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

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]

1. Department of Brain Development, Child Mind Institute, New York, USA
2. Johns Hopkins University, Baltimore, Maryland, USA
3. Nathan Kline Institute for Psychiatric Research, Orangeburg, New York, USA

*ting.xu@childmind.org

Abstract

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 the individual difference 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.

Main

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 (e.g., demographic, behavioral, cognitive, psychiatric)[1–8]. An inherent assumption of this shift is that the measures employed are reliable — i.e., 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 Reliability eXplorer (ReX), an open-source tool designed to facilitate quantification and optimization process by addressing a critical gap in studying individual differences: the failure to take into account the component variances of reliability (i.e., within- and between-subject 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-subject variance may be inflated, as the contributions of within-subject variances to its measurement are rarely considered. The following features in ReX are intended to help address these challenges:

Feature 1. Evaluation and visualization tools to identify the impact of variations and their contributions to reliability

Previous efforts in studying individual differences commonly focus on between-subject variation of observations and treat this as the true inter-individual difference[9]. Within-subject variation is, in contrast, often overlooked or misinterpreted when studying inter-individual differences in brain function, in particular in cross-sectional studies. For example, metabolomic or biological/psychological changes in 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 inter-individual differences which are contaminated with the within-individual variations as the true individual difference can compromise brain-behavior association discovery across individuals. Deciphering sources of variation both within- and between-subjects is central to interpreting individual differences in these scenarios. In ReX, we formally construct the variation space and provide a visualization module (Figure 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 are the "true" between-individual variation rather than the observed between-individual variation. Using the variation space, it's easier to differentiate within- from the between-individual variation and examine which factors (e.g. analytic methods, experiment designs, subject 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 presented the theoretical variation field map (Figure 1a) and examples of a wide range of behavioral measurements (details in Methods). We demonstrated the between-individual variation in Human Connectome Project (HCP) behavioral battery is

not consistent across task domains (Figure 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 (Figure 1c)[11].

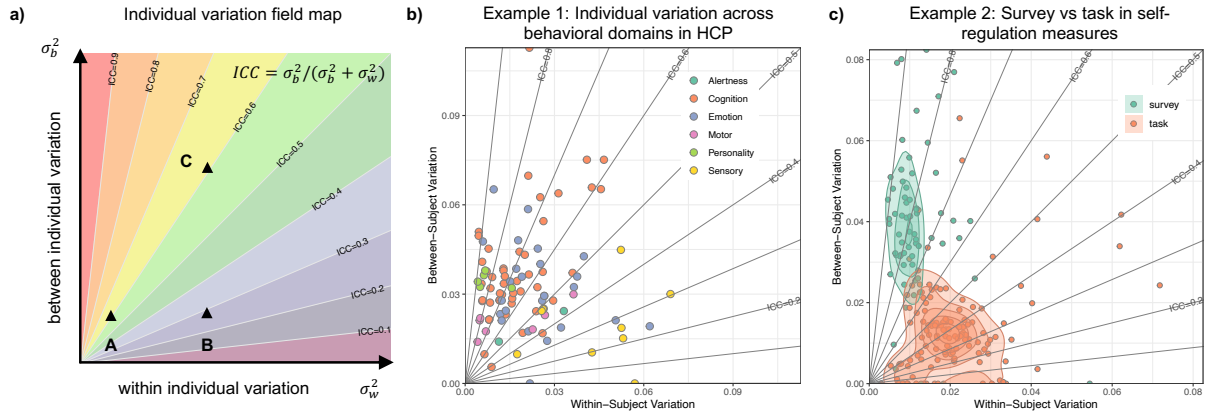


Figure 1. Theoretical individual variation field map in ReX and its applications. **a)** The two-dimension theoretical individual variation field map characterizes within- and between-individual variability and the likelihood of individual characterization (quantified by the intraclass correlation [ICC] reliability). **b)** Within- and between-individual variations of National Institutes of Health (NIH) 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.

Feature 2. Optimization module via Gradient Flow Map (GFM).

In addition to the variation calculation module, the novel feature of ReX is the GFM, which indicates the most efficient direction to improve the reliability in measuring individual differences. As shown in Figure 1a, reducing within-individual variation (i.e. from point B to A) and increasing between-individual variation (i.e. from point B to C) can improve reliability to the same extent (i.e. delta ICC = 0.3). 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 a relatively small within-individual variation (x-axis) and large between-individual variation (y-axis) (e.g. Figure 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 (e.g. point C in Figure 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 (i.e. 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. As shown in Figure 2d, when comparing the performance of two different measures (e.g. pipelines in measuring brain functional connectivity, details in Supplementary), the change of the variation in x and y cannot fully determine whether it improves reliability (Figure 2e, the 1st, 3rd quadrant). 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 (Figure 2c). The resultant GFM (Figure 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 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 (examples in Methods). 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.

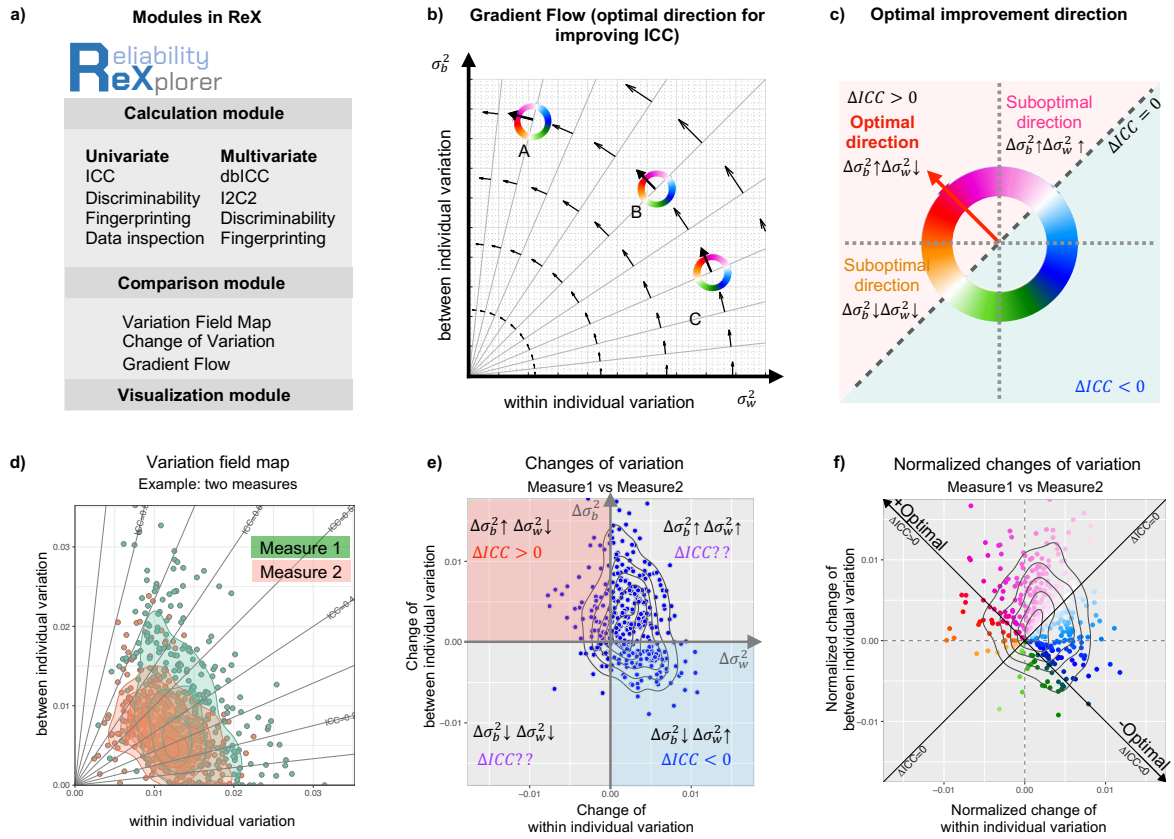


Figure 2. Gradient flow map in ReX and its application example. **a)** ReX modules. **b)** The theoretical gradient flow map (i.e. the first derivative of the ICC) captures the most efficient way to improve reliability. **c)** Normalized changes of variation as compared to the optimal direction for improving ICC. **d)** Example of within- and between- individual variability of two different measures. **e)** Change of variation can not fully determine whether it improves reliability. **f)** Normalized change of variation to the optimal direction reveals whether one measure displays higher or lower reliability than the other.

Feature 3. Differing formulations for reliability.

To accommodate the needs of a broad range of designs, ReX offers users a range of parametric and non-parametric methods for both univariate and multivariate reliability. Intraclass correlation (ICC) formations including one-way random, two-way random, and two-way mixed models using the linear mixed model (LMM in 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 utilizes 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), the two-way mixed model for single measure consistency ICC(3,1) and average measure consistency ICC(3,k)[13]. In LMM, the random factors and residuals are assumed to be independent. Users can specify the confounding variables as covariates in the model (e.g. age, sex). The parametric and non-parametric multivariate formulations of reliability implemented in ReX were recently developed in the imaging field including the distance-based ICC (dbICC)[14], the image intraclass correlation coefficient (I2C2)[15], discriminability[16], and identification rate (i.e. fingerprinting)[17]. It's important to note that reliability is a prerequisite and the upper bound for validity. Yet, it doesn't 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 (see Methods and Supplementary Video 1).

Example applications

To demonstrate the utility of ReX, we include six example applications (see Methods and Extended Data Figure 1). Application 1 shows the differential contributions of within- and between-subject 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. The resulting visualizations from ReX are included to make obvious how the tool allows users to intuitively interpret the results easily. Of note, ReX can be applied to any repeated measure dataset, though the power and effect size of the reliability will depend on the data quality and quantity. It's recommended to also calculate the power of the reliability and consider tradeoffs 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, cost of the approach (e.g. scan time, collecting rare patients, etc.) to select measures needs to be considered for assessing individual differences and reliability[18–20].

Summary

Recognizing the growing need for techniques to guide optimization efforts for the measurement of individual differences, we proposed the reliability gradient flow map to quantify the optimization efforts of measuring reliability and individual variations. We develop ReX that 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.

Methods

ReX follows the classical test theory and provides visualization of theoretical variation field map, gradient flow map, as well as the reliability-validity relationship map to facilitate understanding of 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 (i.e. $Reliability = \sigma_T^2/(\sigma_T^2 + \sigma_E^2)$). 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 inter-individual variance. In ReX, we use two-dimensional space (i.e. variation field map) to formulize the true between-individual variation (y-axis) and within-individual variation (x-axis). The visualization of this field map (Figure 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 Ti and the unwanted score Tu that is not related to the trait of interest (i.e. contaminants relative to the trait of interest).

$$\sigma_T^2 = \sigma_{Ti}^2 + \sigma_{Tu}^2$$

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].

$$Validity = \sigma_{Ti}^2/(\sigma_{Ti}^2 + \sigma_{Tu}^2 + \sigma_E^2)$$

$$Reliability = (\sigma_{Ti}^2 + \sigma_{Tu}^2)/(\sigma_T^2 + \sigma_{Tu}^2 + \sigma_E^2)$$

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 Tr, validity equals reliability. If there is a signal but not related to the trait, validity is lower than reliability (see the theoretical plot in Supplementary Video 1 or on 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 doesn't 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].

The reliability models

ReX includes multiple intraclass correlation (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 linear mixed model (LMM)[12,13].

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

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

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

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

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

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

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

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

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

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

$$dbICC = 1 - \frac{MSDw}{MSDb},$$

where MSD_w is the mean within-individuals distances and 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

$$I2C2 = 1 - \frac{trace(Ku)}{trace(Ko)},$$

where

$$trace(Ko) = \frac{1}{\Sigma_i J-1}\sum_i\sum_j\sum_v(X_{ij}(v) - \underline{X..}(v))^2,$$

and

$$trace(Ku) = \frac{1}{\Sigma_i (J-1)}\sum_i\sum_j\sum_v(X_{ij}(v) - \underline{X_{.i}})^2.$$

Here $\underline{X..}(v)$ is the average over all subjects and all repetitions *J for* each variable *v*. The $\underline{X_{.i}}$ is the average over all repetitions *j* for each subject *i* and variable *v*. The non-parametric 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 (i.e. fingerprinting) is the proportion of subjects whose within-individual similarities over all repetitions are higher than all the between-individual similarities[17].

The Gradient Flow Map (GFM)

In the theoretical field map, one can recognize that both decrease of x and the increase of y can improve reliability. However, the contribution of within- (Δx) and between-individual variance (Δy) to the increase in reliability is not the same. 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 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 Figure 2b the optimal direction of a measure at (x_0, y_0) is always perpendicular to vector (x_0, y_0). When $x_0=y_0$ (i.e. reliability=0.5), the optimal direction (slope=-1, angle of the slope=3/4π) 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 (Figure 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 the (Δx, Δy) by a relative angle to the x=y line.

$$Normalized\ \Delta x = cos(\theta)\Delta x - sin(\theta)\Delta y$$
$$Normalized\ \Delta x = sin(\theta)\Delta x + cos(\theta)\Delta y$$
$$where\ \theta = 1/4\pi - arctan(y_0/x_0)$$

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

Example applications.

Here, we provide six application examples using publicly available data in both brain and behavioral studies to demonstrate the utility of ReX. The sample codes are available on GitHub (https://github.com/TingsterX/Reliability_Explorer/tree/main/application_examples).

Application 1-2. Understanding differences in reliability among tools based upon their component variance. When attempting to understand the performance of differing tools for assessing individual differences in behavior and cognition, comparisons of their component variances can be informative. Here, we demonstrate the utility of the ReX computation and visualization modules in informing our understanding of reliability differences commonly noted in i) NIH Toolbox measures of neurological and behavioral functions (http://www.nihtoolbox.org) using the Human Connectome Project (HCP) dataset (https://www.humanconnectome.org)[10], and ii) survey and task performance, using self-regulation measures from a previous study[11]. Using just a few clicks or a few command lines in ReX (Figure S1-2) can easily analyze the data and generate figures - highlighting the turnkey nature of the tool in visualizing such relationships.

Application 1. Motivated by findings that cognitive measures in brain-wise association studies show greater prediction accuracy than personality and emotion measures (cognition: r=0.4-0.6, personality: r=0.1-0.2, emotion: r=0.1-0.2)[24], we assess the reliability across these behavioral domains to explore to what extent the theoretical prediction can be attenuated[25]. Specifically, we pooled the HCP S1200 release and retest release data together (N=46, 32F, 2 visits). Age and sex are added as covariance in the ReX two-way mixed model. The original data are available via the HCP (https://www.humanconnectome.org/study/hcp-young-adult) with a restricted license. The result (Figure S1) showed that personality and cognitive measures have greater reliability (cognition ICC=0.69±0.12, personality ICC=0.83±0.09) than emotion and sensory tasks (emotion ICC=0.56±0.20). These findings suggest that the observed brain-wise association for cognition and personality is likely to reflect the theoretical association that brain-cognition association is stronger than the brain-personality association. However, when examining the association in emotion and sensory, researchers need to consider a larger sample size as they are relatively less reliable[18]. ReX also provides the estimation of the within- and between-individual variation, as well as the variation field map. The field map visualization allows users to intuitively attribute the difference to lower within-subject variance and higher between-subject variance for personality and cognitive measures.

Application 2. To make a decision of selecting the self-report survey or behavioral task as a behavioral measure for examining self-regulation, we re-analyzed the self-regulation test-retest data from the previously published study[11]. This dataset includes 150 participants who completed at least one self-regulation test twice within a year. The data are available at https://github.com/IanEisenberg/Self_Regulation_Ontology. Consistent with the previous study, the survey exhibited higher reliability (Figure S2). Using variation field maps generated by ReX, users easily get the idea of how higher reliability in the survey data attributes to lower within-subject variance and higher between-subject variance.

Applications 3-6. Comparing the pipelines, preprocessing strategy, data collection requirements, and data aggregation strategies for measuring brain functional connectivity.

The reliability of the neuroimaging data is the main challenge in the discovery of associations between individual differences in brain structure or function and behavioral phenotypes. All levels from the experiment designs, data collection, preprocessing, and downstream analytic strategies have an impact on the data. Quantification and comparison of the reliability and component variances can facilitate the selection of a better solution in studying individual differences of brain function. Here, we used functional magnetic resonance imaging (fMRI) test-retest dataset to demonstrate the utility of the gradient flow map from the ReX comparison module in comparing i) different minimal preprocessing pipelines (Application 3), ii) downstream nuisance removal strategies (Application 4), iii) the amount of data to acquire (10min vs. 30min, Application 5), and vi) data collection strategies (i.e. a single long scan versus multiple short scans, Application 6). Two publicly available fMRI datasets are used for Applications 3-6.

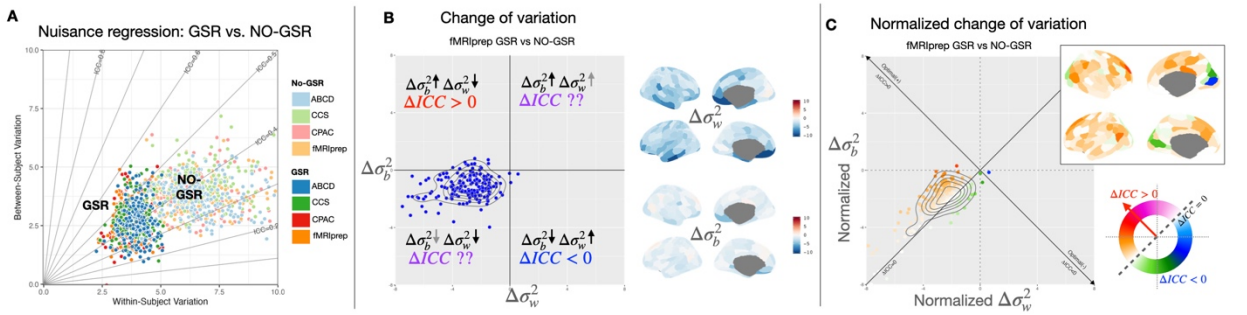
HNU dataset. The Hangzhou Normal University (HNU) test-retest dataset is publicly available from the Consortium for Reliability and Reproducibility (CoRR)[26]. The dataset consists of 30 healthy participants (15M, age=24 ± 2.41 years) who were each scanned for 10 sessions within a month (10 min per session x 10 sessions per subject). To reduce the dimensionality of the data, Schaefer parcellation (n=200) was applied and the preprocessed data was averaged within each parcel[27]. The data and preprocessing details were listed in our recent study[28]. Briefly, we configurated four conventional fMRI pipelines (fMRIPrep, ABCD, CCS, and C-PAC default) in Configurable Pipeline for the Analysis of Connectomes (C-PAC) and preprocessed this test-retest dataset in C-PAC. The preprocessed data was used in Application examples 3-5.

HCP dataset. The Human Connectome Project (HCP) dataset was preprocessed with the HCP minimal preprocessing pipeline[29]. We included 170 unrelated participants with low head motion (mean framewise displacement [FD] < 0.25mm) from the S1200 release that was analyzed in a previous study[19]. Similarly, to reduce the dimension of the data, the preprocessed data was averaged within each parcel based on Glasser et al. [30]. The preprocessed data was used in Application examples 6.

Application 3. This application example (Figure S3) aims to examine the influence of fMRI pipelines on individual variation and reliability. The HNU data was minimally preprocessed in our recent study using four conventional fMRI pipelines including fMRIprep, ABCD, CCS, and C-PAC default pipelines[28,31–34]. We carried out the pair-wise comparison between pipelines in ReX based on 30min data per participant. There are substantial changes in between-individual variation while the within-individual variation remains

relatively small. The results showed that each pipeline has a certain degree of improvement compared to the others. Yet, no single pipeline provides the best improvement across the whole brain and imbalance across different brain regions. This finding suggests the importance of examining the variability of the analytic tools in fMRI data processing.

Application 4. This application example (Extended Data Figure 1) aims to assess the impact of a combination of different factors on brain connectivity. One factor is fMRI preprocessing pipelines (e.g. fMRIprep, ABCD, CCS, and C-PAC)[28,31–34], and the other is global signal regression (GSR) which is the most controversial nuisance removal step. We used the HNU dataset in Application 3 and further applied the downstream preprocessing including nuisance regression, bandpass filtering (0.01-0.1Hz), and smoothing (FWHM=6mm along the surface). In the nuisance regression step across four fMRI pipelines, we conducted two different regression strategies - one adding the global signal in the regression (GSR) and one without GSR. To summarize the high-dimension connectivity at each of the 200 parcels, the multivariate dbICC method was used here. Across pipelines, both within- and between-individual variances are relatively similar while GSR has a large impact, reducing both within- and between-individual variation (Extended Data Figure 1a). Take the fMRIprep pipeline as an example, GSR reduces the within-individual variation in the medial frontal lobe (Extended Data Figure 1b). However, the gradient field map and the corresponding surface map demonstrated that the change of the variation does not necessarily improve the reliability in the medial frontal lobe (Extended Data Figure 1c). Instead, the lateral frontal lobe and the temporoparietal junction with reduced within- and between-individual variation yielded improved reliability (Extended Data Figure 1c).



Extended Data Figure 1. Application 4 – fMRI preprocessing pipelines comparison. 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.

Application 5. In this application (Figure S4), we examined the reliability of different amounts of fMRI data to address a key question in fMRI data requirement - to what extent the reliability can be improved by acquiring more data and how within- and between-individual variations contribute to the improvement. The demo data is the HNU dataset used in Application 3 and was minimally preprocessed in our recent study using four conventional fMRI pipelines including fMRIprep, ABCD, CCS, and C-PAC default pipeline. We demonstrated the increase in the amount of data required per participant leads to a decrease in within-individual variation across all the functional connectivity (Figure S4). It also increased the between-

individual variation for a large proportion of the functional connectivities. Overall, requiring more data per participant improved the reliability in the optimal direction yet not the same across brain regions.

Application 6. How to obtain the desired amounts of data per participant is one of the challenges in the fMRI design. The same amount of data can be collected within a single long scan or multiple shorter scans data concatenation. Our previous study has demonstrated that multiple shorter scans improve the reliability of functional connectivity compared to a long single scan with the same amount of data in total[19]. In this application example, we made use of the data (HCP test-retest dataset) from our previous study and examined such data aggregation strategies[19]. We calculated the normalized change of the within- and between-individual variance in ReX for each functional connectivity. Compared to a single long scan (14 min), concatenating two shorter scans (2 x 7 min) reduced the within-individual variation and yielded higher reliability (Figure S5) for most of the functional connectivities, in particular the within-network connectivities.

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)[29]. Self-regulation data are available on Github (https://github.com/IanEisenberg/Self_Regulation_Ontology)[11]. HNU data are available on Consortium for Reliability and Reproducibility (CoRR: https://fcon_1000.projects.nitrc.org/indi/CoRR/html/index.html) [26]. Application data and code are available on Github (https://github.com/TingsterX/Reliability_Explorer/tree/main/application_examples).

Code availability

ReX is implemented using multiple R packages (lme4, dplyr, ggplot2, scales, stats, reshape2, shinybusy, colorspace, RColorBrewer). The toolbox is available on GitHub (https://github.com/tingsterx/reliability_explorer), with a web-based R/Shiny application on DockerHub (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 DockerHub.

Acknowledgments

We thank Xinhui Li for organizing the preprocessed HNU data from different pipelines. This work is supported by gifts from Joseph P. Healey, Phyllis Green, and Randolph Cowen to the Child Mind Institute and the National Institutes of Health fundings (RF1MH128696 to TX, R24MH114806 and 5R01MH124045 to MPM, Additional grant support for JTV comes from R01MH120482 (to Theodore D. Satterthwaite, MPM) and funding from Microsoft Research.

Author Contributions

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

Reference

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* vol. 160 140–151 Preprint at https://doi.org/10.1016/j.neuroimage.2017.03.064 (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. *PsyArXiv*. (2022) doi:10.31234/osf.io/zq37h.
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. U. S. A.* **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).
21. Steyer, R., Smelser, N. J. & Jena, D. Classical (psychometric) test theory. International encyclopedia of the social & behavioral sciences 3, 1955–1962 (2001).
22. Kline, T. J. B. Psychological Testing: A Practical Approach to Design and Evaluation. (SAGE Publications, 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).
24. Ooi, L. Q. R. et al. Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI. Neuroimage 263, 119636 (2022).
25. Marek, S. et al. Publisher Correction: Reproducible brain-wide association studies require thousands of individuals. Nature 605, E11 (2022).
26. Zuo, X.-N. et al. An open science resource for establishing reliability and reproducibility in functional connectomics. Sci Data 1, 140049 (2014).
27. Schaefer, A. et al. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cereb. Cortex 28, 3095–3114 (2018).
28. Li, X. et al. Moving Beyond Processing and Analysis-Related Variation in Neuroscience. bioRxiv 2021.12.01.470790 (2022) doi:10.1101/2021.12.01.470790.
29. Glasser, M. F. et al. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage 80, 105–124 (2013).
30. Glasser, M. F. et al. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178 (2016).
31. Feczko, E. et al. Adolescent Brain Cognitive Development (ABCD) Community MRI Collection and Utilities. bioRxiv 2021.07.09.451638 (2021) doi:10.1101/2021.07.09.451638.
32. Xu, T., Yang, Z., Jiang, L., Xing, X.-X. & Zuo, X.-N. A Connectome Computation System for discovery science of brain. Sci Bull. Fac. Agric. Kyushu Univ. 60, 86–95 (2015).
33. Esteban, O. et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods 16, 111–116 (2019).
34. Craddock, C. et al. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac). Front. Neuroinform. 42, (2013).

Supplementary Materials

Demo data in ReX Shiny App.

The demo data was from the publicly available Human Connectome Project (HCP) which has been preprocessed with the HCP minimal preprocessing pipeline (Glasser et al., 2013). The HCP includes 45 participants who were scanned twice for the entire HCP protocol as the HCP test-retest dataset[11]. We included 31 (23F, age=30.16±3.32) subjects who completed all the test-retest fMRI scans with low head motions (mean framewise displacement <= 0.25mm). In addition to the HCP minimal preprocessing, we further applied the nuisance regression (including Friston's 24 head motion parameters, mean signal from the white matter, mean signal from the cerebrospinal fluid, linear and quadratic trends), bandpass filtering (0.01-0.1Hz) and smoothing (FWHM=6mm along the surface). In the global signal regression (GSR) pipeline, we also included the global signal in the nuisance regression step. Finally, to reduce the dimension of the data, we used the Glasser parcellation and averaged the preprocessed data within each of the 360 parcels (Glasser et al., 2016). The signal of one parcel (labeled as 31pv_L) in the posterior cingulate cortex (PCC) was used as the seed to calculate the functional connectivity.

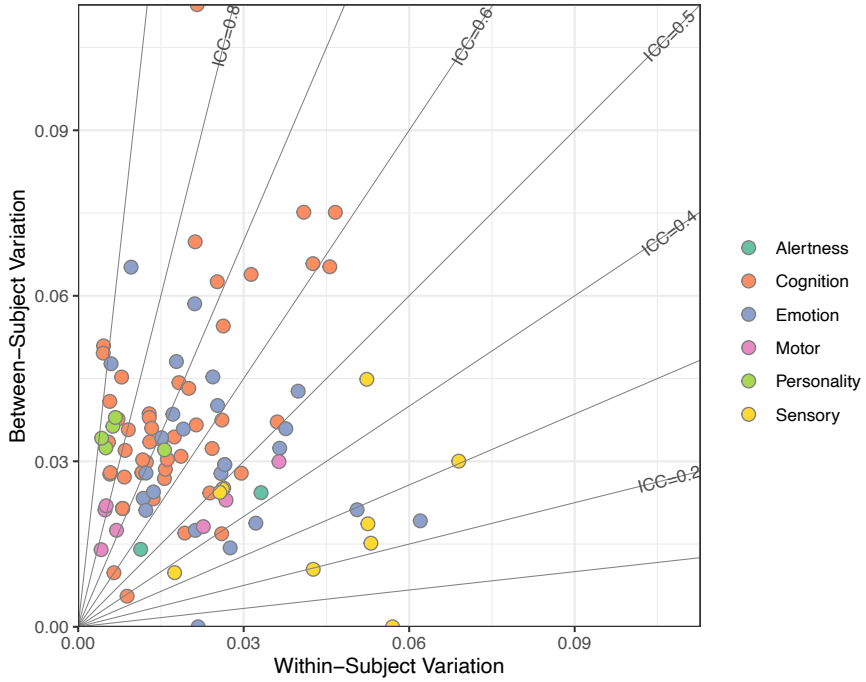
Supplementary Figures

R version

```
library(ReX)
subID   <- as.matrix(df["Subject"])
session <- as.matrix(df["session"])
cov     <- df[,c("Age", "Gender")]
# Calculate the individual variation and ICC using the 2-way mixed model
df <- data.frame(ReX::lme_ICC_2wayM(data, subID, session, cov))
# Plot the variation field map
ReX::rex_plot.var.field.n(df, group.name = "Domain",
                  size.point = 3, plot.density=FALSE, show.contour = FALSE)
```

Command-line version using docker

```
# Calculate the individual variation and ICC using 2-way mixed model
docker run --rm -v /data/demo1:/demo tingsterx/rex:latest rex_command_calc.R \
 -i /demo/data_Cog.csv -o /demo/result_Cog \
 -u ICC3 -s Subject -r session -d 5:57 -c Age,Gender
# Plot the variation field map
docker run --rm -v /data/demo1:/demo tingsterx/rex:latest rex_command_plotN.R -i
"/demo/result_Cog.csv","/demo/result_Emo.csv","/demo/result_Per.csv","/demo/result_Alt.csv","/demo/res
ult_Mot.csv","/demo/result_Sen.csv" --group_name Domain --divice png -o /demo/variation_field_map
```



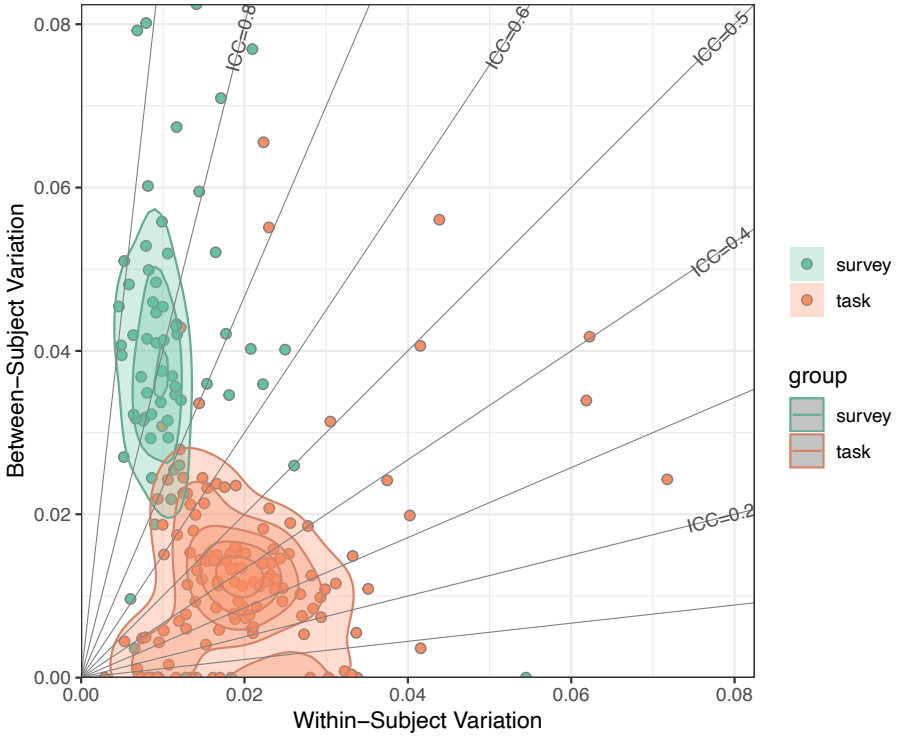
Supplementary Figure 1. Application 1 - an example of ReX code for calculating within- and between-individual variation for NIH Toolbox measures using the HCP retest dataset.

R version

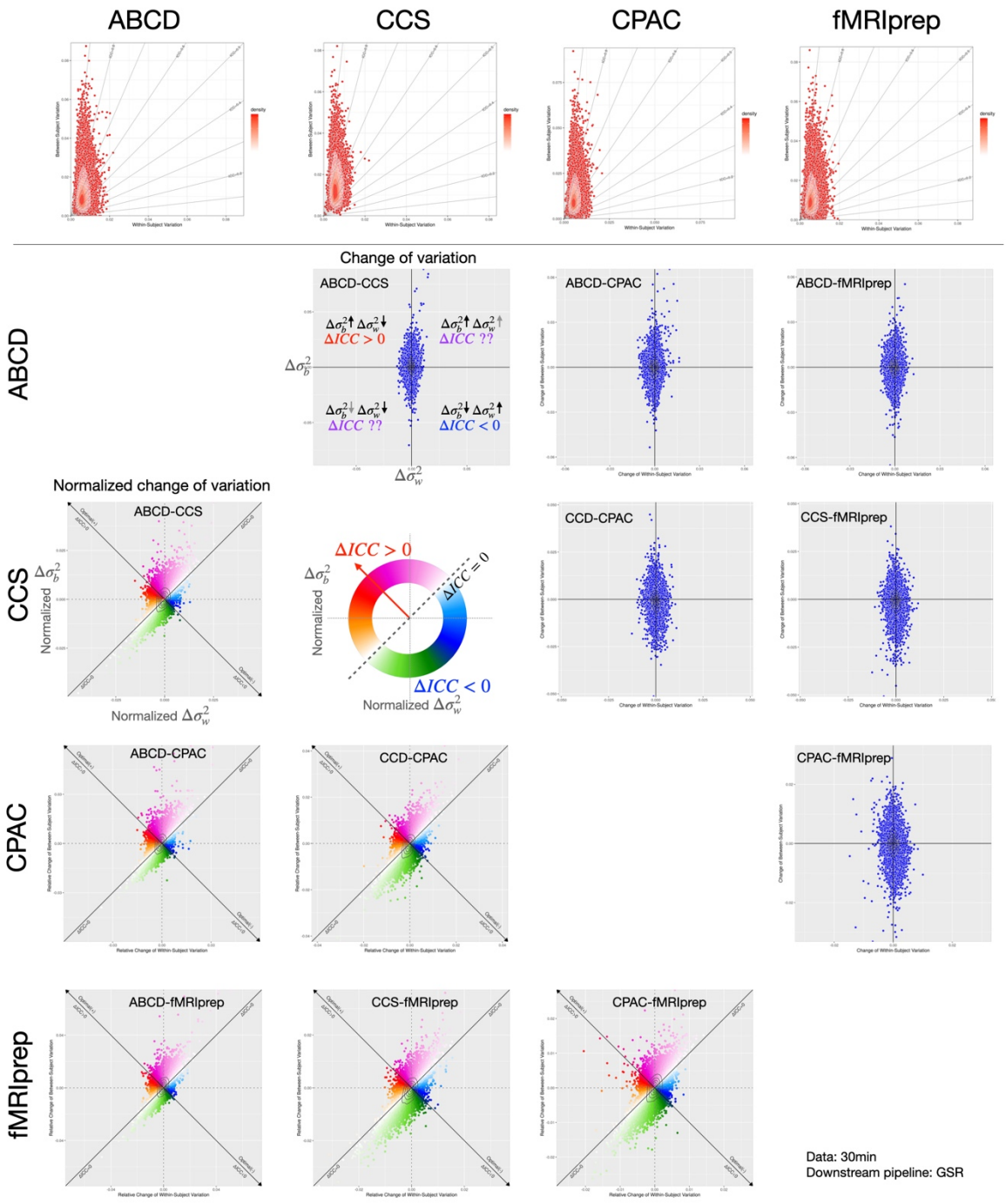
```
library(ReX)
subID   <- as.matrix(df[,"subID"])
session <- as.matrix(df[,"session"])
# Calculate the individual variation and ICC using 2-way mixed model
df_icc <- data.frame(lme_ICC_2wayM(data, subID, session))
# Plot the variation field map
rex_plot.var.field.n(df_icc, group.name = 'measure')
```

Command-line version using docker

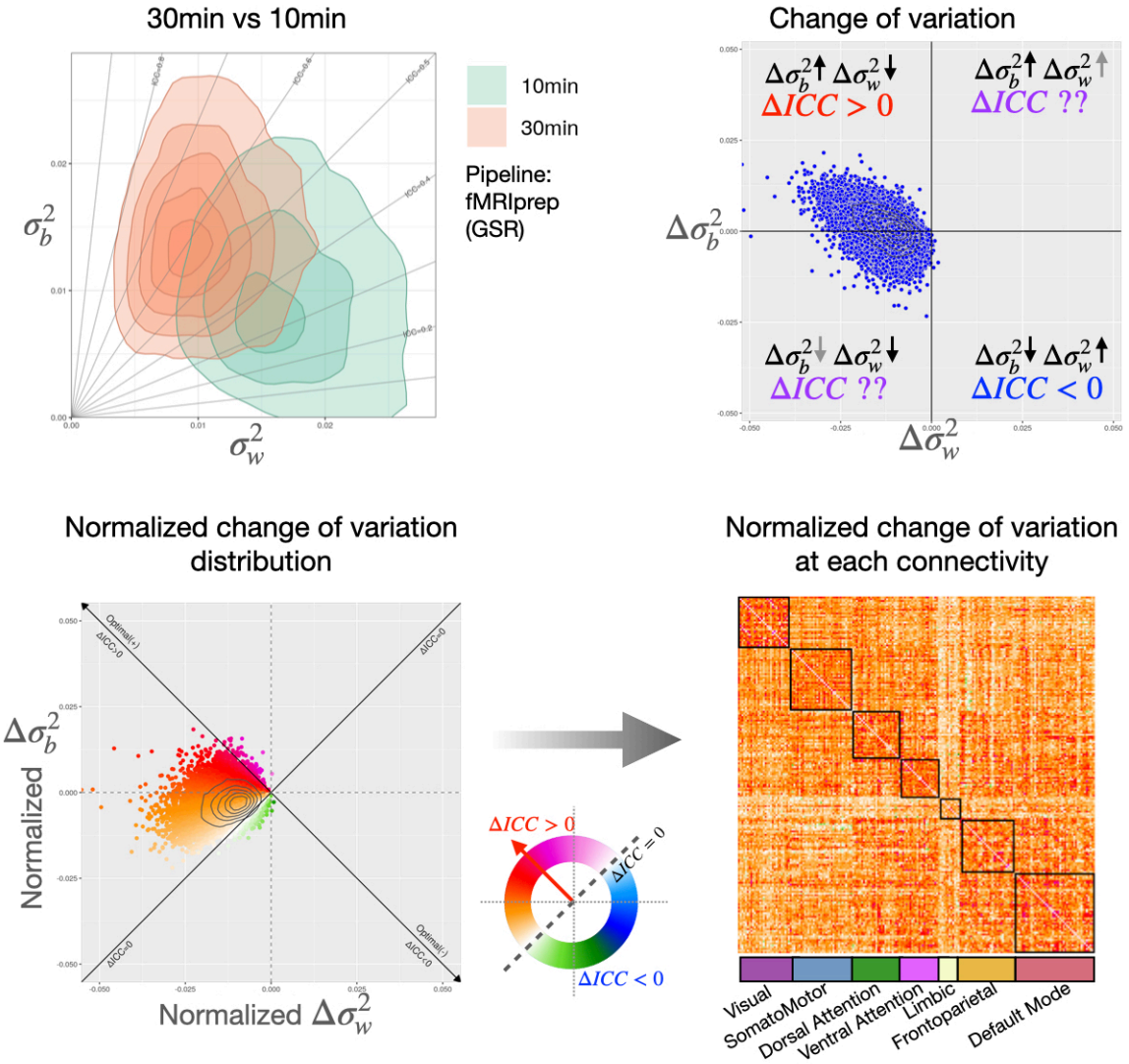
```
# Calculate the individual variation and ICC using 2-way mixed model
docker run --rm -v /data/demo2:/demo tingsterx/rex:latest rex_command_calc.R \
  -i /demo/data_task.csv -o /demo/result_task -u ICC3 -s subID -r session -d 3:130
docker run --rm -v /data/demo2:/demo tingsterx/rex:latest rex_command_calc.R \
  -i /demo/data_surv.csv -o /demo/result_surv -u ICC3 -s subID -r session -d 3:66
# Plot the variation field map
docker run --rm -v /data:/data tingsterx/rex:latest rex_command_plotN.R \
  -i "/demo/result_task_ReX_2wayM.csv","/demo/result_surv_ReX_2wayM.csv" \
  --divice png -o /demo/variation_field_map
```



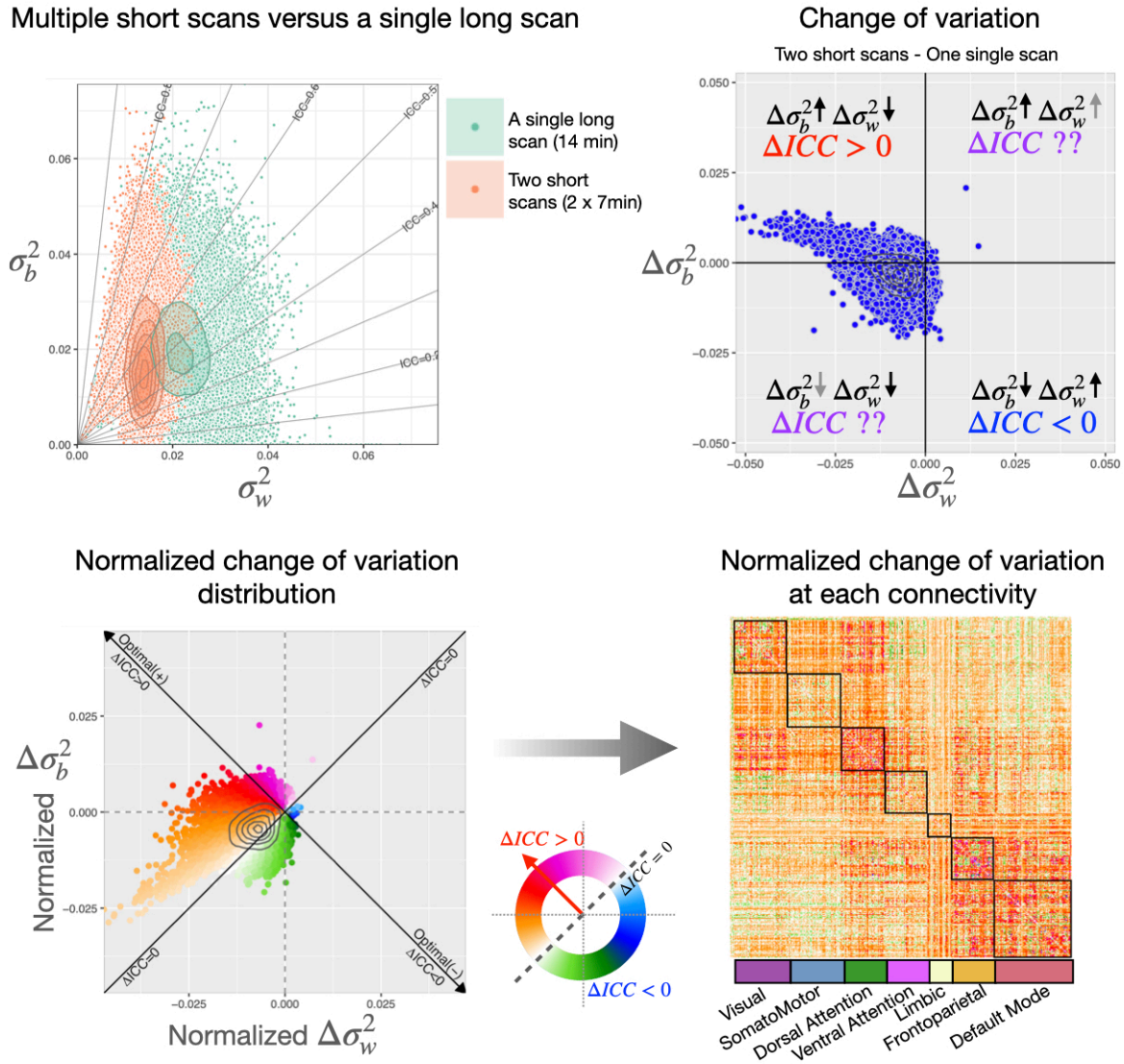
Supplementary Figure 2. Application 2 - an example of ReX code and output figure for calculating within- and between-individual variation for task and survey measures using Enkavi dataset.



Supplementary Figure 3. Results for Application 3 for comparing preprocessing packages (ABCD, CCS, CPAC, fMRIPrep) using the HNU dataset (N=30, 2 subsets, 30min per subset). The pipeline configurations are available at https://github.com/XinhuiLi/PipelineAgreement. The top panel is the variation field map indicating the within- and between-individual variation of functional connectivities for each pipeline. The upper triangle plots in the lower panel are the absolute change of variation between pipelines while the lower triangle plots are the normalized change of variation between pipelines.



Supplementary Figure 4. Results for Application 5 to determine the impact of amount of data in data collection (10 min vs 30 min). Upper left: the distribution of within- and between-individual variation for 30 min and 10 min data. Upper right: the change of variation comparing 30 min versus 10 min data. Lower left: the normalized change of variation comparing 30 min versus 10 min data. Lower right: the normalized change of variation for each connectivity.



Supplementary Figure 5. Results for Application 6 to determine how to obtain the desired amount of data (i.e. a single long scan versus multiple short scans). Upper left: the distribution of within- and between-individual variation for two data aggregation strategies. Upper right: the change of variation comparing data collected from two-shorter scans versus a long single scan. Lower left: the normalized change of variation comparing data collected from two-shorter scans versus a long single scan. Lower right: the normalized change of variation for each connectivity.