A guide to the measurement and interpretation of fMRI test-retest reliability

Stephanie Noble[a*], Dustin Scheinost[a,b,c,d], R. Todd Constable[a,c,d,e]

[a] *Department of Radiology and Biomedical Imaging, Yale School of Medicine*
[b] *Department of Statistics and Data Science, Yale University*
[c] *Child Study Center, Yale School of Medicine*
[d] *Department of Biomedical Engineering, Yale School of Medicine*
[e] *Department of Neurosurgery, Yale School of Medicine*

** Corresponding author*

Abstract

The test-retest reliability of functional neuroimaging data has recently been a topic of much discussion. Despite early conflicting reports, converging reports now suggest that test-retest reliability is poor for standard univariate measures—namely, voxel- and region-level task-based activation and edge-level functional connectivity. To better understand the implications of these recent studies requires understanding the nuances of test-retest reliability as commonly measured by the intraclass correlation coefficient (ICC). Here we provide a guide to the measurement and interpretation of test-retest reliability in functional neuroimaging and review major findings in the literature. We highlight the importance of making choices that improve reliability so long as they do not diminish validity, pointing to the potential of multivariate approaches that improve both. Finally, we discuss the implications of recent reports of low test-retest reliability in the context of ongoing work in the field.

Keywords: fMRI; test-retest reliability; task-based activation; functional connectivity; intraclass correlation coefficient

Highlights

- Context is needed to appreciate recent reports of poor univariate fMRI reliability
- We provide a guide to measuring and interpreting fMRI test-retest reliability via ICC
- We discuss factors that influence reliability, highlighting better multivariate than univariate reliability
- Test-retest reliability is not the only goal and validity must also be considered
- Recent findings should motivate better practices but do not invalidate existing work

Introduction

The ability to attain similar results given repeated measurements—or test-retest reliability—is a desirable quality of a measure. Converging evidence has demonstrated that univariate brain measures derived from functional magnetic resonance imaging (fMRI) show poor test-retest reliability, whether these measures reflect voxel- or region-level task-based activation or edge-level functional connectivity [1,2]. Furthermore, as will be discussed, a number of studies have highlighted factors that can improve reliability. However, the implications of this body of work are nuanced and often misunderstood. It is crucial to understand how we measure and interpret test-retest reliability because this informs the interpretation of existing research and the choices we should make moving forward. Drawing from lessons learned in our recent review and meta-analysis of test-retest reliability of functional connectivity [1] alongside other important work, we will provide a guide to measurement of test-retest reliability in fMRI, highlight major findings, and provide context needed to facilitate their interpretation.

A primer on measuring fMRI test-retest reliability with the intraclass correlation coefficient

Test-retest reliability is a Measurement Theory concept that quantifies the stability of a measure under repeated measurements [3]. This is important since it not only informs how precisely we can characterize an object, but also how precisely we are able to measure associations with other variables of interest. As will be discussed at the end of this section, reliability is considered to be complementary to *validity*, broadly defined as the "relevance of a measuring instrument for a particular purpose" [3].

While there are many measures of test-retest reliability [4], it is most commonly measured in fMRI using the *Intraclass Correlation Coefficient (ICC)*. A theoretical discussion of the properties of ICC and statistical inference can be found in [5,6]; here we will only highlight a few key features. In human-oriented research, the ICC is typically defined as the proportion of total measured variance (e.g., variability between people, sessions, etc.) that can be attributed to variability between people: $ICC = \rho(X, X') = \frac{\sigma^2_{person}}{\sigma^2_{total}}$. Three forms of ICC are commonly used: ICC(1,k) reflects "*absolute agreement*" and does not explicitly include *facets* (i.e., sources of

error; [7]) in the model; ICC(2,k) reflects “absolute agreement” and includes a random facet (i.e., levels reflect a random sample of possible levels); and ICC(3,k) (a.k.a., Cronbach’s alpha) reflects “*consistency*” and includes a fixed facet (i.e., levels reflect all possible levels of interest). The reliability of average measures may be estimated by choosing k>1. The ICC is similar to the Pearson’s correlation between repeated measurements, except that the Pearson’s correlation is invariant to linear transformations between measurements (i.e., reflects fit to y=mx+b); in contrast, consistency ICCs are only invariant to translation (i.e., reflects fit to y=x+b) and agreement ICCs are affected by both translation and scaling (i.e., reflects fit to y=x) [6,8]. This follows the assumption that repeated measurements be of the same *class*, and thus reflect the same population variance (and, for “agreement” ICC, mean). Incidentally, ICC can also be seen as being sensitive to inter-subject discriminability since it increases with more distance between- and less distance within-subjects. A common historical rule of thumb categorizes ICC as poor <0.4, fair 0.4–0.59, good 0.6–0.74, excellent ≥0.75 [9].

Despite its straightforward appearance, the ICC can be estimated a number of ways. The following is a selection of choices one can make in estimating the ICC:

- How many facets (sources of error) to include, if any (for >2 facets, see Generalizability Theory [7]);
- Whether to model facets as random (ICC(2,k)) or fixed (ICC(3,k));
- Whether to estimate ICC for average measurements (a “Decision Study” permits exploring combinations of facets [7]);
- Which variance estimation procedure to use (note that the standard ANOVA has several limitations, including negative estimates commonly set to 0; for alternatives, see [7,10];
- How precisely to model the error structure (for structural equation modeling procedures, see [11,12]);
- Whether to include covariates of no interest [13];

and more. Since it can be difficult to navigate the many choices, ICC(2,1) may be an ideal starting point; most univariate fMRI ICC studies include repeated measurements over time, which can introduce systematic error across subjects, and most study findings will be relevant to single rather than average measures. The choice of random or fixed facets can be tricky, so it is useful to consider whether the investigator wishes to generalize beyond the measured levels of the facet in the given study (e.g., whether a multisite study should inform other studies using other sites or whether they only care about informing future experiments performed at

that same sites). If one is unsure about desired generalization, it may help to start by considering random facets because this is more conservative (i.e., over-estimates facet contribution) and in practice yields similar results as fixed facets for fMRI [10]. Special consideration is due for time-related facets (e.g., session). In a standard non-nested model, the time-related variance component reflects changes over time shared across individuals, which are likely to be minimal at non-development timescales [14]. However, this does not reflect individual-specific changes over time; that variability is typically included in the residual variance component, which is typically very large [15].

Typically, reliability of the image is reported as the average of univariate ICC coefficients across the brain, but a few multivariate ICCs may also be used, including “multivariate generalizability” [7] and the I2C2, which pools variance components across the image [16]. We have not observed an ICC based on the multivariate distance, but related measures of discriminability include “fingerprinting” [17] and the a metric called “discriminability” [18].

The interpretation of the ICC should be considered alongside, and secondary to, *validity*. It is often stated that reliability provides an upper bound for validity, which refers to the fact that the correlation between observed variables cannot exceed their within-variable correlations [19,20]. Specifically, the observed correlation $\rho(X,Y)$ (between observed variables *X* and *Y*) can be described as depending on the true correlation $\rho(X,Y)$ (between the true value of the variables X_T and Y_T) and the within-measure correlations ($\rho(X,X')$ and $\rho(Y,Y')$, i.e., reliability coefficients):

$$\rho(X,Y)=\rho(X_T,Y_T)\cdot\sqrt{\rho(X,X')\cdot\rho(Y,Y')}\ .$$

Thus improving reliability in isolation can increase the observed correlation, thus attenuating the sample size required to detect an effect [21,22] (although correcting for this seeming attenuation is not recommended in practice for many reasons [23]). There are a few ways to think of validity in this context. Defining validity as $\rho(X,Y)$ and treating Y as ground truth is analogous to the popularly used *criterion validity*. One can also think of the true correlation in the absence of test-retest error $\rho(X_T,Y_T)$as reflecting “true” criterion validity. It has also been argued that this is “validation” and that “validity” should be reserved for ontological claim that an existing attribute causally affects a measure [24]. For the latter two interpretations, validity can indeed be present despite low reliability: think of a noisy thermometer that gives very different results at each measurement for a person but the correct result averaged across 100

measurements. And clearly high reliability does not imply validity: think of a thermometer that always registers zero. However, both cases are associated with low utility (and low observed criterion validity) for a single test. In summary, both reliability and validity are essential for a high quality measure, although the relationship between the two can vary depending on the definitions used.

Factors influencing test-retest reliability of fMRI

Using data amassed across the past decade, recent meta-analyses have underscored the poor test-retest reliability of univariate fMRI—that is, at the voxel and region level for task-based activation [2] and at the edge level for resting-state functional connectivity [1]. For context, structural MRI measures exhibit relatively high reliability compared with functional MRI, reflecting the expected immutability of brain structure and the expected higher amount of state and/or noise fluctuations in rest [2,25]. However, the literature points towards a number of factors that can increase reliability. In the following, we highlight a selection of references from the recent literature and most recent available literature reviews of task-based activation [26] and functional connectivity [1]. We limit our discussion to these commonly used fMRI units of analysis that have recently drawn attention in the literature but that other fMRI measures are actively investigated (e.g., [27] [28]).

Both task-based activation and functional connectivity reliability have been found to increase with the following factors: shorter test-retest intervals [26,29,30], task type [26,31] (task > rest for functional connectivity; basic > complex tasks for activation), locations with larger and significant effects (although typically a small association [26,29,32,33], locations in cortex rather than subcortex [2,15,34], and non-clinical populations [35,36]; meanwhile, minimal effects on functional connectivity reliability were observed with the use of multiple harmonized sites and scanners [37,38] (see [39] for more about increasing reliability with harmonization; see [40] for details on how to harmonize scanners) although site effects become more prominent with multivariate measures [37,38].

While univariate reliability is low in fMRI, multivariate reliability is substantially greater. Near perfect discriminability has been observed on the basis of the multivariate pattern of connectivity by comparing correlations between connectomes (“fingerprinting”) [17] and distances between connectomes (“discriminability”) [18], even with poor univariate ICC in the

same data [15,41]. The I2C2, which pools univariate estimates of variance across the image, also shows substantially higher reliability of connectivity than ICC [16]. In this vein, second-order network measures may show greater reliability than first-order measures, with greater dissociation after global signal regression [42]. Task-based activations have also been shown to exhibit greater reliability when combined across multiple areas or within a multivariate model [43].

Some findings have been reported specifically within either the activation or functional connectivity literatures. Activation studies reveal more task-relevant effects: reliability has been shown to improve for block rather than event-related designs and target-nontarget rather than task-rest contrasts [44]. Note that the latter effects may be small [2]—especially if the nontarget condition results in similar activation as the target [45]. In general, reliability is expected to depend on a number of task-specific factors, including the task itself. Functional connectivity has been shown to improve with more within-subject data [30,46], eyes open, awake, and active recordings [47,48], no artifact correction (a highly variable and complicated result; cf. [1,49] and **Implications**), within-network location [15], averaging over longer rather than shorter intervals within a given dataset [15] [31], no task regression for task data [31], full rather than partial correlation-based connectivity with shrinkage [50], and younger rather than older adult populations [51]; meanwhile, minimal effects were observed with slice timing correction [49]. We expect that many of the factors listed here that improve reliability in functional connectivity studies also improve reliability in activation studies. For example, we expect reliability increases with scan duration for not only connectivity but also activation [52], although this was not found in a recent activation meta-analysis [2].

Finally, although we have not yet observed investigations of reliability across the lifespan for connectivity or activation, we have observed age-related differences in reliability of connectivity including lower ICCs in infants [53,54] and older adults [51] compared with younger adults as well as differences in spatial distributions of reliability between children and adults [55]. Thus, we hypothesize that reliability will follow an inverted U-shape (i.e., smaller at young age, peaking at young adulthood, and decreasing in older adults) and exhibit changes in spatial distribution with age.

Implications

Evidence amassed across the field points to the low test-retest reliability of univariate fMRI data, as well as to a number of factors that can affect its reliability. What does this mean for existing research and how should we move forward? The low reliability of these fundamental levels of analysis in fMRI can clearly impair our ability to detect effects (see **A primer on measuring fMRI test-retest reliability**). This is particularly problematic for fMRI, where typical effect sizes are likely small to moderate [56-58]. Thus, it is reasonable to consider ways we can improve reliability (see **Factors influencing test-retest reliability of fMRI**). At the same time, we urge the reader to temper an unduly pessimistic interpretation of these findings by considering the following.

Much existing work relies on group-level inference, which can be robust in the absence of high individual-level test-retest reliability [59]. While increasing the number of subjects from a population for a test does not change the expected value of the ICC or effect size, it will change the power to detect an effect [22,60] along with the precision of the ICC and effect size estimate [41,61] (see [58] for an illustration of low precision leading to inflation in fMRI). Thus, a study with a large sample size can have substantial power and precision to detect effects even if observed effects are small due to low test-retest reliability. An illustrative example is the robust group-level activation in bilateral amygdala for both emotion tasks in Fig 4 of [2] despite low reliability, likely due to a combination of relatively large group size and magnitude of the underlying effect. Still, one must bear in mind the limited generalizability of group effects to the individual [59] and the possible attenuation of observed effects due to low reliability (although even small true effects can be meaningful [62]).

In addition, the low reliability frequently reported pertains to univariate measures, yet multivariate analyses commonly used in modern neuroimaging are substantially more reliable (see **Factors influencing test-retest reliability of fMRI**), reflecting greater stability in the pattern of the image and/or pooling elements. As such, it may be helpful to think of reliability of univariate fMRI measures as establishing the lower bound on reliability of fMRI [43]. In addition, recent work has underscored the poor power of mass univariate analyses for activation [63] and connectivity [57,58], stemming from small univariate effect sizes and multiple testing correction requirements. Multivariate approaches tend to show larger effect sizes [58] and are therefore recommended for better reliability, power, and precision.

Finally, it is important to recognize that reliability is not the same as validity. Choices that improve reliability may not improve validity—and can even do the opposite, depending on the

definition of validity employed. Thus, it is worth proceeding with caution when making choices just to improve reliability. For example, recall that reliability generally increases with longer scans and decreases with artifact removal. Acquiring longer scans may be a reasonable decision that better captures an individual as they move through different states [15]. However, deciding to retain artifacts in an effort to increase reliability can be more problematic; the increased reliability could be due to unwanted but reliable artifacts like motion [64,65], and retaining artifacts is associated with poorer outcome measures [49,66]. It can be a simple choice to remove a nuisance variable that is known to block measurement of desired associations, but it can be tricky to understand and properly remove fMRI confounds when their removal decreases desired associations [67]; either way, the decision to retain artifacts should not be based solely on an expected improvement in reliability. A related problem is that low reliability due to high within-subject variability may be meaningful or it could reflect noise. This is often difficult to disentangle in complex living systems, where both sources of variability are expected. In summary, both reliability and validity are desirable, and we should not strive for one without the other.

Conclusion

With increasing interest in reproducible findings there has been increasingly critical interest in the reliability of fMRI, typically measured by the ICC. Recent surveys of the functional connectivity and task-based activation literatures point towards the low reliability of univariate measures, as well as a number of factors that influence reliability. However, study design and analysis decisions should not be made based on reliability alone; instead, one should seek to also understand the impact of any decision of validity. Notably, there is growing evidence that multivariate approaches improve both reliability and validity, and thus offer a promising avenue for future research. A better understanding of fMRI reliability and validity alongside adoption of best practices in the field [68] will enable us to be better positioned to achieve both.

Acknowledgments

This work was supported by the National Institute of Mental Health under award numbers K00MH122372 (S.N.), R01MH121095 (R.T.C., D.S.), and P50MH115716 (R.T.C.).

Conflict of Interest

None declared.

CReDiT Statement

Stephanie Noble: Conceptualization, Writing – Original Draft, Writing – Review & Editing. Dustin Scheinost: Writing – Review & Editing. Todd Constable: Writing – Review & Editing.

Annotated references

** 1. Noble S, Scheinost D, Constable RT: **A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis**. *Neuroimage* 2019, **203**:116157.

The authors present a review and meta-analysis of test-retest reliability (ICC) of functional connectivity. Highlights include a low meta-analytic estimate of edge-level ICC and an overview of factors that influence reliability.

** 2. Elliott ML, Knodt AR, Ireland D, Morris ML, Poulton R, Ramrakha S, Sison ML, Moffitt TE, Caspi A, Hariri AR: **What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis**. *Psychological Science* 2020:0956797620916786.

The authors present a meta-analysis and original studies of test-retest reliability (ICC) of task-based activation. Highlights include a low meta-analytic estimate of ICC and low estimate of ICC in a priori regions for original studies.

* 18. Bridgeford EW, Wang S, Yang Z, Wang Z, Xu T, Craddock C, Dey J, Kiar G, Gray-Roncal W, Coulantoni C: **Eliminating accidental deviations to minimize generalization error with applications in connectomics and genomics**. *bioRxiv* 2020:802629.
The authors introduce a “discriminability” metric that uses multivariate distance to quantify test-retest reliability. They demonstrate desirable properties of this approach and assess the influence of functional connectivity processing steps on discriminability.

** 43. Kragel P, Han X, Kraynak T, Gianaros PJ, Wager T: **fMRI can be highly reliable, but it depends on what you measure**. 2020.

The authors demonstrate that multivariate models of task-based activation show substantial reliability in contrast to recent reports of poor univariate reliability.

** 58. Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, Donohue MR, Foran W, Miller RL, Feczko E: **Towards Reproducible Brain-Wide Association Studies**. *bioRxiv* 2020.

The authors demonstrate that univariate effect sizes in functional connectivity are low, resulting in high error rates. They also highlight the larger effect size for multivariate measures.

References

1. Noble S, Scheinost D, Constable RT: **A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis**. *Neuroimage* 2019, **203**:116157.
2. Elliott ML, Knodt AR, Ireland D, Morris ML, Poulton R, Ramrakha S, Sison ML, Moffitt TE, Caspi A, Hariri AR: **What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis**. *Psychological Science* 2020:0956797620916786.
3. Lavrakas PJ: *Encyclopedia of survey research methods*: Sage Publications; 2008.
4. Müller R, Büttner P: **A critical discussion of intraclass correlation coefficients**. *Statistics in medicine* 1994, **13**:2465-2476.
5. Shrout PE, Fleiss JL: **Intraclass correlations: uses in assessing rater reliability**. *Psychol Bull* 1979, **86**:420-428.
6. McGraw KO, Wong SP: **Forming inferences about some intraclass correlation coefficients**. *Psychological methods* 1996, **1**:30.
7. Webb NM, Shavelson RJ, Haertel EH: **4 Reliability coefficients and generalizability theory**. *Handbook of statistics* 2006, **26**:81-124.
8. Koo TK, Li MY: **A guideline of selecting and reporting intraclass correlation coefficients for reliability research**. *Journal of chiropractic medicine* 2016, **15**:155-163.
9. Cicchetti DV, Sparrow SA: **Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior**. *Am J Ment Defic* 1981, **86**:127-137.
10. Chen G, Taylor PA, Haller SP, Kircanski K, Stoddard J, Pine DS, Leibenluft E, Brotman MA, Cox RW: **Intraclass correlation: Improved modeling approaches and applications for neuroimaging**. *Hum Brain Mapp* 2018, **39**:1187-1206.
11. Brandmaier AM, Wenger E, Bodammer NC, Kühn S, Raz N, Lindenberger U: **Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED)**. *Elife* 2018, **7**:e35718.
12. Cooper SR, Jackson JJ, Barch DM, Braver TS: **Neuroimaging of individual differences: A latent variable modeling perspective**. *Neuroscience & Biobehavioral Reviews* 2019, **98**:29-46.
13. Zuo XN, Xu T, Jiang L, Yang Z, Cao XY, He Y, Zang YF, Castellanos FX, Milham MP: **Toward reliable characterization of functional homogeneity in the human brain:**

preprocessing, scan duration, imaging resolution and computational space. *Neuroimage* 2013, **65**:374-386.
14. Gratton C, Laumann TO, Nielsen AN, Greene DJ, Gordon EM, Gilmore AW, Nelson SM, Coalson RS, Snyder AZ, Schlaggar BL: **Functional brain networks are dominated by stable group and individual factors, not cognitive or daily variation**. *Neuron* 2018, **98**:439-452. e435.
15. Noble S, Spann MN, Tokoglu F, Shen X, Constable RT, Scheinost D: **Influences on the test–retest reliability of functional connectivity MRI and its relationship with behavioral utility**. *Cerebral Cortex* 2017, **27**:5415-5429.
16. Shou H, Eloyan A, Lee S, Zipunnikov V, Crainiceanu AN, Nebel NB, Caffo B, Lindquist MA, Crainiceanu CM: **Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (I2C2)**. *Cogn Affect Behav Neurosci* 2013, **13**:714-724.
17. Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, Papademetris X, Constable RT: **Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity**. *Nature neuroscience* 2015.
18. Bridgeford EW, Wang S, Yang Z, Wang Z, Xu T, Craddock C, Dey J, Kiar G, Gray-Roncal W, Coulantoni C: **Eliminating accidental deviations to minimize generalization error with applications in connectomics and genomics**. *bioRxiv* 2020:802629.
19. Spearman C: **Reprinted: The proof and measurement of association between two things (2010)**. *International Journal of Epidemiology* 1904, **39**:1137-1150.
20. Muchinsky PM: **The correction for attenuation**. *Educational and psychological measurement* 1996, **56**:63-75.
21. Zimmerman DW, Zumbo BD: **Resolving the issue of how reliability is related to statistical power: adhering to mathematical definitions**. *Journal of Modern Applied Statistical Methods* 2015, **14**:5.
22. Zuo X-N, Xu T, Milham MP: **Harnessing reliability for neuroscience research**. *Nature human behaviour* 2019, **3**:768-771.
23. Winne PH, Belfry MJ: **Interpretive problems when correcting for attenuation**. *Journal of Educational Measurement* 1982:125-134.
24. Borsboom D, Mellenbergh GJ, Van Heerden J: **The concept of validity**. *Psychological review* 2004, **111**:1061.
25. Cannon TD, Sun F, McEwen SJ, Papademetris X, He G, van Erp TG, Jacobson A, Bearden CE, Walker E, Hu X: **Reliability of neuroanatomical measurements in a multisite longitudinal study of youth at risk for psychosis**. *Human brain mapping* 2014, **35**:2424-2434.
26. Bennett CM, Miller MB: **How reliable are the results from functional magnetic resonance imaging?** *Annals of the New York Academy of Sciences* 2010, **1191**:133-155.
27. Xu T, Opitz A, Craddock RC, Wright MJ, Zuo X-N, Milham MP: **Assessing variations in areal organization for the intrinsic brain: from fingerprints to reliability**. *Cerebral Cortex* 2016, **26**:4192-4211.
28. Cao H, McEwen SC, Forsyth JK, Gee DG, Bearden CE, Addington J, Goodyear B, Cadenhead KS, Mirzakhanian H, Cornblatt BA: **Toward leveraging human connectomic data in large consortia: generalizability of fmri-based brain graphs across sites, sessions, and paradigms**. *Cerebral Cortex* 2019, **29**:1263-1279.
29. Shehzad Z, Kelly AM, Reiss PT, Gee DG, Gotimer K, Uddin LQ, Lee SH, Margulies DS, Roy AK, Biswal BB, et al.: **The resting brain: unconstrained yet reliable**. *Cereb Cortex* 2009, **19**:2209-2229.

30. Birn RM, Molloy EK, Patriat R, Parker T, Meier TB, Kirk GR, Nair VA, Meyerand ME, Prabhakaran V: **The effect of scan length on the reliability of resting-state fMRI connectivity estimates**. *Neuroimage* 2013, **83**:550-558.
31. Cho JW, Korchmaros A, Vogelstein JT, Milham M, Xu T: **Impact of Concatenating fMRI Data on Reliability for Functional Connectomics**. *BioRxiv* 2020.
32. Stirnberg R, Huijbers W, Brenner D, Poser BA, Breteler M, Stöcker T: **Rapid whole-brain resting-state fMRI at 3 T: Efficiency-optimized three-dimensional EPI versus repetition time-matched simultaneous-multi-slice EPI**. *Neuroimage* 2017, **163**:81-92.
33. Aron AR, Gluck MA, Poldrack RA: **Long-term test–retest reliability of functional MRI in a classification learning task**. *Neuroimage* 2006, **29**:1000-1006.
34. Shah LM, Cramer JA, Ferguson MA, Birn RM, Anderson JS: **Reliability and reproducibility of individual differences in functional connectivity acquired during task and resting state**. *Brain Behav* 2016, **6**:e00456.
35. Blautzik J, Keeser D, Berman A, Paolini M, Kirsch V, Mueller S, Coates U, Reiser M, Teipel SJ, Meindl T: **Long-term test-retest reliability of resting-state networks in healthy elderly subjects and with amnestic mild cognitive impairment patients**. *J Alzheimers Dis* 2013, **34**:741-754.
36. Manoach DS, Halpern EF, Kramer TS, Chang Y, Goff DC, Rauch SL, Kennedy DN, Gollub RL: **Test-retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects**. *American Journal of Psychiatry* 2001, **158**:955-958.
37. Badhwar A, Collin-Verreault Y, Orban P, Urchs S, Chouinard I, Vogel J, Potvin O, Duchesne S, Bellec P: **Multivariate consistency of resting-state fMRI connectivity maps acquired on a single individual over 2.5 years, 13 sites and 3 vendors**. *NeuroImage* 2020, **205**:116210.
38. Noble S, Scheinost D, Finn ES, Shen X, Papademetris X, McEwen SC, Bearden CE, Addington J, Goodyear B, Cadenhead KS, et al.: **Multisite reliability of MR-based functional connectivity**. *Neuroimage* 2017, **146**:959-970.
39. Friedman L, Stern H, Brown GG, Mathalon DH, Turner J, Glover GH, Gollub RL, Lauriello J, Lim KO, Cannon T, et al.: **Test-retest and between-site reliability in a multicenter fMRI study**. *Hum Brain Mapp* 2008, **29**:958-972.
40. Pomponio R, Erus G, Habes M, Doshi J, Srinivasan D, Mamourian E, Bashyam V, Nasrallah IM, Satterthwaite TD, Fan Y: **Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan**. *NeuroImage* 2020, **208**:116450.
41. Pannunzi M, Hindriks R, Bettinardi RG, Wenger E, Lisofsky N, Martensson J, Butler O, Filevich E, Becker M, Lochstet M: **Resting-state fMRI correlations: from link-wise unreliability to whole brain stability**. *Neuroimage* 2017, **157**:250-262.
42. Braun U, Plichta MM, Esslinger C, Sauer C, Haddad L, Grimm O, Mier D, Mohnke S, Heinz A, Erk S: **Test–retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures**. *Neuroimage* 2012, **59**:1404-1412.
43. Kragel P, Han X, Kraynak T, Gianaros PJ, Wager T: **fMRI can be highly reliable, but it depends on what you measure**. 2020.
44. Bennett CM, Miller MB: **fMRI reliability: influences of task and experimental design**. *Cognitive, Affective, & Behavioral Neuroscience* 2013, **13**:690-702.
45. Infantolino ZP, Luking KR, Sauder CL, Curtin JJ, Hajcak G: **Robust is not necessarily reliable: From within-subjects fMRI contrasts to between-subjects comparisons**. *NeuroImage* 2018, **173**:146-152.

46. Mejia AF, Nebel MB, Shou H, Crainiceanu CM, Pekar JJ, Mostofsky S, Caffo B, Lindquist MA: **Improving reliability of subject-level resting-state fMRI parcellation with shrinkage estimators**. *Neuroimage* 2015, **112**:14-29.
47. Zou Q, Long X, Zuo X, Yan C, Zhu C, Yang Y, Liu D, He Y, Zang Y: **Functional connectivity between the thalamus and visual cortex under eyes closed and eyes open conditions: A resting-state fMRI study**. *Human brain mapping* 2009, **30**:3066-3078.
48. Wang J, Han J, Nguyen VT, Guo L, Guo CC: **Improving the Test-Retest Reliability of Resting State fMRI by Removing the Impact of Sleep**. *Front Neurosci* 2017, **11**:249.
49. Shirer WR, Jiang H, Price CM, Ng B, Greicius MD: **Optimization of rs-fMRI Pre-processing for Enhanced Signal-Noise Separation, Test-Retest Reliability, and Group Discrimination**. *Neuroimage* 2015, **117**:67-79.
50. Mejia AF, Nebel MB, Barber AD, Choe AS, Pekar JJ, Caffo BS, Lindquist MA: **Improved estimation of subject-level functional connectivity using full and partial correlation with empirical Bayes shrinkage**. *NeuroImage* 2018, **172**:478-491.
51. Song J, Desphande AS, Meier TB, Tudorascu DL, Vergun S, Nair VA, Biswal BB, Meyerand ME, Birn RM, Bellec P, et al.: **Age-related differences in test-retest reliability in resting-state brain functional connectivity**. *PLoS One* 2012, **7**:e49847.
52. Gordon EM, Laumann TO, Gilmore AW, Newbold DJ, Greene DJ, Berg JJ, Ortega M, Hoyt-Drazen C, Gratton C, Sun H: **Precision functional mapping of individual human brains**. *Neuron* 2017, **95**:791-807. e797.
53. Wang Y, Hinds W, Duarte CS, Lee S, Monk C, Wall M, Canino G, Milani ACC, Jackowski A, Mamin MG: **Intra-session test-retest reliability of functional connectivity in infants**. *bioRxiv* 2020.
54. Dufford A, Noble S, Gao S, Scheinost D: **Low Infant Functional Connectome-based Identification Accuracy Across the First Year of Life**. In prep.
55. Mueller S, Wang D, Fox MD, Pan R, Lu J, Li K, Sun W, Buckner RL, Liu H: **Reliability correction for functional connectivity: Theory and implementation**. *Hum Brain Mapp* 2015, **36**:4664-4680.
56. Cremers HR, Wager TD, Yarkoni T: **The relation between statistical power and inference in fMRI**. *PloS one* 2017, **12**:e0184923.
57. Noble S, Scheinost D: **The constrained network-based statistic: a new level of inference for neuroimaging**. *Medical Image Computing and Computer Assisted Intervention* 2020.
58. Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, Donohue MR, Foran W, Miller RL, Feczko E: **Towards Reproducible Brain-Wide Association Studies**. *bioRxiv* 2020.
59. Fröhner JH, Teckentrup V, Smolka MN, Kroemer NB: **Addressing the reliability fallacy in fMRI: Similar group effects may arise from unreliable individual effects**. *Neuroimage* 2019, **195**:174-189.
60. Sullivan GM, Feinn R: **Using effect size—or why the P value is not enough**. *Journal of graduate medical education* 2012, **4**:279-282.
61. Shoukri MM, Asyali M, Donner A: **Sample size requirements for the design of reliability study: review and new results**. *Statistical Methods in Medical Research* 2004, **13**:251-271.
62. Funder DC, Ozer DJ: **Evaluating effect size in psychological research: Sense and nonsense**. *Advances in Methods and Practices in Psychological Science* 2019, **2**:156-168.

63. Noble S, Scheinost D, Constable RT: **Cluster failure or power failure? Evaluating sensitivity in cluster-level inference**. *Neuroimage* 2020, **209**:116468.
64. Yan C-G, Cheung B, Kelly C, Colcombe S, Craddock RC, Di Martino A, Li Q, Zuo X-N, Castellanos FX, Milham MP: **A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics**. *Neuroimage* 2013, **76**:183-201.
65. Couvy-Duchesne B, Blokland GA, Hickie IB, Thompson PM, Martin NG, de Zubicaray GI, McMahon KL, Wright MJ: **Heritability of head motion during resting state functional MRI in 462 healthy twins**. *Neuroimage* 2014, **102 Pt 2**:424-434.
66. Li J, Kong R, Liegeois R, Orban C, Tan Y, Sun N, Holmes AJ, Sabuncu MR, Ge T, Yeo BT: **Global signal regression strengthens association between resting-state functional connectivity and behavior**. *NeuroImage* 2019, **196**:126-141.
67. Pervaiz U, Vidaurre D, Woolrich MW, Smith SM: **Optimising network modelling methods for fMRI**. *NeuroImage* 2020, **211**:116604.
68. Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafo MR, Nichols TE, Poline JB, Vul E, Yarkoni T: **Scanning the horizon: towards transparent and reproducible neuroimaging research**. *Nat Rev Neurosci* 2017, **18**:115-126.