or longitudinal (in)stability, variance. We offered suggestions for expanding the Intra-Class Effect Decomposition approach used here into future investigations. Further detailed mapping of the reliability and longitudinal stability of structural brain measures over the lifespan should facilitate improving the efficiency and accuracy of developmental cognitive neuroscience.



## References

Anand, C., Brandmaier, A. M., Lynn, J., Arshad, M., Stanley, J. A., & Raz, N. (2022). Test-retest and repositioning effects of white matter microstructure measurements in selected white matter tracts. *Neuroimage: Reports*, *2*(2), 100096. https://doi.org/10.1016/j.ynirp.2022.100096

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. https://doi.org/10.1037/met0000077

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

Bentler, P. M. (1990). Comparitive Fix Indexed in Structural Models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bethlehem, R. A. I., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., Astle, D. E., Auyeung, B., Ayub, M., Bae, J., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S. A.,

Benegal, V., … Alexander-Bloch, A. F. (2022). Brain charts for the human lifespan. *Nature*, *604*(7906), 525–533. https://doi.org/10.1038/s41586-022-04554-y

Blasi, A., Lloyd-Fox, S., Johnson, Mark. H., & Elwell, C. (2014). Test–retest reliability of functional near infrared spectroscopy in infants. *Neurophotonics*, *1*(2), 025005. https://doi.org/10.1117/1.NPh.1.2.025005

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). Jossey-Bass.

Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Hertzog, C., & Lindenberger, U. (2015). LIFESPAN: A tool for the computer-aided design of longitudinal studies. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00272

Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Lindenberger, U., & Hertzog, C. (2018). Precision, Reliability, and Effect Size of Slope Variance in Latent Growth Curve Models: Implications for Statistical Power Analysis. *Frontiers in Psychology*, *9*, 294. https://doi.org/10.3389/fpsyg.2018.00294

Brandmaier, A. M., Wenger, E., Bodammer, N. C., Kühn, S., Raz, N., & Lindenberger, U. (2018). Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED). *ELife*, *7*. https://doi.org/10.7554/eLife.35718

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C. A., Wager, T. D., Banich, M. T., Speer, N. K., Sutherland, M. T., Riedel, M. C., Dick, A. S., Bjork, J. M., Thomas, K. M., … Dale, A. M. (2018). The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, *32*, 43–54. https://doi.org/10.1016/j.dcn.2018.03.001

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.

Compton, W. M., Dowling, G. J., & Garavan, H. (2019). Ensuring the Best Use of Data: The Adolescent Brain Cognitive Development Study. *JAMA Pediatrics*, *173*(9), 809. https://doi.org/10.1001/jamapediatrics.2019.2081

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology*, *8*(SEP), 1–16. https://doi.org/10.3389/fpsyg.2017.01482

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Furby, L. (1970). How we should measure 'change': Or should we? *Psychological Bulletin*, *74*(1), 68–80. https://doi.org/10.1037/h0029382

Deary, I. J., Pattie, A., & Starr, J. M. (2013). The Stability of Intelligence From Age 11 to Age 90 Years: The Lothian Birth Cohort of 1921. *Psychological Science*, *24*(12), 2361–2368. https://doi.org/10.1177/0956797613486487

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, *31*(7), 792–806.

Feldstein Ewing, S. W., Bjork, J. M., & Luciana, M. (2018). Implications of the ABCD study for developmental neuroscience. *Developmental Cognitive Neuroscience*, *32*, 161–164. https://doi.org/10.1016/j.dcn.2018.05.003

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fleiss, J. L. (1986). *Design and Analysis of Clinical Experiments*. Wiley.

Gawronski, B., Deutsch, R., & Banse, R. (2011). Response interference tasks as indirect measures of automatic associations. In K. Klauer, C. Stahl, & A. Voss (Eds.), *Cognitive methods in social psychology* (Issue 1, pp. 78–123). Guilford.

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Hagler, D. J., Hatton, SeanN., Cornejo, M. D., Makowski, C., Fair, D. A., Dick, A. S., Sutherland, M. T., Casey, B. J., Barch, D. M., Harms, M. P., Watts, R., Bjork, J. M., Garavan, H. P., Hilmer, L., Pung, C. J., Sicat, C. S., Kuperman, J., Bartsch, H., Xue, F., … Dale, A. M. (2019). Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *NeuroImage*, *202*, 116091. https://doi.org/10.1016/j.neuroimage.2019.116091

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. (2020). *Theoretically Informed Generative Models Can Advance the Psychological and Brain Sciences: Lessons from the Reliability Paradox* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/xr7y3

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., & Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, *32*(1), 180–194. https://doi.org/10.1016/j.neuroimage.2006.02.051

Healthy Brain Study Consortium, Aarts, E., Akkerman, A., Altgassen, M., Bartels, R., Beckers, D., Bevelander, K., Bijleveld, E., Davidson, E. B., Boleij, A., Bralten, J., Cillessen, T., Claassen, J., Cools, R., Cornelissen, I., Dresler, M., Eijsvogels, T., Faber, M., Fernández, G., … Willemsen, A. (2021). Protocol of the Healthy Brain Study: An accessible resource for understanding the human brain and how it dynamically and

individually operates in its bio-social context. *PLOS ONE*, *16*(12), e0260952. https://doi.org/10.1371/journal.pone.0260952

Hertzog, C., & Nesselroade, J. R. (2003). Assessing Psychological Change in Adulthood: An Overview of Methodological Issues. *Psychology and Aging*, *18*(4), 639–657. https://doi.org/10.1037/0882-7974.18.4.639

Hussey, I., & Hughes, S. (2018). *Hidden invalidity among fifteen commonly used measures in social and personality psychology*. https://doi.org/10.31234/osf.io/7rbfp

Karch, J. D., Filevich, E., Wenger, E., Lisofsky, N., Becker, M., Butler, O., Mårtensson, J., Lindenberger, U., Brandmaier, A. M., & Kühn, S. (2019). Identifying predictors of within-person variance in MRI-based brain volume estimates. *NeuroImage*, *200*, 575–589. https://doi.org/10.1016/j.neuroimage.2019.05.030

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

Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A.-L., de Mooij, S. M. M., Moutoussis, M., Goodyer, I. M., Bullmore, E., Jones, P. B., Fonagy, P., Lindenberger, U., & Dolan, R. J. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience*, *33*, 99–117. https://doi.org/10.1016/j.dcn.2017.11.007

Kievit, R. A., Davis, S. W., Mitchell, D. J., Taylor, J. R., Duncan, J., & Henson, R. N. A. (2014). Distinct aspects of frontal lobe structure mediate age-related differences in fluid intelligence and multitasking. *Nature Communications*, *5*(1), Article 1. https://doi.org/10.1038/ncomms6658

Kievit, R. A., & Simpson-Kent, I. L. (2020). *It's About Time: Towards a Longitudinal Cognitive Neuroscience of Intelligence*. 19. https://doi.org/10.31234/osf.io/n2yg7

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Laboratory for Computational Neuroimaging. (n.d.). *FreeSurfer software suite*.

Li, X., Ai, L., Giavasis, S., Jin, H., Feczko, E., Xu, T., Clucas, J., Franco, A., Sólon Heinsfeld, A., Adebimpe, A., Vogelstein, J. T., Yan, C.-G., Esteban, O., Poldrack, R. A., Craddock, C., Fair, D., Satterthwaite, T., Kiar, G., & Milham, M. P. (2021). *Moving Beyond Processing and Analysis-Related Variation in Neuroscience* [Preprint]. Neuroscience. https://doi.org/10.1101/2021.12.01.470790

Lindberg, D. M., Stence, N. V., Grubenhoff, J. A., Lewis, T., Mirsky, D. M., Miller, A. L., O'Neill, B. R., Grice, K., Mourani, P. M., & Runyan, D. K. (2019). Feasibility and Accuracy of Fast MRI Versus CT for Traumatic Brain Injury in Young Children. *Pediatrics*, *144*(4), e20190419. https://doi.org/10.1542/peds.2019-0419

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Lord, F. M. (1956). The Measurement of Growth. *Educational and Psychological Measurement*, *16*, 421–437.

Magistro, D., Takeuchi, H., Nejad, K. K., Taki, Y., Sekiguchi, A., Nouchi, R., Kotozaki, Y., Nakagawa, S., Miyauchi, C. M., Iizuka, K., Yokoyama, R., Shinada, T., Yamamoto, Y., Hanawa, S., Araki, T., Hashizume, H., Sassa, Y., & Kawashima, R. (2015). The Relationship between Processing Speed and Regional White Matter Volume in

Healthy Young People. *PLOS ONE*, *10*(9), e0136386. https://doi.org/10.1371/journal.pone.0136386

McEvoy, L. K., Smith, M. E., & Gevins, A. (2000). Test–retest reliability of cognitive EEG. *Clinical Neurophysiology*, *111*(3), 457–463. https://doi.org/10.1016/S1388-2457(99)00258-8

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Mowinckel, A. M., & Vidal-Piñeiro, D. (2019). *Visualisation of Brain Statistics with R-packages ggseg and ggseg3d*.

Muetzel, R. L., Mous, S. E., van der Ende, J., Blanken, L. M. E., van der Lugt, A., Jaddoe, V. W. V., Verhulst, F. C., Tiemeier, H., & White, T. (2015). White matter integrity and cognitive performance in school-age children: A population-based neuroimaging study. *NeuroImage*, *119*, 119–128. https://doi.org/10.1016/j.neuroimage.2015.06.014

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 92–105). American Psychological Association. https://doi.org/10.1037/10099-006

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

Noble, S., Spann, M. N., Tokoglu, F., Shen, X., Constable, R. T., & Scheinost, D. (2017). Influences on the Test–Retest Reliability of Functional Connectivity MRI and its Relationship with Behavioral Utility. *Cerebral Cortex*, *27*(11), 5415–5429. https://doi.org/10.1093/cercor/bhx230

Oertzen, T. (2010). Power equivalence in structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *63*(2), 257–272. https://doi.org/10.1348/000711009X441021

Parsons, S. (2022). Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2020.2577

Parsons, S., Kievit, R., & Brandmaier, A. M. (2022). *ICED: IntraClass Effect Decomposition* (0.0.1). https://github.com/sdparsons/ICED

Parsons, S., Kruijt, A., & Fox, E. (2019). Psychological Science needs a standard practice of reporting the reliability of cognitive behavioural measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

Poulton, R., Moffitt, T. E., & Silva, P. A. (2015). The Dunedin Multidisciplinary Health and Development Study: Overview of the first 40 years, with an eye to the future. *Social Psychiatry and Psychiatric Epidemiology*, *50*(5), 679–693. https://doi.org/10.1007/s00127-015-1048-8

Rapuano, K. M., Conley, M. I., Juliano, A. C., Conan, G. M., Maza, M. T., Woodman, K., Martinez, S. A., Earl, E., Perrone, A., Feczko, E., Fair, D. A., Watts, R., Casey, B. J., & Rosenberg, M. D. (2022). An open-access accelerated adult equivalent of the ABCD

Study neuroimaging dataset (a-ABCD). *NeuroImage*, *255*, 119215. https://doi.org/10.1016/j.neuroimage.2022.119215

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*(1), 1–12. https://doi.org/10.1037/a0018326

Rosseel, Y. (2012). lavaan: An R package for Structural Equation Modelling. *Journal Of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rouder, J., & Haaf, J. M. (2018). *A Psychometrics of Individual Differences in Experimental Tasks*. https://doi.org/10.31234/osf.io/f3h2k

Saragosa-Harris, N. M., Chaku, N., MacSweeney, N., Guazzelli Williamson, V., Scheuplein, M., Feola, B., Cardenas-Iniguez, C., Demir-Lira, E., McNeilly, E. A., Huffman, L. G., Whitmore, L., Michalska, K. J., Damme, K. S., Rakesh, D., & Mills, K. L. (2022). A practical guide for researchers and reviewers using the ABCD Study and other large longitudinal datasets. *Developmental Cognitive Neuroscience*, *55*, 101115. https://doi.org/10.1016/j.dcn.2022.101115

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*(2), 199–223. https://doi.org/10.1037/1082-989X.1.2.199

Schnack, H. G., van Haren, N. E. M., Brouwer, R. M., Evans, A., Durston, S., Boomsma, D. I., Kahn, R. S., & Hulshoff Pol, H. E. (2015). Changes in Thickness and Surface Area of the Human Cortex and Their Relationship with Intelligence. *Cerebral Cortex*, *25*(6), 1608–1617. https://doi.org/10.1093/cercor/bht357

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* (pp. xiii, 137). Sage Publications, Inc.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72. https://doi.org/10.2307/1412159

Srivastava, S. (2018). *Sound Inference in Complicated Research: A Multi-Strategy Approach* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/bwr48

Taylor, B. K., Frenzel, M. R., Eastman, J. A., Wiesman, A. I., Wang, Y.-P., Calhoun, V. D., Stephen, J. M., & Wilson, T. W. (2020). Reliability of the NIH toolbox cognitive battery in children and adolescents: A 3-year longitudinal examination. *Psychological Medicine*, 1–10. https://doi.org/10.1017/S0033291720003487

Thomas, D. R., & Zumbo, B. D. (2012). Difference Scores From the Point of View of Reliability and Repeated-Measures ANOVA: In Defense of Difference Scores for Data Analysis. *Educational and Psychological Measurement*, *72*(1), 37–43. https://doi.org/10.1177/0013164411409929

Trefler, A., Sadeghi, N., Thomas, A. G., Pierpaoli, C., Baker, C. I., & Thomas, C. (2016). Impact of time-of-day on brain morphometric measures derived from T1-weighted magnetic resonance imaging. *NeuroImage*, *133*, 41–52. https://doi.org/10.1016/j.neuroimage.2016.02.034

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1), 1–26. https://doi.org/10.1037/met0000107

von Rhein, D., Mennes, M., van Ewijk, H., Groenman, A. P., Zwiers, M. P., Oosterlaan, J., Heslenfeld, D., Franke, B., Hoekstra, P. J., Faraone, S. V., Hartman, C., & Buitelaar, J.

(2015). The NeuroIMAGE study: A prospective phenotypic, cognitive, genetic and MRI study in children with attention-deficit/hyperactivity disorder. Design and descriptives. *European Child & Adolescent Psychiatry*, *24*(3), 265–281. https://doi.org/10.1007/s00787-014-0573-4

Walhovd, K. B., Fjell, A. M., Westerhausen, R., Nyberg, L., Ebmeier, K. P., Lindenberger, U., Bartrés-Faz, D., Baaré, W. F. C., Siebner, H. R., Henson, R., Drevon, C. A., Strømstad Knudsen, G. P., Ljøsne, I. B., Penninx, B. W. J. H., Ghisletta, P., Rogeberg, O., Tyler, L., Bertram, L., & Lifebrain Consortium. (2018). Healthy minds 0–100 years: Optimising the use of European brain imaging cohorts ("Lifebrain"). *European Psychiatry*, *50*, 47–56. https://doi.org/10.1016/j.eurpsy.2017.12.006

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 Reliability Coefficients and Generalizability Theory. In *Handbook of Statistics* (Vol. 26, pp. 81–124). Elsevier. https://doi.org/10.1016/S0169-7161(06)26004-8

Wenger, E., Polk, S. E., Kleemeyer, M. M., Weiskopf, N., Bodammer, N. C., Lindenberger, U., & Brandmaier, A. M. (2021). *Reliability of quantitative multiparameter maps is high for MT and PD but attenuated for R1 and R2* in healthy young adults* [Preprint]. Neuroscience. https://doi.org/10.1101/2021.11.10.467254

Winkler, A. M., Greve, D. N., Bjuland, K. J., Nichols, T. E., Sabuncu, M. R., Håberg, A. K., Skranes, J., & Rimol, L. M. (2018). Joint Analysis of Cortical Area and Thickness as a Replacement for the Analysis of the Volume of the Cerebral Cortex. *Cerebral Cortex*, *28*(2), 738–749. https://doi.org/10.1093/cercor/bhx308

Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of*

*Mathematical and Statistical Psychology*, *51*(2), 343–351.

https://doi.org/10.1111/j.2044-8317.1998.tb00685.x

Zuo, X.-N., Xu, T., & Milham, M. P. (2019). Harnessing reliability for neuroscience research.

*Nature Human Behaviour*. https://doi.org/10.1038/s41562-019-0655-x