nature methods

Brief Communication

https://doi.org/10.1038/s41592-023-01901-3

ReX: an integrative tool for quantifying and optimizing measurement reliability for the study of individual differences

Received: 1 November 2021

Accepted: 28 April 2023

Published online: 01 June 2023

Check for updates

Ting Xu[1] ✉, Gregory Kiar[1], Jae Wook Cho[1], Eric W. Bridgeford[2], Aki Nikolaidis[1], Joshua T. Vogelstein[2] & Michael P. Milham[1,3]

Characterizing multifaceted individual differences in brain function using neuroimaging is central to biomarker discovery in neuroscience. We provide an integrative toolbox, Reliability eXplorer (ReX), to facilitate the examination of individual variation and reliability as well as the effective direction for optimization of measuring individual differences in biomarker discovery. We also illustrate gradient flows, a two-dimensional field map-based approach to identifying and representing the most effective direction for optimization when measuring individual differences, which is implemented in ReX.

Over the past decade, research into individual differences has become a central focus in the brain-imaging community. Researchers have shifted from looking at average effects within and between groups to relating individual variation in brain organization and function to genetic and phenotypic variables (for example, demographic, behavioral, cognitive, psychiatric)[1–8]. An inherent assumption of this shift is that the measures employed are reliable, that is, they will detect differences that are stable over time as well as across instruments, settings and analysts; these are necessary conditions for valid and reproducible brain-wise association research. Not surprisingly, as crises related to reproducibility have plagued the imaging field and the scientific community more broadly, researchers have revisited this assumption and begun the arduous task of quantifying and optimizing measurement reliability for individual difference research in the neuroscience community.

Here, we present ReX, an open-source tool designed to facilitate the quantification and optimization process by addressing a critical gap in studying individual differences: the failure to take into account the component variances of reliability (that is, within- and between-individual variance). The majority of reliability studies in the neuroimaging literature tend to treat reliability as a unitary construct rather than a ratio. This approach is problematic, as it overlooks the differential contributions of its component variances, which may be more readily mapped to a specific design or procedural optimizations being considered. Compounding the challenge at hand is that, when the experiment paradigms are cross-sectional, estimates of between-individual variance may be inflated, as the contributions of within-individual variances to its measurement are rarely considered. In this paper, we describe three features in ReX to help address these challenges.

First, ReX provides evaluation and a visualization module to identify the impact of variations and their contributions to reliability. Previous efforts in studying individual differences commonly focus on between-individual variation of observations and treat this as the true interindividual difference[9]. Within-individual variation is, by contrast, often overlooked or misinterpreted when studying interindividual differences in brain function, in particular in cross-sectional studies. For example, metabolomic or psychological changes over hours, days or weeks within an individual can alter the brain and mental states. Together with noise, these within-individual variations are embedded in the observed behavioral or brain connectome data. Treating the observed interindividual differences, which are contaminated with within-individual variation, as the true individual difference can compromise brain–behavior association discovery across individuals. Deciphering sources of variation both within and between individuals is central to interpreting individual differences in these scenarios. In ReX, we formally construct the variation space and provide a visualization module (Fig. 1a) using the 'true' between-individual variation σ_b^2 (y axis) against the within-individual variation σ_w^2 (x axis). Here, σ_b^2 is the 'true' between-individual variation rather than the observed between-individual variation. Using the variation space, it is easier to

[1]Department of Brain Development, Child Mind Institute, New York, NY, USA. [2]Johns Hopkins University, Baltimore, MD, USA. [3]Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, USA. ✉e-mail: ting.xu@childmind.org

https://doi.org/10.1038/s41592-023-01901-3

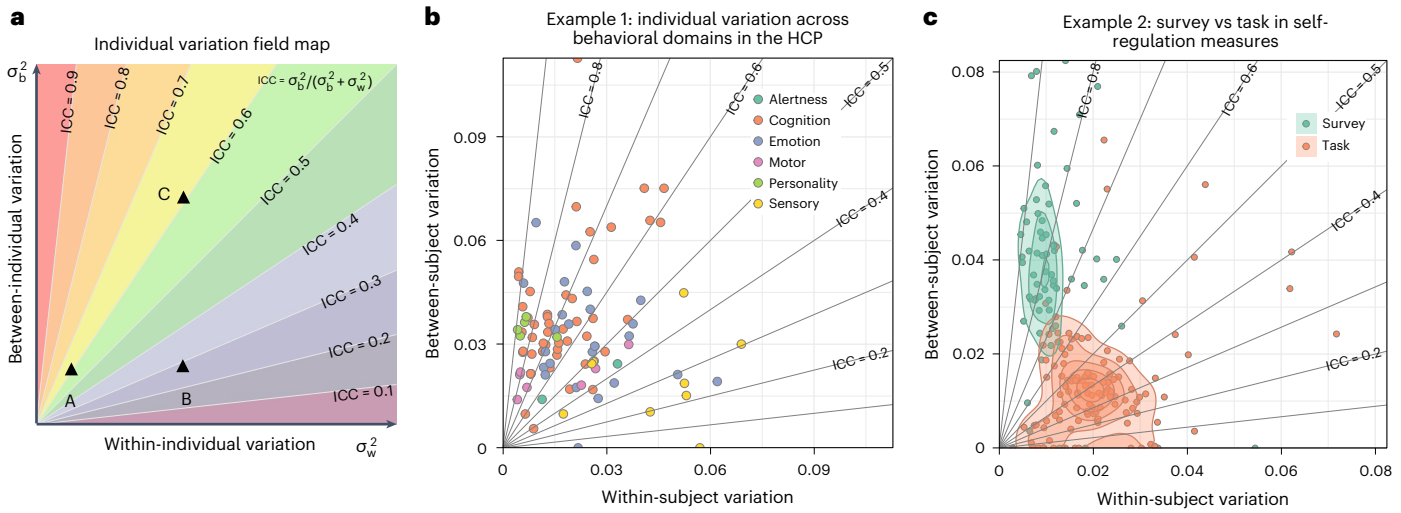


Fig. 1 | Theoretical individual variation field map in ReX and its applications. **a**, The two-dimension theoretical individual variation field map characterizes within- and between-individual variability and the likelihood of individual characterization (quantified by the ICC reliability). **b**, Within- and between-individual variations of National Institutes of Health Toolbox measures included in the HCP. **c**, Within- and between-individual variations of self-regulation measures using self-report surveys and behavioral tasks. Each dot represents self-regulation measures from one task or one survey.

differentiate within-individual from between-individual variation and examine which factors (for example, analytic methods, experiment designs, individual traits, state, etc.) influence the component variation separately or in combination. To illustrate the utility of the variation space of ReX in understanding the individual difference, we present the theoretical variation field map (Fig. 1a) and examples of a wide range of behavioral measurements. We demonstrate that the between-individual variation in the Human Connectome Project (HCP) behavioral battery is not consistent across task domains (Fig. 1b)[10]. Personality and cognition tasks show less within-individual variation, while the variation of emotion and sensory tasks attributes more to within-individual variation. In studying self-regulation measures, the variation space facilitates the comparison of selecting self-report surveys to behavioral tasks in characterizing individual differences (Fig. 1c)[11].

Second, we developed Gradient Flow Map (GFM) for reliability optimization. In addition to the variation calculation module, another feature of ReX is the GFM, which indicates the most efficient direction to improve the reliability in measuring individual differences. Reducing within-individual variation (that is, from point B to A) and increasing between-individual variation (that is, from point B to C) can improve reliability to the same extent (that is, change in intraclass correlation (ICC) = 0.3, Fig. 1a). However, the contribution of changes in within- and between-individual variation for improving reliability is not the same. The decrease in within-individual variation is more efficient (from point B to A) than the increase in between-individual variation (from point B to C). In general, if a measure has relatively small within-individual variation (x axis) and large between-individual variation (y axis) (for example, Fig. 2b, point A), the reduction in x improves ICC more than a similar increment in y. On the other hand, if a measure is relatively high in x but low in y (for example, point C in Fig. 2b), the most efficient direction to improve the reliability is to increase the between-individual variation. Such optimal direction for improving reliability can be calculated as the first derivative of the reliability, the ratio of the true between-individual variation to the total variation (that is, ICC). The improvement of x and y that is closest to this optimal direction is more likely to improve the reliability the most under the Gaussian assumption. When comparing the performance of two different measures (for example, pipelines in measuring brain functional connectivity, Fig. 2d and Supplementary Note), the change of the variation in x and y cannot fully determine whether it improves reliability (Fig. 2e, first and third quadrants). Thus, ReX normalizes the change of within- and between-individual variation as compared to the optimal direction and visualizes such normalized changes using a standard color map (Fig. 2c). The resultant GFM (Fig. 2f) provides a straightforward answer to whether the change of within- and between-individual variation from one pipeline to the other improves reliability as well as to whether the improvement is in the most efficient direction. Using the GFM can support multifaceted applications to facilitate comparing and optimizing possible analytic strategies and experiment designs. Of note, ReX determines how much the approaches that have been tested align with the most efficient direction to improve reliability. Interpreting new approaches requires collecting repeated-measure datasets or available estimated within- and between-individual variations.

Third, to accommodate the needs of a broad range of designs, ReX offers users a range of parametric and nonparametric methods for both univariate and multivariate reliability. ICC formations including one-way random, two-way random and two-way mixed models using the linear mixed model (LMM in the R package lme4) with the restricted maximum likelihood (ReML) estimation method[12]. Compared to the traditional ANOVA-based method, LMM allows missing data in the sample and ReML avoids negative ICC values. Specifically, ReX uses the one-way random model for single-measure ICC(1, 1) and average-measure ICC(1, k); the two-way random model for single-measure agreement ICC(2, 1) and average-measure agreement ICC(2, k); and the two-way mixed model for single-measure consistency ICC(3, 1) and average-measure consistency ICC(3, k)[13]. In the LMM, the random factors and residuals are assumed to be independent. Users can specify the confounding variables as covariates in the model (for example, age, sex). The parametric and nonparametric multivariate formulations of reliability implemented in ReX were recently developed in the imaging field including the distance-based ICC (dbICC)[14], the image ICC coefficient (I2C2)[15], discriminability[16] and identification rate (that is, fingerprinting)[17]. It is important to note that reliability is a prerequisite and the upper bound for validity. However, it does not imply validity. Depending on the trait of interest, the validity of the same measurement may vary. Optimizations for reliability need to be complemented by those focused on the validity of the specific trait (Supplementary Video 1).

https://doi.org/10.1038/s41592-023-01901-3

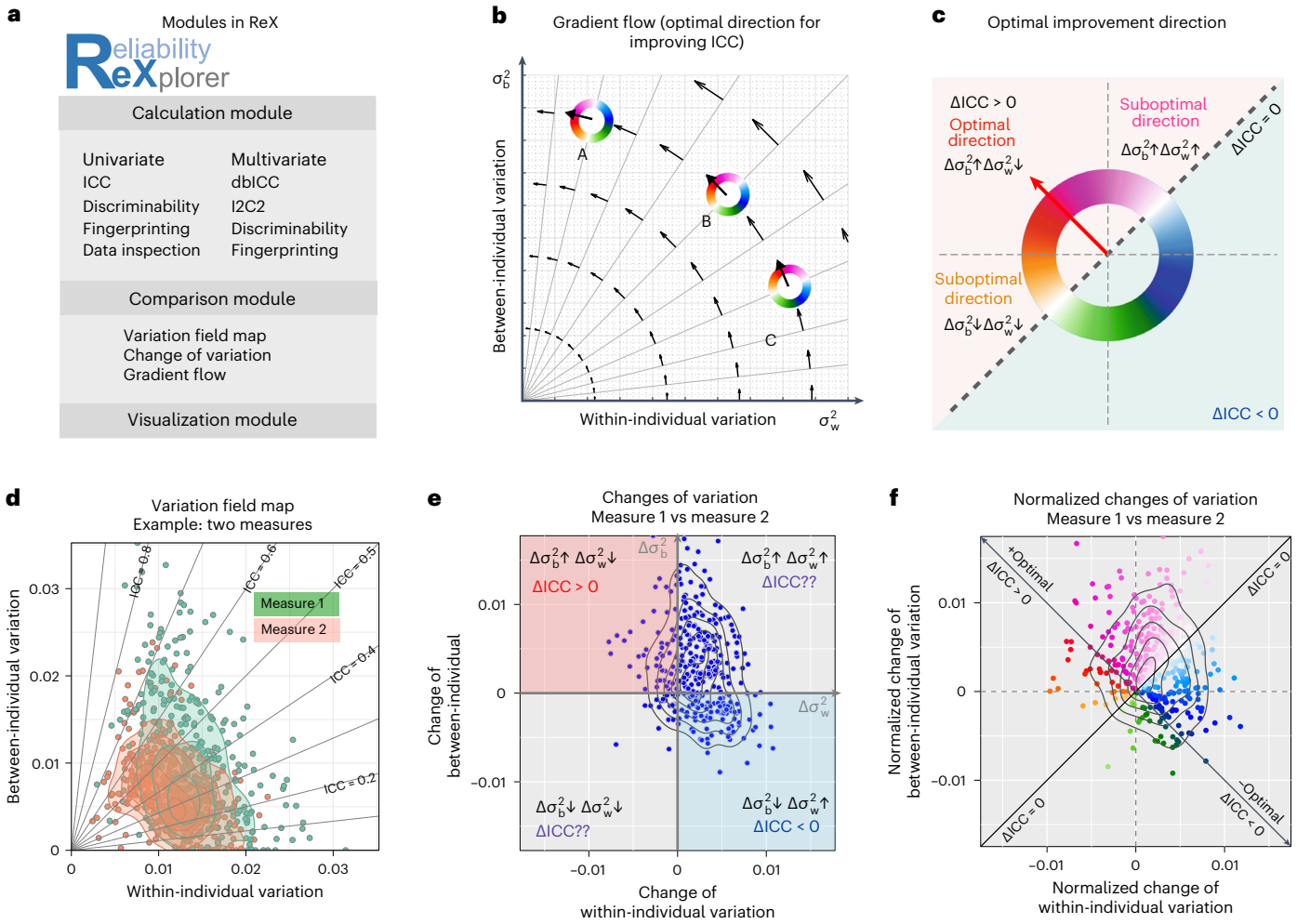


Fig. 2 | GFM in ReX and its application example. a, ReX modules. **b**, The theoretical GFM (that is, the first derivative of the ICC) captures the most efficient way to improve reliability. **c**, Normalized changes of variation as compared to the optimal direction for improving ICC. **d**, Example of within- and between-individual variability of two different measures. **e**, Change of within- and between-individual variation. **f**, Normalized change of variation to the optimal direction reveals whether one measure displays higher or lower reliability than the other.

To demonstrate the utility of ReX, we include six example applications (Supplementary Note). Application 1 shows the differential contributions of within- and between-individual variances to reliability across behavioral assessments. The remaining applications (2–6) use ReX to facilitate the optimal selection of experimental choices in behavioral measures, neuroimaging data-preprocessing pipelines, the amount of data required and data-aggregation strategies (Extended Data Fig. 1 and Supplementary Figs. 1–5). The resulting visualizations from ReX are included to make obvious how the tool allows users to intuitively interpret results easily. Of note, ReX can be applied to any repeated-measure dataset, although the power and effect size of the reliability will depend on the data quality and quantity. It is recommended to also calculate the power of the reliability and consider trade-offs of selecting the data-collecting and -preprocessing strategies[18–20]. In addition, the optimal direction in ReX is the theoretical direction that improves reliability, which might not be the most practical direction. In practice, the cost of the approach (for example, scan time, collecting rare patients, etc.) to select measures needs to be considered for assessing individual differences and reliability[18–20].

Recognizing the growing need for techniques to guide optimization efforts for the measurement of individual differences, we proposed the reliability GFM to quantify the optimization efforts of measuring reliability and individual variations. We develop ReX, which integrates reliability concepts, calculation, optimization and visualization to bridge the gap between establishing reliability and measuring individual variations. We hope that ReX will help calculate and compare reliabilities across experiments and analytic methods to facilitate studying individual differences in neuroscience and psychology.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-023-01901-3.

References

1. Seghier, M. L. & Price, C. J. Interpreting and utilising intersubject variability in brain function. *Trends Cogn. Sci.* **22**, 517–530 (2018).
2. Dubois, J. & Adolphs, R. Building a science of individual differences from fMRI. *Trends Cogn. Sci.* **20**, 425–443 (2016).
3. Barch, D. M. et al. Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* **80**, 169–189 (2013).

4. Finn, E. S. et al. Can brain state be manipulated to emphasize individual differences in functional connectivity? *NeuroImage* **160**, 140–151 (2017).
5. Lebreton, M., Bavard, S., Daunizeau, J. & Palminteri, S. Assessing inter-individual differences with task-related functional neuroimaging. *Nat. Hum. Behav.* **3**, 897–905 (2019).
6. Van Horn, J. D., Grafton, S. T. & Miller, M. B. Individual variability in brain activity: a nuisance or an opportunity? *Brain Imaging Behav.* **2**, 327–334 (2008).
7. Palminteri, S. & Chevallier, C. Can we infer inter-individual differences in risk-taking from behavioral tasks? *Front. Psychol.* **9**, 2307 (2018).
8. Genon, S., Eickhoff, S. B. & Kharabian, S. Linking interindividual variability in brain structure to behaviour. *Nat. Rev. Neurosci.* **23**, 307–318 (2022).
9. Hsu, S., Poldrack, R., Ram, N. & Wagner, A. D. Observed correlations from cross-sectional individual differences research reflect both between-person and within-person correlations. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/zq37h (2022).
10. Van Essen, D. C. et al. The WU-Minn Human Connectome Project: an overview. *NeuroImage* **80**, 62–79 (2013).
11. Enkavi, A. Z. et al. Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proc. Natl Acad. Sci. USA* **116**, 5472–5477 (2019).
12. Chen, G. et al. Intraclass correlation: improved modeling approaches and applications for neuroimaging. *Hum. Brain Mapp.* **39**, 1187–1206 (2018).
13. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**, 155–163 (2016).
14. Xu, M., Reiss, P. T. & Cribben, I. Generalized reliability based on distances. *Biometrics* **77**, 258–270 (2021).
15. Shou, H. et al. Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (I2C2). *Cogn. Affect. Behav. Neurosci.* **13**, 714–724 (2013).
16. Bridgeford, E. W. et al. Eliminating accidental deviations to minimize generalization error and maximize replicability: applications in connectomics and genomics. *PLoS Comput. Biol.* **17**, e1009279 (2021).
17. Finn, E. S. et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* **18**, 1664–1671 (2015).
18. Zuo, X.-N., Xu, T. & Milham, M. P. Harnessing reliability for neuroscience research. *Nat. Hum. Behav.* **3**, 768–771 (2019).
19. Cho, J. W., Korchmaros, A., Vogelstein, J. T., Milham, M. P. & Xu, T. Impact of concatenating fMRI data on reliability for functional connectomics. *NeuroImage* **226**, 117549 (2021).
20. Noble, S., Scheinost, D. & Constable, R. T. A guide to the measurement and interpretation of fMRI test–retest reliability. *Curr. Opin. Behav. Sci.* **40**, 27–32 (2021).

https://doi.org/10.1038/s41592-023-01901-3

Methods

ReX follows the classical test theory and provides visualization of the theoretical variation field map GFM as well as the reliability–validity relationship map to facilitate understanding the relationship between validity, reliability and its component individual variance. The details of each map are introduced as follows.

The variation field map, reliability and validity

In classical test theory, the observed score (X) from each person obtained using a measurement contains a true score (T) and an error score (E)[21]. Reliability is defined as the ratio of true score variance σ_T^2 to the observed score variance σ_X^2, which is the sum of the variance of true scores and the variance of error scores (that is, Reliability $= \frac{\sigma_T^2}{\sigma_T^2+\sigma_\varepsilon^2}$). In practice, a true score is always compounded with error. In the study of individual differences, within-individual variance (as the error term) is embedded in the observed interindividual variance. In ReX, we use two-dimensional space (that is, variation field map) to formulize the true between-individual variation (y axis) and within-individual variation (x axis). The visualization of this field map (Fig. 1a) along with the contour line of reliability allows the users to intuitively interpret the theoretical contribution of each variance component to reliability.

It is worth noting that reliability is a necessary prerequisite for validity but is not sufficient. The true score T of measurement here refers to the consistent score over tests of an individual. It contains a valid score for the trait of interest T_i and the unwanted score T_u that is not related to the trait of interest (that is, contaminants relative to the trait of interest).

$$\sigma_T^2 = \sigma_{T_\mathrm{i}}^2 + \sigma_{T_\mathrm{u}}^2$$

In test theory, validity is defined as the proportion of variation in the trait of interest to the total variation of the observed score[22].

$$\text{Validity} = \frac{\sigma_{T_\mathrm{i}}^2}{\sigma_{T_\mathrm{i}}^2 + \sigma_{T_\mathrm{u}}^2 + \sigma_E^2}$$

$$\text{Reliability} = \frac{\sigma_{T_\mathrm{i}}^2 + \sigma_{T_\mathrm{u}}^2}{\sigma_{T_\mathrm{i}}^2 + \sigma_{T_\mathrm{u}}^2 + \sigma_E^2}$$

Depending on the trait of interest, validity may vary for the same measurement. In other words, the validity of a measurement can be different in examining different traits, while the reliability always remains the same (for example, using cortical thickness to measure age and IQ). When the true score T equals the trait score T_r, validity equals reliability. If there is a signal but it is not related to the trait, validity is lower than reliability (Supplementary Video 1; GitHub, https://github.com/TingsterX/Reliability_Explorer/blob/main/reliability_and_validity/reliability_and_validity.md). In summary, reliability is the upper bound for validity. It does not imply validity, but it is a prerequisite for validity. Depending on the trait of interest, validity of the specific trait needs to be considered in optimizations for reliability[18,23].

Reliability models

ReX includes multiple ICC models for univariate reliability estimation implemented in one-way random (equation (1)), two-way random (equation (2)) and two-way mixed (equation (3)) models using the LMM[12,13].

$$y = \mu_0 + \lambda_i + \epsilon_{ij},\ \lambda_i \sim \mathcal{N}(0,\ \sigma_\lambda^2), \epsilon_{ij} \sim \mathcal{N}(0,\ \sigma_\epsilon^2) \quad (1)$$

$$y = \mu_0 + \lambda_i + \alpha_j + \epsilon_{ij},\lambda_i \sim \mathcal{N}(0,\ \sigma_\lambda^2), \alpha_j \sim \mathcal{N}(0,\ \sigma_\alpha^2),\ \epsilon_{ij} \sim \mathcal{N}(0,\ \sigma_\epsilon^2) \quad (2)$$

$$y = \mu_0 + \lambda_i + \alpha_j + \epsilon_{ij},\ \alpha_j \text{ is the fixed effect},\ \lambda_i \sim \mathcal{N}(0,\ \sigma_\lambda^2),\ \epsilon_{ij} \sim \mathcal{N}(0,\ \sigma_\epsilon^2) \quad (3)$$

In equations (1–3), the term $i = 1, 2, \ldots, n$ indexes individual repetitions, $j = 1, 2, \ldots, k$ indexes test–retest repetitions, μ is the intercept represents the group average, and $\mathcal{N}$ is a normal distribution. λ_i is the random effect in equations (1–3) and represents the differences at the i-th individual level so that its variance σ_λ^2 indicates between-individual variation. The error term ϵ_{ij} represents the differences across tests of each individual, and its variance indicates within-individual variance. The random effect and the error term are assumed to be independent (that is, orthogonal). The absolute agreement of single-rater ICC(1, 1) and the absolute agreement of multiple-rater ICC(1, k) are estimated using equation (1) as follows.

$$\mathrm{ICC}(1,1) = \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_\epsilon^2},\ \mathrm{ICC}(1,k) = \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_\epsilon^2 k^{-1}}$$

The absolute agreement of single-rater ICC(2, 1) and the absolute agreement of multiple-rater ICC(2, k) are estimated using equation (2) as follows.

$$\mathrm{ICC}(2,1) = \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_\alpha^2 + \sigma_\epsilon^2},\ \mathrm{ICC}(2,k) = \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + (\sigma_\alpha^2 + \sigma_\epsilon^2)k^{-1}}$$

The consistency of single-rater ICC(3, 1) and the consistency of multiple raters are estimated using equation (3) as follows.

$$\mathrm{ICC}(3,1) = \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_\epsilon^2},\ \mathrm{ICC}(3,k) = \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_\epsilon^2 k^{-1}}$$

ReX also provides parametric and nonparametric multivariate formulations of reliability that were recently developed in the imaging field, namely, dbICC[14], the I2C2[15], discriminability and the identification rate (that is, fingerprinting)[17] as well as univariate nonparametric generalizations when appropriate (that is, discriminability and identification rate). The parametric reliability (dbICC) is based on the Euclidean distance estimated by

$$\mathrm{dbICC} = 1 - \frac{\mathrm{MSD_w}}{\mathrm{MSD_b}},$$

where $\mathrm{MSD_w}$ is the mean within-individual distances and $\mathrm{MSD_b}$ is the mean between-individual distances of the observed score(s) for all variable(s) of interest. The parametric multivariate reliability (I2C2) is estimated by

$$\mathrm{I2C2} = 1 - \frac{\mathrm{trace}(K_\mathrm{u})}{\mathrm{trace}(K_\mathrm{o})},$$

where

$$\mathrm{trace}\,(K_\mathrm{o}) = \frac{1}{\Sigma_i J - 1}\sum_i\sum_j\sum_v (X_{ij}(v) - X_{\cdot\cdot}(v))^2,$$

and

$$\mathrm{trace}(K_\mathrm{u}) = \frac{1}{\Sigma_i (J-1)}\sum_i\sum_j\sum_v (X_{ij}(v) - X_{\cdot i})^2.$$

Here $X_{\cdot\cdot}(v)$ is the average over all individuals and all repetitions J for each variable v. $X_{\cdot i}$ is the average over all repetitions j for each individual i and variable v. The nonparametric reliability indices (discriminability and fingerprinting) are both estimated by comparing the observed within-individual distance to the observed

between-individual distance. Discriminability is the fraction of times that observed within-individual similarity is greater than the between-individual similarities[16]. Identification rate (that is, fingerprinting) is the proportion of individuals whose within-individual similarities over all repetitions are higher than all the between-individual similarities[17].

The Gradient Flow Map

In the theoretical field map, one can recognize that both decreases in x and increases in y can improve reliability. However, the contribution of within-individual (Δx) and between-individual (Δy) variance to the increase in reliability is not the same. If a measure has a relatively small x and large y, the reduction in x improves reliability more than the same increment in y. On the other hand, if a measure is relatively high in x but low in y, an increase in y improves reliability more than the same reduction in x. In theory, the most efficient direction to improve reliability can be calculated as the first derivative of the reliability, which is $(-y(y+x)^{-2}, x(y+x)^{-2})$. As shown in Fig. 2b, the optimal direction of a measure at (x_0, y_0) is always perpendicular to the vector (x_0, y_0). When $x_0 = y_0$ (that is, reliability = 0.5), the optimal direction (slope = −1, angle of the slope = (3/4)π) in the x axis and the y axis is the same ($|\Delta x| = |\Delta y|$). In ReX, we use this optimal direction when $x = y$ as the reference to normalize the relative change of Δx and Δy (Fig. 2c). Specifically, let (x_0, y_0) be the estimated within- and between-individual variance of a measure. The change of (x_0, y_0) to (x_1, y_1) is Δx and Δy. The relative Δx and Δy can be calculated by rotating $(\Delta x, \Delta y)$ by a relative angle to the $x = y$ line.

$$\text{Normalized } \Delta x = \cos(\theta)\Delta x - \sin(\theta)\Delta y$$

$$\text{Normalized } \Delta x = \sin(\theta)\Delta x + \cos(\theta)\Delta y,$$

$$\text{where } \theta = \frac{1}{4}\pi - \arctan(\frac{y_0}{x_0}).$$

In ReX, we use a standard circular color map (Fig. 2c) to visualize the angle of the normalized changes of x and y. The darker red and magenta represent Δx and Δy improved reliability, while darker blue and green represent Δx and Δy decreased reliability from (x_0, y_0) to (x_1, y_1). The light color indicates that the change is less close to the optimal direction.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data used in application examples are available from public repositories. HCP data are available on ConnectomeDB (https://www.humanconnectome.org/study/hcp-young-adult)[10]. Self-regulation data are available on GitHub (https://github.com/IanEisenberg/Self_Regulation_Ontology)[11]. HNU data are available from the Consortium for Reliability and Reproducibility (https://fcon_1000.projects.nitrc.org/indi/CoRR/html/index.html)[19]. Application data and code are available on GitHub (https://github.com/TingsterX/Reliability_Explorer/tree/main/application_examples). Source data are provided with this paper.

Code availability

ReX is implemented using multiple R packages (lme4, dplyr, ggplot2, scales, stats, reshape2, shinybusy, colorspace, RColorBrewer). The toolbox is available under a GNU version 3 license on GitHub (https://github.com/tingsterx/reliability_explorer), with a web-based R–Shiny application on Docker Hub (tingsterx:reliability_explorer) and shinyapps.io: https://tingsterx.shinyapps.io/ReliabilityExplorer. Docker images of the command line version (tingsterx:rex) used in this paper are available on Docker Hub.

References

21. Steyer, R., Smelser, N. J. & Jena, D. Classical (psychometric) test theory. *In International Encyclopedia of the Social & Behavioral Sciences* **Vol. 3**, 1955–1962 (2001).
22. Kline, T. J. B. *Psychological Testing: a Practical Approach to Design and Evaluation* (SAGE, 2005).
23. Noble, S., Scheinost, D. & Constable, R. T. A decade of test–retest reliability of functional connectivity: a systematic review and meta-analysis. *NeuroImage* **203**, 116157 (2019).

Acknowledgements

We thank X. Li for organizing the preprocessed HNU data from different pipelines. This work is supported by gifts from J.P. Healey, P. Green and R. Cowen to the Child Mind Institute and National Institutes of Health funding (RF1MH128696 to T.X., R24MH114806 and 5R01MH124045 to M.P.M.). Additional grant support for J.T.V. comes from R01MH120482 (to T.D. Satterthwaite, M.P.M.), and he has funding from Microsoft Research.

Author contributions

T.X. conceptualized and developed the software. T.X. and J.W.C. prepared the data. T.X. wrote the original draft with input from M.P.M., G.K. and J.T.V. All authors reviewed, edited and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

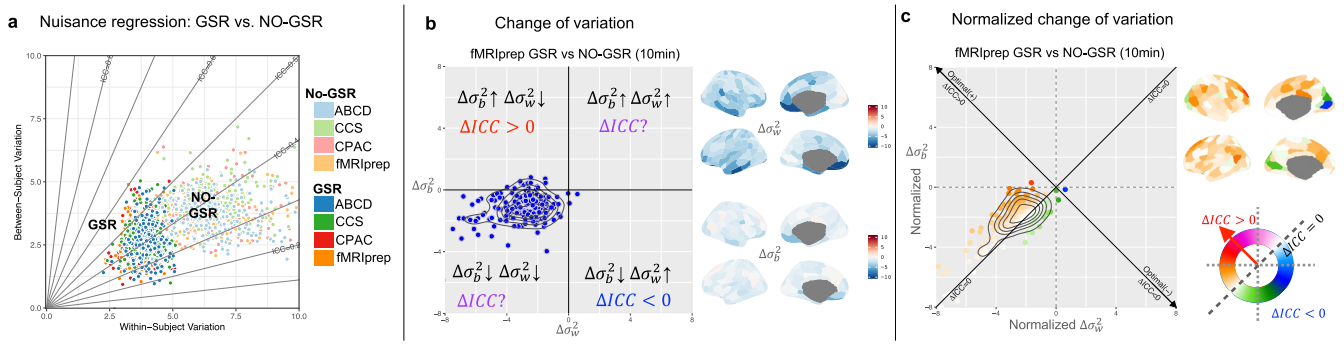
Extended data is available for this paper at https://doi.org/10.1038/s41592-023-01901-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-023-01901-3.

Correspondence and requests for materials should be addressed to Ting Xu.

Peer review information *Nature Methods* thanks Ye Tian and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editor: Nina Vogt, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

https://doi.org/10.1038/s41592-023-01901-3



Extended Data Fig. 1 | Results for Application 4 to compare the impact of global signal regression in multiple fMRI preprocessing pipelines at the parcel level. a) The within- and between-individual variance of GSR and No-GSR results from four pipelines. **b**) The change of within- and between-individual variance comparing GSR versus No-GSR results of the fMRIprep pipeline. **c**) The normalized change of the within- and between-individual variance comparing GSR versus No-GSR results of the fMRIprep pipeline.

nature portfolio

Corresponding author(s): Ting Xu

Last updated by author(s): Mar 17, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	N/A. No data is collected in the current study. Data used are available from previous published studies. See Data Availability Statement
Data analysis	The ReX toolbox is available on GitHub (https://github.com/TingsterX/Reliability_Explorer) and Dockerhub (tingsterx/rex:v1.0.1). R 4.2.1 dplyr 1.0.10 lme4 1.1.30 ggplot2 3.3.6 RColorBrewer 1.1.3 scales 1.2.1 stats 4.2.1 reshape2 1.4.4 colorspace 2.0.3 shinybusy 0.3.1 I2C2 0.2.4 The fMRI data were preprocessed using Configurable Pipeline for the Analysis of Connectomes (C-PAC) pipeline v1.8.2 ABCD, CCS, fMRIPrep pipelines used in the current study are the C-PAC comparable version implemented in C-PAC v1.8.2 Pipeline code is available on GitHub: https://github.com/FCP-INDI/C-PAC/releases/tag/v1.8.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data used in application examples are available from public repositories. HCP dataset is available on ConnectomeDB (https://www.humanconnectome.org/study/hcp-young-adult). Self-regulation dataset is available on Github (https://github.com/IanEisenberg/Self_Regulation_Ontology). HNU dataset is available on Consortium for Reliability and Reproducibility (CoRR: https://fcon_1000.projects.nitrc.org/indi/CoRR/html/index.html). Application data are available on Github (https://github.com/TingsterX/Reliability_Explorer/tree/main/application_examples).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Sex was reported for each of datasets used in the current study. 1) HCP test-retest dataset: N=46, 32 Female. 2) Self-regulation test-retest dataset: N=150, 92 Female. 3) HNU test-retest dataset: N=30, 15 Female. 4) HCP unrelated participants from S1200 release: N=170, 92 Female.
Population characteristics	1) HCP test-retest dataset: N=46, 32 F, mean age = 30.20, std = 3.36) 2) Self-regulation test-retest dataset (N=150, 92 F, mean age = 34.15, std=7.38) 3) HNU test-retest dataset (N=30, 15 F, mean age = 24, std = 2.41) 4) HCP unrelated participants from S1200 release (N=170, 92F, mean age 28.5, std=3.47)
Recruitment	No subjects were recruited for this study. All data comes from public repositories.
Ethics oversight	All data used in the current study were collected from previous study with IRB approval at their original institutions under informed consent from participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	There are four previously published test-retest datasets used for application examples. They are 1) HCP test-retest dataset (N=46, 2 sessions), 2) self-regulation test-retest dataset (N=150, 2 sessions), 3) HNU test-retest dataset (N=30, 10 sessions) and (4) HCP unrelated participants from S1200 release (N=170, 2 sessions). Each has 0.95, 0.99, 0.84, 0.99 power to achieve a moderate (ICC=0.5) reliability (alpha=0.05, two tails).
Data exclusions	For HCP test-retest and self-regulation datasets, participants who didn't complete both test and retest sessions were excluded. For HNU and HCP S1200 datasets, participants with the higher head motion (framewise displacement ≥0.25 mm) were excluded.
Replication	The containerized toolbox is provided on Dockerhub. Details see "Code availability" session
Randomization	Participants were not divided into separate experimental groups
Blinding	Participants were not divided into separate experimental groups

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Magnetic resonance imaging

Experimental design

Design type	Resting state fMRI
Design specifications	No specifications
Behavioral performance measures	None

Acquisition

Imaging type(s)	Structural (T1w) and functional MRI
Field strength	3 Tesla
Sequence & imaging parameters	HCP dataset: TR = 720 ms, TE = 33 ms, Flip angle = 52 degree, multi-slice factor N=8, FOV=20.8cm(A-P) x 18cm(R-L). See details http://fmri.ucsd.edu/Howto/3T/HCP.html HNU dataset: TR = 2000mx, TE = 30ms, flig angle = 90 degree. See details: http://fcon_1000.projects.nitrc.org/indi/CoRR/html/hnu_1.html
Area of acquisition	whole brain scan

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	The Configurable Pipeline for the Analysis of Connectomes (C-PAC, https://fcp-indi.github.io/docs/latest/user/index). In addition to the C-PAC default pipeline, Adolescent Brain Cognitive Development (ABCD), Connectome Computational System (CCS), and fMRIPrep pipelines were also employed using C-PAC.
Normalization	ANTs was used for the non-linear registration for ABCD, CPAC, fMRIPrep pipelines. FSL Fnirt was used for CCS pipeline.
Normalization template	MNI152 template
Noise and artifact removal	Data with global signal regression (GSR) and without GSR were examined, Details see Application 3-6
Volume censoring	No volume censoring was performed

Statistical modeling & inference

Model type and settings	No model type or setting required
Effect(s) tested	The test-retest reliability of the functional connectivity from pipelines

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)	The timeseries were averaged first within each parcel to calculate the parcel-wise connectivity matrix.
Correction	Not applicable

Models & analysis

n/a	Involved in the study
☐	☒ Functional and/or effective connectivity
☒	☐ Graph analysis
☒	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity: Pearson correlation was used to calculate the functional connectivity