# Beyond the intraclass correlation: A hierarchical modeling approach to test-retest assessment

Gang Chen*[a], Daniel S. Pine[b], Melissa A. Brotman[c], Ashley R. Smith[b], Robert W. Cox[a], and Simone P. Haller[c]

[a]Scientific and Statistical Computing Core, National Institute of Mental Health, USA
[b]Section on Development and Affective Neuroscience, National Institute of Mental Health, USA
[c]Neuroscience and Novel Therapeutics Unit, Emotion and Development Branch, National Institute of Mental Health, USA

## Abstract

The concept of *test-retest reliability* indexes the repeatability or consistency of a measurement across time. High reliability is critical for any scientific study, specifically for the study of individual differences. Evidence of poor reliability of commonly used behavioral and functional neuroimaging tasks is mounting. These reports have called into question the adequacy of using even the most common, well-characterized cognitive tasks with robust population-level effects, to measure individual differences. Here, we lay out a hierarchical framework that estimates reliability as a correlation divorced from trial-level variability, and show how reliability tends to be underestimated under the conventional intraclass correlation framework. In addition, we examine how reliability estimation diverges between the modeling frameworks and assess how different factors (e.g., trial and subject sample sizes, relative magnitude of cross-trial variability) impact reliability estimates across the different frameworks. This work highlights that a large number of trials (e.g., greater than 100) may be required to achieve reasonably precise reliability estimates. We reference the tools of **TRR** and **3dLMEr** for the community to apply trial-level models to behavior and neuroimaging data.

## 1 Introduction

The concept of test-retest reliability originated from the notion of inter-rater reliability, i.e., the measurement of agreement or consistency across different observers (Shrout and Fleiss, 1979). All statistics compress and extract information from data; test-retest reliability captures the degree of agreement or consistency across multiple measurements (rather than observers) of the same quantity (reaction time (RT), BOLD response, personality traits) under similar circumstances. Traditionally, reliability is assessed through the statistical metric of the intraclass correlation coefficient (ICC). In the current context we define test-retest reliability as a property of individual differences, and consider ICC a conventional statistical measure of reliability.

Assessment of reliability is critical for almost all data collected in scientific studies. Here we focus on one specific type of data structure common to behavioral and neuroimaging studies: subjects perform an experiment with a task manipulation (i.e., a task with different conditions aimed to probe specific processes). Each experimental condition is instantiated with many trials. For example, an experimenter might evaluate

*Corresponding author. E-mail address: gangchen@mail.nih.gov

cognitive interference using the Stroop task (i.e., RT slowing due to conflicting information; MacLeod, 1991). The task includes two conditions, one where the name of the color (e.g., "blue", "green") and the color of the printed word match (i.e., congruent condition) and one where there is a mismatch between the print color and the color word (i.e., incongruent condition). The contrast between the two conditions renders an RT difference score ("contrast value") per subject that indexes the ability to inhibit cognitive interference. If this task is completed twice by the same subjects, reliability across these two measures can be computed. Usually this measurement takes the form of ICC which represents the fraction of the total variance that can be attributed to inter-individual differences in the participants' ability to ignore interfering information.

Assessment of reliability has always been part of rigorous questionnaire development; much less psychometric scrutiny has been applied to behavioral and imaging tasks until recently. Most concerningly, the ICC estimates, especially those derived from contrast or subtraction scores (Meyer et al., 2017; Infantolino et al., 2018), now being reported appear unacceptably low for behavioral (around 0.5 or below; Hedge et al., 2018) and imaging tasks (less than 0.4; Elliott et al., 2020), casting doubt on their utility in studies of individual differences. For neuroimaging tasks specifically, the extent to which common contrasts exhibit poor ICCs in the primary brain regions of interest has been troubling. Task-based fMRI has been used for some time to examine associations between individual differences as part of an important search for brain biomarkers of risk and disease. High reliability is a critical requirement for these usages.

Two aspects of the Stroop task example above are noteworthy and typical in modern test-retest reliability assessments of behavioral and imaging tasks: i) the experimenter seeks to assess the reliability of a contrast between the point estimates of two conditions, and ii) many trials are used as instantiations of each task condition. While trials are clearly an important source of variance, these are rarely included in the model structure and certainly have not occupied a place in traditional reliability calculations via ICC. More broadly, modeling trial-level variance in general has not been widely practiced in neuroimaging.

Through the adoption of a hierarchical framework, we show that, just as in behavioral studies (Rouder et al., 2019), trial-level variability is relatively large for many common neuroimaging experiments. We demonstrate that, if we explicitly account for trial-level effects in a hierarchical model, reliability estimates are higher compared to those derived from conventional ICCs. We systematically explore how reliability estimates from the hierarchical solution compare to ICC derived estimates, demonstrating that differences emerge (with ICCs lower than those from the hierarchical framework) when the number of trials is low and/or cross-trial variability is high; the number of subjects has surprisingly little impact. We detail other advantages of the hierarchical framework including information on estimation uncertainty.

### 1.1 Classical definition of ICC

In all comparisons with the ICC, we focus on the most common type, ICC(3,1), which quantifies the consistency (i.e., rather than absolute agreement) of an effect of interest between two sessions when the same group of subjects is measured twice (Shrout and Fleiss, 1979). Returning to our introductory example, suppose that the effect of interest is the contrast between incongruent and congruent conditions of the Stroop task. The investigator typically recruits $n$ subjects who perform the Stroop task in two sessions while their RT is collected. Suppose that each of the two conditions is represented by $m$ trials as exemplars in the experiment. The investigator typically follows the conventional two-level analytical pipeline. First, the original subject-level data $y_{crst}$ ($c = 1, 2;\ r = 1, 2;\ s = 1, 2, ..., n;\ t = 1, 2, .., m$) from the $s$-th subject during $r$-th repetition for the $t$-th trial under the $c$-th condition are averaged across trials to obtain the condition-level effect estimates $\widehat{y}_{crs\cdot}$ that are followed by contrasting the two conditions,

$$y_{rs} = \widehat{y}_{1rs\cdot} - \widehat{y}_{2rs\cdot} \tag{1}$$

Then, at the population level, the condensed data $y_{rs}$ are fed into a condition-level model (CLM) under a two-way mixed-effects ANOVA or LME framework with a Gaussian distribution,

$$\begin{aligned} &y_{rs}|a_r, \tau_s, \sigma_e \sim \mathcal{N}(a_r + \tau_s,\ \sigma_e^2); \\ &\tau_s|\ \widetilde{\sigma}_\tau \sim \mathcal{N}(0,\ \widetilde{\sigma}_\tau^2); \\ &r = 1, 2;\ \ s = 1, 2, ..., n; \end{aligned} \tag{2}$$

where $a_r$ represents the population-level effect during the $r$-th repetition, which is considered to be a "fixed" effect under the conventional statistical framework; $\tau_s$ is a "random" effect associated with $s$-th subject. The ICC under (2) is defined as

$$\text{ICC(3,1)} = \frac{\widetilde{\sigma}_\tau^2}{\widetilde{\sigma}_\tau^2 + \sigma_e^2}. \tag{3}$$

Such a two-level analytical pipeline, averaging across trials and contrasting between conditions followed by the ICC computation, is typically adopted for behavioral and neuroimaging data analysis. The residual variability $\sigma_e$ may appear to capture the within-subject cross-repetition variability; however, as we will detail later, "hidden" variability remains embedded in $\sigma_e$.

The classical ICC can be viewed from two statistical perspectives. First, the total variance under the model (2) is partitioned into two components, *cross-subject variability* $\widetilde{\sigma}_\tau^2$ and *within-subject variability* $\sigma_e^2$; thus, the ICC formulation (3) directly reveals the amount of cross-subject variability relative to the total variability. Alternatively, ICC can be viewed as the Pearson correlation of the subject-level effects between the two repetitions,

$$\text{ICC(3,1)} = \text{Corr}(y_{1s},\ y_{2s}) = \frac{\text{Cov}(y_{1s},\ y_{2s})}{\sqrt{var(y_{1s})\ var(y_{2s})}} = \frac{\widetilde{\sigma}_\tau^2}{\widetilde{\sigma}_\tau^2 + \sigma_e^2}. \tag{4}$$

The ICC formulation is unique: unlike other generic *inter*class correlation where two variables are usually from two different classes (e.g., height and weight), ICC always involves two measures of the same class (e.g., the same type of effects $y_{1s}$ and $y_{2s}$). Further, homoscedasticity is implicitly assumed in the sense that the random variables $y_{1s}$ and $y_{2s}$ share the same variance across repetitions as shown in (4). This is distinct from a Pearson correlation in which the two variables are usually heteroscedastic.

The computation of conventional ICCs is relatively straightforward. Using "Study 1" of a publicly available dataset of the Stroop task (Hedge et al., 2018), we obtain a modest ICC(3,1) = 0.49 (Table 1) for the subtraction value between conditions. The evidence for population-level effects is quite strong with a Stroop effect of $a_1 = 81$ ms (95% uncertainty interval (72, 91)) and $a_2 = 59$ ms (95% uncertainty interval (49, 69)) for the first and second repetition, respectively. As a comparison, the two conditions separately (i.e., no subtraction value) show a higher reliability with an ICC(3,1) of 0.72 and 0.69 for congruent and incongruent condition, respectively.

## 1.2 Separation between population effects and test-retest reliability

It is conceptually important to differentiate the effects at different hierarchical levels (e.g., population and subject). Population-level effects are of general interest as researchers hope to generalize from specific samples to a hypothetical population. Population-level effects are captured through terms (usually called "fixed" effects) such as condition-level effects $a_r$ at the population level in the LME model (2). In contrast, lower-level (e.g., subject, trial) effects are mostly of no interest to the investigator since the samples (e.g., subjects and trials) are simply adopted as representatives of a hypothetical population pool, and are dummy-coded and expressed in terms such as subject-level effects $\tau_s$ in the LME model (2).

| Model | RT effect | Population Effects (ms) | | | Data Variability (ms) | | $R_v$ ($R_u$) | TRR |
|---|---|---|---|---|---|---|---|---|
| | | repetition | mean | 95% interval | subject | residual | | |
| model (2) | congruent | session 1 | 639 | (617, 660) | $\widetilde{\sigma}_\tau$: 64 | $\sigma_e$: 39 | 4.1 (0.93) | 0.72 |
| | | session 2 | 607 | (585, 629) | | | | |
| CLM | incongruent | session 1 | 720 | (700, 744) | $\widetilde{\sigma}_\tau$: 71 | $\sigma_e$: 47 | 3.0 (0.96) | 0.69 |
| | | session 2 | 666 | (641, 691) | | | | |
| ICC(3,1) | incongruent vs congruent | session 1 | 81 | (72, 91) | $\widetilde{\sigma}_\tau$: 23 | $\sigma_e$: 24 | 12.6 (0.61) | 0.49 |
| | | session 2 | 59 | (49, 69) | | | | |
| LME (5) | congruent | session 1 | 639 | (617, 660) | $\sigma_{\tau_1}$: 71 | $\sigma_0$: 300 | - | 0.78 |
| | | session 2 | 607 | (584, 629) | $\sigma_{\tau_2}$: 74 | | | |
| TLM | incongruent | session 1 | 720 | (693, 746) | $\sigma_{\tau_1}$: 89 | $\sigma_0$: 250 | - | 0.73 |
| | | session 2 | 666 | (643, 689) | $\sigma_{\tau_2}$: 77 | | | |
| LME (12) | incongruent vs congruent | session 1 | 81 | (71, 92) | $\sigma_{\tau_1}$: 27 | $\sigma_0$: 276 | - | 1.0 |
| | | session 2 | 59 | (50, 68) | $\sigma_{\tau_2}$: 18 | | | |
| TLM | average of 2 conditions | session 1 | 679 | (656, 703) | $\sigma_{\tau_1}$: 79 | | - | 0.74 |
| | | session 2 | 636 | (614, 659) | $\sigma_{\tau_2}$: 75 | | | |
| BML (17) | incongruent vs congruent | session 1 | 36 | (28, 42) | $\sigma_{\tau_1}$: 7 | $\sigma_0$: 75 | - | 0.71 |
| | | session 2 | 27 | (22, 33) | $\sigma_{\tau_2}$: 7 | | | |
| exGaussian | average of 2 conditions | session 1 | 670 | (656, 684) | $\sigma_{\tau_1}$: 34 | | - | 0.68 |
| | | session 2 | 643 | (631, 659) | $\sigma_{\tau_2}$: 28 | | | |

Table 1: Test-retest reliability (TRR) estimated through condition- and trial-level modeling for the Study 1 data of Stroop task from Hedge et al. (2018). Reaction time (in milliseconds) was recorded from 47 subjects who completed two sessions of the Stroop task. With 240 trials per condition, RT ranged from 1 to 30830 ms; all data were used here without censoring. The variability ratio $R_v$ (defined in formulas (8) and (14)) and underestimation rate $R_u$ (defined in formulas (7) and (13)) are indicators for the degree of ICC underestimation.

It is important to recognize the dissociation between population effects and test-retest reliability. From the modeling perspective, the relationship between the population and subject level can loosely described as "crossed" or "orthogonal": the subject-specific effects $\tau_s$ in the LME model (5) are "perpendicular to" the population-level effects of overall average $a_r$ per repetition in the sense that the former are fluctuations relative to the latter. Even though cross-subject variability is the focus in the context of test-retest reliability, a popular misconception is that strong population-level effects are necessarily associated with high reliability as long as sample sizes are large. Fröhner et al. (2019) demonstrated the dissociation between the strength of population-level effects and the extent to which individual differences were reliable. Fig. 1 illustrates four extreme scenarios on a two-dimensional continuous space of population-level effects and reliability: one may have strong population-level effects with high (Fig. 1A) or low (Fig. 1B) reliability; contrarily, it is also possible to have weak population-level effects accompanied by high (Fig. 1C) or low (Fig. 1D) reliability.

# 2 Methods: assessing test-retest reliability through trial-level modeling

## 2.1 Motivation for adopting a new framework

There are several limitations of the ICC framework that we aim to address within a hierarchical model. First, it can be difficult to obtain a measure of ICC uncertainty. An approximation for the variance of ICC estimate is available (Shoukri et al., 2004) that could provide a confidence interval, but it requires a large subject sample size because of the reliance on the asymptotic property. Even though bootstrapping could be adopted to find a quantile interval, the approach is rarely utilized in practice due to its computational cost especially for large datasets in neuroimaging. Second, the LME framework assumes a Gaussian distribution and, when this assumption is violated (e.g., skewed data, outliers), parameter estimation through the optimization of a nonlinear objective function can become unstable or singular. Third, it is not possible to

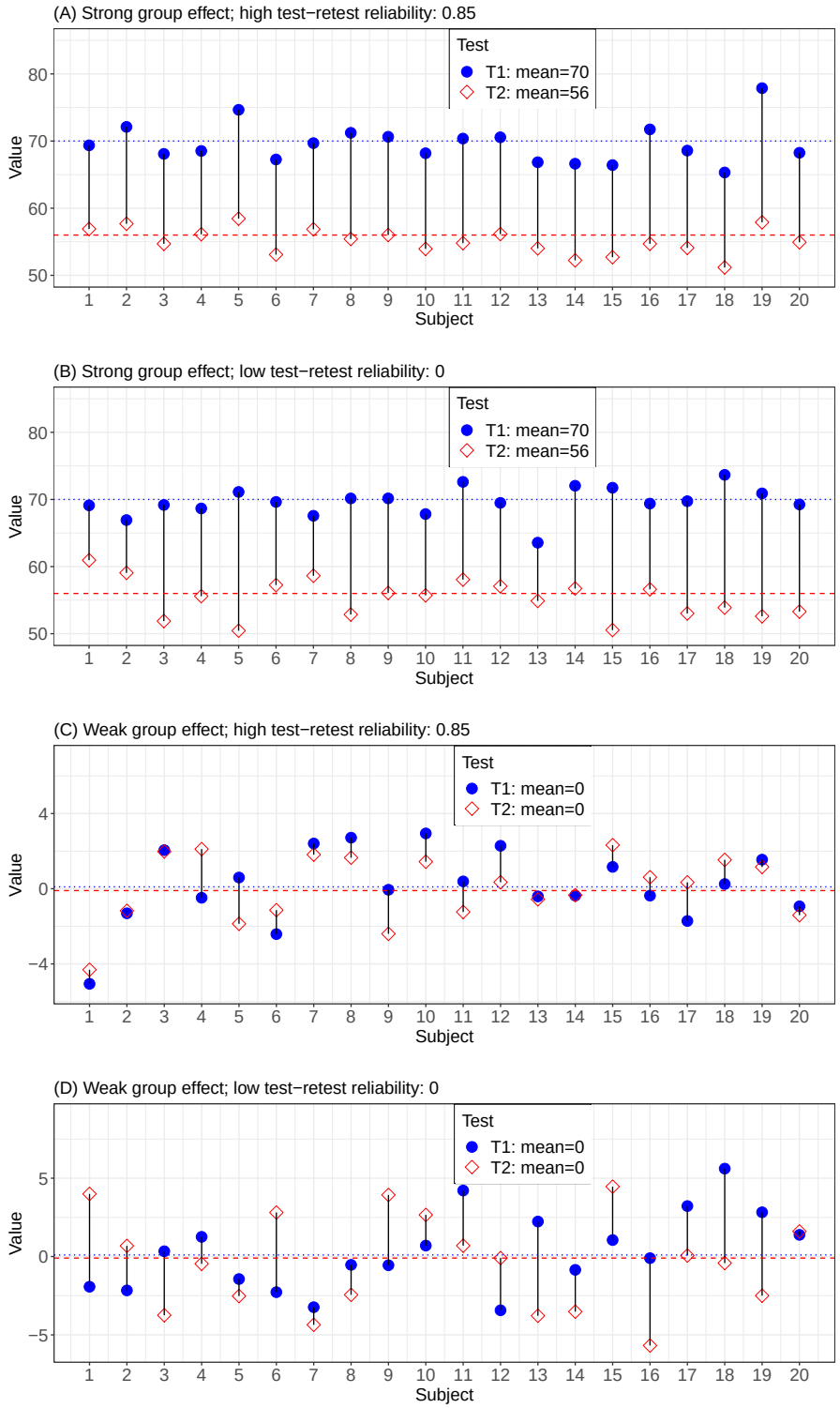



Figure 1: Separation between test-retest reliability and population-level effects. Four scenarios are illustrated to demonstrate that the reliability is not necessarily tied to the strength of population effects. Hypothetical data were randomly drawn from a bivariate Gaussian distribution: 20 subjects completed a depression screening at two separate tests (test T1: blue filled circle; test T2: red empty diamond).The population effects are easy to observe (colored horizontal lines). However, it is harder to assess the reliability over the two sessions, which can be observed by assessing the proportion of subjects for which the two testing scores are on the same side (above or below) of their respective population average. With strong population effects from a severely depressed group, reliability can be high (A) or low (B); on the other hand, weak population effects from a control group may correspond to high (C) or low (D) reliability. $\mathcal{N}(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, 49\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix})$ for each of the 20 subjects with (A) $\mu_1 = 70$, $\mu_2 = 56$, $\rho = 0.85$, (B) $\mu_1 = 70$, $\mu_2 = 56$, $\rho = 0$, (C) $\mu_1 = \mu_2 = 0$, $\rho = 0.85$, and (D) $\mu_1 = \mu_2 = 0$, $\rho = 0$.

integrate measurement error within the conventional platform. Under specific circumstances, the input data may contain sampling errors. For instance, BOLD response as the effect of interest in neuroimaging is not

directly collected, but instead estimated from a time series regression model. Including measurement error can calibrate and improve the model fit. For example, instead of censoring at an artificial threshold, outliers can be more effectively handled through down-weighting based on their large uncertainty (Chen et al., 2020). This modeling strategy has been widely adopted to achieve higher efficiency and robustness in traditional meta-analysis (Viechtbauer, 2005) as well as in neuroimaging population-level analysis (Worsley et al., 2002; Woolrich et al., 2004; Chen et al., 2012). Although these approaches exist, the most common pipelines still use subject-level effect estimates for the population-level model without considering the associated uncertainty information and condition-level effects are estimated through one regressor per condition, ignoring cross-trial variability. Lastly, we are specifically interested in accounting for cross-trial variability in our modeling. Trials are related to condition-level effects the same way participants are related to population-level effects. While trials are expected to provide robust estimation for condition-level effects, at the same time, variability associated with trials are largely of no interest and in practice, the data are typically collapsed across trials and further flattened across conditions as illustrated in the data reduction formula (1). As a result, cross-trial variability does not have a place in the model and the associated uncertainty is neither accounted for nor propagated in the conventional formula (2) in the ICC computation.

## 2.2 LME framework for a single condition

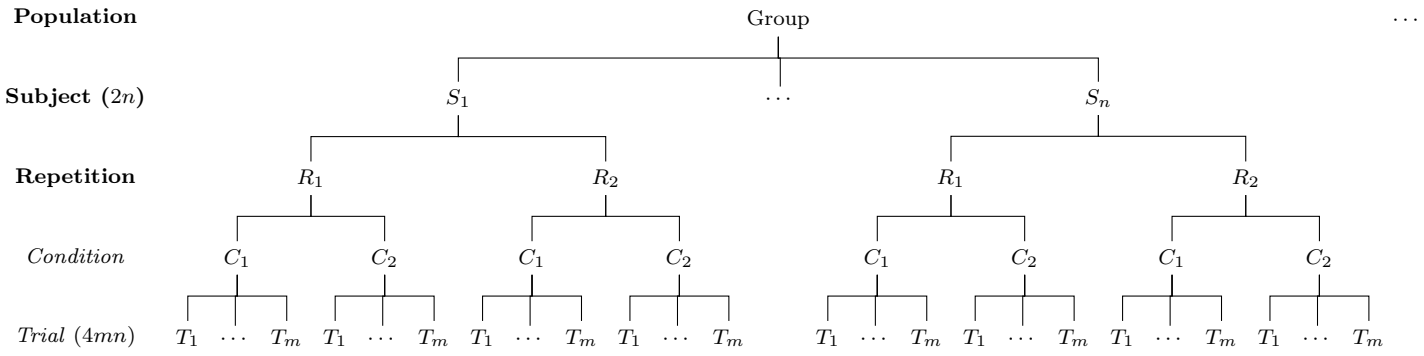



Figure 2: Hierarchical structure of test-retest data. Assume that, in a study with two repetitions, a group of $n$ subjects are recruited to perform a task (e.g., Stroop) of two conditions (e.g., congruent and incongruent), and each condition is instantiated with $m$ trials. The collected data are structured across a hierarchical layout of 5 levels (population, subject, repetition, condition and trial) with total $n \times 2 \times 2 \times m = 4mn$ data points at the trial level compared to $2n$ across-condition contrasts at the subject level.

We start with simply accommodating trial-level effects in reliability estimation for a single condition in the LME framework. There are five levels involved in the hierarchical structure of a reliability dataset (Fig. 2): population, subject, repetition, condition and trial. As opposed to the common practice of collapsing the two lower levels (condition and trial), we directly utilize the trial-level effect estimates $y_{rst}$ of the condition (Chen et al., 2020), where $r$, $s$ and $t$ index repetitions, subjects and trials ($r = 1, 2;\ s = 1, 2, ..., n;\ t = 1, 2, .., m$). Specifically, we expand the LME model (2) and accommodate the trial-level effects $y_{rst}$ as below,

$$\begin{aligned}
&y_{rst}|a_r, \tau_{rs}, \sigma_0 \sim \mathcal{N}(a_r + \tau_{rs},\ \sigma_0^2);\\
&(\tau_{1s},\ \tau_{2s})^T|\rho, \sigma_{\tau_1}, \sigma_{\tau_2} \sim \mathcal{N}(\mathbf{0}_{2\times 1},\ \boldsymbol{R}_{2\times 2});\\
&\boldsymbol{R} = \begin{bmatrix} \sigma_{\tau_1}^2 & \rho\sigma_{\tau_1}\sigma_{\tau_2} \\ \rho\sigma_{\tau_1}\sigma_{\tau_2} & \sigma_{\tau_2}^2 \end{bmatrix};\\
&s = 1, 2, ...n;\ r = 1, 2;\ t = 1, 2, .., m;
\end{aligned} \tag{5}$$

where $a_r$, as in (2), represents the population-level effect during the $r$-th repetition, $\tau_{rs}$ characterizes the subject-level effects during the $r$-th repetition, $\sigma_0$ captures the cross-trial variability, and $\boldsymbol{R}$ is the variance-

covariance matrix for the subject-level effects $\tau_{rs}$ between the two repetitions. Usually $a_r$ are termed as the population-level (or "fixed") intercepts while $\tau_{rs}$ are the varying (or "random") intercepts across subjects.

The parameter $\rho$ captures the correlation between the two repetitions for the subject-level effects $\tau_{1s}$ and $\tau_{2s}$; thus, $\rho$ represents the test-retest reliability of individual differences while the trial-level variability is explicitly characterized through $\sigma_0$ in the model (5). The first line in the formulation (5) specifies the effect decomposition while the second line specifies the distributional assumptions of the terms. All the parameters including the reliability measure $\rho$, cross-subject varibilities $\sigma_{\tau_1}$ and $\sigma_{\tau_2}$, and cross-trial variability $\sigma_0$ are numerically estimated through restricted maximum likelihood (Pinheiro and Bates, 2000). The inclusion of cross-trial variability $\sigma_0$ into the LME model (5) precludes formulating reliability as a variance ratio as traditionally done in the formula (3). The reliability estimates for the Stroop task dataset based on the LME model (5) are shown in Table 1.

The relationship between ICC and the reliability metric $\rho$ under the LME model (5) can be directly characterized. As is evident from Table 1, the conventional ICC formulation tends to yield lower reliability estimates of a single condition compared to the trial-level computations. For comparison, we temporarily assume homoscedasticity between the two repetitions: $\sigma_{\tau_1} = \sigma_{\tau_2} = \sigma_\tau$. The specific amount of underestimation can be revealingly expressed as a linear relationship (Appendix A),

$$\text{ICC(3,1)} = R_u \rho, \tag{6}$$

where the underestimation rate

$$R_u = \frac{1}{1 + \frac{1}{m} R_v^2} \tag{7}$$

is dependent on the trial sample size $m$ and the variability ratio $R_v$ (the magnitude of cross-trial variability relative to cross-subject variability)

$$R_v = \frac{\sigma_0}{\sigma_\tau}. \tag{8}$$

Two aspects of the ICC underestimation are noteworthy First, the trial sample size $m$ plays a crucial role; specifically, the degree of ICC underestimation decreases as the trial sample size increases. In contrast, the subject sample size $n$ on average does not impact the ICC underestimation, which might be surprising. Second, the underestimation also depends on the variability ratio $R_v$. If cross-trial variability is roughly the same or smaller than its cross-subject counterpart (i.e., $\sigma_0 \lessapprox \sigma_\tau$) with a reasonable trial sample size $m$ or if $\sqrt{m} \gg R_v$, the underestimation is negligible: $R_u \approx 1 - \frac{1}{m} R_v^2 \approx 1$.

A correction could be applied to ICC estimates post-hoc to adjust for trial-level variance if the variability ratio $R_v$ were known under the conventional framework (2). Specifically, the ICC formulation (3) could be adjusted to (Appendix A)

$$\text{ICCa} = \frac{\widetilde{\sigma}_\tau^2}{\widetilde{\sigma}_\tau^2 + \sigma_e^2 - \frac{1}{m}\sigma_0^2}, \tag{9}$$

where $\widetilde{\sigma}_\tau$ and $\sigma_e$ are the subject-level and residual variability under the conventional ICC formulation (2) while $\sigma_0$ represents the cross-trial variability under the TLM-based LME formulation (5). As test-retest reliability is supposed to capture the correlation between the two condition-level effects ($\tau_{1s}$ and $\tau_{2s}$ in the model (5)) without the interference of trial-level fluctuations, it is desirable to explicitly account for the trial-level variability and to exclude it from the formulation. In the same vein, the adjustment formula (8) accomplishes the same goal of removing the contamination of cross-trial variability through the term $\frac{1}{m}\sigma_0^2$. However, the adjusted ICC (9) is of little practical use because to estimate $\sigma_0$, one would have to resort to the

TLM-based LME formulation (5); under that circumstance, one might as well directly obtain the reliability through TLM rather than going back to CLM to make the adjustment.

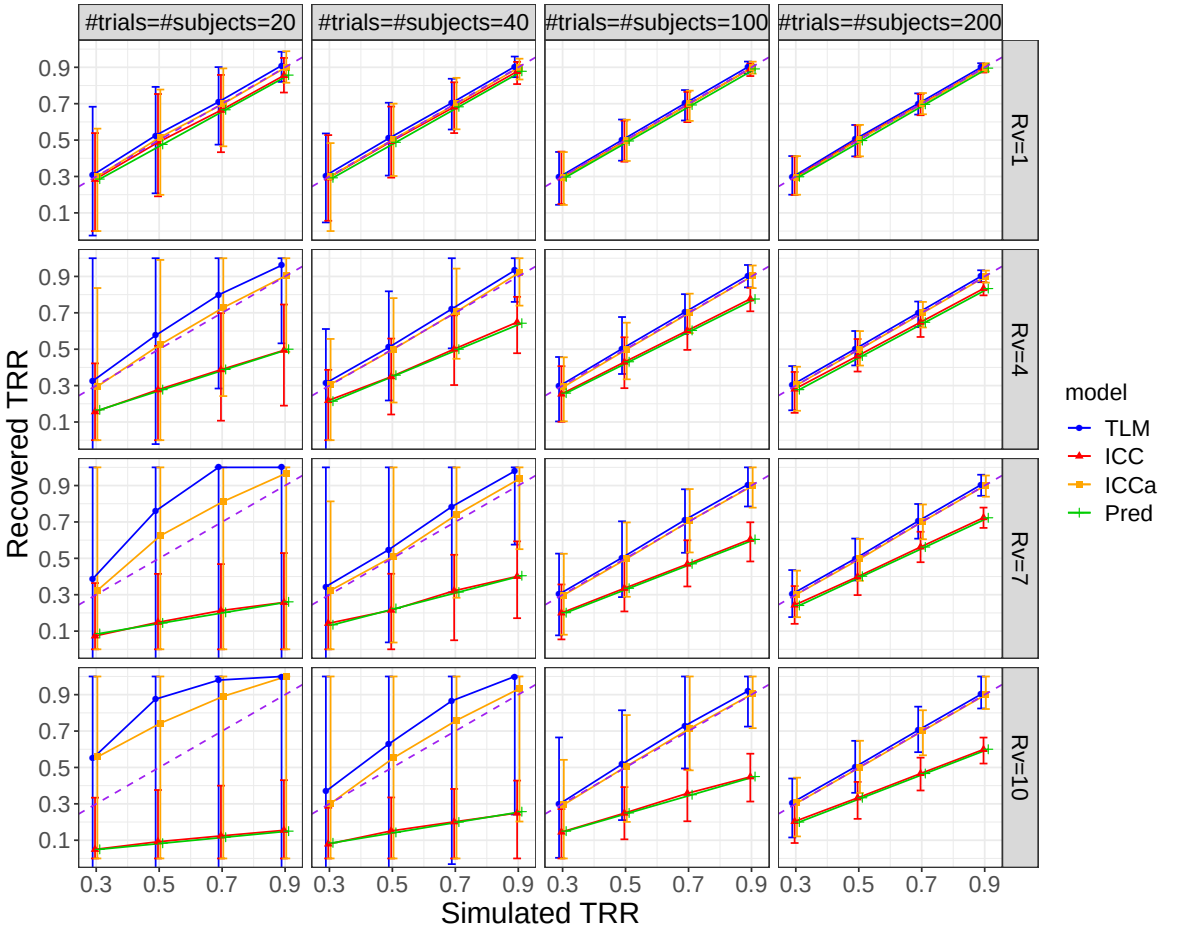
The ICC underestimation can be illustrated on the experimental data from the Stroop task (Hedge et al., 2018). The ICC(3,1) formulation resulted in values of 0.72 and 0.69 for the congruent and incongruent condition, respectively (Table 1). With the TLM-based LME model (5), we obtain a reliability estimate of 0.78 and 0.73 for each condition. In contrast, we could, per the formulation (8), obtain adjusted reliability estimates of 0.78 and 0.72 for the two conditions which are largely consistent with the results from the TLM-based LME. The mild amount of ICC underestimation with $R_u = 0.93$ or 0.96 was due to a large trial sample size ($m = 240$) and moderate variability ratios of $R_v = 4.1$ and 3.0.

### 2.3 Model comparisons through simulations

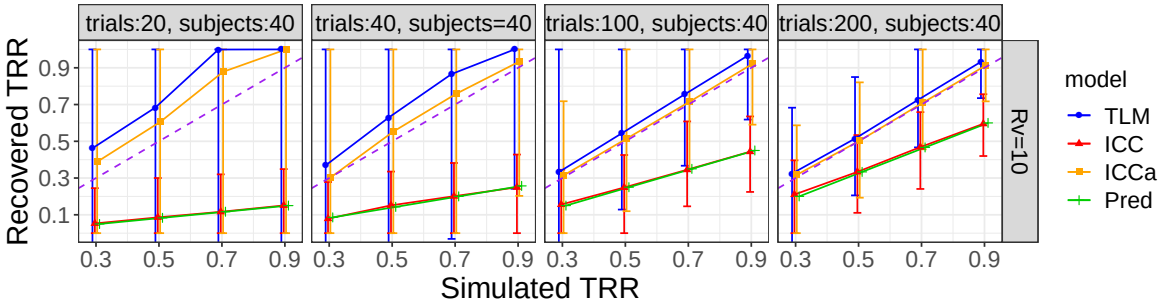
To comprehensively compare the two modeling frameworks for a single condition, we adopt simulations. The numerical schemes for the simulations are presented in Appendix B with results shown in Fig. 3. The population-level effects $a_r$ under both CLM and TLM were robustly recovered (not shown). Findings and implications can be summarized as follows:

1) **Systematical ICC underestimation is confirmed.** Simulations also confirmed the dependency of the underestimation on trial sample size $m$ (Fig. 3B) and variability ratio $R_v$. When cross-trial variability is much larger than cross-subject variability ($R_v \gg 1$) the underestimation is largest. When $R_v \lessapprox 1$ (top row, Fig. 3A), the simulated reliability values were successfully recovered for both formulations. When $R_v > 1$ (second to fourth row, Fig. 3A), the larger the ratio $R_v$, the larger the ICC underestimation; as predicted, a linear relationship at the attenuation rate $R_u$ in (6) (green lines, Fig. 3A,B,C) emerged between ICC (red triangles) and the underlying reliability.
2) **ICC underestimation could be adjusted.** Once the cross-trial variability component $\frac{\sigma_0^2}{m}$ is explicitly accounted for in the ICC formulation as shown in (9), the adjusted ICCa performs very well; the adjustment is quite successful when the sample size for both subjects and trials is 40 or above (second to fourth row, Fig. 3A).
3) **Subject sample size, on average, has no impact on ICC underestimation**. As shown in Fig. 3C, the degree of ICC underestimation (green line) does not depend on the number of subjects $n$. However, as the subject number $n$ increases, the precision of both reliability estimates improve.
4) **Achieving high precision of reliability estimates can be difficult.** The uncertainty associated with reliability estimates is monotonically associated with the reliability magnitude, variability ratio $R_v$, and the sample sizes $m$ and $n$ for subjects and trials. Specifically, uncertainty reduces as (a) $m$ or $n$ increases, (b) reliability increases, or (c) $R_v$ becomes smaller. While both sample sizes influence reliability precision, with increases in $m$ or $n$ (from first to fourth column, Fig. 3) resulting in narrower error bars, the trial sample size $m$ plays a more substantial role than the subject sample size $n$. Moreover, large sample sizes for both subjects and trials (e.g., 100 or more when $R_v \geq 10$) may be required to achieve a reliability estimate with a reasonable precision, but the impact of the trial sample size $m$ is slightly more impactful than the subject sample size $n$ (Fig. 3B vs. Fig. 3C).
5) **Reliability is not associated with population-level effects.** The two sets of population-level parameters, $(a_1,\ a_2) = (0,\ 0)$ and $(1.0,\ 0.9)$, describe a scenario with no population effects and one with strong population effects. For both scenarios, simulation parameters were successfully recovered from the TLM-based LME formulation (5) and the two scenarios rendered very similar reliability patterns (note, only $(a_1,\ a_2) = (0,\ 0)$ is shown in Fig. 3).
6) **LME modeling is susceptible to numerical instability.** The performance of the TLM-based LME formulation is acceptable but suffers from numerical failures with a slight overestimation of reliability when cross-trial variability is much larger than cross-subject variability ($R_v \gg 1$) (second to fourth row,

(A) Equal sample size for trials and subjects: $m = n$




(B) Fixed sample size for subjects: $n = 40$




(C) Fixed sample size for trials: $m = 40$

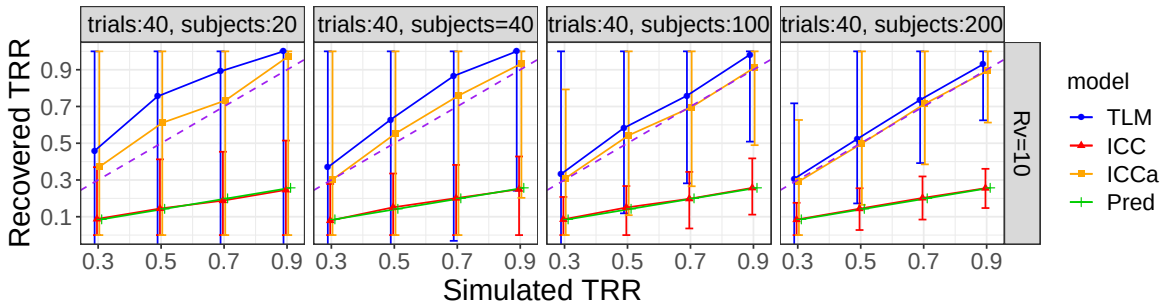



Figure 3: Simulation results for a single condition. The four columns correspond to the sample sizes of subjects and trials while the four rows code the variability ratio $R_v = \frac{\sigma_0}{\sigma_\tau}$. The $x$- and $y$-axis are the simulated and recovered reliability, respectively. Each data point is the median among the 1000 iterative simulations with the error bar showing the 90% highest density interval. The dashed purple diagonal line indicates an idealized retrieval of the parameter value based on the data-generating model (5). The green line shows the theoretically predicted ICC per the formula (6).

Fig. 3A). In fact, the larger the ratio $R_v$, the more severe the overestimation. Overestimation via the

LME model occurred when the numerical optimizer was trapped at the boundary of $\rho = 1$ with the convergence failure leading to a reliability estimate of 1.0. When $R_v > 10$, numerical instability worsens substantially, consistent with the result from the Stroop task data in Table 1.

## 2.4 LME framework for a contrast between two conditions

Now we extend the LME formulation to the more common scenario of a contrast between two conditions. Suppose that we obtain the trial-level effects $y_{crst}$ of two conditions, where the four indices $c$, $p$, $s$ and $t$ code conditions, subjects, repetitions and trials. The LME formulation would read:

$$\begin{aligned} &y_{crst}|\mu_{crs}, \sigma_0 \sim \mathcal{N}(\mu_{crs},\ \sigma_0^2); \\ &c = 1, 2;\ r = 1, 2;\ s = 1, 2, .., n;\ t = 1, 2, .., m; \end{aligned} \tag{10}$$

where $\mu_{crs}$ is the $s$-th subject's effect under the $c$-th condition during the $r$-th repetition, and $\sigma_0$ captures the within-subject, within-repetition cross-trial variability. When the contrast is of interest, we need to parameterize the subject-level effects $\mu_{crs}$. Different methods exist for factor parameterization; we opt to quantify the two conditions through the following indicator,

$$I_c = \begin{cases} \frac{1}{2}, & \text{if } c = 1; \\ -\frac{1}{2}, & \text{if } c = 2. \end{cases} \tag{11}$$

The subject-level effects $\mu_{crs}$ are further integrated into the LME formulation (10) as below,

$$\begin{aligned} &\mu_{crs} = a_r + b_r I_c + \tau_{rs} + \lambda_{rs} I_c; \\ &(\tau_{1s},\ \tau_{2s})^T \sim \mathcal{N}(\mathbf{0}_{2\times 1},\ \boldsymbol{R}^{(0)}_{2\times 2});\ (\lambda_{1s},\ \lambda_{2s})^T \sim \mathcal{N}(\mathbf{0}_{2\times 1},\ \boldsymbol{R}^{(1)}_{2\times 2}); \\ &\boldsymbol{R}^{(0)} = \begin{bmatrix} \sigma^2_{\tau_1} & \rho_0\sigma_{\tau_1}\sigma_{\tau_2} \\ \rho_0\sigma_{\tau_1}\sigma_{\tau_2} & \sigma^2_{\tau_2} \end{bmatrix};\ \boldsymbol{R}^{(1)} = \begin{bmatrix} \sigma^2_{\lambda_1} & \rho_1\sigma_{\lambda_1}\sigma_{\lambda_2} \\ \rho_1\sigma_{\lambda_1}\sigma_{\lambda_2} & \sigma^2_{\lambda_2} \end{bmatrix}; \\ &c = 1, 2;\ s = 1, 2, ...n;\ r = 1, 2. \end{aligned} \tag{12}$$

The variance-covariance matrices $\boldsymbol{R}^{(0)}$ and $\boldsymbol{R}^{(1)}$ characterize the relatedness across subjects between the two repetitions for the two parameter sets of $(\tau_{1s},\ \tau_{2s})^T$ and $(\lambda_{1s},\ \lambda_{2s})^T$, respectively. The assumption of independence between two parameter sets is discussed in Appendix C.

The correlation $\rho_1$ embedded in the matrix $\boldsymbol{R}^{(1)}$ is the reliability for the contrast between the two conditions. With the two conditions coded through the indicator variable $I_c$ in (11), the contrast effects correspond to the slope terms (i.e., $b_r$ and $\lambda_{rs}$) under the LME framework (12). Specifically, $b_r$ is the population-level contrast while $\lambda_{rs}$ codes the subject-specific contrast effects. In parallel to the single condition in (5) with the intercept interpretation, here $a_r$ and $\tau_{rs}$ are also intercepts: the former is the population-level average between the two conditions during the $r$-th repetition while the latter indicates the condition average for the $s$-th subject during the $r$-th repetition.[1] Just as the matrix $\boldsymbol{R}$ in (5) for a single condition, the matrix $\boldsymbol{R}^{(1)}$ characterizes the relationships of subject-level contrasts $\lambda_{rs}$ between the two repetitions; thus, reliability can be similarly retrieved from the LME model (12) as the correlation $\rho_1$ between the two varying slopes $\lambda_{1s}$ and $\lambda_{2s}$. For neueroimaging data analysis, whole-brain voxel-wise reliability for a contrast under the TLM-based LME framework (12) can be performed using the program **3dLMEr** (Chen et al., 2013) in AFNI.

Should one account for the correlation structure for the likelihood distribution in the LME model (12)? In other words, instead of a single standard deviation (dispersion or scaling) parameter $\sigma_0$ in the model (12), one

[1] Instead of using the indicator variable $I_c$ in (11) to represent the two conditions, one may adopt dummy coding (as in Haines et al. (2020)): one condition is coded as 1 while the other as 0 (reference condition). When dummy coding, the slopes in the LME model (12) still correspond to the contrast; however, the intercepts are associated with the reference condition.

may specify a $2\times2$ variance-covariance matrix between the two sessions as in Haines et al. (2020) with another parameter added to capture the reliability in the residuals between the two sessions. One could further argue that another correlation might be added to account for the relationship between the two conditions in the likelihood distribution, resulting in $4\times4$ block-diagonal variance-covariance matrix. However, such a variance-covariance structure likely means that effects in the hierarchical data structure are not well accounted for, this should not be necessary if the model is reasonably specified. Rather, it might be worth tuning and comparing alternative models to improve model fit. An extra bonus of the LME framework (12) is that the reliability for the average between the two conditions is also readily available. That is, the correlation $\rho_0$ embedded in the matrix $\boldsymbol{R}^{(0)}$ for cross-subject varying intercepts $\tau_{rs}$ captures the reliability for the average effect.

How does cross-trial variability impact the ICC computation for a condition-level contrast? With the homoscedasticity assumption $\sigma_{\lambda_1} = \sigma_{\lambda_2} = \sigma_\lambda$, the extent of underestimation via the conventional ICC for a contrast is updated from the single effect case (7) to (Appendix D)

$$R_u = \frac{1}{1+\frac{2}{m}R_v^2} \tag{13}$$

with the variability ratio $R_v$ similarly defined as before,

$$R_v = \frac{\sigma_0}{\sigma_\lambda}. \tag{14}$$

Likewise, one could adjust the original ICC by removing the impact of the cross-trial variance (Appendix D),

$$\text{ICCa} = \frac{\widetilde{\sigma}_\lambda^2}{\widetilde{\sigma}_\lambda^2 + \sigma_e^2 - \frac{2}{m}\sigma_0^2}. \tag{15}$$

The number 2 occurs, unlike in (7) and (9) for the individual conditions, as there is twice the amount of data involved in the cross-trial variability for a contrast.

The reliability estimation for the condition-level contrast of the experimental Stroop data can be compared across different frameworks (Table 1). The ICC(3,1) value for the contrast was estimated as 0.49. For the corresponding TLM-based LME, a reliability estimate of 1.0 was retrieved due to a singularity problem with the parameter $\rho_1$ numerically degenerating at the boundary. With $\sigma_0 = 0.276$, we adjusted the reliability to be 1.18 per (15). This uninterpretable value originated as adjusted cross-trial variability $\frac{2}{m}\sigma_0^2 = 6.35\times10^{-4}$ was larger than $\sigma_e^2 = 5.53\times10^{-4}$, another incidence of a numerical singularity issue when estimating reliability through the TLM-based LME in (12). This may indicate that assumptions were violated. Nevertheless, the closeness to the parameter boundary is an indication of a reliability estimate close to 1, which suggests a substantial underestimation from the ICC formulation.

Simulations were also adopted to explore the various aspects of model performance for the condition-level contrast (Appendix E). The result patterns (Fig. 9) were largely similar to those for a single condition effect (Fig. 3). However, the degree to which the ICC falls below the reliability estimated via the TLM-based LME is more severe than for either of the two conditions or for the average effect between the two conditions. The magnitude of a contrast effect is much smaller than that of a condition or the average between conditions (see the mean column in Table 1). Thus, a contrast is associated with a larger variability ratio $R_v$, which significantly attenuates the ICC estimate relative to the TLM-based LME estimate. As shown in Table 1, the ratio $R_v$ for the contrast is about three times larger than that for single conditions.

## 2.5 Extension of the LME framework to BML

The extension from LME to BML is relatively straightforward. As apparent from numerical simulations and from applications to the Stroop data, the LME framework is ill-suited for reliability estimation: information on

the precision and distribution of reliability estimates is not available; distribution assumption (e.g., Gaussian) can be violated; sampling errors cannot be accommodated; and numerical failures occur. We now integrate the LME formulations (5) and (12) into the BML framework. Specifically, we convert the two LME models with little modification for a single condition,

$$\begin{aligned}
&y_{rst}|\mu_{rs}, \sigma_0 \sim \mathcal{N}(\mu_{rs},\ \sigma_0^2);\\
&\mu_{rs} = a_r + \tau_{rs};\\
&(\tau_{1s},\ \tau_{2s})^T \sim \mathcal{N}(\mathbf{0}_{2\times 1},\ \boldsymbol{R}_{2\times 2})\\
&\boldsymbol{R} = \begin{bmatrix} \sigma_{\tau_1}^2 & \rho\sigma_{\tau_1}\sigma_{\tau_2} \\ \rho\sigma_{\tau_1}\sigma_{\tau_2} & \sigma_{\tau_2}^2 \end{bmatrix};\\
&r = 1, 2;\ s = 1, 2, ...n;\ t = 1, 2, .., m;
\end{aligned} \tag{16}$$

and a contrast between two conditions,

$$\begin{aligned}
&y_{crst}|\mu_{crs}, \sigma_0 \sim \mathcal{N}(\mu_{crs},\ \sigma_0^2);\\
&\mu_{crs} = a_r + b_r I_c + \tau_{rs} + \lambda_{rs} I_c;\\
&(\tau_{1s},\ \tau_{2s})^T \sim \mathcal{N}(\mathbf{0}_{2\times 1},\ \boldsymbol{R}^{(0)}_{2\times 2});\ (\lambda_{1s},\ \lambda_{2s})^T \sim \mathcal{N}(\mathbf{0}_{2\times 1},\ \boldsymbol{R}^{(1)}_{2\times 2});\\
&\boldsymbol{R}^{(0)} = \begin{bmatrix} \sigma_{\tau_1}^2 & \rho_0\sigma_{\tau_1}\sigma_{\tau_2} \\ \rho_0\sigma_{\tau_1}\sigma_{\tau_2} & \sigma_{\tau_2}^2 \end{bmatrix};\ \boldsymbol{R}^{(1)} = \begin{bmatrix} \sigma_{\lambda_1}^2 & \rho_1\sigma_{\lambda_1}\sigma_{\lambda_2} \\ \rho_1\sigma_{\lambda_1}\sigma_{\lambda_2} & \sigma_{\lambda_2}^2 \end{bmatrix};\\
&c = 1, 2;\ s = 1, 2, ...n;\ r = 1, 2.
\end{aligned} \tag{17}$$

These BML models can be further extended to resolve the issues associated with the LME framework. For example, numerical instabilities and convergence issues would be largely avoided via modeling improvements including accommodating the data through various distributions such as Student's $t$, exponentially modified Gaussian (exGaussian), log-normal, etc. More importantly, instead of providing point estimates without uncertainty information, the reliability estimation for $\rho$ in (16), $\rho_0$ and $\rho_1$ in (17) can be expressed by its whole posterior distribution. Furthermore, sampling errors can be readily incorporated into the BML framework. For neuroimaging data, trial-level effects are usually not directly measured, but instead estimated through subject-level time series regression. Thus, it is desirable to include the standard errors of the trial-level effect estimates into the reliability formulation. With the hat notation for effect estimate $\widehat{y}$ and its standard error $\widehat{\sigma}$, we broaden the two BML models (16) and (17), respectively, to,

$$\widehat{y}_{rst}|\mu_{rs}, \widehat{\sigma}_{rst}, \sigma_0 \sim \mathcal{N}(\mu_{rs},\ \widehat{\sigma}_{rst}^2 + \sigma_0^2), \tag{18}$$

$$\widehat{y}_{crst}|\mu_{rs}, \widehat{\sigma}_{crst}, \sigma_0 \sim \mathcal{N}(\mu_{rs},\ \widehat{\sigma}_{crst}^2 + \sigma_0^2). \tag{19}$$

The modeling capability and advantages of the Bayesian framework can be illustrated through model comparisons using the Stroop data. The LME framework heavily relies on the assumption of a Gaussian distribution as a prior and its properties such as convenient inferences through standard statistics such as Student's $t$. However, this convenience comes at a cost: when the Gaussian prior is ill-suited, estimates can be inaccurate. Extending on the work by Haines et al. (2020), who applied three likelihood distributions including Gaussian, log-normal and shifted log-normal to the data, we additionally considered Student's $t$ and exGaussian, because of their ability to handle skewed data and outliers. Moreover, instead of assuming a diagonal matrix $\boldsymbol{R}^{(0)}$ for the variance-covariance structure of the varying intercepts $\tau_{rs}$ as in Haines (2020), we adopted a generic $\boldsymbol{R}^{(0)}$ in the BML model (17). In the Stroop dataset, the reliability estimation based on the Gaussian prior had the worst fit among the five priors (Fig. 4A-E), as also reported by Haines et al. (2020). As exGaussian can accommodate skewed and outlying data such as reaction time which is lower-bounded,

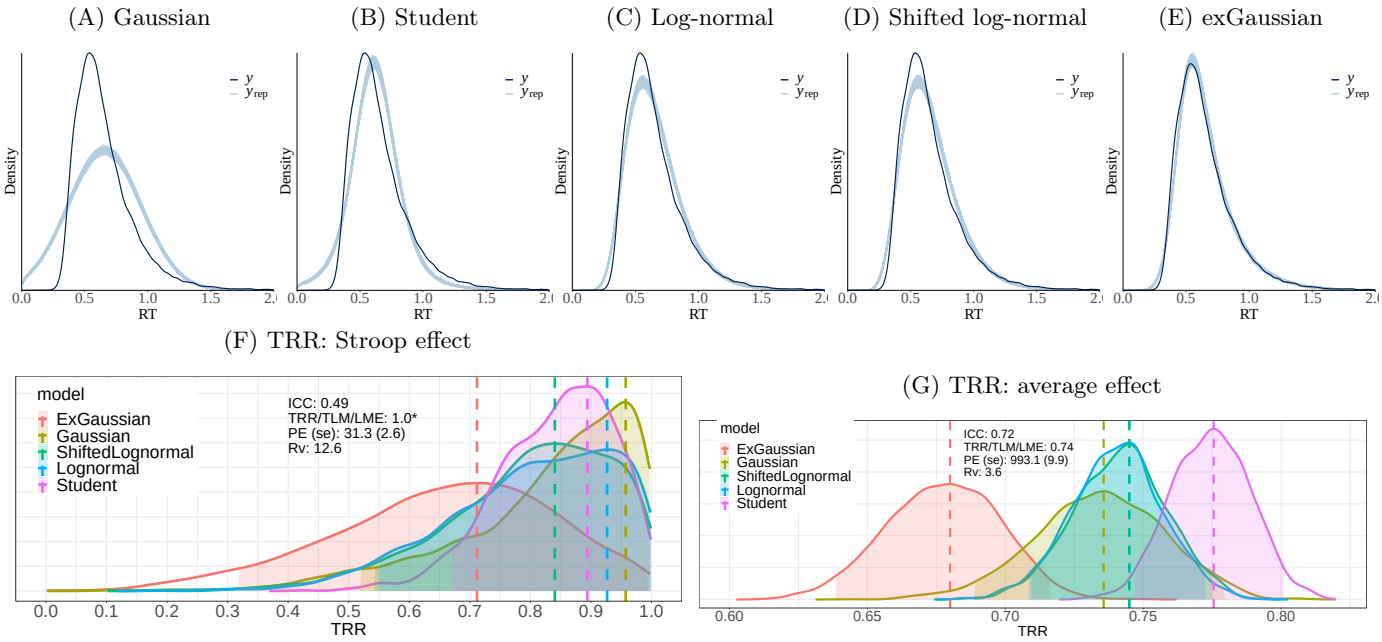



Figure 4: Model comparisons among five potential RT distributions for a Stroop dataset (Hedge et al., 2018). (A-E) Visual comparisons are illustrated with each of the five panels showing the posterior predictive density (light blue line) that is composed of 200 sub-curves each of which corresponds to one draw from the posterior distribution. Overlaid is the raw data (solid black curves with linear interpolation). Overlaps between the solid black curve and the light blue line indicate how well the respective model fits the raw data. (F) Reliability distributions for the Stroop effect are shown for the five distributions. The reliability distributions were based on 2000 draws from MCMC simulations for each BML model. The dashed vertical line indicates the mode (peak) of each reliability distribution. Gaussian likelihood rendered the highest reliability with a mode of 0.96 but also exhibited the poorest fit to the data. The exGaussian distribution provided the lowest reliability with a mode of 0.71 while achieving the best fit among the five distributions. (G) Reliability distributions for the average between congruent and incongruent conditions are shown for the five distributions. Unsurprisingly, their uncertainty is much narrower than the contrast (F).

it outperformed all alternatives per the information criterion through leave-one-out cross-validation. This is important, considering the amount of effort devoted to data cleaning in common practice.

The rich information that can be derived from Bayesian modeling is worth noting. Unlike a simple point estimate under the conventional framework, we empirically construct the posterior distribution for a parameter through Monte Carlo simulations. The singularity problem (Table 1) that we encountered with the LME model (12) is not an issue in the BML model (17). With a mode of 0.71 and a 95% highest density interval of [0.32, 0.99] for reliability estimate, it is clear that the ICC = 0.49 is substantially lower than the estimate derived through the BML.

We provide the program **TRR** for reliability estimation under BML. The two BML models, one for a single condition through (16) (or (18)) and the other for a contrast through (17) (or (19)), are implemented in the program **TRR** through Markov Chain Monte Carlo (MCMC) simulations using Stan (Carpenter et al., 2017) and the R package brms (Bürkner, 2017). Each Bayesian model is specified with a likelihood function, followed by priors for lower-level effects (e.g., trial, subject). The hyperpriors employed for model parameters (e.g., population-level effects, variances in prior distributions) are detailed in Appendix F. The program **TRR**[2] is publicly available as part of the AFNI suite and can be used to estimate reliability for behavior and region-level neuroimaging data. Runtime ranges from minutes to hours depending on the amount of data.

[2]https://rb.gy/uakhmj

# 3 BML modeling of reliability applied to a neuroimaging dataset

## 3.1 Data description

**Modified Eriksen Flanker Task** Analysis in the current report used a subset of the subjects in Smith et al. (2020): 24 adults (>18 years; age: $26.81 \pm 6.36$) and 18 youth (<18 years; age: $14.01 \pm 2.48$). Subjects performed a modified Eriksen Flanker task (Eriksen and Eriksen, 1974) with 432 experimental trials during fMRI scanning in each of two separate sessions $53.5 \pm 11.8$ days apart. Participants were asked to identify, via button press, the direction of a central arrow, flanked by two arrows on either side. On half of the trials, the arrows were congruent with the center arrow (i.e., pointing in the same direction as the center arrow) and on the other half the arrows were incongruent with the center arrow (i.e., flanking arrows were pointing the opposite direction as the center arrow). The two trial types were randomized across the task with 108 additional fixation-only trials per session for a total of 540 trials per session. On each trial, a jittered fixation at a variable interval (300-600 ms) appeared on the screen followed by the Flanker arrows at a fixed time of 200 ms. The trial ended with a blank response screen of 1700 ms. The task was completed in four runs with three blocks per run to provide intermittent performance feedback to maximize commission errors. Stimulus presentation and jitter orders were optimized and pseudorandomized using the make_random_timing.py program in AFNI. Details regarding image acquisition and pre-processing are in Appendix G.

**Subject-level Analysis** At the subject level, we analyzed the data with a time series model with regressors time-locked to stimulus onset reflecting trial type (incongruent, congruent) and error condition (correct, commission, omission). Regressors were created with a gamma variate for the hemodynamic response. The effects of interest at the condition level were two main contrasts: Cognitive Conflict (incongruent correct responses vs. congruent correct responses) and Error (incongruent commission errors vs. incongruent correct responses). All 42 participants were included in the conflict contrast, but only 27 participants had sufficient ($\geq 20$) commission errors in the incongruent condition to be included in the error contrast. For the conflict contrast, there were a total of 32005 trials across the two sessions of the task, which corresponds to $350 \pm 36$ incongruent trials and $412 \pm 19$ congruent trials across sessions per subject. For the subset of participants included into the error contrast, there were a total of 11366 trials available across both sessions, which corresponds to $331 \pm 28$ incongruent correct trials and $90 \pm 27$ incongruent commission errors per subject. We analyzed the subject-level fMRI data at the whole-brain level as well as at the region level using 12 Regions-Of-Interest (ROIs). We compare two reliability modeling approaches: a conventional CLM with regressors created at the condition level and TLM with trial-level regressors.

**Region-of-interest (ROI) Selection** Seed coordinates for ROIs were selected using Neurosynth term-based meta-analyses with the terms "cognitive control" and "error", the two main population-level effects of interest. Additionally, to derive ROIs outside the main population-level effects, we also selected peak coordinates from term-based meta-analyses for the term "visual" and "default mode". To select a reasonable number of peak coordinates, all four $z$-value maps (uniformity test for "cognitive control" and "error", association test for the "visual" and "default mode" map), FDR-corrected to 0.01, were further thresholded to $z$-value of 10. Spheres with a 6-mm radius (57 voxels) were created for each of the 12 sets of peak coordinates derived from the surviving clusters. Six (6) spheres were derived from the "cognitive control" and "error" maps respectively, 4 spheres were derived from the "default mode" map and 2 spheres from the "visual" map. Spheres derived from the "default mode" and "visual" map were used for both conflict and error data for a total of 12 ROIs for each contrast.

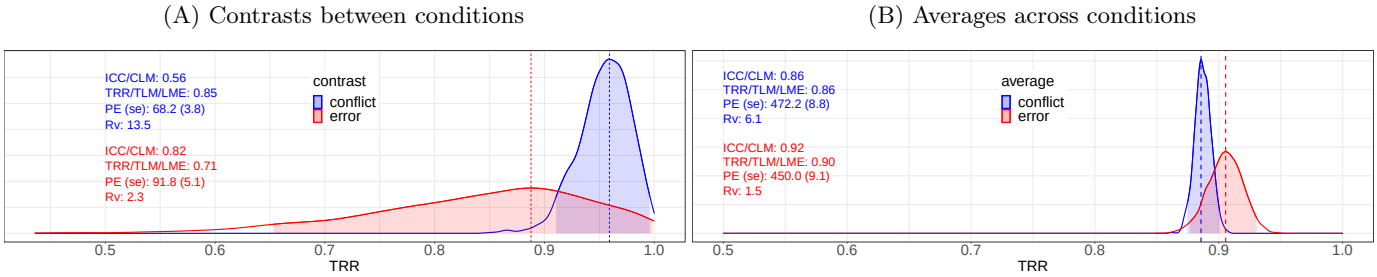



Figure 5: Reliability distributions of behavioral data (RT) from Flanker task. The reliability estimates for the contrast (A) and average (B) between the two conditions were based on 2000 draws from MCMC simulations of the BML model with an exGaussian likelihood and are shown here as a kernel density estimate which smooths the posterior samples. Each dashed vertical line indicates the mode (peak) of the perspective reliability distribution, and the shaded area shows the 95% highest density interval. The conflict reliability distribution was much more concentrated while the error reliability was relatively diffuse. The magnitude of the variability ratio $R_v$ is a proxy to assess ICC underestimation. Population effects (PE) and standard errors (se) in milliseconds are shown for reference.

### 3.2 Test-retest reliability estimation for behavioral data

Reliability estimates for the RT data of the Flanker task were relatively high.[3] The RT data were examined for the two main contrasts of interest: conflict (i.e., incongruent correct responses vs. congruent correct responses) and error (i.e., incongruent commission errors vs. incongruent correct responses). Trials with omissions were not considered. The RT values ranged within [4, 1669] and [9, 1686] ms for the conflict and error subset, respectively. As expected, the average RT values under each condition rendered relatively precise reliability estimates (Fig. 5B), and there was more convergence between conventional ICC and BML for these estimates (conflict: ICC(3,1)=0.86, BML=0.86 (mode); error:ICC(3,1)=0.92, BML=0.90 (mode)). The reliability estimates for the RT contrast via the conventional ICC were 0.56 and 0.82 for conflict and error contrast, respectively. Under the BML framework, five distributional candidates were considered using the model (17): Gaussian, Student's $t$, log-normal, shifted log-normal and exGaussian. Model comparisons and validations for each of the two RT contrasts (conflict and error) indicated that, similar to the RT data of the Stroop task (Fig. 4), the exGaussian distribution was the best fit alongside the Student's $t$. Using the BML framework, the reliability for the conflict contrast, estimated as the mode of the reliability distribution, showed higher than estimated by the ICC, at around 0.95. There was less discrepancy for the error contrast between the two approaches with a mode of 0.88 through BML. However, for the error contrast, the BML model revealed that larger uncertainty was associated with the reliability estimation for this contrast (Fig. 5A). The difference between the two frameworks was largely negligible except for the conflict contrast, which is also indicated by the associated variability ratio.

### 3.3 Whole-brain reliability estimation for neuroimaging data

As the BML model is not computationally feasible at the whole brain voxel-level, two modeling frameworks were adopted for whole brain analysis: ICCs using the program **3dICC** (Chen et al., 2018) with condition-level modeling (2) and the LME model (12) using the program **3dLMEr** (Chen et al., 2013) with trial-level modeling. Each of the two models was applied to the two effects (the average and the contrast between the two conditions) for conflict and error data for two sessions and 42/27 subjects respectively, resulting in a total of eight analyses. Input for the ICC computation are condition-level contrast while for the LME computation trial-level conflict and error data was comprised of 32005 and 11366 three-dimensional volumes.

Individual differences were largely reliable among most voxels for the average between the two conditions as assessed via ICC using condition-level modeling (Fig. 6A,B). Similarly, reliability was in the moderate to high range for the average between the two conditions using TLM-based LME (Fig. 6C,D). Both approaches

[3]The reliability analysis scripts used in this paper are available at https://github.com/afni-gangc/TRR

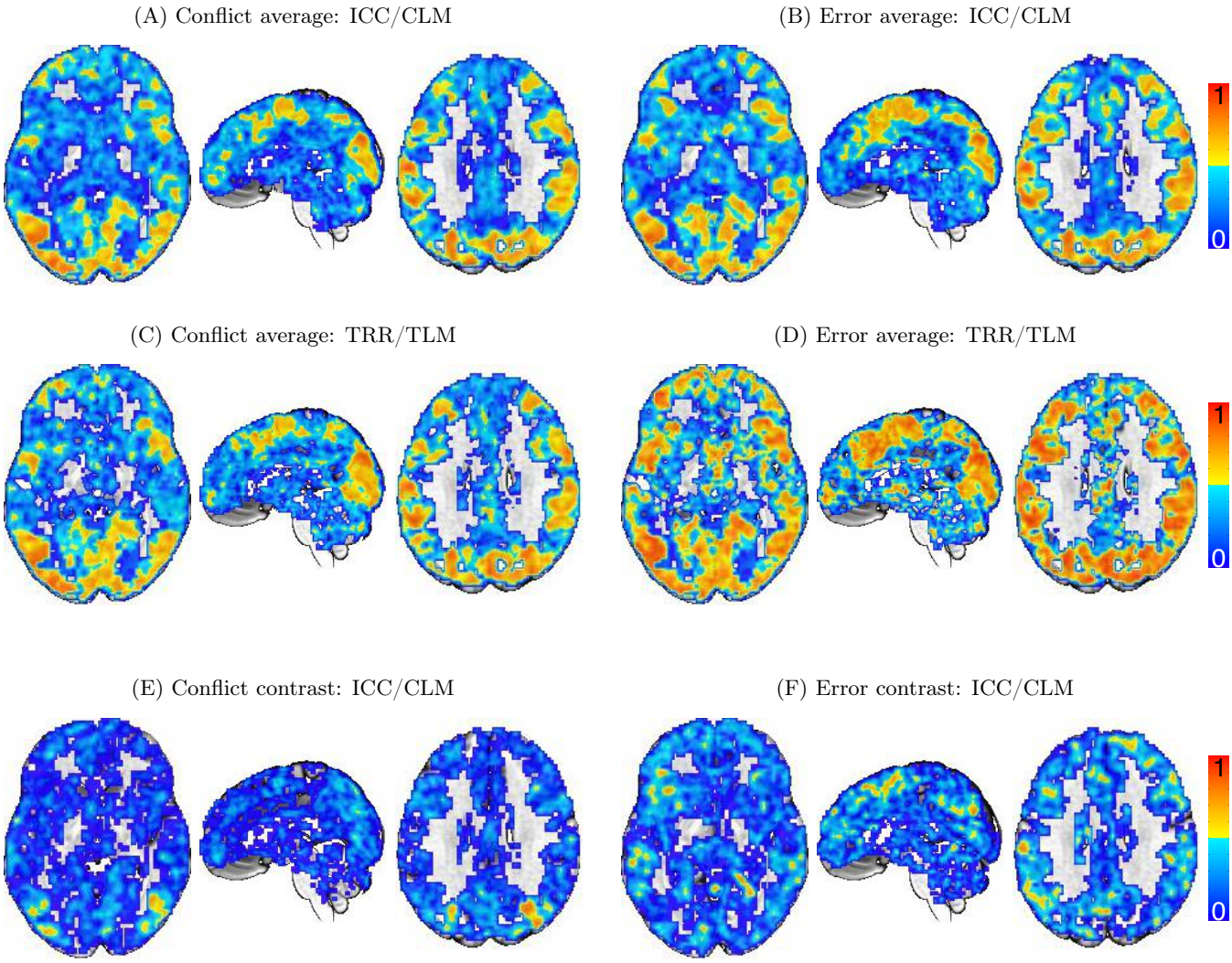



Figure 6: Whole-brain voxel-level reliability estimates for the fMRI Flanker dataset. The ICC values for the average conflict effect (A) showed negligible underestimation compared to the reliability estimated based on TLM/LME (C). In contrast, the ICC values for the average error effect (B) showed a moderate amount of underestimation compared to the reliability estimation (D) based on TLM/LME. The ICC values for the conflict (E) and error contrast (F) were much smaller than their average counterparts. Note that the TLM-based approach numerically failed for both contrasts at most voxels in the brain and the results are thus not shown. The three slices of axial ($Z = 0$), sagittal ($Y = 14$) and axial ($Z = 28$) planes are oriented in the neurological convention (right is right) in the MNI space.

rendered similar results for the average of conflict responses (Fig. 6, A vs C), while small but noticeable underestimation from ICC emerged for the average of the error responses (Fig. 6, B vs D). The divergence in estimates in the error contrast is likely due to a smaller number of trials for the error data, which is consistent with simulation results in Fig. 3B. Contrasts yielded lower estimates overall, with ICCs in the poor to adequate range (ICC(3,1) $< 0.4$, with the exception of primary visual, parietal and motor regions), which is consistent with previous work (e.g., Elliott et al., 2020). ICCs were higher for the error than the conflict contrast in several regions. For the contrasts, numerical failures occurred for most voxels in the TLM-based LME and thus results are not shown. Thus, in the next section we explore the comparison between CLM/ICC and TLM on the region level via the BML.

### 3.4 Region-based reliability estimation for neuroimaging data

Reliability estimation at the region level was performed through BML. The BML model (19) was adopted using the program **TRR** with the trial-level effect estimates from each subject as input. A Student's $t$-distribution was utilized for cross-trial variability to account for potential outlying values (Chen et al., 2020).

Standard errors of trial-level effects from the subject level were also incorporated into the BML model to improve robustness. The runtime was about 1.5 hours for each ROI through 4 Markov chains (CPUs) on a Linux computer (Fedora version 29) with AMD Opteron® 6376 at 1.4 GHz.

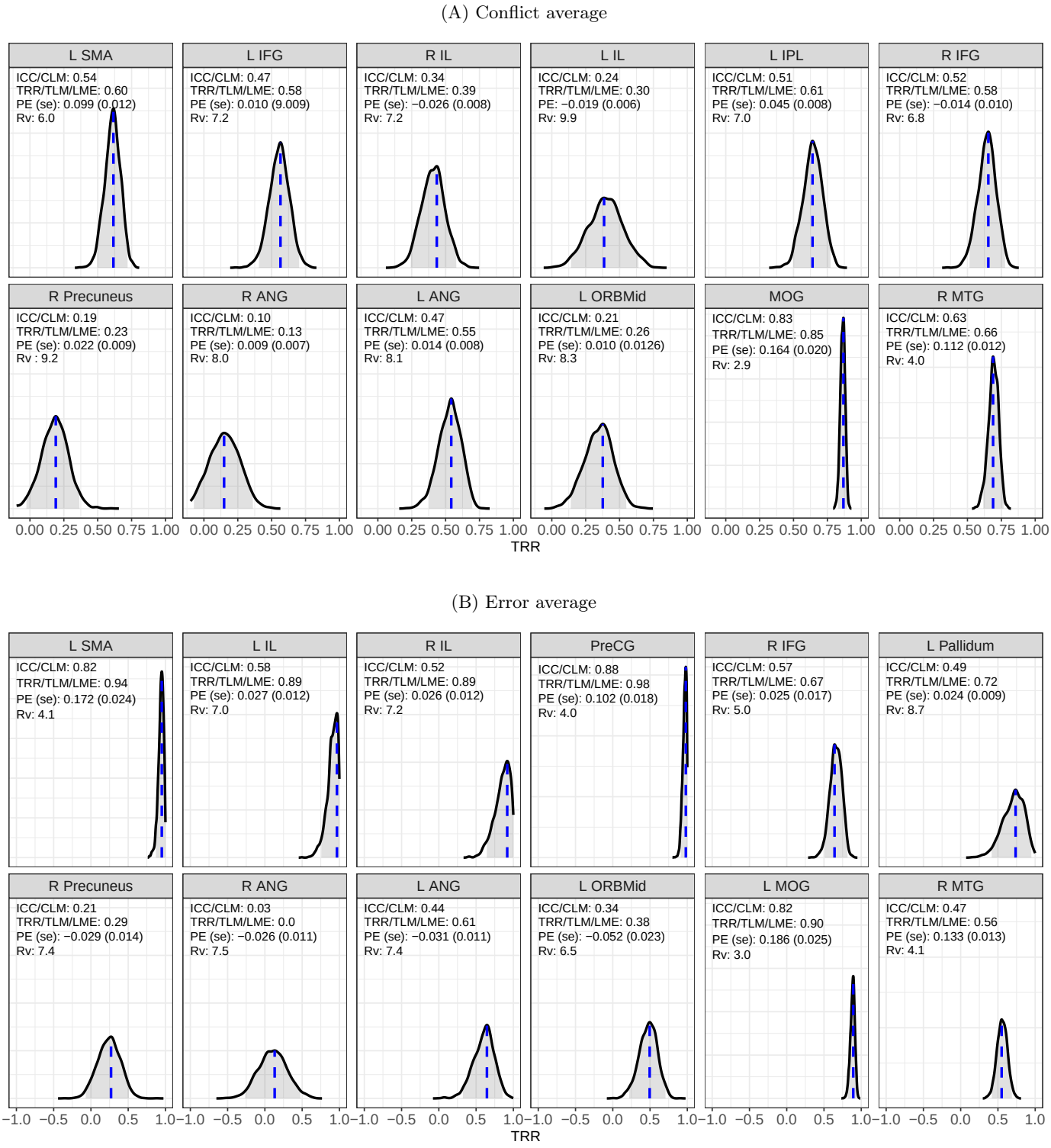
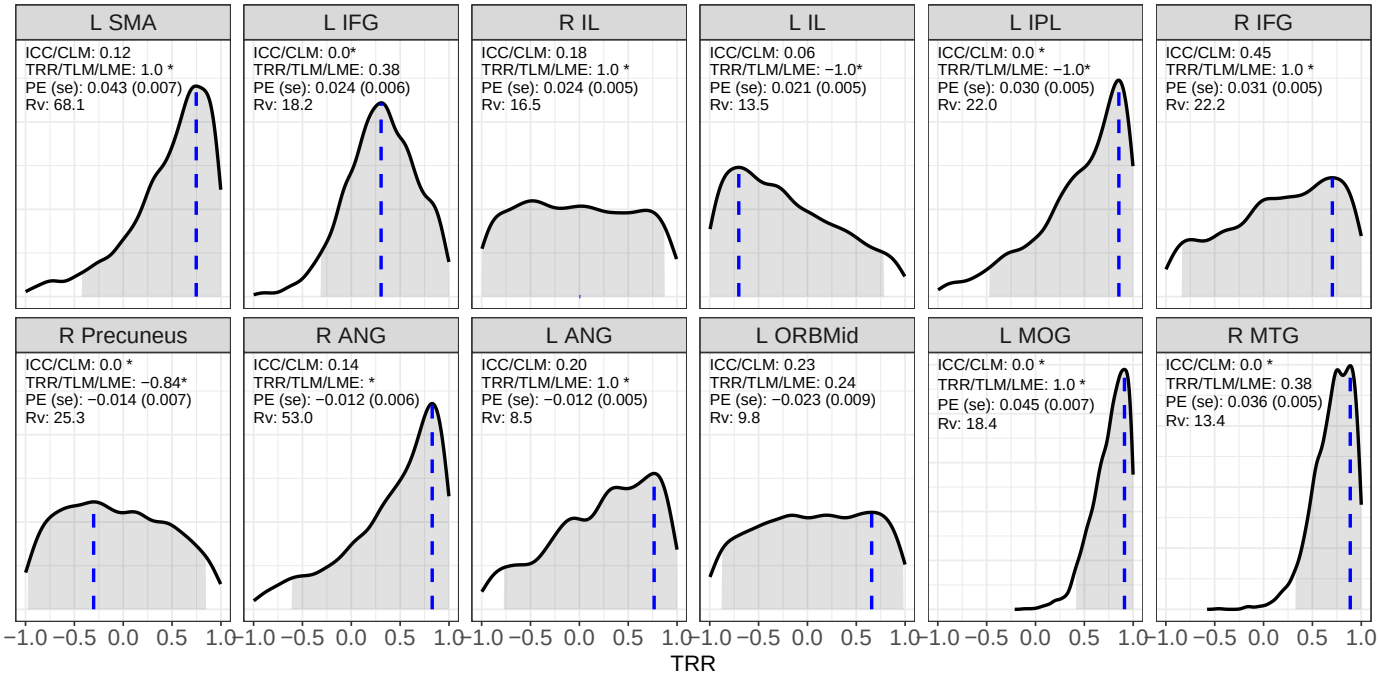



Figure 7: Reliability distributions for the average between the two conditions at 12 regions. The reliability estimates for the average effects of conflict (A) and error (B) were obtained using the program **TRR** based on 2000 draws from MCMC simulations of the BML model with a Student's $t$-distribution. Each blue vertical line indicates the mode (peak) of the reliability distribution, and the shaded area shows the 95% highest density interval. Four quantities are shown in each region's density plot: conventional ICC based on condition-level modeling (CLM), reliability estimated through the linear mixed-effects (LME) framework using trial-level modeling (TLM) at the subject level, population effect (PE) estimate plus its standard error (se), and variability ratio $R_v$. Asterisk (*) indicates numerical problem of either singularity or convergence failure under LME. Abbreviations: R: right, L: left, SMA: supplementary motor area, IFG: inferior frontal gyrus, IL: insula lobe, IPL: inferior parietal lobule, PreCG: precentral gyrus, MOG: middle occipital gyrus, MTG: middle temporal gyrus, ANG: angular gyrus, ORBmid: middle orbital gyrus.

(A) Conflict contrast




(B) Error contrast

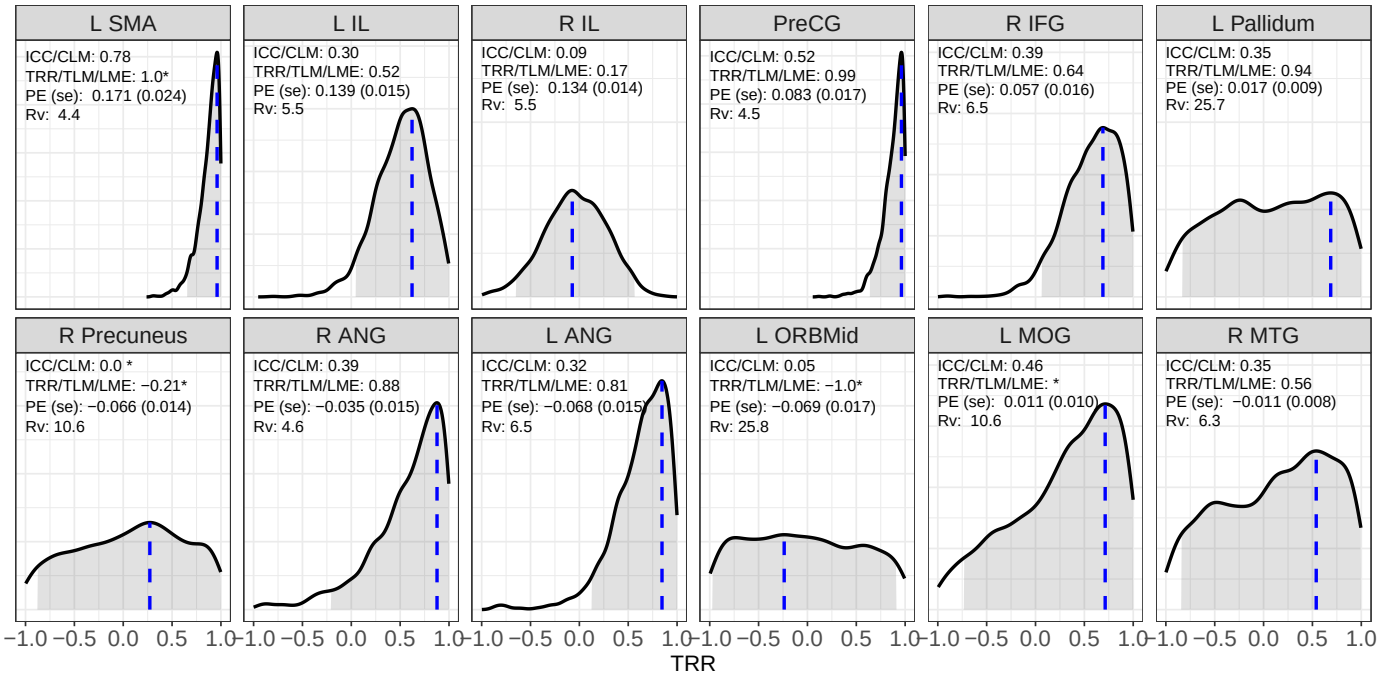



Figure 8: Reliability distributions for the two contrasts of the Flanker task for ROIs. The reliability estimates for the conflict (A) and error (B) contrast were obtained using the program **TRR** based on 2000 draws from MCMC simulations of the BML model with a Student's $t$-distribution. TEach blue vertical line indicates the mode (peak) of the respective reliability distribution, and the shaded area shows the 95% highest density interval. Four quantities are shown in each region's density plot: conventional ICC based on condition-level modeling (CLM), reliability estimated through the linear mixed-effects (LME) framework using trial-level modeling at the subject level, population effect (PE) estimate plus standard error (se), and variability ratio $R_v$. Asterisk (*) indicates numerical problem of either singularity or convergence failure under LME. Abbreviations: R: right, L: left, SMA: supplementary motor area, IFG: inferior frontal gyrus, IL: insula lobe, IPL: inferior parietal lobule, PreCG: precentral gyrus, MOG: middle occipital gyrus, MTG: middle temporal gyrus, ANG: angular gyrus, ORBmid: middle orbital gyrus.

Most regions exhibited adequate to excellent reliability of the average between the two conditions for both conflict and error with a few exceptions (R precuneus and R angular gyrus). There was some variability in the precision of these estimates. For the contrasts, large variations existed across regions in terms of

both magnitude of the reliability estimate and precision (Figs. 7,8). Some regions showed moderate to high reliability with relatively high precision; some regions exhibited high reliability with a wide range of uncertainty; others were difficult to assess as their reliability distributions were very diffuse. For regions with such substantial amount of uncertainty, an estimate through the ICC would be largely meaningless. The following regions demonstrated high reliability with relatively high precision: left MOG and right MTG for the conflict contrast (Fig. 8A); left SMA and PreCG for the error contrast (Fig. 8B). Some regions had reasonably high reliability but with moderate to poor precision (e.g., left SMA, left IPL, right angular gyrus for conflict contrast; left IL, right IFG, left and right IFG, MOG and right MTG for error contrast). As previously demonstrated, the variability ratio can be used as a proxy for the degree of underestimation via ICCs, and provides an explanation as to why such heterogeneity exists. Ten (10) out of the 12 ROIs had a substantial range of uncertainty associated with their reliability estimate for both conflict and error contrasts (Fig. 8). These were for the most part associated with a large variability ratios. It is for this reason that we advocate that reliability should not simply be expressed as a point estimates through the conventional ICC.

### 3.5 Implications for practice

The advancements through the BML are substantial for any researcher hoping to assess reliability for their task-based imaging data. It is specifically the additional precision information that is crucial for the researcher. The current version of the Flanker task was selected as a task that, with its two contrasts, covered several scenarios in the literature: The conflict contrast, with its 432 experimental trials per session likely represents the upper range of trial-level repetitions. The error contrast with a minimum of 20 trials in the incongruent commission error condition has just barely enough trials to derive robust estimates in a conventional analysis. Additionally, the number of participants for the conflict contrast (42) was almost double of that of the error contrast (27). The variability ratio was generally larger for the conflict than the error contrast, both in the behavioral and imaging data (albeit variable across regions). Half of the trials in both contrasts overlap (i.e., incongruent correct responses are used across both contrasts). Hence, commission error responses must be driving the increased reliability in the error contrast compared to the congruent correct trials included in the conflict contrast. Hence, despite relatively few instantiations in the task, commission error responses in the error contrast may generate a more consistent signal within person that distinctly differs between individuals. Alternatively, it may be that, despite involving the same visual display, the discrepancy between the psychological processes engaged in the error contrast is more pronounced compared to the conflict contrast where differences in cognitive demands are more subtle. Interestingly, the error contrast of RT showed large uncertainty around reliability estimates compared to the conflict contrast, which had a more narrow reliability distribution. No such visible differences in uncertainty emerged between the two contrasts in the region-level analysis of the imaging data. It is plausible that while behaviorally errors are few and inconsistent, generated via different mechanisms (i.e., short error RTs due to premature, anticipatory responses or flanking distractors, long error RTs due to attention lapses) all of which can occur in the same individual, these errors result in a common, intra-individually relatively consistent in inter-individually distinct error signal in the brain. This tentative explanation highlights the need to more systematically use this rich and informative method through BML for further exploration of reliability across different cognitive processes and experimental designs. Most importantly, our initial results provides a more nuanced assessment of task fMRI reliability than recent reports: reliability variability is large, even within the same contrast across regions, both in magnitude and precision of reliability estimates; understanding the source of trial-level variability may provide an important avenue to improve reliability in future work.

# 4 Discussion

High reliability of measurement is of great importance. Here, we adopt a new modeling framework for reliability estimation in behavioral and imaging tasks that (a) characterizes the data hierarchy including individual trials, (b) handles outliers and skewness in a principled way, and (c) properly quantifies uncertainty. Our investigation illustrates the cost of data reduction and the benefits of constructing an adaptive model that preserves the hierarchical integrity of the data (see Chen et al., 2020 for a trial-level modeling approach at the population level). Several key findings emerged in the step-by-step process by which we built the reliability model formulation: (a) reliability is conceptually distinct from population effects; (b) the conventional ICC underestimates reliability in cases where trial-level variability is large relative to between-subject variability; (c) trial-level variability is surprisingly large in both behavioral and neuroimaging data; and (d) a large number of trials is important to adequately assess reliability. We will discuss each of these in turn.

## 4.1 The relationship between population-level effects and test-retest reliability

We demonstrate that population effects are not necessarily tied to the reliability of individual differences. As illustrated in Fig. 1, all four possible scenarios can occur: large population effects can have strong or weak reliability, and the absence of population-level effects does not preclude high reliability. Researchers usually choose to examine the reliability of those regions that exhibit strong population effects; strong evidence of population-level effects may pique the researcher's interest in exploring the reliability of individual differences. However, limiting the search space to population level effects may be too narrow. The common practice of dichotomizing results into "significant" and "not significant" via multiple testing adjustment specifically exacerbates this problem (Chen et al., 2021). This can mean missed opportunities to estimate the reliability for regions involved in the task or regions that, while not showing involvement "on average", are relied upon by a subset of individuals. For example, the statistical evidence associated with the four regions of the Flanker dataset (R Precuneus, R ANG, L ANG and L ORBMid) was not strong enough per the currently adopted criteria for multiple testing adjustment, and would not have been part of the current reliability exploration if the selection had been solely based on the statistical strength at the population level.

## 4.2 BML as an adaptive solution for reliability estimation

Recent reports using the ICC metric suggests that the reliability of common cognitive and fMRI tasks is largely inadequate. Issues of reliability are now front and center in the field, and have motivated new modeling approaches for behavioral and imaging data with multivariate approaches recently highlighted as an avenue with more promising test-retest reliability (e.g., Noble et al., 2021). However, we show that in terms of univariate approaches, hierarchical modeling approaches accounting for trial-level variability may, under specific circumstances, improve modeling and associated reliability. Furthermore, the underestimation of reliability by the conventional ICC is linearly associated with the magnitude of reliability, with its rate depending on two factors: the trial sample size and the relative magnitude of cross-trial to cross-subject variability. The larger the relative cross-trial variability, the more severe the underestimation. Therefore, a reliability estimate, when implicitly "contaminated" by cross-trial variability, is sensitive to trial sample size. As a result, the reliability estimates reported in the literature, even on the same task, may not necessarily be comparable across experiments, leading to a portability problem (Rounder and Haaf, 2019).

Hierarchical models such as LME and BML through TLM allows the researcher to disentangle trial-level effects from other sources of variance, arguably providing a more accurate assessment of reliability. In other words, even though trial-level effects are of no interest, accounting for them in the model structure disentangles the cross-trial variability from the cross-subject variability and allows the accurate estimation of the correlation structure across repetitions. With a data structure involving five levels (Fig. 2), a hierarchical

model closely follows the underlying data generative process and simultaneously incorporates all the information available through proper regularization. As a result, one effectively gains high predictive accuracy for reliability estimation.

The BML framework is well-suited for reliability estimation. The adoption of the Bayesian formulation is mostly not intended to inject prior information, but to overcome several limitations of the LME approach through Monte Carlo simulations. Despite their accommodation of multilevel data structure, LME models can encounter numerical difficulties. In addition, they only provide point estimates with no easy access to uncertainty and have a limited flexibility of handling deviations from a Gaussian distribution such as data skewness and outliers. BML can overcome theses LME-associated issues. For example, rather than censoring data through brute force or arbitrary thresholding, BML resorts to a principled approach and adapts to the data through a wide variety of distributions (see Fig. 4). The information contained in posterior distributions is another benefit, showing subtle differences in distributions and allowing for straightforward interpretations. Despite a large amount of noise unaccounted for in fMRI data and a high variability ratio, the modeling investment is worthwhile in revealing reliability values above 0.9 with our experimental data (Fig. 4).

Model comparison and validation are important in improving the accuracy of reliability estimation. We note that a Gaussian assumption adopted usually as default may not always be an appropriate choice. The choice of assumption reflects prior knowledge about the specific data generating mechanism (i.e., bounded, skewed) and realities about variability in measurement (e.g., outliers). Hence, we recommend that, unlike in the conventional framework where this choice is largely epistemological, any assumption adopted as a prior or likelihood should be calibrated by the data and further cross-validated though, for instance, posterior predictive checks. Even for data with substantially less complexity than imaging data, such as RT data, a consensus has not been reached in terms of what distribution is a "default" best fit, although most investigations have leaned towards adopting either an exGaussian or a shifted log-normal distribution. This is because even for this 'simple' measure, complexity arises through mechanisms such as linear increases of uncertainty with increasing means (Wagenmakers and Brown, 2007).

### 4.3 High cross-trial variability

Large cross-variability is the root cause for ICC underestimation, numerical failures in LME modeling and poor precision of reliability estimates. Only in scenarios where cross-trial variability is roughly the same order of magnitude as, or smaller than, the cross-subject variability, the conventional ICC and the reliability estimates based on trial-level modeling will yield similar results (7) and (13). However, experimental data from both behavioral and neuroimaging investigations point to a much larger cross-trial variability than cross-subject variability. Specifically, the variability ratio $R_v$ may reach up to 10 for simple effects and go beyond 20 for contrasts. Trial-level effects fluctuate substantially and their sequence appears to some extent without a clear pattern (Chen et al., 2020). Such randomness occurs across brain regions within the same subjects as well as across subjects. In other words, a large proportion of cross-trial fluctuations cannot be accounted for by habituation, fatigue or sensitization. At the same time, these trial-level fluctuations are not purely random: there is a high degree of bilateral synchronization within the same subject (Chen et al., 2020). Some of this variance may be accounted for by behavioral measures such as reaction time (i.e., momentary lapses in attention) or more systematic stimulus ratings that can be modeled through trial-level modulation analysis at the subject level. It is also plausible that brain regions may constantly undergo some intrinsic fluctuations - external stimuli or tasks may be surprisingly small constrains stacked on top of large intrinsic neuronal activities. Additionally, cross-trial variability might stem from suboptimal modeling in subject-level time series regression due to (i) a substantial amount of confounding effects not properly accounted for and (ii) poor characterization of varying hemodynamic response across trials, condition, regions and subjects through a fixed basis function. However, large cross-trial variability also exists in behavioral tasks although

to slightly lesser extent (Rouder et al., 2019). Thus, suboptimal modeling likely is not the main (or only) culprit; understanding the origins of this variability remains an important future endeavor.

### 4.4 Importance of trial sample size

The number of trials plays a crucial role in achieving a reasonable precision of reliability estimation. The central limit theorem is pivotal to many modeling frameworks including the conventional ICC. However, the asymptotic property of unbiasedness and Gaussianity both rely on large sample sizes, which cannot necessarily be met or easily predetermined in practice. In fact, ensuring a large enough sample size especially in terms of trials for reliability estimation is an important issue that has received far too little attention.

First, we showed how strongly ICC estimation accuracy depends on the number of trials. It has been generally assumed that subject sample size might help in achieving high statistical efficiency (i.e., standard error) of test-retest reliability (Shoukri et al., 2004). However, while subject sample size plays a role in stability and precision of estimates, trial number has a much stronger impact on reliability estimation and efficiency, as shown in the formulas (7) and (13) as well as simulation results in Figs. 3 and 9.

Second, as cross-trial variability can be high, trial sample size is pivotal in determining the precision of reliability estimates. A substantial number of trials may be required to achieve a reasonable precision. As shown in our investigation, the uncertainty of reliability estimation, as represented by the standard error or highest density interval, often remains wide when cross-trial variability is large. Uncertainty depends on four factors (Figs. 3 and 9): reliability magnitude, variability ratio, trial and, to a lesser extent, subject sample size. Among the four factors, only the sample sizes can be easily manipulated. Also, as shown in experimental results (Figs. 4, 5, 7 and 8), the precision of reliability estimation varies substantially across brain regions or between single conditions and contrasts. To dissolve those diffuse posterior distributions of reliability estimates, a few hundred or even more trials may have to be adopted. In practice, such experimental designs may not be feasible due to their time burden on the subject and financial burden on the experimenter.

### 4.5 Difficulty of obtaining high reliability precision for a contrast

For effects of a single condition or the average among two or more conditions, it is relatively easy to achieve reasonable reliability precision. Empirical data indicate that cross-trial variability $\sigma_0$ is larger than cross-subject variability $\sigma_{\tau_r}$ with a variability ratio $R_v$ less than 10. Thus, it is possible, with a sizeable trial sample size (possibly larger than what is typically adopted in the field), to obtain reliability estimation within a small or moderate amount of uncertainty (Figs. 4, 5 and 7). Consequentially, one may be able to estimate reliability under a linear mixed-effects framework through trial-level modeling (Fig. 6C,D).

In comparison, a reasonably high precision of reliability estimation for a contrast might be harder to attain. Cross-trial variability $\sigma_0$ measures the trial-level fluctuations per condition (and per subject) regardless of the research focus on a single condition or a contrast. However, when a contrast is of interest, cross-subject variability $\sigma_\tau$ measures the fluctuations relative to the contrast, as characterized in the parameters $\lambda_{rs}$ in formulations (12) and (17). The contrast between two conditions is usually a few times smaller in magnitude. For example, suppose that the magnitude of the BOLD response in the congruent and incongruent condition is 1.0% and 0.8% at a brain region, respectively. Their contrast of 0.2% would be 4-5 times smaller than the magnitude of each condition alone. Yet, cross-trial variability $\sigma_0$ remains roughly the same regardless of the effect (i.e., contrast, a single condition or cross-condition average). Thus, the relative magnitude of cross-trial variability would be much larger for the contrast than a single condition, leading to a sizeable variability ratio $R_v$. In other words, the cross-subject effects $\lambda_{rs}$ often will get dwarfed by cross-trial variability $\sigma_0$, resulting in a large uncertainty for the reliability estimation of a contrast. The variability ratio $R_v$ can differ substantially across regions and may be so large in some regions that the requirement for trial sample sizes may become

practically unfeasible. Nevertheless, some brain regions could still achieve high reliability estimation with a reasonable precision (Fig. 8) such as the PreCG and L SMA in the error contrast. In general, we recommend the adoption of the BML framework for its flexibility to closely characterize the data structure. Additionally, even though difficult to achieve a high precision for reliability estimates, the resulting posteriors from a BML model would likely still encapsulate the distribution shape regardless of its centrality or diffusivity.

### 4.6 Two types of generalizability

Scientific investigation strives to gain knowledge through legitimate generalization. With limited samples and properly built models, one draws broad conclusions that extend far beyond specific experiments. From a statistical perspective, generalization is made possible through inferences regarding the observed or a hypothetical population based on the data at hand. Two types of generalizability are relevant in the current context concerning reliability: population-level effects and reliability of individual differences.

Population-level effects generalize through measurement units of subjects and trials. Specifically, population-level effects usually lie at the top of the data hierarchy and are modeled as fixed effects under the conventional LME framework. They are relatively intuitive and easy to visualize as the horizontal lines illustrated in Fig. 1. For example, the population-level contrast between incongruent and congruent conditions is the main focus in the experimental designs of the Flanker and Stroop task; examining group differences in these population-level activation patters is a common research goal. Typical sample size for subjects and trials is below 100. In this context, cross-subject and cross-trial fluctuations are considered noise and nuisance.

Test-retest reliability concerns a different type of generalizability: the consistency of individual differences that can be examined as trait-like measures, behavioral (e.g., RT) or BOLD response. Unlike population-level effects that are assumed to be "fixed" in a statistical model, reliability is characterized by subject-level effects that vary across subjects and are termed as "random" effects under the LME framework. Cross-subject fluctuations are expected to be consistent and systematic - such a pattern is the second type of generalizability that is characterized as the correlation of subject-level fluctuations across repetitions. The generalizability of reliability lies in the reference of subject-level effects relative to their associated population-level effects. For example, subject-specific effects characterize the relative variations around the population effects. A high reliability of individual differences in a Flanker task experiment means that subjects with a larger inhibition effect relative to the population average are expected to show a similar pattern when the experiment is repeated. Due to their smaller effect size compared to population effects, subject-level effects and reliability are much more subtle and require visual detection through close inspection of within-subject similarity using population effects as references (dots and diamond in Fig. 1).

### 4.7 Limitations

BML modeling remains limited to region-level reliability estimation. Currently, it is not possible to apply the BML framework at the whole brain level. Because long chains of iterations are required to obtain stable numerical simulations under the Bayesian framework, the computational cost of BML is usually high for large datasets. Thus, its application is currently limited to behavioral and region-based neuroimaging data.

Trial-level modeling requires careful experimental designs. When each trial is modeled separately at the subject level, the risk of high correlations or multicollinearity may arise among the regressors. To avoid such potential issue, trial sequence and timing can be randomized to reduce multicollinearity using tools such as RSFgen in AFNI or optseq[4]. However, even if statistically separable, trial-level effect estimates can be unreliable, and are largely limited to the common approach of presuming a fixed-shape hemodynamic response for most experimental designs. As a substantial amount of variability exists across tasks, brain

[4]https://surfer.nmr.mgh.harvard.edu/optseq/

regions, subjects and even trials, such a presumption might misidentify trial-level effect magnitude, resulting in compromised reliability estimation.

## 5 Conclusion

The conventional intraclass correlation, when adopted for datasets with many trials making up a condition of interest, could underestimate the reliability to varying extents. Reliability accuracy via ICC is lost in two ways: first, its formulation fails to accurately partition the hierarchical structure embedded in the data; second, its uncertainty information can be difficult to obtain. We recommend that reliability be estimated through subject-level correlation across repetitions via a Bayesian multilevel model that maps the data structure as accurate as possible. We offer two publicly available programs, **TRR** and **3dLMEr**, for reliability estimation. In addition, we suggest that reliability be reported with either a full posterior distribution or a mode combined with its highest density interval. A large number of trials might be required to achieve acceptable amount of reliability estimation uncertainty especially for subtle effects such as a contrast between two conditions.

## Acknowledgments

The research and writing of the paper were supported (GC and RWC) by the NIMH and NINDS Intramural Research Programs (ZICMH002888) of the NIH/HHS, USA. Data collection was supported (DSP) by the NIMH Intramural Research Program (ZIAMH002781). Our work was inspired by the modeling platforms of Haines et al. (2020) and Rouder and Haaf (2019). We are appreciative of the technical support from the Stan (Carpenter et al., 2017) and R (R Core Team, 2019) communities. Most of the modeling work was performed in Stan through the R packages brms (Bürkner, 2018) and lme4 (Bates et al., 2015). The figures were generated with the R package ggplot2 (Wickham, 2009). This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). We thank the anonymous reviewers for their critical reading with thoughtful suggestions that helped improve, clarify and contextualize our manuscript.

# Appendices

## A ICC underestimation for a single condition

We seek to derive the conventional ICC under the LME framework through trial-level modeling. It is worth noting that, under the LME formulation (5), ICC can be conceptualized as the correlation between the two cross-trial averages at the subject level $\overline{y}_{rs\cdot} \sim \mathcal{N}(a_r + \tau_{rs},\ \frac{1}{m}\sigma_0^2)$ $(r = 1, 2)$ with the homoscedastity assumption between the two repetitions $\sigma_{\tau_1} = \sigma_{\tau_2} = \sigma_\tau$. With the notations

$$\xi_s = \frac{1}{2}(\tau_{1s} + \tau_{2s}),\ \eta_s = \frac{1}{2}(\tau_{1s} - \tau_{2s}),$$

we have

$$\begin{aligned}
&\overline{y}_{1s\cdot} \sim \mathcal{N}(a_1 + \xi_s + \eta_s,\ \frac{1}{m}\sigma_0^2),\ \overline{y}_{2s\cdot} \sim \mathcal{N}(a_2 + \xi_s - \eta_s,\ \frac{1}{m}\sigma_0^2),\\
&\mathrm{Var}(\xi_s) = \frac{1}{2}(1+\rho)\sigma_\tau^2,\ \mathrm{Var}(\eta_s) = \frac{1}{2}(1-\rho)\sigma_\tau^2,\ \mathrm{Cov}(\xi_s,\ \eta_s) = \frac{1}{4}(\mathrm{Var}(\tau_{1s}) - \mathrm{Var}(\tau_{2s})) = 0,\\
&\mathrm{Var}(\overline{y}_{1s\cdot}) = \mathrm{Var}(\xi_s) + \mathrm{Var}(\eta_s) + \mathrm{Cov}(\xi_s,\ \eta_s) + \frac{1}{m}\sigma_0^2 = \sigma_\tau^2 + \frac{1}{m}\sigma_0^2,\\
&\mathrm{Var}(\overline{y}_{2s\cdot}) = \mathrm{Var}(\xi_s) + \mathrm{Var}(\eta_s) - \mathrm{Cov}(\xi_s,\ \eta_s) + \frac{1}{m}\sigma_0^2 = \sigma_\tau^2 + \frac{1}{m}\sigma_0^2,\\
&\mathrm{Cov}(\overline{y}_{1s\cdot},\ \overline{y}_{2s\cdot}) = \mathrm{Var}(\xi_s) - \mathrm{Var}(\eta_s) = \rho\sigma_\tau^2.
\end{aligned} \tag{20}$$

Through the notations

$$R_v = \frac{\sigma_0}{\sigma_\tau},\ R_u = \frac{1}{1 + \frac{1}{m}R_v^2},$$

it becomes clear that ICC can be directly expressed as the function of $\rho$

$$\mathrm{ICC(3,1)} = \frac{\mathrm{Cov}(\overline{y}_{1s\cdot},\ \overline{y}_{2s\cdot})}{\sqrt{\mathrm{Var}(\overline{y}_{1s\cdot})\ \mathrm{Var}(\overline{y}_{2s\cdot})}} = \frac{\rho\sigma_\tau^2}{\sigma_\tau^2 + \frac{1}{m}\sigma_0^2} = \frac{1}{1 + \frac{1}{m}R_v^2}\rho = R_u\rho.$$

The variability ratio $R_v$ characterizes the magnitude of cross-trial variability $\sigma_0$ relative to the cross-subject variability $\sigma_\tau$, and the parameter $R_u$ encapsulates the rate of ICC underestimation. It is quite revealing that the extent of ICC underestimation depends on two factors, the trial sample size $m$ and the relative magnitude of cross-trial variability $R_v$.

The ICC underestimation can be conceptually corrected. Under the homoscedasticity assumption, the derivations (20) indicate that $\mathrm{Var}(\overline{y}_{1s\cdot}) = \mathrm{Var}(\overline{y}_{2s\cdot}) = \sigma_\tau^2 + \frac{1}{m}\sigma_0^2$. That is, the variability of cross-trial averages is composed of two components, one associated with cross-subject variability $\sigma_\tau$ and the other with cross-trial variability $\sigma_0$. Thus, if the cross-trial variability $\sigma_0$ is known, we could restore the accuracy of ICC by removing the trial-related variance component from the denominator of the ICC formulation (3),

$$\mathrm{ICCa} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2 - \frac{1}{m}\sigma_0^2}.$$

# B Simulations with a single condition

Simulations were conducted for reliability with a single condition. Below are the manipulation parameters for the two models of condition-level (2) and trial-level modeling (5) with two repetitions of data collection:

1) homoscedasticity with fixed scaling parameters across sessions: $\sigma_{\tau_1} = \sigma_{\tau_2} = \sigma_\tau = 1$,
2) 4 different reliability values: $\rho = 0.3$, 0.5, 0.7 and 0.9,
3) 4 different sample sizes of subjects: $n = 20$, 40, 100 and 200,
4) 4 different sample sizes of trials per repetition: $m = 20$, 40, 100 and 200,
5) 4 different ratios of cross-trial relative to cross-subject variability: $R_v = \frac{\sigma_0}{\sigma_\tau} = 1$, 4, 7 and 10,
6) 2 different sets of population-level effects across the two repetitions: $(a_1,\ a_2) = (0,\ 0)$ and $(1.0,\ 0.9)$,
7) 3 different approaches to assessing cross-session reliability:
   (a) conventional ICC estimated with ANOVA/LME (2) through aggregation across trials,[5]

[5]The aggregation step more accurately reflects the typical preprocessing in behavior data than in neuroimaging. The following two pipelines are not strictly commutative in fMRI data analysis: (a) obtain condition-level effect estimates from time series regression with one regressor per condition, and (b) perform time series regression with one regressor per trial and then obtain the condition-level effect estimate by averaging the trial-level regression coefficients. The aggregation step in our simulations follows the latter pipeline as a rough approximation for the former.

(b) reliability $\rho$ estimated through LME (5),

(c) conventional ICC adjusted by removing the cross-trial variability $\frac{\sigma_0^2}{m}$.

Each of these $4 \times 4 \times 4 \times 4 \times 2 \times 3 = 1536$ combinations was simulated 1000 times, using the function *lmer* in the *R* package lme4 (Bates et al., 2015) with the following iterative steps.

i) For each subject $s$, obtain subject-level effects during the two repetitions through random sampling: $\begin{bmatrix} \tau_{1s} \\ \tau_{2s} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}_{2\times 1},\ \boldsymbol{R}_{2\times 2})$, where $\boldsymbol{R} = \begin{bmatrix} \sigma_{\tau_1}^2 & \rho\sigma_{\tau_1}\sigma_{\tau_2} \\ \rho\sigma_{\tau_1}\sigma_{\tau_2} & \sigma_{\tau_2}^2 \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, $s = 1, 2, .., n$.

ii) Draw the simulated data under LME (5):
$y_{rst} \sim \mathcal{N}(a_r + \tau_{rs},\ \sigma_0^2),\ r = 1, 2;\ s = 1, 2, .., n;\ t = 1, 2, .., m.$

iii) Solve the two LME models of (2) and (5).

iv) Recover the simulated parameters including the three reliability estimates. Specifically, the conventional ICC is obtained through the formula (3) while the reliability is estimated through $\rho$ in (5). In addition, the adjusted ICC is obtained by removing the cross-trial variability $\frac{\sigma_0^2}{m}$ through the formula (9).

# C Correlation structure among the varying intercepts and varying slopes in the LME model (12)

With two conditions and a $2 \times 2$ factorial structure, we denote $\mu_{crs}$ as the $s$-th subject's condition-level effects ($c = 1, 2;\ r = 1, 2;\ s = 1, 2, ..., n$) and assume that the four effects associated with each subject follow a quad-variate Gaussian distribution,

$$\begin{aligned} &(\mu_{11s},\ \mu_{12s},\ \mu_{21s},\ \mu_{22s})^T \sim \mathcal{N}(\mathbf{0}_{4\times 1},\ \boldsymbol{P}_{4\times 4}), \\ &\boldsymbol{P} = \text{diag}(q_{11}, q_{12}, q_{21}, q_{22})\ \boldsymbol{C}\ \text{diag}(q_{11}, q_{12}, q_{21}, q_{22}), \\ &s = 1, 2, ..., n, \end{aligned} \tag{21}$$

where $\boldsymbol{P}$ and $\boldsymbol{C}$ are the variance-covariance and correlation matrices for the four effects. With the symmetry assumptions $\text{corr}(\mu_{c1s},\ \mu_{c2s}) = \pi$ $(c = 1, 2)$, $\text{corr}(\mu_{1rs},\ \mu_{2rs}) = \theta$ $(r = 1, 2)$ and $\text{corr}(\mu_{11s},\ \mu_{22s}) = \text{corr}(\mu_{12s},\ \mu_{22s}) = \eta)$, the correlation matrix $\boldsymbol{C}$ is of the following structure,

$$\boldsymbol{C} = \begin{bmatrix} 1 & \pi & \theta & \eta \\ \pi & 1 & \eta & \theta \\ \theta & \eta & 1 & \pi \\ \eta & \theta & \pi & 1 \end{bmatrix}, \tag{22}$$

where the correlations $\theta$, $\eta$ and $\pi$ are such that the correlation matrix $\boldsymbol{C}$ is positive semi-definite.

Now we derive the variance-covariance structure of the quad-variate $(\tau_{1s}, \tau_{2s}, \lambda_{1s}, \lambda_{2s})^T$ under the LME formulation (12). With the indicator variable $I_c$ defined in (11), the four variables are the varying intercepts and slopes and can be expressed as $\tau_{rs} = \frac{1}{2}(\mu_{1rs} + \mu_{2rs})$ and $\lambda_{rs} = \mu_{1rs} - \mu_{2rs}$ $(r = 1, 2)$. Furthermore, the correlation matrix for the quad-variate $(\tau_{1s}, \tau_{2s}, \lambda_{1s}, \lambda_{2s})^T$ is of the the following block diagonal form that justifies the independence assumption between the varying intercepts and varying slope in the LME formulation (12),

$$\begin{bmatrix} 1 & \rho_0 & 0 & 0 \\ \rho_0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho_1 \\ 0 & 0 & \rho_1 & 1 \end{bmatrix}. \tag{23}$$

We obtain the correlation $\rho_0$ between the two varying intercept components as

$$\rho_0 = \frac{\text{cov}(\tau_{1s},\ \tau_{2s})}{\sqrt{\text{var}(\tau_{1s})\ \text{var}(\tau_{2s})}} = \frac{\text{cov}(\frac{1}{2}(\mu_{11s}+\mu_{21s}),\ \frac{1}{2}(\mu_{12s}+\mu_{22s}))}{\sqrt{\text{var}(\frac{1}{2}(\mu_{11s}+\mu_{21s}))\ \text{var}(\frac{1}{2}(\mu_{12s}+\mu_{22s}))}} = \frac{\pi+\eta}{1+\theta}, \tag{24}$$

and the correlation $\rho_1$ between the two varying slope components as

$$\rho_1 = \frac{\text{cov}(\lambda_{1s},\ \lambda_{2s})}{\sqrt{\text{var}(\lambda_{1s})\ var(\lambda_{2s})}} = \frac{\text{cov}(\mu_{11s}-\mu_{21s},\ \mu_{12s}-\mu_{22s})}{\sqrt{\text{var}(\mu_{11s}-\mu_{21s})\ \text{var}(\mu_{12s}-\mu_{22s})}} = \frac{\pi-\eta}{1-\theta}. \tag{25}$$

The correlation of 0s in (23) can be similarly derived as in (24) and (25).

# D ICC underestimation for a contrast between two conditions

The extent of ICC underestimation follows a similar derivation for the case of a contrast between two conditions as that of a single condition. Based on the distribution assumption $y_{crst}|a_r, b_r, \tau_{rs}, \lambda_{rs}, \sigma_0 \sim \mathcal{N}(a_r + b_r I_c + \tau_{rs} + \lambda_{rs} I_c,\ \sigma_0^2)$ in the LME model (12) and the homoscedasticity assumption $\sigma_{\lambda_1} = \sigma_{\lambda_2} = \sigma_\lambda$, we have

$$\begin{aligned}
&\overline{y}_{1rs\cdot} - \overline{y}_{2rs\cdot}|b_r, \lambda_{rs}, \sigma_0 \sim \mathcal{N}(b_r + \lambda_{rs},\ \frac{2}{m}\sigma_0^2),\ r = 1, 2;\\
&\text{Var}(\overline{y}_{11s\cdot} - \overline{y}_{21s\cdot}) = \text{Var}(\overline{y}_{21s\cdot} - \overline{y}_{22s\cdot}) = \sigma_\lambda^2 + \frac{2}{m}\sigma_0^2;\\
&\text{Cov}(\overline{y}_{11s\cdot} - \overline{y}_{21s\cdot},\ \overline{y}_{21s\cdot} - \overline{y}_{22s\cdot}) = \text{Cov}(\lambda_{1s},\ \lambda_{2s}) = \rho_1\sigma_\lambda^2.
\end{aligned}$$

Plugging the above results into the definition of ICC for the contrast, we immediately see the amount of ICC underestimation,

$$\text{ICC(3,1)} = \frac{\text{Cov}(\overline{y}_{11s\cdot} - \overline{y}_{21s\cdot},\ \overline{y}_{21s\cdot} - \overline{y}_{22s\cdot})}{\sqrt{\text{Var}(\overline{y}_{11s\cdot} - \overline{y}_{21s\cdot})\ \text{Var}(\overline{y}_{21s\cdot} - \overline{y}_{22s\cdot})}} = \frac{\rho_1\sigma_\lambda^2}{\sigma_\lambda^2 + \frac{2}{m}\sigma_0^2} = \frac{1}{1+\frac{2}{m}R_v^2}\rho_1 = R_u\rho_1$$

where $R_v = \frac{\sigma_0}{\sigma_\lambda}$ is the variability ratio and $R_u = \frac{1}{1+\frac{2}{m}R_v^2}$ is the underestimation rate.

The ICC underestimation can be conceptually corrected. If the cross-trial variability $\sigma_0$ is known, we could restore the accuracy of ICC by removing the trial-related variance component $\sigma_0^2$ from the denominator of the ICC formulation,

$$\text{ICCa} = \frac{\widetilde{\sigma}_\lambda^2}{\widetilde{\sigma}_\lambda^2 + \sigma_e^2 - \frac{2}{m}\sigma_0^2}.$$

The ICC underestimation for the average effect between the two conditions can be similarly derived. In fact, all the formulas remain the same as long as we replace the symbols $b_r$, $\lambda$ and $\rho_1$ by $a_r$, $\tau$ and $\rho_0$.

# E Simulations of trial-level LME modeling for a contrast

Simulations for the reliability with a contrast under the LME model (12) are similar to the situation with a single condition but with a slightly more complexity. With the correlation matrix $\boldsymbol{C}$ in (22) across the four condition-level effects $(\mu_{11s},\ \mu_{12s},\ \mu_{21s},\ \mu_{22s})^T$ and the homoscedasticity assumption in (21): $q_{11} = q_{21} = q_{12} = q_{22} = 1$, the cross-session reliability in the LME model (12) for the average and contrast between the two conditions can be simulated per the formulas (24) and (25). Below are the manipulation parameters:

1) 4 simulated reliability values $\rho_1 = 0.3$, 0.5, 0.7 and 0.9 that, respectively, correspond to four sets of correlation structure $\boldsymbol{C}$ in (22): $(\pi,\ \theta,\ \eta) = (0.7,\ 0.5,\ 0.55)$, $(0.7,\ 0.5,\ 0.45)$, $(0.7,\ 0.5,\ 0.35)$, $(0.7,\ 0.5,\ 0.25)$.
2) 4 different numbers of subjects: $n = 20$, 40, 100 and 200,

3) 4 different numbers of trials per session: $m$ = 20, 40, 100 and 200,
4) 4 different ratios of cross-trial variability $\sigma_0$ relative to cross-subject variability $\sigma_\lambda$: $\frac{\sigma_0}{\sigma_\lambda}$ = 1, 4, 7 and 10,
5) 2 different sets of population-level effects across the two sessions, $(a_{11},\ a_{12},\ a_{21},\ a_{22})$ = (0, 0, 0, 0) and (1.0, 0.9, 0, 0).
6) 3 different approaches to assessing cross-session reliability:
   (a) conventional ICC based ANOVA/LME (2) through aggregation across trials and conditions;
   (b) cross-session reliability $\rho_1$ based on the LME formulation (12);
   (c) conventional ICC adjusted by removing the cross-trial variability $\frac{2\sigma_0^2}{m}$ per formula (15).

Each of these $4 \times 4 \times 4 \times 4 \times 2 \times 3 = 1536$ combinations was simulated 1000 times. During each simulation, data were randomly drawn through the following steps:

1. For each subject $p$, obtain subject-level effects during the two sessions: $(\mu_{11s},\ \mu_{12s},\ \mu_{21s},\ \mu_{22s})^T \sim \mathcal{N}((a_{11},\ a_{12},\ a_{21},\ a_{22})^T,\ \boldsymbol{P})$, where $\boldsymbol{P} = \boldsymbol{C}$.
2. Draw simulated data per the LME formulation (10): $y_{crst} \sim \mathcal{N}(\mu_{crs},\ \sigma_0^2),\ c = 1, 2;\ r = 1, 2;\ s = 1, 2, .., n;\ t = 1, 2, .., m$.
3. Solve the two LME models, (2) and (12), using the function *lmer* in the *R* package lme4.
4. Recover the simulated parameters including the three reliability estimates. Specifically, the conventional ICC is obtained through the formula (3) while the cross-session reliability $\rho_1$ is estimated through the LME model (12). In addition, adjusted ICC is obtained per formula (15).

The simulation results largely follow a similar pattern to the situation for a single condition (Fig. 9).

# F Hyperpriors adopted for BML modeling

The prior distribution for all lower-level (e.g., subject) effects considered here is Gaussian, as specified in the respective model; for example, see the distribution assumptions in the BML models (16, 17, 18, 19). In addition, prior distributions (usually called hyperpriors) are needed for three types of model parameters in each model: (a) population effects or location parameters (e.g., population-level intercept and slopes), (b) standard deviations or scaling parameters for lower-level effects, and (c) various parameters such as the covariances in a variance-covariance matrix and the degrees of freedom in Student's $t$-distribution. Noninformative hyperpriors are adopted for population-level effects. In contrast, weakly-informative priors are utilized for standard deviations of lower-level parameters such as varying intercepts and slopes at the subject level, and such hyperpriors include a Student's half-$t(3, 0, 1)$ or a half-Gaussian $\mathcal{N}_+(0, 1)$ (a Gaussian distribution with restriction to the positive side of the respective distribution). For variance-covariance matrices, the LKJ correlation prior (Lewandowski et al., 2009) is used with the shape parameter taking the value of 1 (i.e., jointly uniform over all correlation matrices of the respective dimension). Lastly, the standard deviation $\sigma$ for the residuals utilizes a half Cauchy prior with a scale parameter depending on the standard deviation of the input data. The hyperprior for the degrees of freedom, $\nu$, of the Student's $t$-distribution is $\Gamma(2, 0.1)$. The consistency and full convergence of the Markov chains were confirmed through the split statistic $\hat{R}$ being less than 1.1 (Gelman et al., 2013). The effective sample size (or the number of independent draws) from the posterior distributions based on Markov chain Monte Carlo simulations was more than 200 so that the quantile (or compatibility) intervals of the posterior distributions could be estimated with reasonable accuracy.

# G Flanker Image acquisition and preprocessing

**Image Acquisition** Neuroimaging data were collected on a 3T GE Scanner using a 32-channel head coil. After a sagittal localizer scan, an automated shim calibrated the magnetic field to decrease signal dropout from a susceptibility artifact. Four functional runs, each consisting of 170 whole-brain (forty-two 3-mm axial

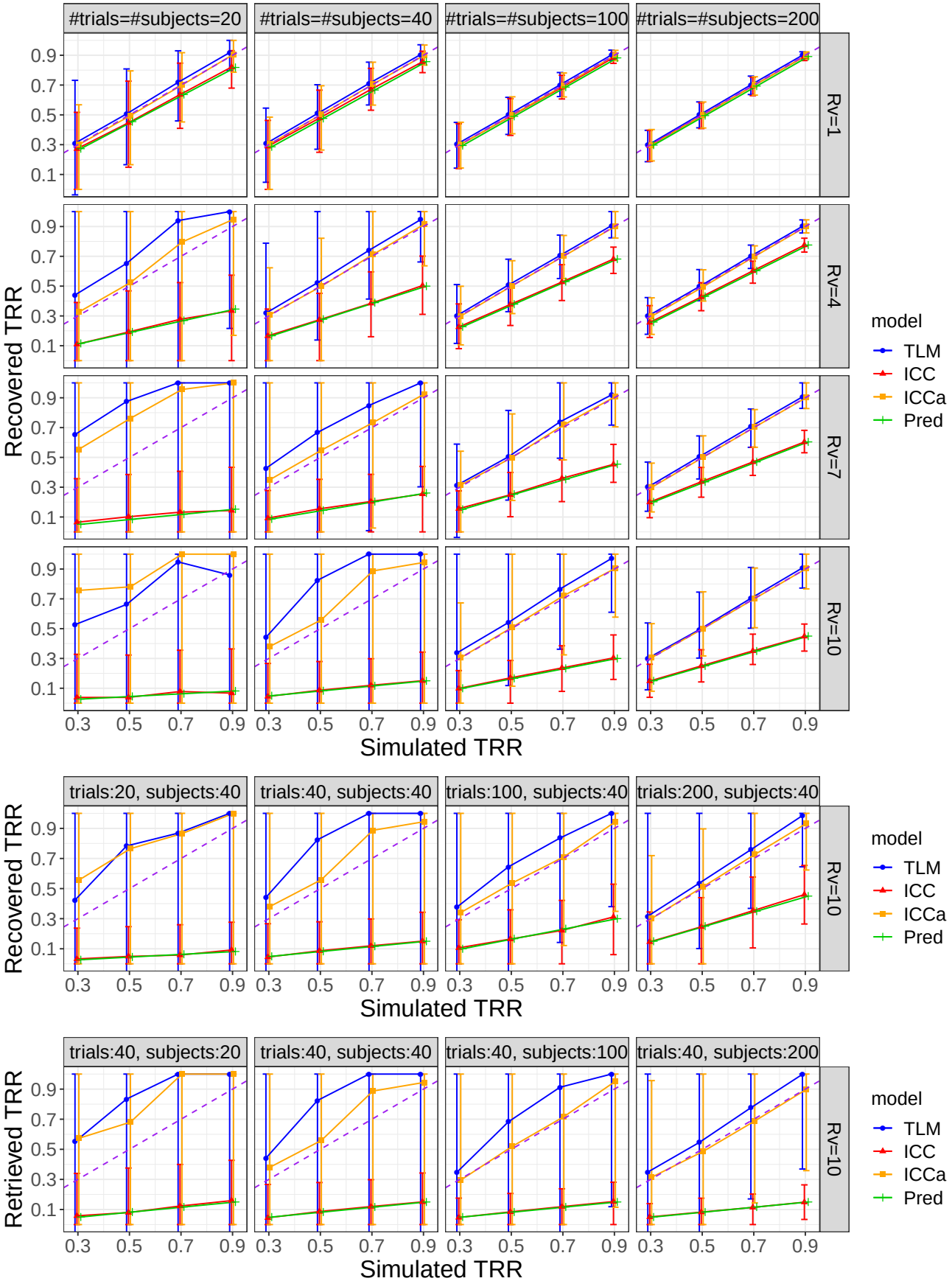



Figure 9: Simulation results for a contrast. The four columns correspond to the sample size of subjects and trials while the four rows are the varying standard deviation ratios of $\frac{\sigma_0}{\sigma_\lambda}$. The $x$- and $y$-axis are the simulated and recovered reliability, respectively. Each data point is the median among the 1000 simulations with the error bar showing the 90% highest density interval. The dashed purple diagonal line indicates an idealized retrieval of the parameter value based on the data-generating model (12). The green line shows the theoretically predicted ICC per the formula (6).

slices) T*2-weighted echoplanar images were acquired at the following specifications: TR = 2s, TE = 25, flip angle = 60°, 24 field of view, 96 × 96 matrix. The first 4 images from each run were discarded to ensure that longitudinal magnetization equilibrium was reached. A structural MPRAGE sequence in the sagittal direction was acquired for co-registration with the functional data using the following specifications: TI/TE = 425/min full, 1-mm slices, flip angle = 7°, 256 × 256 matrix.

**Image Preprocessing** Neuroimaging data were analyzed using AFNI version 20.3.00 (http://afni.nimh.nih.gov/afni/; Cox, 1996) with standard preprocessing including despiking, slice-timing correction, distortion correction, alignment of all volumes to a base volume with minimum outliers, nonlinear registration to the MNI template, spatial smoothing with a 6.5mm FWHM kernel, masking, and intensity scaling. Final voxel size was 2.5 × 2.5 × 2.5 mm. We excluded any pair of successive TRs in which the sum head displacement (Euclidean norm of the derivative of the translation and rotation parameters) between those TRs exceeded 1 mm. TRs in which more than 10% of voxels were outliers were also excluded. Participants' datasets were excluded if the average motion per TR after censoring was greater than 0.25 mm or if more than 15% of TRs were censored for motion or outliers. In addition, 6 head motion parameters were included as nuisance regressors in individual-level models.

## References

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67, 1–48.

Box, G.E.P., 1976. Science and Statistics. Journal of the American Statistical Association 71, 791–799.

Bürkner, P.-C., 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software 80, 1–28.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: A Probabilistic Programming Language. Journal of Statistical Software 76, 1–32.

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., ..., Orr, C. A. (2018). The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. Developmental cognitive neuroscience, 32, 43-54.

Chen, G., Saad, Z.S., Nath, A.R., Beauchamp, M.S., Cox, R.W., 2012. FMRI group analysis combining effect estimates and their variances. NeuroImage 60, 747–765.

Chen, G., Saad, Z.S., Britton, J.C., Pine, D.S., Cox, R.W., 2013. Linear mixed-effects modeling approach to FMRI group analysis. NeuroImage 73, 176–190.

Chen, G., Taylor, P.A., Haller, S.P., Kircanski, K., Stoddard, J., Pine, D.S., Leibenluft, E., Brotman, M.A., Cox, R.W., 2018. Intraclass correlation: Improved modeling approaches and applications for neuroimaging. Human Brain Mapping 39, 1187–1206.

Chen, G., Padmala, S., Chen, Y., Taylor, P.A., Cox, R.W., Pessoa, L., 2020. To pool or not to pool: Can we ignore cross-trial variability in FMRI? NeuroImage 117496.

Chen, G., Taylor, P.A., Stoddard, J., Cox, R.W., Bandettini, P.A., Pessoa, L., 2021. Dichotomous thinking and informational waste in neuroimaging. bioRxiv 2021.05.09.443246.

Cox, R.W. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages, Computers and Biomedical Research, 29: 162-73.

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

Eriksen, B. A., Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception and psychophysics, 16(1), 143-149.

Fröhner, J.H., Teckentrup, V., Smolka, M.N., Kroemer, N.B., 2019. Addressing the reliability fallacy in fMRI: Similar group effects may arise from unreliable individual effects. NeuroImage 195, 174–189.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis, 3rd Edition. ed. Chapman and Hall/CRC, Boca Raton.

Haines, N., Kvam, P.D., Irving, L.H., Smith, C., Beauchaine, T.P., Pitt, M.A., Ahn, W.-Y., Turner, B., 2020. Learning from the Reliability Paradox: How Theoretically Informed Generative Models Can Advance the Social, Behavioral, and Brain Sciences (preprint). PsyArXiv.

Hedge, C., Powell, G., Sumner, P., 2018. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behav Res 50, 1166–1186.

Lewandowski, D., Kurowicka, D., Joe, H., 2009. Generating random correlation matrices based on vines and extended onion method. Journal of Multivariate Analysis 100, 1989–2001.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. Psychological bulletin, 109(2), 163.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1, 30–46.

Noble, S., Scheinost, D., Constable, R.T., 2021. A guide to the measurement and interpretation of fMRI test-retest reliability. Current Opinion in Behavioral Sciences 40, 27–32.

Pinheiro, J., Bates, D., 2000. Mixed-Effects Models in S and S-PLUS, 1st ed. 2000. Corr. 3rd printing 2002 edition. ed. Springer, New York.

Ratcliff, R., 1979. Group reaction time distributions and an analysis of distribution statistics. Psychological Bulletin 446–461.

Rouder, J.N., Haaf, J.M., 2019. A psychometrics of individual differences in experimental tasks. Psychon Bull Rev 26, 452–467.

Rouder, J., Kumar, A., Haaf, J.M., 2019. Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail. PsyArXiv.

Shoukri, M., Asyali, M., Donner, A., 2004. Sample size requirements for the design of reliability study: Review and new results. Statistical Methods in Medical Research 13, 251–271.

Shrout, P. E., Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological bulletin, 86(2), 420.

Smith, A. R., White, L. K., Leibenluft, E., McGlade, A. L., Heckelman, A. C., Haller, S. P., ..., Pine, D. S. (2020). The heterogeneity of anxious phenotypes: neural responses to errors in treatment-seeking anxious and behaviorally inhibited youths. Journal of the American Academy of Child and Adolescent Psychiatry, 59(6), 759-769.

Viechtbauer, W., 2005. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. Journal of Educational and Behavioral Statistics 30, 261–293.

Wagenmakers, E.-J., Brown, S., 2007. On the linear relation between the mean and the standard deviation of a response time distribution. Psychological Review 114, 830–841.

Westfall, J., Nichols, T.E., Yarkoni, T., 2017. Fixing the stimulus-as-fixed-effect fallacy in task fMRI. Wellcome Open Res 1.

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis, Use R! Springer-Verlag, New York.

Woolrich, M.W., Behrens, T.E.J., Beckmann, C.F., Jenkinson, M., Smith, S.M., 2004. Multilevel linear modelling for FMRI group analysis using Bayesian inference. NeuroImage 21, 1732–1747.