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* (TRR) indexes the repeatability or consistency of a measurement across time. High TRR of measures is critical for any scientific study, specifically for the study of individual differences. Evidence of poor TRR of commonly used behavioral and functional neuroimaging tasks is mounting (e.g., Hedge et al., 2018; Elliot et al., 2020). These reports have called into question the adequacy of using even the most common, well-characterized cognitive tasks with robust population-level task effects, to measure individual differences. Here, we demonstrate the limitations of the intraclass correlation coefficient (ICC), the classical metric that captures TRR as a proportional variance ratio. Specifically, the ICC metric is limited when characterizing TRR of cognitive tasks that rely on many individual trials to repeatedly evoke a psychological state or behavior. We first examine when and why conventional ICCs underestimate TRR. Further, based on recent foundational work (Rouder and Haaf, 2019; Haines et al., 2020), we lay out a hierarchical framework that takes into account the data structure down to the trial level and estimates TRR as a correlation divorced from trial-level variability. As part of this process, we examine several modeling issues associated with the conventional ICC formulation and assess how different factors (e.g., trial and subject sample sizes, relative magnitude of cross-trial variability) impact TRR. We reference the tools of **TRR** and **3dLMEr** for the community to apply these models to behavior and neuroimaging data.

1 Introduction

The concept of test-retest reliability (TRR) 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; TRR 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, TRR is assessed through the statistical metric of the intraclass correlation coefficient (ICC). Thus, for definitional clarity, TRR in the current context is defined as a property of individual differences, and ICC is a conventional statistical measure of TRR.

Assessment of TRR 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

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

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 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). Under the conventional ICC formulation, 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, an ICC can be computed as the fraction of the total variance that can be attributed to inter-individual differences in the interference effect.

Assessment of TRR 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 now being reported appear unacceptably low for behavioral (around 0.5 or below; Hedge et al., 2018) and imaging tasks (less than 0.4; Elliot et al., 2020), casting doubt on their utility in studies of individual differences. For imaging tasks specifically, the extent to which common average contrast values 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 TRR is a critical requirement for these usages.

Two aspects of the Stroop task example above are noteworthy and typical in modern TRR 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/or ii) many trials are used as instantiations of each 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 TRR calculations via ICC. Modeling trial-level variance has not been widely practiced in neuroimaging. Ignoring cross-trial variations effectively means that all instantiations (i.e., trials) are presumed to have the same estimate, and precision information (i.e., variance underlying the condition estimate) is not considered at the population level. As we will demonstrate, trial-level variability is surprisingly large in both behavior and task-based imaging data, often dwarfing subject-level variability. This unmodeled variance will contaminate the subject-level variance on which conventional TRR estimates hinge. Specifically, we show that

1) conventional ICCs can substantially underestimate TRR;
2) a single ICC value is misleading due to its failure to capture the estimation uncertainty;
3) the degree of ICC underestimation depends on the number of trials and cross-trial variability;
4) the number of subjects has surprisingly little impact on ICC.

In addition, based on recent foundational work (Rouder and Haaf, 2019; Haines et al., 2020), we aim to

1) build a hierarchical model that more accurately characterizes test-retest data, which accounts for cross-trial variability and estimates TRR as a correlation;
2) show that this model identifies brain regions with TRR as high as 0.9, yet also highlights low precision of TRR estimation with few trial replicates;
3) illustrate that the number of trials is more important in improving TRR precision than the number of subjects;
4) make hierarchical modeling programs **TRR** and **3dLMEr** available for TRR estimation;

In the process of laying out a new hierarchical solution to estimate TRR, we examine modeling issues associated with the conventional ICC framework and highlight one question: what is the source of the large cross-trial variability - often several times the magnitude of cross-subject variability?

1.1 Classical definition of ICC

In all comparisons with 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 linear mixed-effects (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; τ_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 ICC computation, is typically adopted for behavioral and neuroimaging data analysis. The residual variability σ_e may appear to capture the within-subject cross-repetition variability; however, "hidden" variability remains embedded in σ_e, which causes a fundamental problem with the conventional ICC - as detailed later.

The classical ICC can be examined from two statistical perspectives. First, as the data variability under ANOVA/LME (2) through CLM is partitioned into two components, *cross-subject variability* $\widetilde{\sigma}_\tau^2$ and *within-subject variability* σ_e^2, the ICC formulation (3) directly reveals the amount of cross-subject variability relative to the total variability. Second, ICC can be viewed as the Pearson correlation of the subject-level effects between the two repetitions,

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

There are two aspects that make the ICC calculation a unique case under a correlational framework. First, unlike any other generic *inter*class correlation where the two variables are usually from two different classes (e.g., height and weight), ICC always involves two repeated measures in the same class (e.g., the same type of effects y_{1s} and y_{2s}). It is this sameness that renders the name *intra*class correlation. The other aspect is that homoscedasticity is implicitly assumed under the conventional ICC formulation 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 the generic situation of Pearson correlation in which the two variables are usually heteroscedastic.

Model	RT effect	Population Effects (ms)			Data Variability (ms)		VR (UR)	TRR
		repetition	mean	95% interval	subject	residual		
model (2)	congruent	session 1	639	(617, 660)	$\widetilde{\sigma}_\tau$: 64	σ_e: 39	4.1 (0.93)	0.72
CLM		session 2	607	(585, 629)				
	incongruent	session 1	720	(700, 744)	$\widetilde{\sigma}_\tau$: 71	σ_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	σ_e: 24	12.6 (0.61)	0.49
		session 2	59	(49, 69)				
LME (5)	congruent	session 1	639	(617, 660)	σ_{τ_1}: 71	σ_0: 300	-	0.78
TLM		session 2	607	(584, 629)	σ_{τ_2}: 74			
	incongruent	session 1	720	(693, 746)	σ_{τ_1}: 89	σ_0: 250	-	0.73
		session 2	666	(643, 689)	σ_{τ_2}: 77			
LME (12)	incongruent vs congruent	session 1	81	(71, 92)	σ_{τ_1}: 27	σ_0: 276	-	1.0
TLM		session 2	59	(50, 68)	σ_{τ_2}: 18			
	average of 2 conditions	session 1	679	(656, 703)	σ_{τ_1}: 79		-	0.74
		session 2	636	(614, 659)	σ_{τ_2}: 75			
BML (17)	incongruent vs congruent	session 1	36	(28, 42)	σ_{τ_1}: 7	σ_0: 75	-	0.71
exGaussian		session 2	27	(22, 33)	σ_{τ_2}: 7			
	average of 2 conditions	session 1	670	(656, 684)	σ_{τ_1}: 34		-	0.68
		session 2	643	(631, 659)	σ_{τ_2}: 28			

Table 1: 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 (VR, defined in formulas (8) and (14)) and underestimation rate (UR, defined in formulas (7) and (13)) are indicators for the degree of ICC underestimation.

Different model frameworks for the conventional ICC may have some subtle differences and limitations. For example, the ANOVA platform cannot handle missing data, confounding variables and sampling errors. As a result, extended models had been adopted to various scenarios (Chen et al., 2018). Nevertheless, the computation of conventional ICCs is relatively straightforward. Using "Study 1" of a publicly available dataset of Stroop effects (Hedge et al., 2018) as an example, we obtain a modest ICC(3,1) = 0.49 (Table 1) for the subtraction value between conditions even though the evidence for population-level effects is quite strong with a Stroop effect $a_1 = 81$ ms (95% uncertainty interval (72, 91)) and $a_2 = 59$ ms (95% uncertainty interval (49, 69)) during the first and second repetition, respectively. As a comparison, the two conditions individually show a much higher TRR with an ICC(3,1) of 0.72 and 0.69 for congruent and incongruent condition, respectively.

1.2 Separation between population effects and TRR

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 the specific sample effect to a hypothetical population. Population-level effects are captured through terms (usually called "fixed" effects under the conventional statistical framework) such as repetition 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 τ_s in the LME model (2).

Cross-subject variability is the focus in the TRR context. From the modeling perspective, the relationship between the population and subject level can loosely described as "crossed" or "orthogonal": the subject-specific effects τ_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.

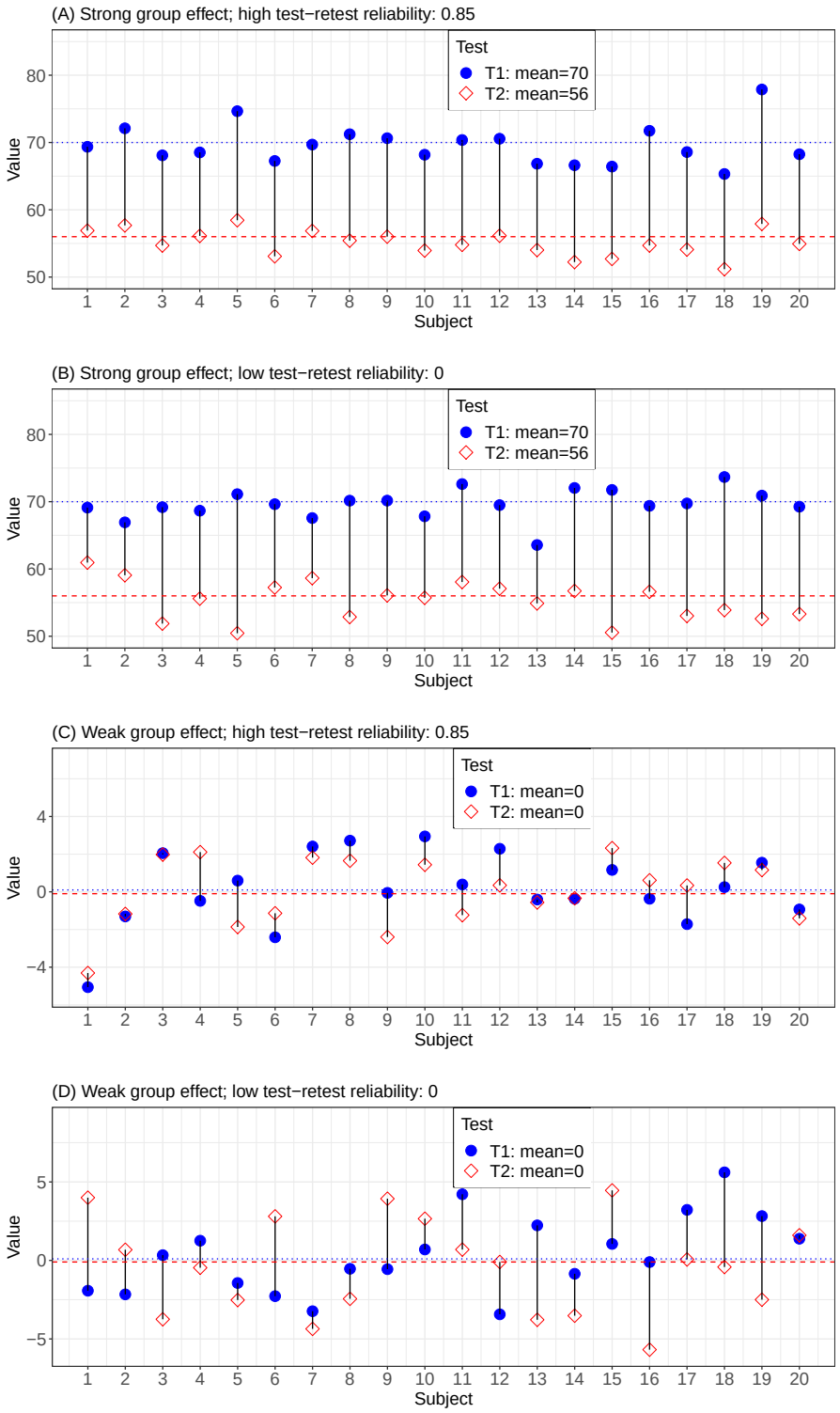


Figure 1: Separation between TRR and population-level effects. Four scenarios are illustrated to demonstrate that TRR is not necessarily tied to the strength of population effects. Hypothetical data were randomly drawn from a bivariate Gaussian distribution: 20 subjects completed a depression screening at two separate time points (Screening 1:T1, blue filled circle and Screening 2:T2, red empty diamond).The population effects are easy to observe (colored horizontal lines); it is harder to assess TRR over the two sessions. TRR 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, TRR 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) TRR. $\mathcal{N}(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix},\ 49\times\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix})$ 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) $\mu_1 = \mu_2 = 0,\ \rho = 0$.

It is crucial to recognize the dissociation between population effects and TRR. A popular misconception is that strong population-level effects are necessarily associated with high TRR as long as sample sizes are large. However, these two types of effects are not conceptually tied to one another as generally assumed. With a hypothetical example, Fig. 1 illustrates four extreme scenarios on a two-dimensional continuous space of population-level effects and TRR: one may have strong population-level effects with high (Fig. 1A) or low (Fig. 1B) TRR; contrarily, it is also possible to have weak population-level effects accompanied by high (Fig. 1C) or low (Fig. 1D) TRR.

1.3 Motivations for extending the conventional ICC

The conventional ICC formulation has been widely utilized. However, a few significant limitations exist with the classic ICC formulation (2). For instance,

1) **Difficulty of obtaining a measure of uncertainty**. In a conventional statistical framework, one usually obtains a ICC estimate through CLM under ANOVA/LME (2). In addition, statistical evidence can be assessed through either converting the ICC value using a Fisher transformation or an F-statistic (McGraw and Wong, 1996; Chen et al., 2018). However, there is no analytical solution to assess the range of uncertainty associated with the ICC estimate. Even though bootstrapping could be adopted to find the quantile interval, the approach is rarely utilized in practice due to its computational cost especially for large datasets in neuroimaging.

2) **Inflexibility to assumption violations and vulnerability to numerical instabilities** . The LME framework assumes a Gaussian distribution. When this assumption is violated (e.g., skewed data, outliers), parameter estimation through the optimization of a nonlinear objective function may become unstable or singular. For example, there may be no clear peak when the objective function is very diffusive, or the numerical solver may get stuck at boundaries (e.g., 0 or 1 for correlation) or a suboptimal peak.

3) **Inability to integrate measurement error**. Under specific circumstances, the input data under ANOVA/LME (2) through CLM may contain sampling errors. For instance, BOLD response as the effect of interest in neuroimaging for TRR is not directly collected, but instead estimated from a time series regression model. Therefore, it would be desirable to incorporate uncertainty information into the subsequent modeling process. However, no straightforward solution is available to achieve the integration through the ANOVA/LME platform.

4) **Inability to integrate cross-trial variability**. The trials that make up a condition are related to condition effect estimates the same way participants are related to population effects. Trials, as a dimension "orthogonal" to subjects, are expected to provide robust estimation for condition-level effects. Yet, at the same time, the trial factor is largely of no interest to the investigator, and thus, in practice, the data is typically collapsed across trials and further flattened across conditions as illustrated in the data reduction step (1). As a result, cross-trial variability does not have a place in the model and the associated uncertainty is neither properly accounted for nor propagated in the ANOVA/LME formulation (2) through CLM for ICC computation.

Some recent modeling endeavors address these limitations. For example, a few ICC variations solve the numerical instabilities and incorporate sampling errors; such methods have been implemented into whole-brain voxel-wise ICC computation through the program 3dICC in AFNI (Chen et al., 2018) for fMRI data analysis. More recently, Haines and colleagues (2020) have used a Bayesian multilevel (BML) framework to accurately characterize the data structure and TRR distribution instead of providing only a point estimate as in the classical ICC.

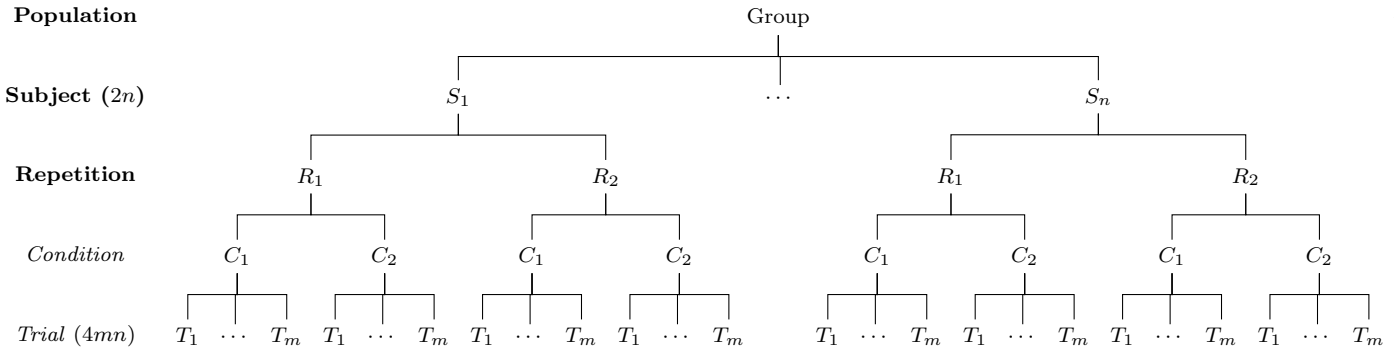


Figure 2: Hierarchical structure of test-retest data. Assume that, in a test-retest study with two repetitions, a group of n subjects are recruited to perform a task (e.g., Stroop) of two conditions (e.g., congruent and incongruent), and each condition is instantiated with m trials. The collected data 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.

The root cause of many issues with the conventional ICC calculation is that the collapsing at the trial and condition level loses the hierarchical integrity of the data structure. The current work extends a previous exploration on trial-level modeling (TLM) (Chen et al., 2020) and is similarly based on the recent significant advances in statistical modeling under a BML framework (Rouder and Haaf, 2019; Haines et al., 2020). Data reduction, as inherent in the step (1) of cross-trial averaging and cross-condition subtraction, is known to produce information loss and potential effect distortions. Specifically, five levels are involved in a TRR dataset and form a hierarchical structure (Fig. 2): population, subject, repetition, condition and trial. The conventional ICC formulation focuses only on the three top levels (population, subject and repetition) and collapses the two lower levels (condition and trial) through averaging across trials and subtraction between the two conditions. Our investigation will maintain the integrity of the hierarchical structure across all five levels. In addition, we will

(a) expand the CLM through ANOVA/LME (2) to explicitly account for cross-trial variability;
(b) assess the extent of underestimation by the conventional ICC;
(c) perform simulations to quantitatively expound the fundamental flaws involved in the conventional ICC;
(d) use a Flanker fMRI dataset with both behavior and neuroimaging data to demonstrate the improved TRR assessment; and
(e) discuss the insights gained from our simulations and modeling applications.

2 Methods: assessing TRR through trial-level modeling

2.1 LME framework for a single condition effect

To share the conceptual evolution and modeling progression, we start with simply accommodating trial-level effects in test-retest reliability estimation. We first focus on a single condition (e.g., congruent or incongruent) before extending the modeling framework to contrasts. The linear mixed-effects (LME) modeling platform is conceptually familiar to most analysts and computationally feasible in terms of numerical simulations; thus, we initially adopt the LME formulation before moving to a BML framework. As opposed to the common practice of acquiring the condition-level effect estimate at the subject level, we obtain the trial-level effect estimates y_{rst} of the condition (Chen et al., 2020), where r, s and t index repetitions, subjects and trials ($r = 1, 2;\ s = 1, 2, ..., n;\ t = 1, 2, .., m$).

We expand the CLM-based LME model (2) and directly 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, τ_{rs} characterizes the subject-level effects during the r-th repetition, σ_0 captures the cross-trial variability, and $\boldsymbol{R}$ is the variance-covariance matrix for the subject-level effects τ_{rs} between the two repetitions. Usually a_r are termed as the population-level intercepts (or "fixed effects" in LME terminology) while τ_{rs} are the varying intercepts (or "random effects") across subjects.

The parameter ρ captures the correlation between the two repetitions for the subject-level effects τ_{1s} and τ_{2s}; thus, ρ represents the TRR with trial-level variability σ_0 directly incorporated into the model structure and explicitly separated from the TRR metric ρ. The inclusion of cross-trial variability σ_0 into the LME model precludes formulating TRR as a variance ratio as traditionally done in the (3). The TRR estimates based on the LME model (5) are shown in Table 1 for the Stroop task dataset. For neuroimaging data analysis, the program **3dLMEr** (Chen et al., 2013) in AFNI can be utilized to compute whole-brain voxel-wise TRR through the TLM-based LME formulation (5).

The conventional ICC tends to underestimate TRR of a single condition estimate. TRR aims to quantify subject-level similarity between two repetitions; the correlation coefficient ρ reflects TRR with cross-trial fluctuations removed from the TRR formulation. The crucial question becomes: How does the conventional ICC compare with the new TRR formulation? To directly compare the conventional ICC with the TRR estimation through the TLM-based LME formulation (5), 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),

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

where the underestimation rate

$$U = \frac{1}{1 + \frac{1}{m}V^2} \tag{7}$$

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

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

Two aspects of the ICC underestimation are noteworthy. First, the trial sample size m plays a crucial role; specifically, the degree of underestimation decreases with the trial sample size. In contrast, the subject sample size n on average does not impact the degree of underestimation, which might be counter-intuitive. Second, the extent of underestimation also depends on the variability ratio 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 $V = \frac{\sigma_0}{\sigma_\tau} \ll \sqrt{m}$, the ICC underestimation is negligible: $U = \frac{1}{1+\frac{1}{m}V^2} \approx 1 - \frac{1}{m}V^2 \approx 1$.

The traditional ICC estimate could, hypothetically, be corrected. If the variability ratio V were known under the CLM-based ANOVA/LME framework (2), one could adjust the ICC formulation (3) 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 σ_e are the subject-level and residual variability under the conventional ICC formulation (2) while σ_0 represents the cross-trial variability under the TLM-based LME model (5). However, the adjusted ICCa (9) is of little practical use because the cross-trial variability σ_0 is part of the residual variability σ_e and remains hidden from the analyst. To be able to estimate σ_0, one would have to resort to the TLM-based LME formulation (5); under that circumstance, one might as well directly obtain TRR through TLM rather than going back to CLM to make the adjustment.

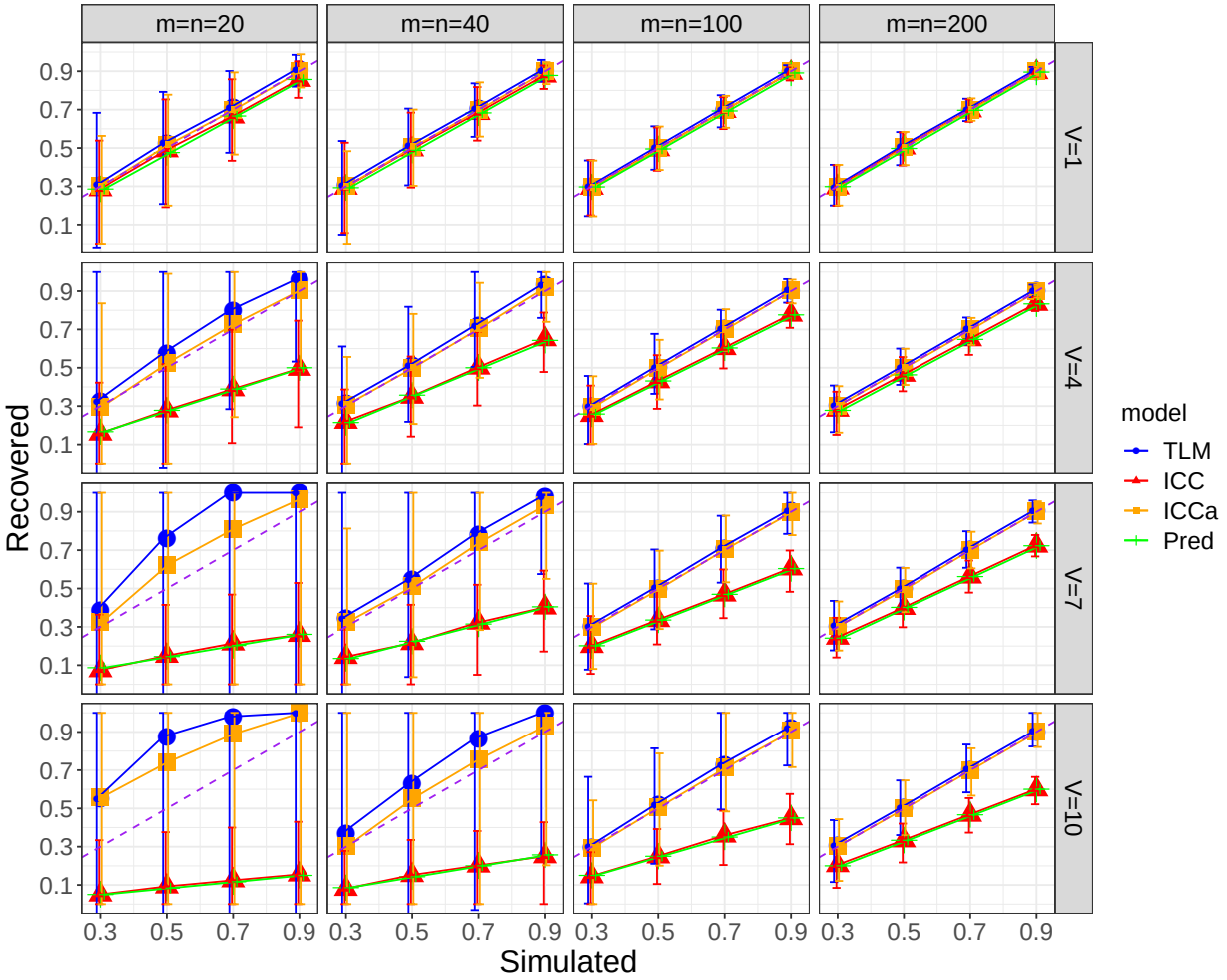
The TRR underestimation via the conventional ICC can be illustrated on the experimental data from the Stroop task (Hedge et al., 2018). We estimated TRR for each of the two conditions (congruent/incongruent) with their conventional ICC(3,1) values of 0.72 and 0.69 for congruent and incongruent condition, respectively (Table 1). Applying the TLM-based LME model (5), we obtained a TRR estimate of 0.78 and 0.73 for each condition, respectively, revealing only a slight underestimation with the conventional ICC. Population effects as well as cross-subject and cross-trial variations, σ_{τ_r} and σ_0 were also estimated. With σ_0 entered into the adjusted formula (9), we achieved the adjusted TRR estimates of 0.78 and 0.72 for the two conditions that are largely consistent with the results from their TLM-based LME estimates. The ICC underestimation was mild with $U = 0.93$ or 0.96 due to a large trial sample size ($m = 240$) and a moderate variability ratio of VR = 4.1 or 3.0.

2.2 Assessing model comparisons through simulations

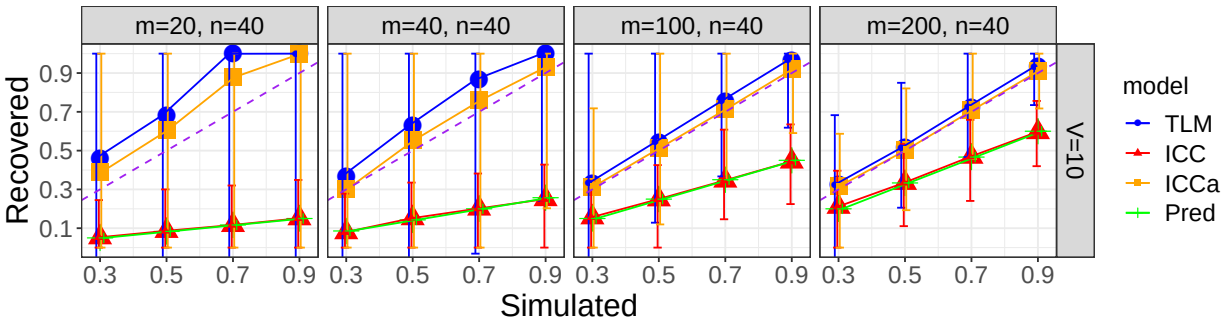
To comprehensively compare the two modeling frameworks for a single condition effect, we use simulations. The numerical schemes regarding the simulations are presented in Appendix B with results shown in Fig. 3. The population-level effects a_r under both CLM and TLM were robustly recovered (not shown). Findings in regard to TRR are summarized below.

1) **ICC underestimation is confirmed.** The amount of underestimation, as shown in the expression (7), depends on two factors: trial sample size m (Fig. 3B) and variability ratio V. ICC can provide reasonable TRR estimation under special circumstances but substantially underestimates TRR when cross-trial variability is much larger than cross-subject variability ($V = \frac{\sigma_0}{\sigma_\tau} \gg 1$). When $V \lessapprox 1$ (top row, Fig. 3A), the simulated TRR values were successfully recovered for both formulations of the conventional ICC and TLM-based LME. When $V > 1$ (second to fourth row, Fig. 3A), the larger the ratio V, the larger the underestimation by the ICC. This observation via simulations is consistent with the experimental results (Table 1) for the Stroop task data. In addition, a linear relationship characterizes the degree of ICC underestimation (red triangles) with simulated TRR values, as theoretically predicted through the attenuation rate U in (6) (green lines, Fig. 3A,B,C).
2) **ICC underestimation could be adjusted.** Once the cross-trial variability component $\frac{\sigma_0^2}{m}$ is explicitly accounted for in the conventional ICC formulation as shown in (9), the adjusted ICCa performs even better than the TLM-based LME model across all settings. In fact, 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 underestimation of the ICC (green line) is the same regardless of the number of subjects. However, as the subject number n increases, the precision of the TRR estimate based on both the conventional ICC and the LME through TLM improves.
4) **It can be challenging to attain a high precision of TRR estimation.** The uncertainty of TRR estimation is monotonically related to TRR magnitude, variability ratio V, as well as the sample sizes

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



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



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

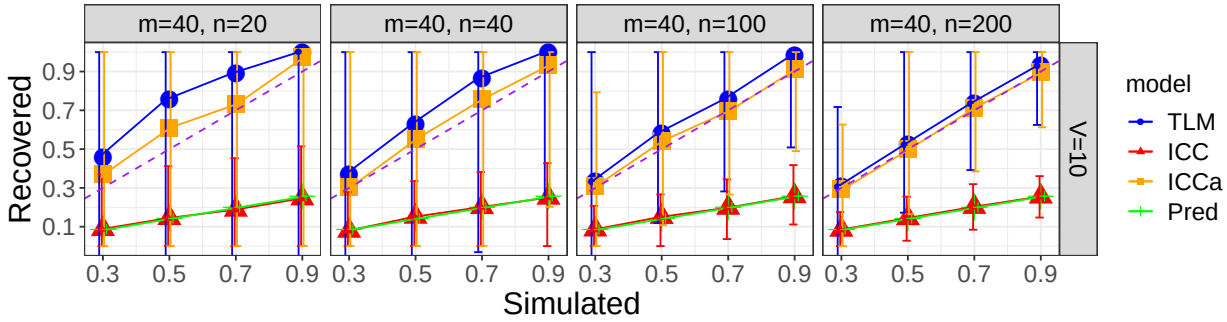


Figure 3: Simulation results for a single condition. The four columns correspond to the sample size of subjects and trials while the four rows code the variability ratio $V = \frac{\sigma_0}{\sigma_\tau}$. The x- and y-axis are the simulated and recovered TRR, respectively. Each data point is the median among the 1000 simulations with the error bar showing the 90% highest density interval. The dashed purple diagonal line indicates a perfect retrieval.

of m and n for subjects and trials. Specifically, uncertainty reduces as (a) the sample size of trials or subjects increases, (b) TRR increases, or (c) the variability ratio $V = \frac{\sigma_0}{\sigma_\tau}$ becomes smaller. Even though sample size influences TRR precision, the trial sample size m plays a more substantial role than the subject sample size n. As the sample size increases (from first to fourth column, Fig. 3), the error bars for TRR estimates narrow. In addition, the precision pattern with a fixed number of subjects n but varying trial sample size m (Fig. 3B) is similar to the one with a fixed number of trials m but varying subject number n (Fig. 3C). Yet, between the two factors, the impact of the trial number m is much larger than the impact of the subject number n; plus, the trial number m impacts on ICC underestimation (Fig. 3B) while the subject number m does not on average (Fig. 3C). The simulations also indicate that a large sample size (e.g., 100 or more when $V = \frac{\sigma_0}{\sigma_\tau} \geq 10$) may be required to achieve a TRR estimate with a reasonable precision.

5) **TRR is dissociated from population-level effects.** Between the two scenarios for our simulations, $(a_1,\ a_2) = (0,\ 0)$ and $(1.0,\ 0.9)$, the first simulates a scenario with no population effects while the latter simulates a scenario with strong population effects with high certainty. Across both scenarios, simulation parameters were successfully recovered from the TLM-based LME formulation (5) and the two scenarios rendered very similar TRR patterns (only the case with $(a_1,\ a_2) = (0,\ 0)$ is shown in Fig. 3), confirming the dissociation between population effects and TRR.
6) **LME modeling is susceptible to numerical instability.** The performance of the TLM-based LME formulation is acceptable under some circumstances but suffers from numerical failures under others. Slight overestimation of TRR occurs when cross-trial variability is much larger than cross-subject variability ($V = \frac{\sigma_0}{\sigma_\tau} \gg 1$) (second to fourth row, Fig. 3A): the larger the variability ratio V, the larger the degree of overestimation. Close examination revealed that the overestimation was caused by the LME model with an almost or near singular fit when the numerical optimizer gets trapped at the boundary of $\rho = 1$ with the convergence failure leading to a TRR estimate of 1.0. When $V > 10$, numerical instability increases substantially, which is consistent with the behavioral data of Stroop task in Table 1.

2.3 LME framework for a contrast between two conditions

Now we extend the single condition effect case to the more common scenario of a contrast between two conditions under the LME framework. 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. A direct LME formulation can be specified as

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

where μ_{crs} is the s-th subject's effect under the c-th condition during the r-th repetition, and σ_0 captures the within-subject within-repetition cross-trial variability. When the condition contrast is of interest, we would have to resort to a derivable approach to parameterizing the subject-level effects μ_{crs}. Many different methods exist for factor parameterization; we opt to dummy-code 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 μ_{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 $(\lambda_{1s},\ \lambda_{2s})^T$. The assumption of independence between two parameter sets is discussed in Appendix C.

The correlation ρ_1 embedded in the matrix $\boldsymbol{R}^{(1)}$ is the TRR 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 λ_{rs}) under the LME framework (12). Specifically, b_r is the population-level condition contrast while λ_{rs} codes the subject-specific contrast effects. In parallel to the single condition effect in (5) with the intercept interpretation, here a_r and τ_{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] The variance-covariance matrix $\boldsymbol{R}^{(1)}$ characterizes the relationships of subject-level contrasts λ_{rs} between the two repetitions. In addition, $\boldsymbol{R}^{(1)}$ has a similar structure to the structure $\boldsymbol{R}$ in (5) for the single condition effect; thus, similarly, TRR can be retrieved from the LME model (12) as the correlation ρ_1 between the two varying slopes λ_{1s} and λ_{2s}. For neueroimaging data analysis, whole-brain voxle-wise TRR for condition 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 σ_0 in the LME model (12), one may specify a 2×2 variance-covariance matrix between the two sessions as in Haines et al. (2020) with another parameter added to capture the TRR in the residuals between the two sessions. One could even further argue that another correlation might be added to account for the relationship between the two conditions in the likelihood distribution, resulting in 4×4 block-diagonal variance-covariance matrix. However, the reason such a variance-covariance structure might occur is mostly due to suboptimal accountability of potential effects in the hierarchical data structure. That is, if the model is reasonably specified, one would not even need to account for this type of structure. Rather, it might be worth tuning and comparing models as alternatives to improve model fit.

An extra bonus of the LME framework for a condition-level contrast is the availability of TRR for the condition average as a byproduct. That is, the correlation ρ_0 embedded in the variance-covariance matrix $\boldsymbol{R}^{(0)}$ for cross-subject varying intercepts τ_{rs} captures the TRR for the average effect. It is worth noting that ρ_0 was assumed to be 0 in Haines et al. (2020). First, the LME formulation (5) with a single condition effect can be considered as a special case of the LME model (12) - equivalent to a condition contrast with the effects from the two conditions being identical. This equivalence would not be true for the model adopted by Haines et al. (2020). Furthermore, the LME framework (12) is more generic and consist between both intercept and slope effects in the sense that TRR is assumed to exist for each of the two conditions, their contrast as well as their average effect. In contrast, the assumption of $\rho_0 = 0$ as in Haines et al. (2020) leads to a degenerate and inadequate situation where TRR is only assumed to exist for the condition contrast but not for each of

[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 serves as the reference condition. Under dummy coding, the slopes in the LME model (12) still correspond to the condition contrast; however, the intercepts are associated with the reference condition.

the two condition nor for their average. To put it differently, the assumption of no TRR with $\rho_0 = 0$ for the average effect equates to capturing interactions without accounting for the associated main effects under an ANOVA framework.

How much cross-trial variability would impact the conventional 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 condition contrast is updated from the single effect case (7) to (Appendix D)

$$U = \frac{1}{1 + \frac{2}{m}V^2} \tag{13}$$

with the variability ratio V similarly defined as before,

$$V = \frac{\sigma_0}{\overline{\sigma}_\lambda}. \tag{14}$$

Likewise, one could adjust the original ICC by removing the cross-trial variance from the denominator (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 occurrence of the number 2, relative to their counterparts (7) and (9) for the single condition case, is due to the double amount of data involved in the cross-trial variability a condition contrast.

Underestimation of contrast TRR via the conventional ICC was once again validated with experimental Stroop data (Table 1). First, the conventional ICC(3,1) value for the contrast was estimated as 0.49 while a TRR estimate of 1.0 was retrieved due to a singularity problem when the parameter ρ_1 numerically degenerates at the boundary. Population effects as well as cross-subject and cross-trial variation, σ_{τ_r} and σ_0 were also estimated. With $\sigma_0 = 0.276$, we adjusted the TRR to be 1.18 per (15). Such an uninterpretable value was due to the fact that the adjusted cross-trial variability $\frac{2}{m}\sigma_0^2 = 6.35 \times 10^{-4}$ was larger than $\sigma_e^2 = 5.53 \times 10^{-4}$, another incidence of numerical singularity issues when estimating TRR through the TLM-based LME in (12) and potentially indicates that assumptions were violated. Nevertheless, the nearness to the parameter boundary is an indication of TRR close to 1, showing the substantial underestimation of the conventional ICC.

Simulations can also be adopted to explore same aspects for the condition contrast case as in a single condition. The degree to which the conventional ICC underestimates TRR for a condition contrast is more severe than for either of the two conditions alone or for the average effect between the two conditions. The extra number of 2 in the underestimation formula in (13) for a condition contrast indicates that cross-trial variability V would result in more severe ICC underestimation for a contrast than a single effect. Furthermore, the magnitude of the difference between conditions is much smaller than that of either condition or the average effect (see the mean column in Table 1). This results in a larger variability ratio V for the contrast, which leads to more severe underestimation. As shown in Table 1, the variability ratio V for the contrast is about three times larger than the ratio for the single condition effects. With a similar but more complex parameter domain, simulation results are presented in Appendix E. Observation patterns were largely similar to those for a single condition effect.

2.4 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 Stroop data, the LME framework is ill-suited for TRR estimation. Several of the motivations for extending the conventional ICC are not helped in this framework: precision information

or TRR distribution unavailable, violation of distribution assumption (e.g., Gaussian), no accommodation of sampling errors, and numerical failures. For our final model, the LME formulations (5) and (12) are integrated into the BML framework. Specifically, with a Gaussian likelihood as an example, we convert the two LME models with little modification to their distributional counterparts of a single effect,

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

and a contrast between two conditions,

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

The BML models can be further extended to solve all issues associated with the LME framework. For example, numerical instabilities and convergence issues would be largely dissolved through modeling improvements such as 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, TRR estimation for ρ in (16), ρ_0 and ρ_1 in (17) can be expressed by its whole posterior distributions. 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 TRR 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,

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

and

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

Lastly, the Gaussian assumption under BML for the response variable can be adaptively relaxed to a large family of distributions such as exGaussian, zero-inflated negative binomial, etc.

Using the Stroop task data, we illustrate advantages of the BML framework. Three likelihood distributions including Gaussian, log-normal and shifted log-normal were applied to the data in Haines et al. (2020) with a conclusion that the shifted log-normal distribution performed the best with the log-normal model a close second. We added two more distributions, Student's t and exGaussian, because of their ability of handling skewed data as well as outliers. In addition, instead of assuming a diagonal matrix $\boldsymbol{R}^{(0)}$ for the variance-covariance structure of the varying intercepts τ_{rs} as in Haines (2020), we adopted a generic $\boldsymbol{R}^{(0)}$ in the BML

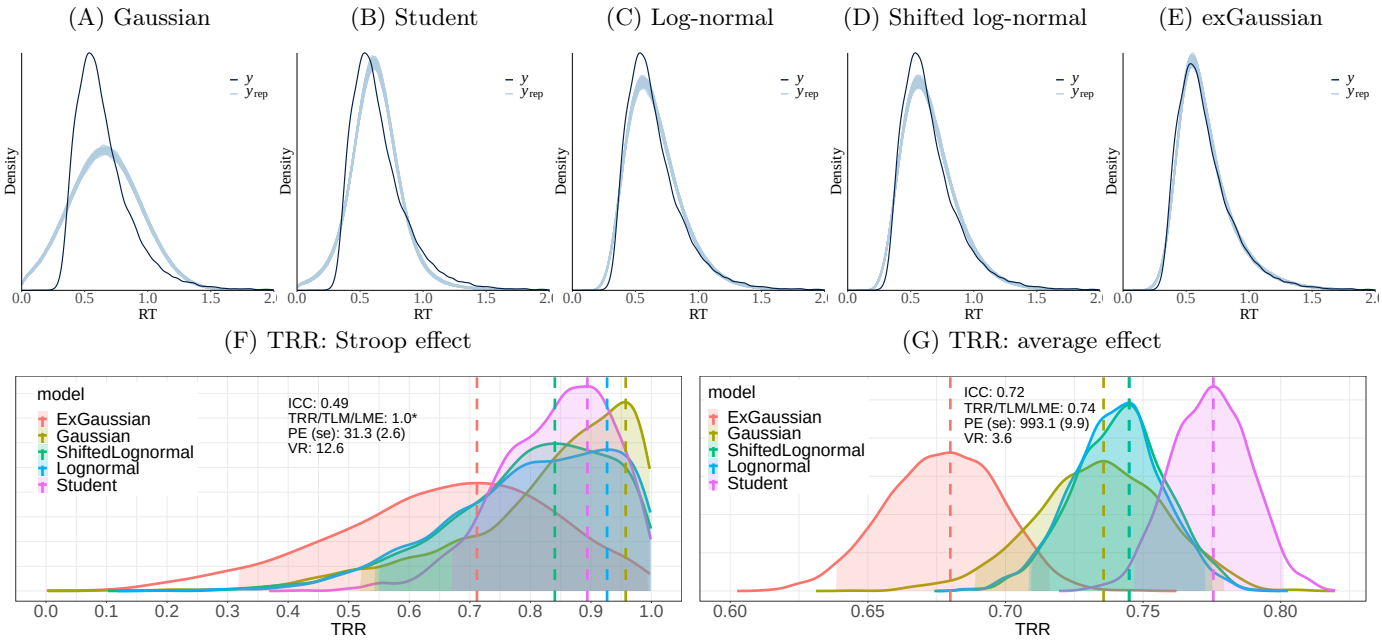


Figure 4: Comparing between five potential RT distributions to characterize the Stroop effect. (A-E) Visual comparisons are illustrated with each of the five panels showing the posterior predictive density (light blue line) that is composed of 200 sub-curves each of which corresponds to one draw from 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) TRR distributions for the Stroop effect in Study 1 from Hedge et al. (2018) are shown for the five likelihood distributions. The TRR estimates were based on 2000 draws from MCMC simulations for each BML model and are shown here as a kernel density that smooths the posterior samples. The dashed vertical line indicates the mode (peak) of each TRR distribution. Gaussian likelihood rendered the highest TRR with a mode of 0.96 but also exhibited the poorest fit to the data. The exGaussian distribution provided the lowest TRR with a mode of 0.71 while achieving the best fit among the five distributions. (G) TRR distributions for the average effect between congruent and incongruent conditions in Study 1 of Hedge et al. (2018) are shown for the five likelihood distributions. Unsurprisingly, their uncertainty is much narrower than the Stroop contrast effect (F).

model (17), leading to largely consistent results with those reported by Haines et al. (2020) with regard to the three common distributions. However, the added exGaussian outperformed all alternatives per the information criterion through leave-one-out cross-validation. The predictive accuracy of the exGaussian model was further confirmed by its best posterior predictive check among the five options (Fig. 4).

The modeling capability of the Bayesian framework can be illustrated through model comparisons. The success of the conventional LME framework heavily relies on the assumption of a Gaussian distribution as a prior, due to its nice properties such as convenient inference making through standard statistics such as Student's t. However, this convenience comes at a cost: when the prior Gaussian is ill-adapted, one may sacrifice accuracy of estimates. In the Stroop dataset, the TRR estimation based on Gaussian prior had the worst fit among the five priors (Fig. 4A-E), as also reported by Haines et al. (2020). In contrast, the exGaussian prior clearly outperformed other options due to its well-known capability in handling skewed and outlying data such as reaction time which is lower-bounded. This is important considering given how variable data cleaning practices are specifically regarding outlying values. The superior performance of exGaussian over the other three distributions (Gaussian, log-normal and shifted log-normal) considered in Haines et al. (2020) illustrates that it might be equally important to fine-tune the model through, for example, prior distributions as opposed to specifying a more complex variance-covariance structure as in Haines et al. (2020) (cf., our adoption of a single dispersion parameter σ_0 for the likelihood).

The rich information from Bayesian modeling is worth noting. The TRR estimation based on the BML model (17) with ExGaussian, together with other distributional assumptions, is presented in Fig. 4. Rather than a simple point estimate under the conventional framework, we empirically construct the posterior distribution for a parameter of interest through Monte Carlo simulations. The singularity problem (Table 1) that

we encountered with the LME model (12) was not an issue with the BML model (17). With a mode of 0.71 and a 95% highest density interval of [0.32, 0.99] for TRR, it is clear that the conventional ICC underestimated TRR (ICC = 0.49). The flexibility of exGaussian also likely provided a more accurate characterization of population-level effects and precision than, for example, the Gaussian assumption (Table 1).

We provide the program **TRR** for test-retest reliability estimation under BML. The BML models for a single condition effect (16, 18) and for a condition contrast (17, 19) are implemented into the program **TRR** through Markov Chain Monte Carlo (MCMC) simulations using Stan (Carpenter et al., 2017) through 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** is publicly available as part of the AFNI suite and can be used to estimate TRR for behavior and region-level neuroimaging data. Runtime ranges from minutes to hours depending on the amount of data.

3 BML modeling of TRR 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 ± 6.36) and 18 youth (<18 years; age: 14.01 ± 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 ± 11.8 days apart. Participants were asked to identify, via button press, the direction a center 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 of the trials 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 brain activity 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 commission errors in the incongruent condition (≥ 20) to be included in the error contrast. For the conflict contrast, there were a total of 32005 observations across two sessions of the Flanker task, which corresponds to approximate 190 trials per condition per session (350 ± 36 incongruent trials and 412 ± 19 congruent trials across sessions) per subject. For this subset of participants included into the error contrast, there were a total of 11366 observations available across both sessions, which corresponds to 331 ± 28 incongruent correct trials and 90 ± 27 incongruent commission errors per condition per subject. We analyzed the subject-level fMRI data at the whole-brain level as well as in a region-based approach using 12 Regions-Of-Interest (ROIs). We compare two 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 independent ROIs were selected using Neurosynth term-based meta-analyses using the terms "cognitive control" and "error", the two main population level effects of interest. Additionally, in order to derive ROIs outside of the main condition-level effects, we also selected peak coordinates from term-based meta-analyses for the term "visual" and "default mode". In order to select a reasonable number of peak coordinates, all four z-valued 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).

3.2 TRR estimation for behavioral data

The RT data from the Flanker task are structured as follows: There were two overlapping subsets of data representing 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.

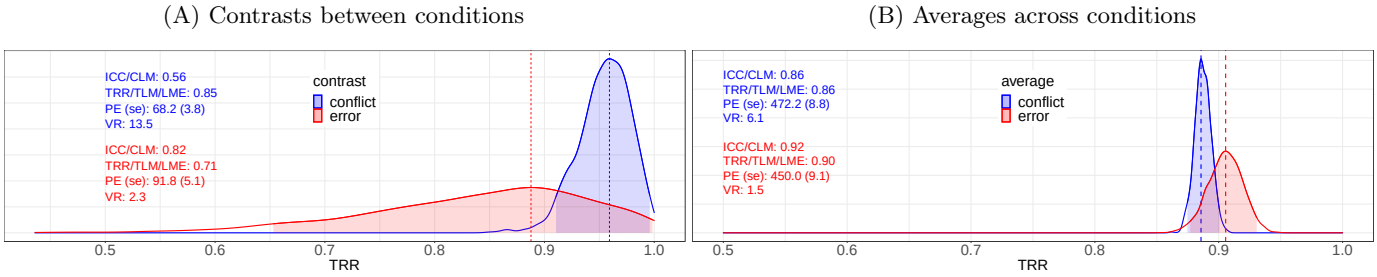


Figure 5: TRR distributions of behavioral data (RT) from Flanker task. The TRR estimates for the contrast (A) and average (B) between the two conditions were based on 2000 draws from MCMC simulations of the BML model with an exGaussian likelihood and are shown here as a kernel density estimate which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

The TRR estimates for the RT data of the Flanker task were relatively high.[2] As expected, the average RT effects rendered more precise estimates, and there was more convergence between conventional ICC and BML for average estimates. The TRR estimates for the RT contrast data 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 (17): Gaussian, Student's t, log-normal, shifted log-normal and exGaussian. Model comparisons and validations for each of the two contrast RT (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 TRR for the conflict contrast, estimated as the mode of the TRR distribution, fell 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 in the BML. However, for the error contrast, the BML revealed that there was larger uncertainty about the TRR estimation (Fig. 5). The point underestimates based on the conventional ICC were largely negligible except for the contrast of conflict, as indicated by the corresponding variability ratio.

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

3.3 Whole-brain TRR estimation for neuroimaging data

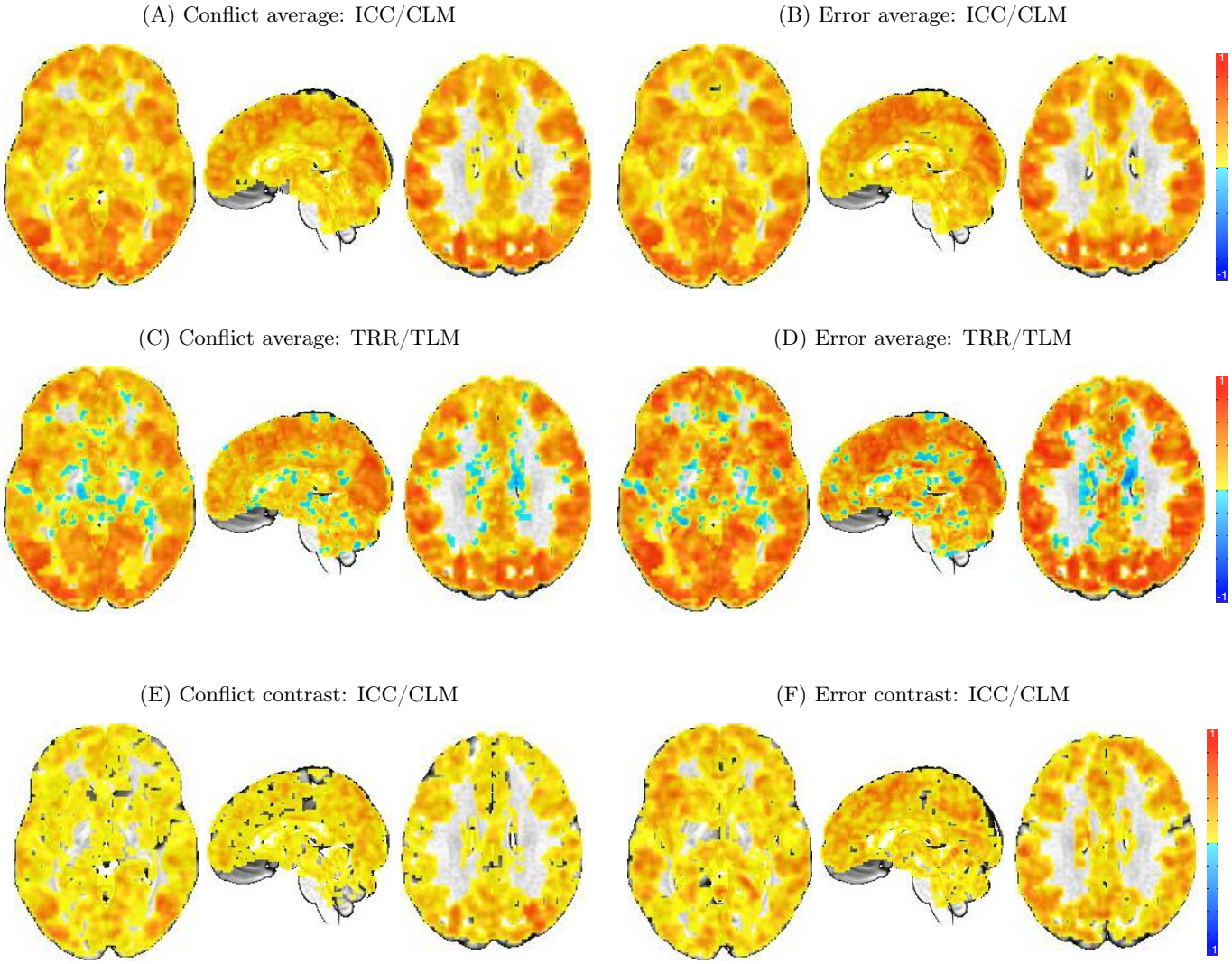


Figure 6: Whole-brain voxel-level TRR estimates for the fMRI Flanker dataset. The conventional ICC values for the average conflict effect (A) showed negligible underestimation compared to the TRR estimated based on TLM (C). In contrast, the conventional ICC values for the average error effect (B) showed a moderate amount of underestimation compared to the TRR estimation (D) based on TLM. The ICC values for the conflict (E) and error contrast (F) were much smaller than their average counterparts. The TLM-based approach numerically failed at most voxels in the brain and is thus not pictured. 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 template space. A small proportion of negative TRR (C,D) shown in white matter and CSF was because of no constraint on the correlation value in the variance-covariance structure. In contrast, the conventional ICC values are lower-bounded by 0 per its definition as a variance ratio. The ICC estimation was performed using 3dICC while the TLM-based TRR estimation was obtained through **3dLMEr**.

As the BML model is not computationally feasible at the whole brain voxel-level, two modeling frameworks were adopted: conventional ICC using the program 3dICC with condition-level modeling (2) and the LME model (12) using the program **3dLMEr** with trial-level modeling. There were a total of eight analyses with each of the two models applied to the two effects (the average and the contrast between the two conditions) for conflict and error data respectively. The input for the ICC computation was composed of a condition-level contrast per session from each of the 42/27 subjects. The trial-level input for TRR estimation had the same structure as the RT data: 32005 and 11366 three-dimensional volumes of trial-level effects for conflict and error subsets, respectively, from two conditions, two sessions and 42/27 subjects.

Individual differences were largely reliable among most voxels for the average effect across both conditions as estimated via the CLM/ICC. Similarly, TRR was in the moderate to high range for the average effect for both conditions when applying the trial-level model in the LME framework (Fig. 6, C vs D). The underes-

timation via the conventional ICC metric (Fig. 6, A vs C) was largely negligible for the average of conflict responses, while small but noticeable for the average of the error responses (Fig. 6, B vs D). For contrasts, ICC estimates were lower, largely below an ICC of 0.4 for both contrasts (with the exception of primary visual, parietal and motor regions). Higher ICCs were noted for the error than the conflict contrast in several regions. The difference in ICC underestimation between conflict and error contrasts is likely due to the larger number of trials for the conflict compared to the error data (i.e, more precise estimates). Based on the ICC results (Fig. 6E,F), the reliability of individual differences for the conflict and error between-condition contrast would be considered poor to adequate, depending on region. Unlike for the average effect (Fig. 6C,D), numerical failures occurred for most voxels in the brain for the contrast analyses and thus results are not shown). Therefore, we explore the comparison between ICC/CLM and TLM on the region level via the BML.

3.4 Region-level TRR estimation for neuroimaging data

TRR 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 so that any potential outlying values can be accounted for (Chen et al., 2020). In addition, the standard errors of the trial-level effects from the subject level were also incorporated into the BML model to improve modeling 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.

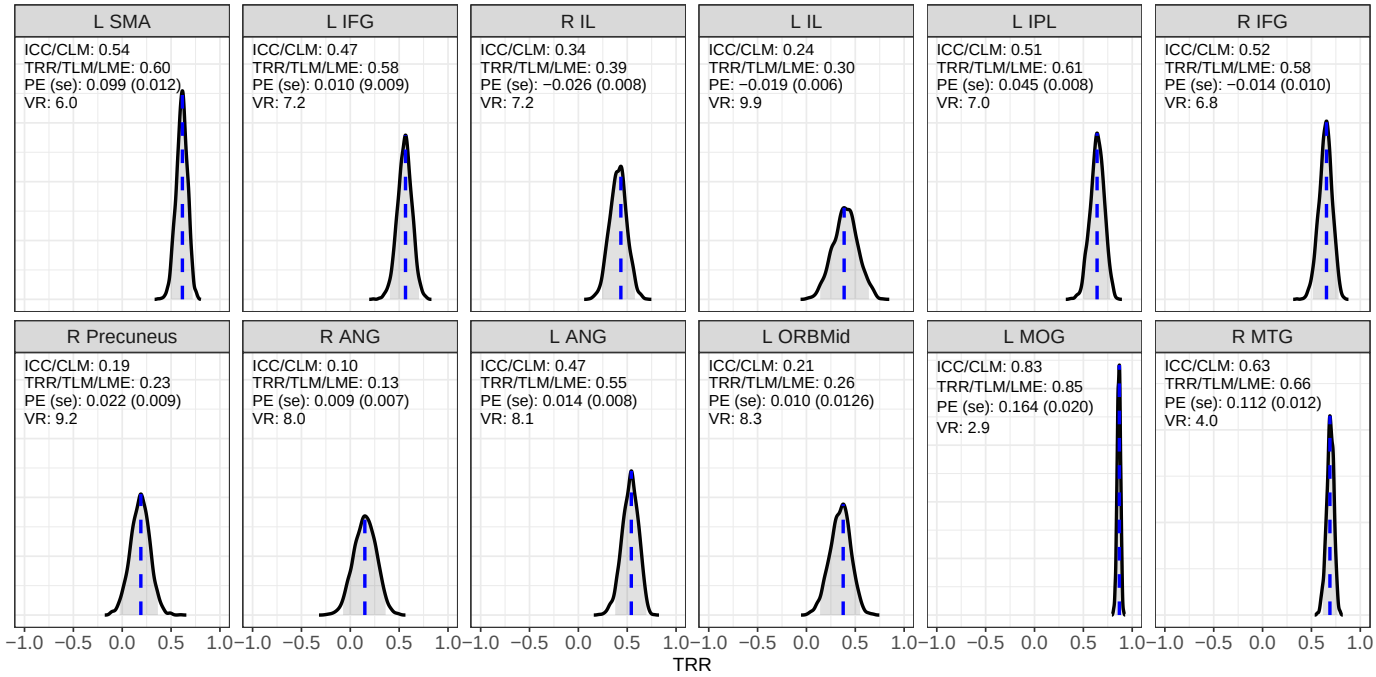
Most regions exhibited adequate to excellent reliability for the average, both conflict and error, with some variability across regions in terms of precision of these estimates. There were only a few exceptions (R precuneus and R angular gyrus). However, for the contrasts, large variations existed across ROIs in terms of both TRR magnitude and precision (Figs. 7,8). Some regions showed moderate to high TRR with relatively high certainty; some regions exhibited high TRR with a wide range of uncertainty; others were difficult to assess as their TRR distributions were very diffusive. 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. 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). The reliability at some regions was so diffusive with a substantial amount of uncertainty that an estimate through the conventional ICC would be pointless.

The variability ratio can be used as a proxy for the degree of underestimation via ICCs. The degree of underestimation, shown as a linear relationship with TRR, through the formula (7) or (13), is derived under the assumption of a Gaussian distribution for cross-trial variability. While this assumption is largely violated for both behavioral data (e.g., RT for Stroop and Flanker tasks) as well as BOLD response, the variability ratio serves a useful metric that provides an explanation as to why such heterogeneity exists across estimates for different regions and contrasts. A large VR is usually associated with substantial underestimation via ICCs and a large amount of TRR estimation uncertainty. It is important to note that 10 out of the 12 ROIs had a substantial range of uncertainty for the TRR for both conflict and error contrasts (Fig. 8) that were for the most part associated with a large variability ratio. It is for this reason, that TRR cannot simply expressed as point estimates through the conventional ICC.

3.5 Implications for practice

The advancements through the BML are substantial for any researcher hoping to assess TRR for their task-based imaging data. It is specifically the additional precision information that is crucial for the researcher to develop a full picture of the task's TRR. 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

(A) Conflict average



(B) Error average

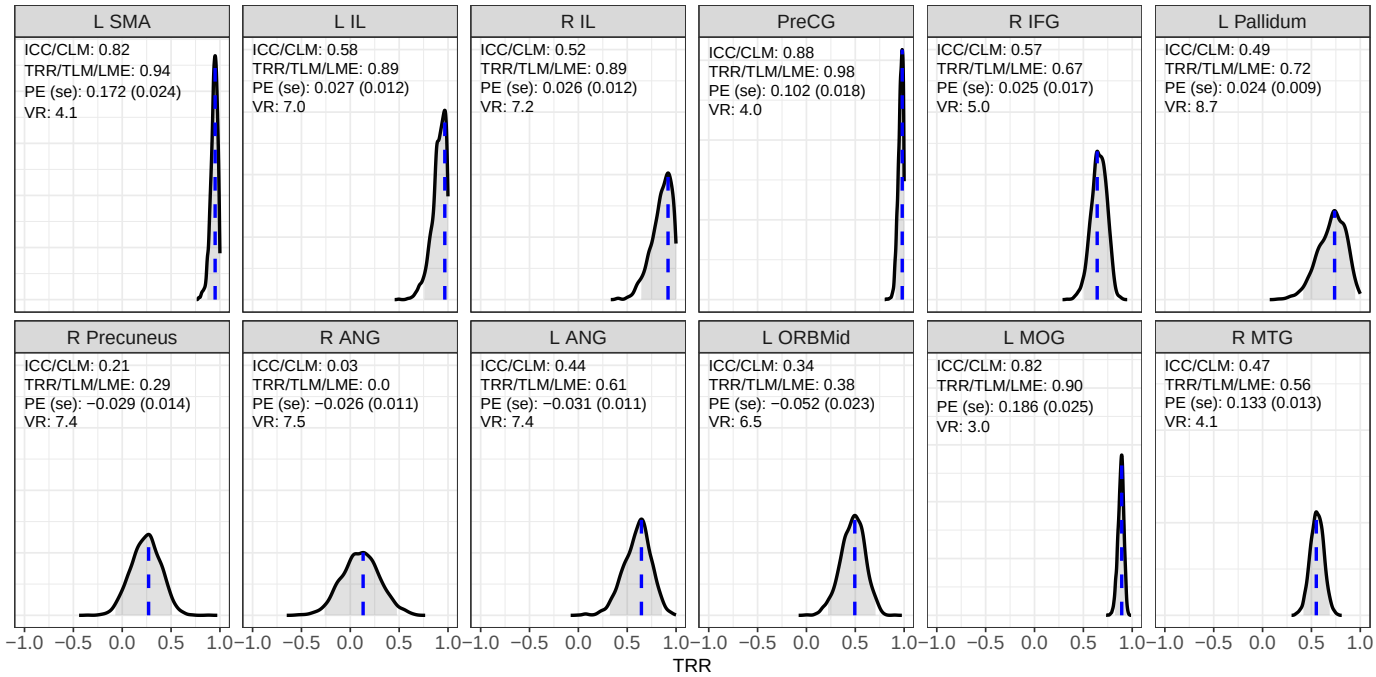
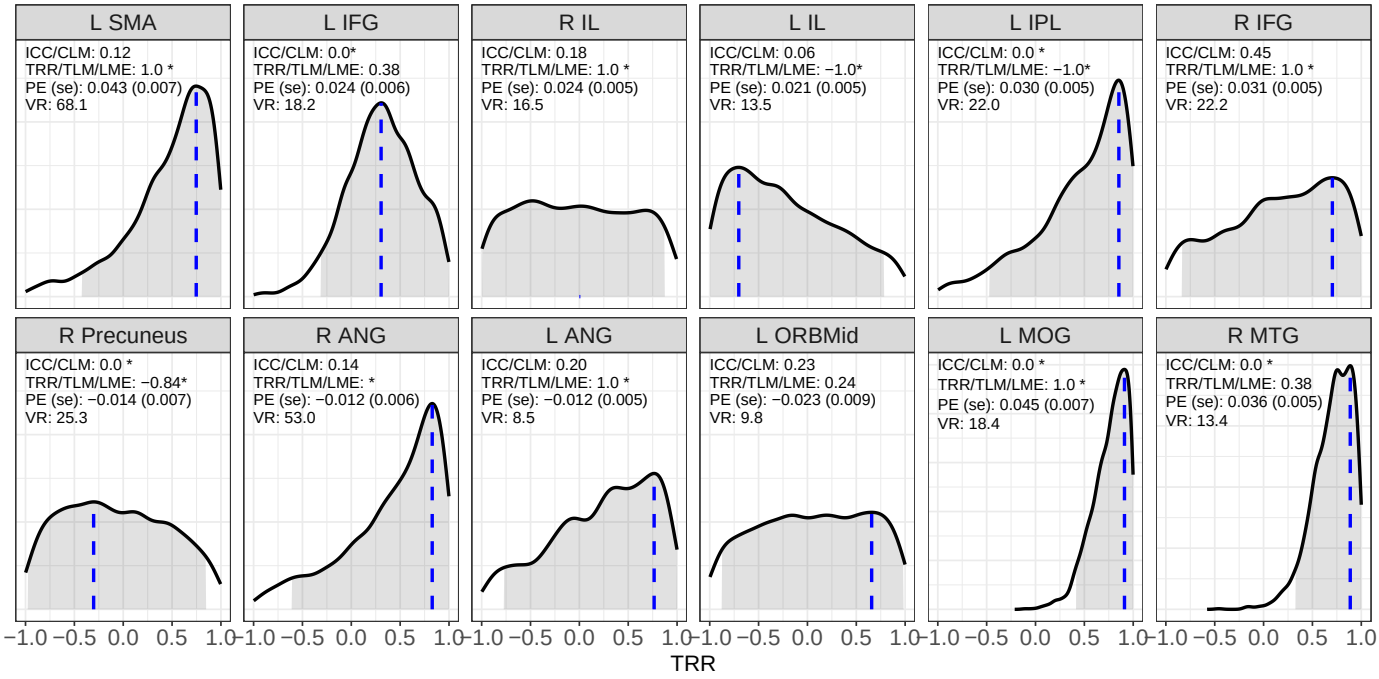


Figure 7: TRR distributions for the average effect across conditions at 12 ROIs. The TRR estimates for the average effect of conflict (A) and error (B) contrast were obtained using the program **TRR** based on 2000 draws from MCMC simulations of the BML model with a Student's t-distribution. The TRR posterior distribution for each ROI is shown here as a kernel density estimate that smooths the posterior samples. Each blue vertical line indicates the mode (peak) of the TRR distribution, and the shaded area shows the 95% highest density interval. Three quanitities are shown for each ROI: conventional ICC, TRR estimated through TLM with LME and variability ratio (VR). 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

with a minimum of 20 trials in the incongruent commission error condition has just barely enough trials to derive summary 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 VR was generally larger in the conflict

(A) Conflict contrast



(B) Error contrast

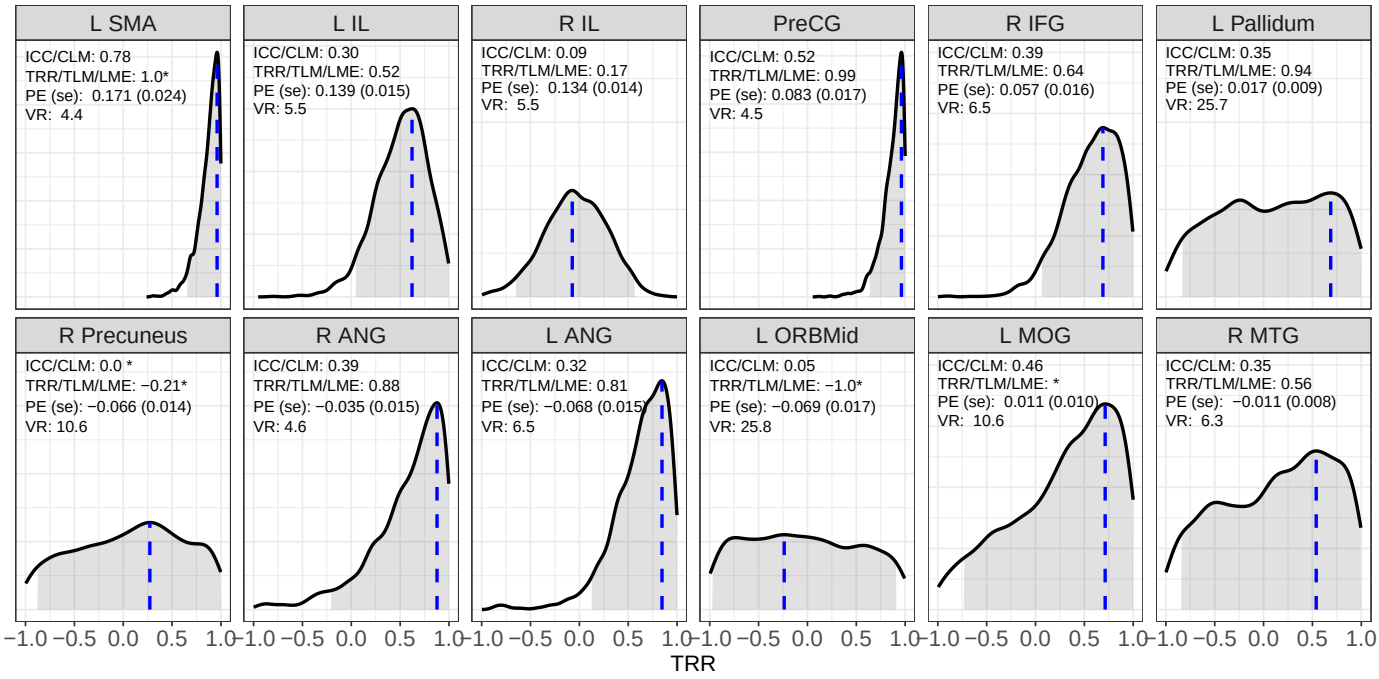


Figure 8: TRR distributions for the two contrasts of the Flanker task for ROIs. The TRR estimates for the conflict (A) and error (B) contrast were obtained using the program **TRR** based on 2000 draws from MCMC simulations of the BML model with a Student's t-distribution. The TRR posterior distribution for each ROI is shown here as a kernel density estimate that smooths the posterior samples. Each blue vertical line indicates the mode (peak) of the respective TRR distribution, and the shaded area shows the 95% highest density interval. Four quantities are shown for each ROI: conventional ICC, TRR estimated through TLM with LME, population-level effect (PE) with its standard error, and variability ratio (VR). 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

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, 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, compared to the correct congruent responses included in the conflict contrast. Interestingly, behaviorally, the error contrast showed a lot of uncertainty around TRR estimates compared to the conflict contrast, which had a more narrow TRR distribution. No such visible difference emerged between the two contrasts in the ROI 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 for further exploration of TRR across different cognitive processes and experimental designs to fully understand how reliably certain processes can be measured and how TRR is impacted by experimental design choices. Most importantly, our initial results provides a more nuanced assessment of task fMRI TRR than recent reports: TRR variability is large, even within the same contrast across regions, both in magnitude and precision of TRR estimates; understanding the source of trial level variability may provide an important avenue to improve TRR in future work.

4 Discussion

High TRR of measurement is of great importance. Here, we adopt a new modeling framework for the estimation of TRR 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. Several key findings emerged in the step-by-step process by which we built the TRR model formulation: (a) TRR is conceptually distinct from population effects; (b) the conventional ICC underestimates TRR in cases where trial-level variability is large relative to between-subject variability; (c) trial-level variability is surprisingly large in both psychometric and neuroimaging data; and (d) a large number of trials is important to adequately assess TRR. We will discuss each of these in turn.

4.1 The relationship between population-level effects and TRR

We demonstrated that population effects are not necessarily tied to the reliability of individual differences. As clearly illustrated in Fig. 1, all four possible scenarios can occur: large population effects can have strong or weak TRR, and the absence of population-level effects does not preclude high TRR. Dependence of high TRR on population effects is often, mistakenly, assumed. In practice, researchers usually choose to examine TRR of those regions that exhibit strong population effects. In other words, 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. This can mean missed opportunities to estimate TRR 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 our ROIs (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 TRR exploration if the selection had been solely based on the statistical strength at the population level.

4.2 BML as an adaptive solution for the underestimation of TRR via ICC

Recent reports using the ICC metric suggests that TRR of common cognitive and fMRI tasks are largely inadequate. Issues of measurement have become front and center in the field; these basic measurement issues have motivated new statistical modeling approaches for behavioral and imaging tasks (Rouder and Haaf, 2019; Haines et al., 2020).

ICC, as defined under the conventional framework of variance ratio under the ANOVA or LME platform, is not equipped to handle data structures that include many trials nested within subject and condition. At least four problems exist with the conventional ICC formulation: 1) failure to accurately characterize the underlying data generating process, 2) underbiased estimation, 3) difficulty of obtaining uncertainty, and 4) contingency on rigid assumptions and occurence of numerical failures. Nevertheless, many experiments involve a large number of trials to increase precision of estimates and estimate small to medium effect sizes. Averaging trial-level effects as commonly practiced in condition-level modeling fails to account for trial-level variability and underestimates reliability. As shown in the current report, the underestimation by the conventional ICC is linearly associated with TRR, and the rate of underestimation depends on two factors: the trial sample size and the relative magnitude of cross-trial variability to cross-subject variability. Specifically, the larger the cross-trial variability relative to cross-subject variability, the more severe the underestimation via the ICC. ICCs are implicitly "contaminated" by cross-trial variability, and are thus sensitive to trial sample size. As a result, ICC values reported in the literature, even on the same task, may have embedded in them underestimation to a different degree, due to different trial sample sizes. Therefore, TRR estimates may not necessarily be comparable across experiments, leading to a portability problem (e.g., Rounder and Haaf, 2019). By definition, ICC as a TRR measure concerns subject-level effects; thus, trial-level effects should be appropriately untangled in effects partitioning.

Hierarchical models such as LME and BML through TLM allows the researcher to disentangle trial-level effects from other sources of variance, providing a more accurate assessment of TRR. 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 at the subject level. 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 regularization and shrinkage. As a result, one effectively gains high predictive accuracy for TRR estimation.

The BML framework is well-suited for TRR 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 estimates and have a limited flexibility of handling deviations from a Gaussian distribution such as data skewness and outliers. BML can further 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. **??**). The full and rich information contained in posterior distributions is another benefit, offering distributional subtleties and 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 TRR values above 0.9 with actual real experimental data (Fig. 4).

4.3 High cross-trial variability

Large cross-variability is the root cause for all the three problems in the current context: ICC underestimation, numerical failures in LME modeling and poor precision of TRR estimates. Experimental data indicates that cross-trial variability is much larger than cross-subject variability. If the former is roughly in the same order of magnitude as, or smaller than, the cross-subject variability, the conventional ICC might be considered a shortcut to approximate TRR as shown in the formulations (7) and (13). However, experimental data from both behavioral and neuroimaging datasets point to a much larger cross-trial variability than cross-subject variability. Specifically, the variability ratio V may reach up to 10 for simple effects and go beyond 20 for contrasts. Such large V values are directly associated with specific issues: underestimation of TRR when the conventional ICC formulation is adopted, numerical convergence issues in LME modeling and a poor precision of TRR estimation under the BML framework. One many suspect that large cross-trial variability might be ascribed to suboptimal modeling in subject-level time series regression due to (i) the substantial amount of confounding effects that are usually not properly accounted for and (ii) poor characterization of varying hemodynamic response across trials, condition, regions and subjects through a prefixed basis function. However, considering the fact that large cross-trial variability also exists in psychometrics to slightly lesser extent (Rouder et al., 2019), suboptimal modeling is less likely the main reason.

The large cross-trial variability in fMRI tasks remains to be fully explored. In a test-retest dataset, there are $2m$ times more trial-level effects than subject-level effects: $4mn$ trial-related terms correspond to $2n$ subject-related terms for a contrast (Fig. 2). Thus, one might expect that cross-trial variability σ_0 would be much smaller than the cross-subject counterpart σ_{τ_r}. On the other hand, trial-level effects fluctuate substantially and their sequences appear to be random 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 simply accounted for by habituation, fatigue or sensitization. At the same time, these trial-level fluctuations are not purely random fluctuations: there is a high degree of bilateral synchronization within the same subject (Chen et al., 2020). In addition, the cross-trial fluctuations are to some extent associated with behavioral measures such as reaction time and stimulus ratings that are typically modeled through trial-level modulation analysis at the subject level. One may hypothesize that the random nature of trial-level fluctuations may be caused by momentary lapses in attention. Alternatively, brain regions may constantly undergo some intrinsic fluctuations - external stimuli or tasks are surprisingly small constrains stacked on top of large intrinsic neuronal activities.

4.4 Importance of trial sample size

Two types of sample size, trials and subjects, are involved in typical psychometric and neuroimaging studies. Per central limit theorem, a reasonably large sample size of an experimental unit is amenable to conventional properties such as a Gaussian distribution. This theorem is pivotal to many modeling frameworks including the conventional ICC formulation. However, the asymptotic property of the unbiasedness and Gaussianity heavily relies on large sample sizes, which cannot necessarily be met or easily predetermined in practice. In fact, the sample size of trials and subjects is a big issue for TRR estimation.

The pivotal role of trial sample size remains largely unrecognized in the field of TRR (Rouder et al., 2019). Surprisingly, the number of trials plays a much more crucial role than that of subjects in the amount of ICC underestimation under the conventional formulation. It has been generally assumed that subject sample size might help in achieving high ICC (Elloitt et al., 2020; Haines et al., 2020). However, the trial number has a much stronger impact on the degree of ICC underestimation, as shown in the underestimation formulas (7) and (13) as well as simulation results in Figs. 3 and 9. The smaller the number of trials, the more severe the ICC underestimation. In contrast, the number of subjects would not, on average, cause any underestimation

of ICC or TRR.

When trial variability is high, trial sample size is pivotal in determining the precision of TRR estimates. A substantial number of trials may be required to achieve reasonable precision. Even though BML addresses risks of underestimation, the uncertainty of TRR estimation as represented by, for example, standard deviation or highest density interval, may remain wide when cross-trial variability is large. The TRR estimation uncertainty depends on four factors (Figs. 3 and 9): TRR magnitude, variability ratio, trial and subject sample size. Among the four factors, only the sample sizes could be experimentally manipulated. Even though the uncertainty of TRR estimates decreases as the sample size increases for both trials and subjects, the impact of trial sample size is much stronger (Fig. 3B,C and Fig. 9B,C). Also, as shown in experimental results (Figs. 4, 5, 8 and 7), the precision of TRR estimation varies substantially across brain regions or between simple effects and contrasts. To dissolve those diffusive posterior distributions of TRR, a few hundred or even more trial samples 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 TRR precision for a contrast

For average condition effects it is relatively easy to achieve reasonable TRR precision. For effects of a single condition or the average among two or more conditions, empirical data indicate that cross-trial variability σ_0 is larger than cross-subject variability σ_{τ_r} with a variability ratio 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 TRR estimation within a small or moderate amount of uncertainty (Figs. 4, 5 and 7). Consequentially, one may be able to estimate TRR under LME through TLM (Fig. 6C,D). While the program **TRR** can perform TRR estimation for both behavior and region-based neuroimaging data, the program **3dLMEr** can be adopted for neuroimaging whole-brain voxel-wise TRR estimation (though uncertainty information is unavailable).

A reasonably high precision of TRR estimation for a contrast might be harder to attain under some circumstances for the following reason: Cross-trial variability σ_0 measures the trial-level fluctuations per condition (and per subject) regardless of the research focus on a single condition or a condition contrast. However, cross-subject variability σ_τ for a condition contrast measures the fluctuations regarding the contrast, not for the individual condition effects nor at the trial level, as characterized in the parameters λ_{rs} in formulations (12) and (17). The contrast between two conditions is usually a few times smaller in effect 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 σ_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 V. In other words, the cross-subject effects λ_{rs} often will get dwarfed by cross-trial variability σ_0, resulting in a large uncertainty for the TRR estimation of a contrast. Another contributing factor in neuroimaging specifically is the cross-region variability; the variability ratio V 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 TRR estimation with a reasonable precision (Fig. 8). In general, we recommend the adoption of the BML framework for its flexibility and adaptivity to closely characterize the data information. Additionally, even though it may be difficult to achieve high precision for TRR 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 is relevant in the current context concerning TRR: population-level effects and reliability of individual differences.

Population-level effects capture the general summarization for the effects of interest. The notion of population effects in experiments such as Stroop and Flanker tasks is directly associated with the representativeness through the measurement unit of, for example, subjects and trials. Such effects usually lie at the top levels of the data hierarchy and are modeled as fixed effects under the conventional LME framework. For example, the population-level contrast between incongruent and congruent conditions is the main focus in experimental designs of inhibition tasks while group difference in terms of developmental trajectory between patients and controls might be a research goal in a longitudinal study. Due to the widespread popularity, population-level effects are relatively intuitive and easy to visualize as the horizontal lines illustrated in Fig. 1. Typical sample size for subjects and trials is below 100. In this context, cross-subject and cross-trial fluctuations are considered noise and nuisance.

TRR concerns a different type of generalizability: the consistency, reliability or conformity of individual differences. From the research perspective, individual differences can be trait-like measures, behavioral (e.g., RT) or BOLD response. Unlike population-level effects that is assumed to be "fixed" in a statistical model, TRR is characterized by subject-level effects that vary across subjects and are termed as "random" effects under the LME framework. Thus, it is no surprise that the cross-subject variability relative to the total variation captures the TRR in the classical quantification of ICC. Specifically, the research interest hinges as to how strongly the two or more effects from each measuring unit (e.g., subject) resemble each other. Individual samples such as trials and subjects are expected to vary across specific exemplars, but a high consistent type of variation should have a systematic pattern, and such a pattern is the second type of generalizability that is characterized as the correlation, not average, across the samples. The generalizability of TRR 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 TRR 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). As a result, their detection may require larger sample sizes. In this context, individual differences are captured at the center stage as a correlation while population-level effects are centered out as baselines.

4.7 Violating model assumptions

Understanding assumptions underlying each model is a necessity, and TRR estimation illustrates this point from multiple angles. An assumption that is often overlooked is the one of data distribution. With only two parameters of location and scaling, the Gaussian distribution has many desirable properties such as guaranteed asymptotic convergence with large enough samples and numerical frugality and has been adopted widely in statistical modeling. However, small samples often result in data with skewed or diffusive distributions. Collecting large enough sample sizes to ensure that a Gaussian shape is met is often not an option in practice including for studies assessing TRR. A typical approach is to censor outliers or transform the distribution. In addition to the arbitrariness involved in determining the threshold for censoring, it is only in rare cases that the choice of the adopted model is justified or comparisons are made with other potential candidate models. In contrast, the Bayesian framework is flexible to fitting different distributions and allows the incorporation of measurement errors.

Model comparison and validation are important in improving the accuracy of TRR estimation. For ex-

ample, through model comparisons and validation (Fig. 4), our investigation indicated that an exponentially modified Gaussian distribution is better suited to describe the RT data than its Gaussian and log-normal counterparts, due to its strengths in accommodating data skewness and outliers. This also confirms the adaptivity of the exGaussian distribution for RT data in the literature (Ratcliff, 1979).

Point estimation can be misleading and can cause numerical failures. As a first-order moment, in conventional statistics, an effect estimator for the mean of data samples is usually adopted. The uncertainty associated with this estimate can be reasonably characterized through the standard deviation. Point estimation is so popular because its information extraction is based on the pithy centrality measure through algorithms such as maximum likelihood. However, for parameters such as variance and correlation, their point estimates, based on higher-order moments, may conceptually and computationally encounter serious issues. First, their distributions are not necessarily symmetric nor highly concentrated; thus, a single value from a point estimate may not do justice to accurately expressing the entire distribution and will thus often provide only a limited to distorted summary, especially when the posterior distribution is skewed or diffusive (Fig. 4). Second, uncertainty information is not readily available for these parameters under the conventional framework such ANOVA and LME. Lastly, numerical singularity or convergence failure is a commonplace when 1) the parameter of interest (e.g., correlation, TRR) gets trapped at the boundary, 2) the posterior distribution is overly diffusive, or 3) the variability ratio V is too large. While the conventional ICC point estimate may feel familiar, the BML framework clearly illustrates that it needs revision. Furthermore, assessing statistical evidence for ICC based on Fisher-transformation or F-statistic (Chen et al., 2018) still does not solve the issue of providing a measure of precision. A large degree of uncertainty may manifest in numerical failures in the LME model, necessitating large modifications to existing experimental designs including more trials (and more subjects to a lesser extent).

4.8 Limitations

Currently, for the BML framework, it is not feasible for TRR analysis to be conducted at the whole brain level. Thus, its application is currently limited to behavioral and region-based neuroimaging data. 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. The program **TRR** can be used behavioral or region-level neuroimaging data; however, for whole-brain voxel-level analysis, users are currently limited to the LME approach which can be performed with the program **3dLMEr** for a single conditions or a contrast with large effects. However, only point estimations of TRR are available with no uncertainty information.

A much larger trial sample size might be needed to achieve reasonable TRR precision. Between the two sampling units of participants and stimulus trials, it is the latter that is more efficient in dampening the TRR estimation uncertainty. In typical fMRI studies, the number of trials ranges from 10 to 50, which are much smaller than those of experimental datasets illustrated here. Aligned with a recent assessment in phychometrics (Rouder et al., 2019), we surmise that hundreds of trials might be required for neuroimaging studies to narrow down TRR estimation uncertainty.

Trial-level modeling requires careful experimental designs. As 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 timing can be randomized to reduce multicollinearity using tools such as RSFgen in AFNI or optseq[3]. However, even if statistically separable, trial level effect estimates can be unreliable. To obtain effect estimates at the trial level, one is 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 regions, subjects and even trials, such a presumption might misidentify trial-level effect magnitude, resulting in compromised TRR estimation.

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

5 Conclusion

The conventional intraclass correlation could underestimate the test-retest reliability to varying extents for datasets with many trials making up a condition of interest. TRR accuracy via ICC is lost in two ways: first, the ICC formulation fails to accurately partition the hierarchical structure embedded in the data; second, a single value of test-retest reliability through ICC fails to capture the uncertainty associated with a point estimate. We recommend that TRR 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 TRR estimation. In addition, we suggest that TRR 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 TRR 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).

Appendices

A ICC underestimation for a single condition effect

We seek to derive the conventional ICC under the LME framework through TLM. 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

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

Through the notations

$$V = \frac{\sigma_0}{\sigma_\tau},\ U = \frac{1}{1+\frac{1}{m}V^2},$$

it becomes clear that ICC can be directly expressed as the function of ρ

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

The variability ratio V characterizes the magnitude of cross-trial variability σ_0 relative to the cross-subject variability σ_τ, and the parameter 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, V.

The underestimation of ICC formulation can be conceptually corrected. Under the homoscedasticity assumption, the derivations (20) indicate that $\text{Var}(\overline{y}_{1s\cdot}) = \text{Var}(\overline{y}_{2s\cdot}) = \sigma_\tau^2 + \frac{1}{m}\sigma_0^2$. That is, the variability of cross-trial averages is composed of two components, one associated with cross-subject variability σ_τ and the other with cross-trial variability σ_0. Thus, if the cross-trial variability σ_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),

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

B Simulations for a single condition effect

Simulations were conducted for TRR with a single condition-level effect. Below are the manipulation parameters for the two models of CLM (2) and TLM (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 TRR 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: $V = \frac{\sigma_0}{\sigma_\tau} = 1$, 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,[4]
   (b) test-retest reliability ρ estimated through LME (5),
   (c) conventional ICC adjusted by removing the cross-trial variability $\frac{\sigma_0^2}{m}$.

Each of these $4\times4\times4\times4\times2\times3 = 1536$ 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: $\begin{bmatrix}\tau_{1s}\\ \tau_{2s}\end{bmatrix} \sim \mathcal{N}(\mathbf{0}_{2\times1},\ \boldsymbol{R}_{2\times2})$, 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):
$y_{rst} \sim \mathcal{N}(a_r + \tau_{rs},\ \sigma_0^2),\ r = 1, 2;\ s = 1, 2, .., n;\ t = 1, 2, .., m.$

[4]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.

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

iv) Recover the simulated parameters including the three TRR estimates. Specifically, the conventional ICC is obtained through the ICC formula (3) while the TRR is estimated through ρ 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 2×2 factorial structure, we denote μ_{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 matrix 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 θ, η and π 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 ρ_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 ρ_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 effect. 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

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

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

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

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

The underestimation of ICC formulation can be conceptually corrected as well. If the cross-trial variability σ_0 is known, we could restore the accuracy of ICC by removing the trial-related variance component σ_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 underestimation of ICC 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, λ and ρ_1 by a_r, τ and ρ_0.

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

Simulations for TRR with a condition contrast under the LME model (12) are similar to the situation with a single condition but with a slightly higher complexity. With the correlation structure $\boldsymbol{C}$ in (22) across the four condition-level effects $(\mu_{11s},\ \mu_{12s},\ \mu_{21s},\ \mu_{22s})^T$ and the assumption of homoscedasticity in (21): $q_{11} = q_{21} = q_{12} = q_{22} = 1$, the cross-session reliability or TRR 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 TRR 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 σ_0 relative to cross-subject variability σ_λ: $\frac{\sigma_0}{\sigma_\lambda} = 1$, 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 ρ_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 ρ_1 is estimated through the LME model (12). In addition, adjusted ICC is obtained per formula (15).

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

F Hyperpriors adopted for BML modeling

The prior distribution for all the 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 σ 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, ν, 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 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 at 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

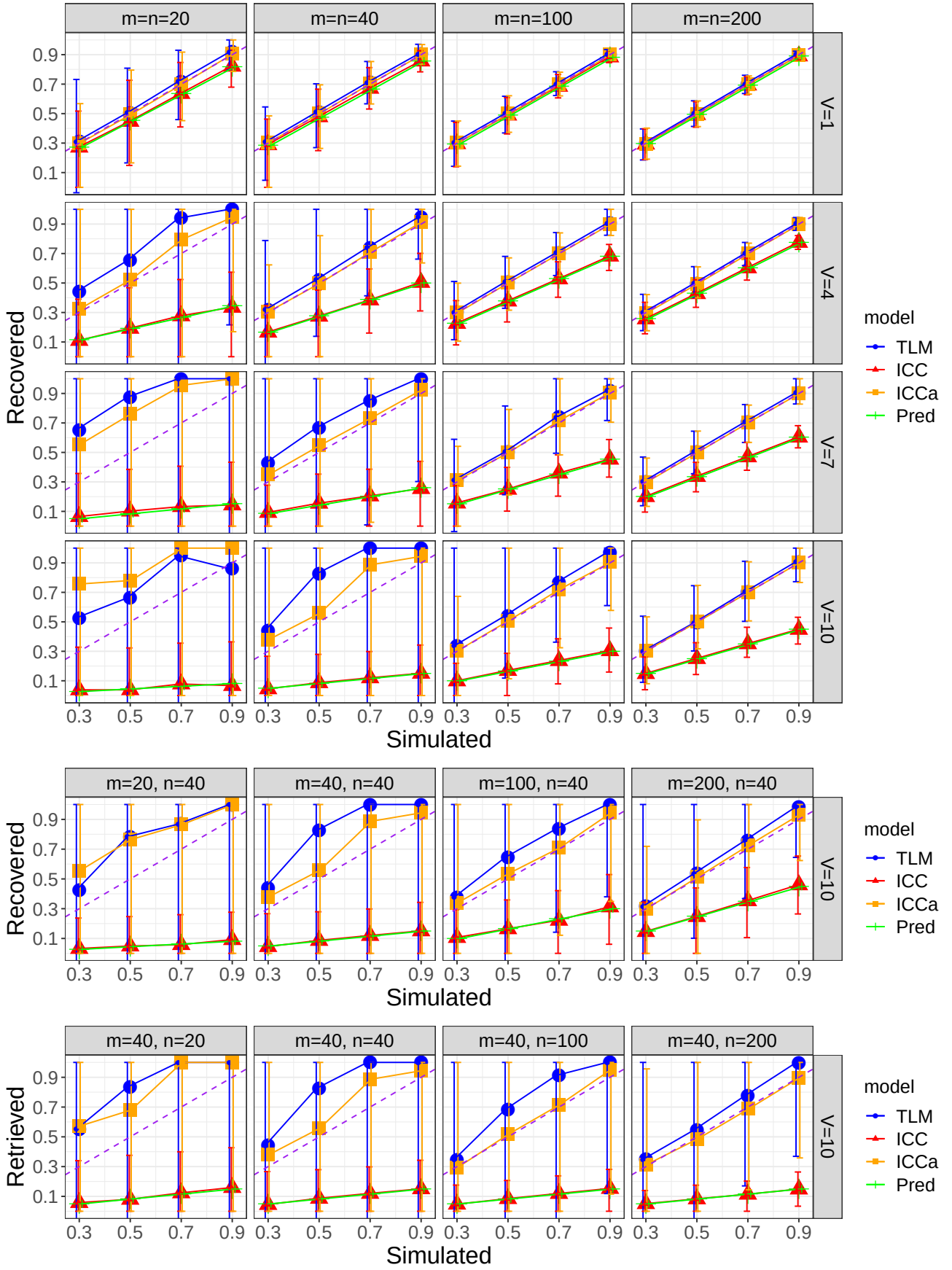


Figure 9: Simulation results for a condition contrast. The four columns correspond to the sample size of subjects and trials while the four rows are the varying standard deviation ratios of $\frac{\sigma_0}{\sigma_\lambda}$. The x- and y-axis are the simulated and recovered TRR, respectively. Each data point is the median among the 1000 simulations with the error bar showing the 90% highest density interval. The dashed purple diagonal line indicates the perfect scenario.

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 $2.5 \times 2.5 \times 2.5$ 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. https://doi.org/10.18637/jss.v067.i01

Box, G.E.P., 1976. Science and Statistics. Journal of the American Statistical Association 71, 791–799. https://doi.org/10.2307/2286841

Bürkner, P.-C., 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software 80, 1–28. https://doi.org/10.18637/jss.v080.i01

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. https://doi.org/10.18637/jss.v076.i01

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., Britton, J.C., Pine, D.S., Cox, R.W., 2013. Linear mixed-effects modeling approach to FMRI group analysis. NeuroImage 73, 176–190. https://doi.org/10.1016/j.neuroimage.2013.01.047

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. https://doi.org/10.1002/hbm.23909

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. https://doi.org/10.1016/j.neuroimage.2020.117496

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.

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. https://doi.org/10.31234/osf.io/xr7y3

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. https://doi.org/10.3758/s13428-017-0935-1

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. https://doi.org/10.1037/1082-989X.1.1.30

Noble, S., Scheinost, D., Constable, R.T., 2019. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. NeuroImage 203, 116157. https://doi.org/10.1016/j.neuroimage.2019.1161

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. https://doi.org/10.3758/s13423-018-1558-y

Rouder, J., Kumar, A., Haaf, J.M., 2019. Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail. PsyArXiv. https://doi.org/10.31234/osf.io/3cjr5

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.

Westfall, J., Nichols, T.E., Yarkoni, T., 2017. Fixing the stimulus-as-fixed-effect fallacy in task fMRI. Wellcome Open Res 1. https://doi.org/10.12688/wellcomeopenres.10298.2

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis, Use R! Springer-Verlag, New York. https://doi.org/10.1007/978-0-387-98141-3