Why is intraclass correlation unreliable as a test-retest reliability metric?

Gang Chen, ...*[a]

[a]Scientific and Statistical Computing Core, National Institute of Mental Health, USA

Abstract

The concept of *test-retest reliability* (TRR) indexes the repeatability or consistency of individual differences across two measures. Adequate TRR is a critical foundation for any scientific study, and 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. However, it has been recently argued that intraclass correlation coefficient (ICC), the classical metric to assess TRR, may be ill-suited to characterize TRR of common cognitive tasks that estimate effects across many trials. Specifically, in the current report, 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 subject sample size has surprisingly little impact on ICC.

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

1) a hierarchical model can more accurately characterize test-retest data;
2) TRR should be characterized as a correlation divorced from cross-trial variability;
3) a hierarchical model identifies brain regions with high TRRs (e.g., 0.9 or above);
4) hierarchical modeling programs TRR and 3dLMEr are available for TRR estimation;
5) the number of trials is more important in improving TRR precision than the number of participants;
6) a high precision of TRR estimation could be hard to attain with common trial sample sizes;
7) a puzzling question remains: why is cross-trial variability much larger than cross-subject variability?

In the process of laying out a new hierarchical solution to estimate TRR, we examine a number of modeling issues associated with the conventional ICC framework.

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. All statistics compress and extract information from data; TRR captures the degree of agreement or consistency across multiple measurements

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

(rather than observers) of the same quantity (RT, BOLD response, personality traits) under similar circumstances. Such agreement or consistency is measured through a correlation coefficient and is traditionally assessed through the metric of the intraclass correlation coefficient (ICC). Thus, for definitional clarity, TRR in the current report is defined as a property of individual difference and ICC is a conventional statistical measure of TRR.

Assessment of TRR is critical. As TRR can be adopted to assess a wide variety of data from clusters, groups or participants, one may quantify the genetic resemblance among relatives with regard to a certain characteristic or traits, or the reliability of BOLD signal across different scanners with data collected from the same group of participants. Here we focus on one specific type of data structure in behavioral and neuroimaging studies: subjects are measured through two repeated sessions under one or two conditions that are instantiated with many trials. Suppose that, for example, an experimenter is evaluating the common cognitive effect of the Eriksen Flanker task (i.e., reaction time slowing due to conflicting information) in a group of children at two time points, one week apart. The task is composed of two conditions, one where flanking arrows call for the same response (congruent condition) and one where flanking arrows call for the opposite response to the target arrow (incongruent/conflict condition) in the center of the visual display. Under the conventional ICC formulation, the contrast between the two conditions renders a reaction time (RT) difference score per child that is indicative of the ability to inhibit a response. As a result, ICC is computed as the fraction of the total variance that can be attributed to inter-individual differences of inhibition 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 whether common tasks should be used to study individual differences, for which they have been used for years, if not decades, in the case of the Flanker task. 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 and surprising to the field. Task-based fMRI is increasingly being used in large scale studies of development and neurodegenerative disease or with the hope of identifying brain biomarkers of disease risk. High TRR of the fMRI indicator is a critical requirement for all of these usages. Additionally, a lot of the fMRI (and behavioral) tasks that show low TRR have been carefully designed to isolate specific cognitive processes – task effects such as the 'Flanker effect' are robust and easy-to-obtain on the population level with a large body of literature to lean on. How is it that these robust task effects have such low TRR?

Two aspects of the Flanker 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 at the trial level effects have not been widely practiced in neuroimaging. Ignoring trial level variance effectively means that all trial effects are presumed to be the same; in addition, precision information of the condition-level point estimates is also abandoned 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 TRR estimation hinges. Here, we will build on previous work to investigate how cross-trial variability can be properly accounted for in TRR studies of task-based imaging and explore the conditions under which this improves TRR estimation. In the step-by-step process of building an adequate TRR modeling formulation in a Bayesian framework, we will assess the relationship between TRR and task/population level effects and detail additional improvements rendered by the final model formulation. In addition, we provide the community with tools that can be utilized to estimate TRR in this new framework

for both behavioral and imaging tasks.

1.1 Classical definition of ICC

In the current report, all conventional ICC models and comparisons to these models will focus on the most common type, ICC(3,1), which quantifies the consistency of an effect of interest between two repetitions (or sessions) when the same group of subjects are measured twice. We will use a hypothetical example to illustrate some basic introductory concepts and for its heuristic expediency. Suppose that the effect of interest is the contrast between incongruent and congruent conditions of Stroop task. (Need to resolve the problem of switching-back-and-forth between Flanker and Stroop.) 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 are 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 as well as neuroimaging data analysis. The residual variability σ_e may appear to capture the within-subject cross-repetition variability; however, as we will detail later, some "dark" variability is embedded in σ_e. It is this "hidden" variability that causes a fundamental problem with the conventional ICC.

The classical ICC can be examined from two different statistical perspectives. First, as the data variation 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 of *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
ICC(3,1)		session 2	666	(641, 691)				
	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
		session 2	607	(584, 629)	σ_{τ_2}: 74			
TLM	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
		session 2	59	(50, 68)	σ_{τ_2}: 18			
TLM	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
		session 2	27	(22, 33)	σ_{τ_2}: 7			
exGaussian	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 went through two sessions of performing Stroop task that included congruent and incongruent conditions. With 240 trials per condition, RT ranged from 1 to 30830 ms, and all data were used here without censoring. The TRR estimate of 1.0 for the contrast between incongruent and congruent condition resulted from a singularity problem with the LME model (12) through TLM when the numerical solver hit the parameter boundary. 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) 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 show a much higher TRR with an ICC(3,1) of 0.72 and 0.69 for congruent and incongruent condition, respectively. As the strong effect of the Stroop task at the population level has been widely investigated and recognized, the unsatisfactory ICC values have been described as paradoxical (Hedge et al., 2018) and have recently motivated new modeling developments (Rouder and Haaf, 2019; Haines et al., 2020).

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 (e.g., instantiated trials, recruited subjects) 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 they 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. As opposed to typical population-level analysis where inferences based on the population-level effects are the ultimate goal, in TRR analysis, the investigator seeks to explore the overall similarity of individual differences across subjects. From the modeling perspective, the relationship between the two levels of pupulation and subject can be loosely described as "orthogonal": the subject-specific effects τ_s in the LME model (5) are "perpendicular to" the population-level effects of overall average effects a_r per repetition in the sense that the former are fluctuations relative to the latter.

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 enough. However, these two types of effects are not conceptually tied together 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 in many fields such as psychometrics and neuroimaging. Recent work has highlighted that TRR of the main contrasts of commonly used tasks as assessed by ICCs is unsatisfactory. For example, cognitive tasks such as Stroop and Flanker are generally considered robust at the population level and expected to show high consistency of individual differences for RT. However, a recent collection of Stroop datasets only registered a lackluster ICC value of around 0.5 (Hedge et al., 2018). Similarly, the ICC performance was even more disappointing for neuroimaging studies with a recent survey of task-related FMRI data reporting ICC values below 0.4 across multiple tasks among many brain regions (Elliot et al., 2020). Resting-state fMRI studies showed equally low reliability estimates with an average ICC below 0.3 (Noble et al., 2019).

A few limitations exist with the classic ICC formulation (2). For instance,

1) **Difficulty of obtaining a measure of uncertainty**. In the vein of conventional statistical framework, one usually obtains a point estimate for an ICC through CLM under ANOVA/LME (2). In addition, statistical evidence can be assessed through either converting the ICC value through Fisher transformation or an F-statistic (McGraw and Wong, 1996; Chen et al., 2018). However, there is no analytical solution to assess the uncertainty range of an 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 heavily relies on the Gaussian assumption. 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.

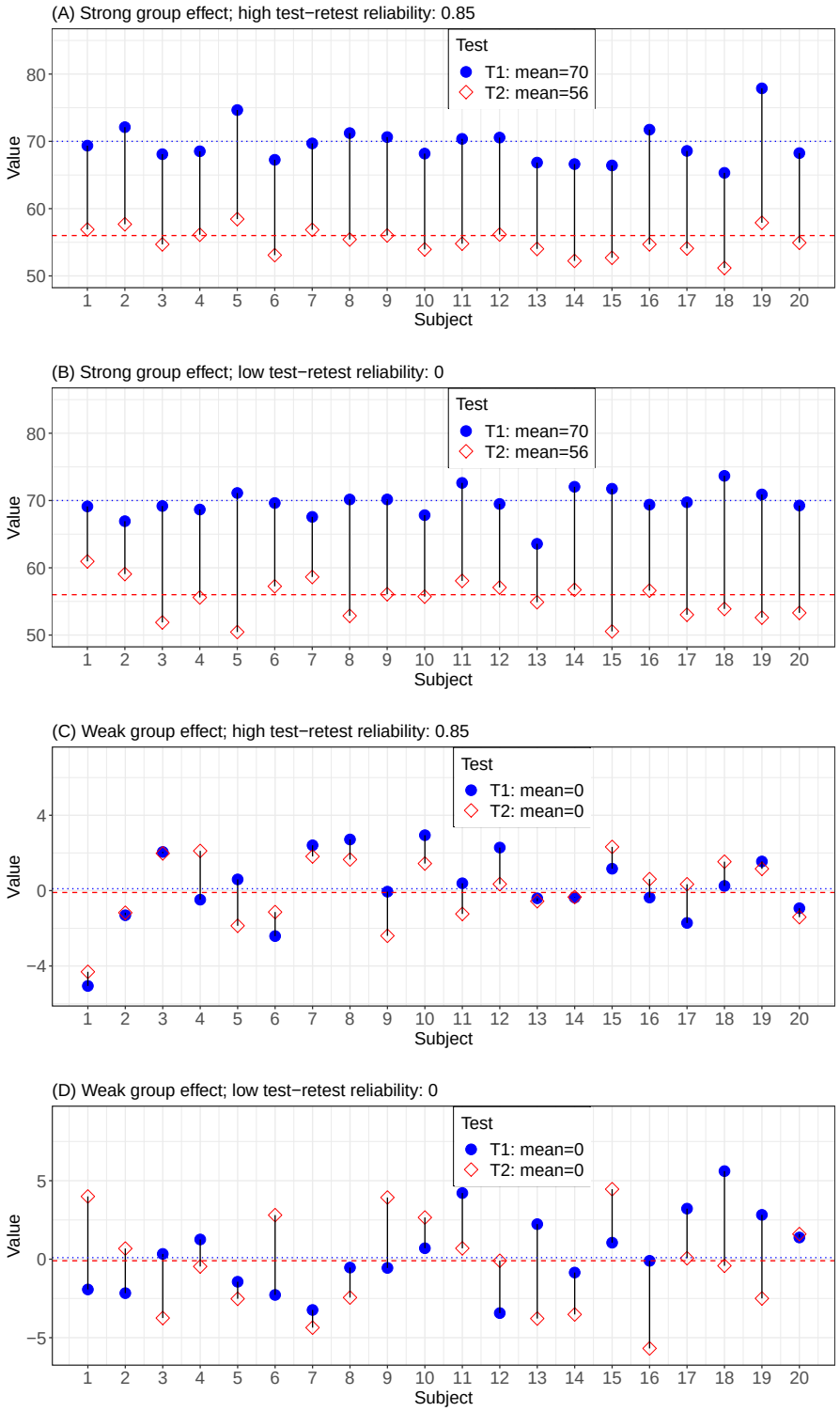


Figure 1: Separation between TRR and population-level effects. Hypothetical data of depression evaluation are the scores from 20 subjects each of which goes through two rounds of testing: T1 (blue filled circle) and T2 (red empty diamond). It is relatively easy to gauge the population effects (colored horizontal lines) but much subtler to assess TRR or the consistency of individual differences. Four extreme cases are illustrated to show that TRR is not necessarily tied to the strength of population effects. 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 TRR. The extent of TRR can be gauged by the proportion of subjects among which the two testing scores from each subject are on the same side (above or below) of their respective population average. That is, with a high TRR, the scores from most subjects during the first (blue filled circles) and second test (red empty diamond and red dashed line) are both either above or below the population average (blue dotted line); the scores are more randomly distributed for a low TRR. However, a high TRR does not necessarily mean that the order of subject-level effects during the two repetitions for most subjects follows the same order as the order of the respective population effects. Data with the two tests (T1 and T2) were randomly drawn from a bivariate Gaussian distribution $\mathcal{N}(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, 49 \times \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix})$ 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) $\mu_1 = \mu_2 = 0,\ \rho = 0$.

3) **Incapability to integrate measurement errors**. 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) **Incapability to integrate cross-trial variability**. The sample size for each condition (i.e., number of trials) serves a purpose similar to that of experimental participants. In this regard, the example in Fig. 1 misses one component in typical psychometrics and neuroimaging studies: trials. The number of trials, as a dimension "orthogonal" to subjects, is 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 the data in practice is typically collapsed along the trial dimension 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 modeling endeavors have been undertaken in recent years to overcome the limitations of the conventional ICC. For example, a few ICC variations have been proposed to solve the numerical instabilities and to 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, a Bayesian multilevel (BML) framework has been adopted to more accurately characterize the data structure (Rounder and Haaf, 2019; Haines et al., 2020) and to provide a full coverage of TRR distribution (Haines et al., 2020) instead of a point estimate as in the classical ICC.

Parallel to the data reduction issue with the conventional ICC, recent work in neuroimaging has revealed a related problem in task-related FMRI modeling practice. When the effect of interest hinges at the condition level, the typical strategy is to create one regressor per condition at the subject level with an implicit assumption that brain responses do not change across trials. With such a complete pooling methodology, the cross-trial variability is fully ignored and dismissed as random noise that is swept under the rug of model residuals. The impact of omitting cross-trial variability in modeling is two-fold (Westfall et al., 2017; Chen et al., 2020). Conceptually, the assumption of no cross-variability leads to the loss of legitimacy of generalizability when the effect estimates at the condition level are carried over to the population level; in other words, the result interpretation would be confined within the specific trials (e.g., particular faces) rather than the associated stimulus category ("face"). Practically, the second consequence of ignoring cross-trial variability at the subject level in the common practice of condition-level modeling may distort effect magnitude as well as its precision at the population level (Chen et al., 2020).

The root cause of the issue with the conventional ICC calculation is that, when collapsing at the trial and condition level, the hierarchical integrity of the data structure is lost. The current work is an extension of a previous report 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 (Rounder and Haaf, 2019; Haines et al., 2020). It has been recognized that data reduction, as inherent in the step (1) of cross-trial averaging and cross-condition subtraction, results in information loss and potentially in effect distortion. 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 (Fig. 2). In addition, we will

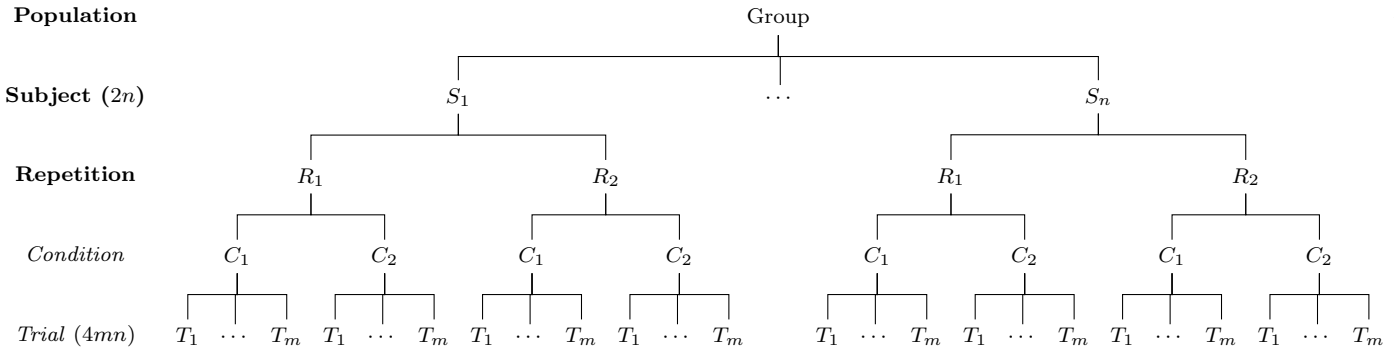


Figure 2: Hierarchical structure of test-retest data. Assume that, in a test-retest study with two repetitions, a group of n subjects are recruited to perform a task (e.g., Stroop) of two conditions (e.g., congruent and incongruent), and each condition is instantiated with m trials. The collected data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

(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 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 with our focus on just one condition, relative to a default condition (e.g., baseline). 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), embodies the population-level effect during the r-th repetition, τ_{rs} represents 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. From the modeling perspective, a_r are the population-level intercepts (or "fixed effects" in the LME terminology) while τ_{rs} are the varying intercepts (or

"random effects") across subjects.

The LME model (5) through trial-level modeling (TLM) provides a pivotal framework and serves as our springboard for a conceptual paradigm shift. The parameter ρ captures the correlation between the two repetitions for the subject-level effects of τ_{1s} and τ_{2s}, thus ρ is conceptually the TRR we strive to characterize when trial-level effects are directly incorporated into the model structure. Even though we no long can formulate the TRR as a variance ratio as traditionally conceptualized in (3), a crucial aspect of the TLM-based LME formulation (5) is the explicit separation of the cross-trial variability σ_0 from the TRR metric ρ.The TRR estimates based on the LME model (5) are shown in Table 1 for the prototypical dataset of Stroop task. 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 explicit expression of the cross-trial variability σ_0 in the LME formulation (5) cannot be overemphasized. The concept of TRR is supposed to quantitatively capture the extent of subject-level similarity between two repetitions; the correlation coefficient ρ is consistent with the fact that cross-trial fluctuations should not be part of the TRR formulation. Thus, the separation of cross-trial variability σ_0 disentangles the cross-trial variability from the correlation ρ embedded in the subject-level effects τ_{rs}. Now that two LME modeling frameworks have been presented, one may wonder: (a) How do they pit against each other? (b) What do their differences reveal?

The conventional ICC tends to underestimate TRR with a single effect. 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; that is, $\sigma_{\tau_1} = \sigma_{\tau_2} = \sigma_\tau$. 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 underestimation severity decreases with the trial sample size. In contrast, the subject sample size n on average does not impact on the underestimation, which might be counter-intuitively at odds with the common perception in the field. 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 underestimated ICC could be hypothetically 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, respectively, under the conventional ICC formulation (2) while σ_0 is the cross-trial variability under the TLM-based LME model (5). Nevertheless, the

adjusted ICCa (9) is of little practical usability because the cross-trial variability σ_0 is shrouded as 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.

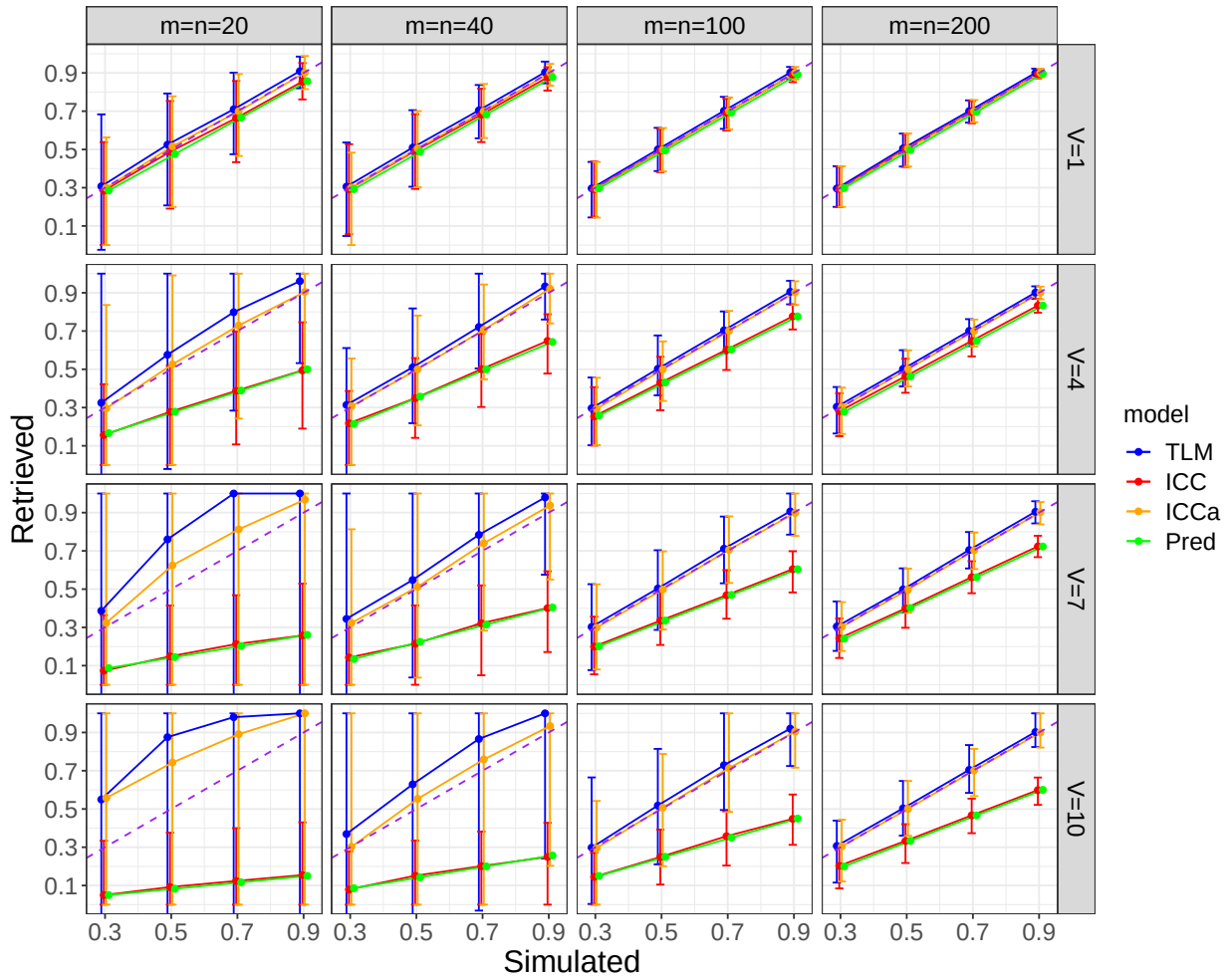
Can the adjustment (9) for the ICC underestimation be validated through real experimental data? We estimated TRR for each of the two conditions through two approaches with the prototypical data of Stroop task. First, the conventional ICC(3,1) values were estimated as 0.72 and 0.69 for congruent and incongruent condition, respectively (Table 1). Applying the TLM-based LME model (5), we obtained a TRR estimate of 0.78 and 0.73 for congruent and incongruent condition, respectively, revealing a slight amount of 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 plugged 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.

2.2 Assessing model comparisons through simulations

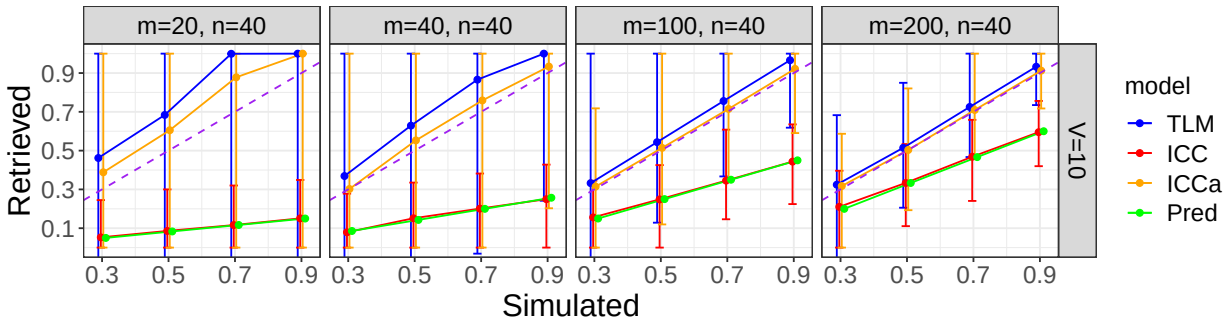
To comprehensively compare the two modeling frameworks for a single effect, we resort to numerical simulations to validate our reasoning in the preceding subsection. The numerical schemes regarding the simulations are presented in Appendix B with results shown in Fig. 3. The population-level effects a_r under both CLM and TLM were robustly retrieved (not shown). Findings in regard to TRR are summarized as 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 may 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 values were successfully retrieved on average from 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 severer the ICC underestimation. This simulation observation is consistent with the result comparison (Table 1) for the prototypical data of Stroop task. In addition, the underestimation rate U is confirmed with the linear relationship (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 in the conventional ICC formulation as shown in (9), the adjusted ICC performs even better across the board than the TLM-based LME model. In fact, the adjustment is relatively 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 ICC underestimation (green line) remains the same regardless of the number of subjects. However, as the subject number n increases, the precision of the TRR estimation based on the conventional ICC and the LME through TLM improves.
4) **The uncertainty of TRR estimation is monotonically related to TRR magnitude, variability ratio V and the sample sizes of m and n for subjects and trials**. Specifically, uncertainty reduces as (a) the sample size of trials or subjects increases, (b) TRR becomes higher, or (c) the variability ratio $V = \frac{\sigma_0}{\sigma_\tau}$ is smaller. Even though sample size is important in achieving a high TRR precision, the trial sample size m plays a more substantive role than the subject sample size n. As the sample size increases (from first to fourth column, Fig. 3), the error bar for TRR estimates become narrower. In addition, the pattern and trend with a fixed number of subjects n but varying trial sample size m (Fig. 3B) is similar to those with a fixed number of trials m but varying subject number n; however, between the two sample sizes, the impact of the trial number m is much larger than the subject number n (plus,

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



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



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

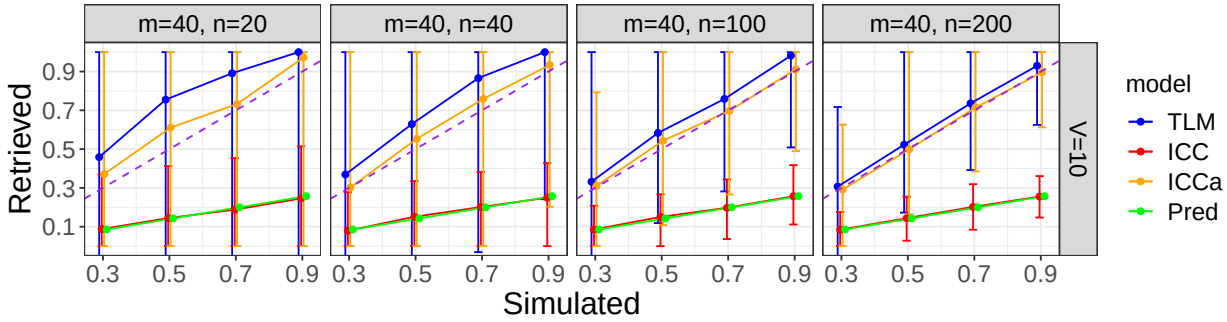


Figure 3: Simulation results for a single condition. The four columns correspond to the sample size of subjects and trials while the four rows are the variability ratios $V = \frac{\sigma_0}{\sigma_\tau}$. The x- and y-axis are the simulated and retrieved TRR, respectively. Each data point is the median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

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 with population-level effects.** Between the two sets of population-level effects for our simulations, $(a_1,\ a_2) = (0, 0)$ and $(1.0, 0.9)$, the former is intended to simulate the situation with no population effects while the latter to explore the case with strong population effects as well as a high certainty. Besides that all simulation parameters were successfully recovered from the TLM-based LME formulation (5), the two scenarios rendered very similar TRR patterns (only the case with $(a_1,\ a_2) = (0, 0)$ is shown in Fig. 3), confirming our earlier assertion regarding the dissociation between population effect and TRR.

6) **The performance of the TLM-based LME formulation is acceptable under some circumstances but suffers from numerical failures under others**. Slight amount of 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 severer the overestimation. Close examinations 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 very 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 generalize the single effect case in the previous subsection to the more applicable scenario of a contrast between two conditions under the LME framework. In real applications, it is a contrast between two conditions, not a single effect, that most experiments usually hinge on. 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. As 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, but we opt to dummy-code the two conditions through the following indicator variable for the convenience of model formulation,

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

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

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

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 $(\lambda_{1s},\ \lambda_{2s})^T$, and the assumption of independence between them 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} code the subject-specific contrast effects. In parallel to the single effect counterparts 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-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 effect case; thus along the vein, 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 improving model fit.

An extra bonus of the LME framework for a condition-level contrast is the availability of TRR for 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 effect can be considered as a special case of the LME model (12) for a condition contrast with the effects from the two conditions being identical; this natural reduction would not be true with 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 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 muddy the conventional ICC computation in the case of a condition-level contrast? With the homoscedasticity assumption $\sigma_{\lambda_1} = \sigma_{\lambda_2} = \sigma_\lambda$, the ICC underestimation rate for a condition contrast is updated from the single effect case (7) to (Appendix D)

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

[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)) in which one condition is coded as 1 while the other serves as the reference or base 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.

with the variability ratio V defined as,

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

Similarly, 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 is due to the double amount of data involved in the cross-trial variability relative to the case of a single condition (i.e., two vs one condition).

The ICC underestimation can be validated through experimental data. Applying the TLM-based LME model (10), we estimated TRR for the condition contrast through two approaches with the prototypical data of Stroop task (Table 1). First, the conventional ICC(3,1) value was estimated as 0.49 while a TRR estimate of 1.0 was retrieved due to a singularity problem when the parameter ρ_1 got trapped 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 $\frac{2}{m}\sigma_0^2 = 6.35 \times 10^{-4}$ is larger than $\sigma_e^2 = 5.53 \times 10^{-4}$, another indication of numerical singularity issue of estimating TRR through the TLM-based LME in (12) and a portending sign of potential assumption violation. Nevertheless, the nearness to the parameter boundary is an indication of TRR close to 1, showing the substantial underestimation of the conventional ICC.

The underestimation of ICC for a condition contrast is more severe than either of the two conditions as well as the average effect between the two conditions. First of all, the extra number of 2 in the underestimation formula in (13) for a condition contrast relative to its counterpart for a single effect in (7) indicates that cross-trial variability V would result in more severe ICC underestimation for a contrast than a single effect. Furthermore, the contrast magnitude is much smaller than that of either condition and the average effect (see the mean column in Table 1). This would lead to a variability ratio V for the contrast a few times larger, further aggravating the underestimation. As shown in Table 1, the variability ratio V for the contrast is about three times larger than that of the single effects, and the conventional ICC estimate of 0.49 for the condition contrast of Stroop task was more substantially underestimated than that for each of the two individual conditions as well as the their average effect.

We hypothesize that TRR estimation for a contrast between two conditions would be much more difficult than that for either condition or their average. As the magnitude of a contrast effect is usually much smaller (the "mean" column in Table 1), the corresponding cross-subject variability σ_{λ_r} is accordingly at least a few times smaller (the "subject" column in Table 1). For example, suppose that the BOLD response under 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 each condition. On the other hand, the cross-trial variability σ_0 remains roughly the same regardless of the effect being a contrast, an individual condition or cross-condition average. Thus, the relative magnitude of cross-trial variability would be much larger, leading to a sizeable variability ratio V. In other words, the cross-subject effects λ_{rs} could get drowned by cross-trial variability σ_0, resulting in a large uncertainty for the TRR estimation of a contrast.

Simulations can also be adopted to explore same aspects for the condition contrast case as in a single condition. With a similar but more complex parameter domain, simulation results are presented in Appendix E, and similar observation patterns transpired as numerated in the previous subsection for a single effect. For example, the role of large cross-trial variability relative to its cross-subject counterpart is also demonstrated through the simulations. More crucially, the precision for TRR estimates was much worse in the simulation

results (wider error bars, Fig. 9) than their counterparts with a single effect, confirming our hypothesis regarding the difficulty of TRR estimation for a contrast.

2.4 Extension of the LME framework to BML

The LME framework is largely unsuited for TRR estimation. Despite its widespread adoption cross many applied fields (including the conventional ICC computation) due to general popularity and computational affordability, its limitations are also conspicuous.

1) **No easy access to the precision information of TRR**. Population-level effects can be estimated with their uncertainty (e.g., standard deviation) under LME. On the other hand, variance for lower-level effects (e.g. subject-specific intercepts and slopes) can also be estimated; however, their uncertainty is usually unavailable. One could resort to the Fisher transformation or an F-statistic (Chen et al., 2018) to reveal the statistical evidence under the conventional ICC framework. However, as the TRR estimation relies on the estimation of variances, it becomes elusive to acquire uncertainty for TRR. Nonparametric approaches such as bootstrapping can be adopted as a workaround solution, but the computational burden is usually heavy.

2) **Unavailability of TRR distribution** The LME framework shares the same methodology of effect estimation under the conventional framework in the sense that a parameter is inferred through a point estimate. Even if uncertainty information is available, no distribution can be conceptually assigned to such a point estimate. As uncertainty information is largely absent for the conventional ICC under LME as illustrated by single ICC values reported in the literature, the concept of TRR distribution practically becomes out of reach under the conventional framework.

3) **Susceptibility to numerical failures**. When the parameter value nears its domain boundary, the numerical scheme may fail to avoid the trap. This is especially a common issue for the correlation parameter of a variance-covariance matrix in the TRR context, and is clearly illustrated with the prototypical RT data of Stroop task (Table 1).

4) **Inflexibility to assumption violations**. Conventional statistical frameworks including LME are mostly conducive to the Gaussian distribution and large sample theory. The TRR estimation may go to rack when (a) sample size is not large enough, (b) the Gaussian assumption is severely violated with, for example, skewed, outlying or diffusive data. Such violations are more common especially with complex parameters such as variance-covariance than ordinary ones (e.g., population mean). Needless to say, outlying effect estimates often occur with both behavioral and neuroimaging data at the trial level, and the typical approach of brute-force censoring leaves some extent of arbitrariness.

5) **No accommodation for sampling errors**. In neuroimaging, trial-level effects are not directly available, and have to be estimated in time series regression solved through generalized least squares. Therefore, the input for TRR estimation contains various extent of uncertainty. However, there is currently no simple numerical approach to incorporating the measurement errors under LME.

It is natural to extend our LME formulations (5) and (12) to the BML framework. Specifically, with a Gaussian likelihood as an example, we convert the two LME models with little modification to their distribu-

tional 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 with all the LME limitations listed above dissolved under BML. For example, numerical instabilities and convergence issues would be largely solved through modeling improvements such as accommodating the data through various distributions such as Student's t, exponentially modified Gaussian, 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 direct measures, but instead estimates 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 exponentially modified Gaussian (exGaussian), zero-inflated negative binomial, etc.

The BML modeling capability can be demonstrated with the prototypical data of the Stroop task. Three likelihood distributions of 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 slightly behind. We added two more distributions, Student's t and exGaussian, with the consideration of their capability of handling skewed and outlying data. In addition, compared to a diagonal matrix $\boldsymbol{R}^{(0)}$ for the variance-covarance structure of the varying intercepts τ_{rs} in the BML model (17), we assume a generic $\boldsymbol{R}^{(0)}$ with results largely consistent with what was reported by Haines et al. (2020) with regard to the three common distributions. However, exGaussian outperformed all alternatives per the information criterion

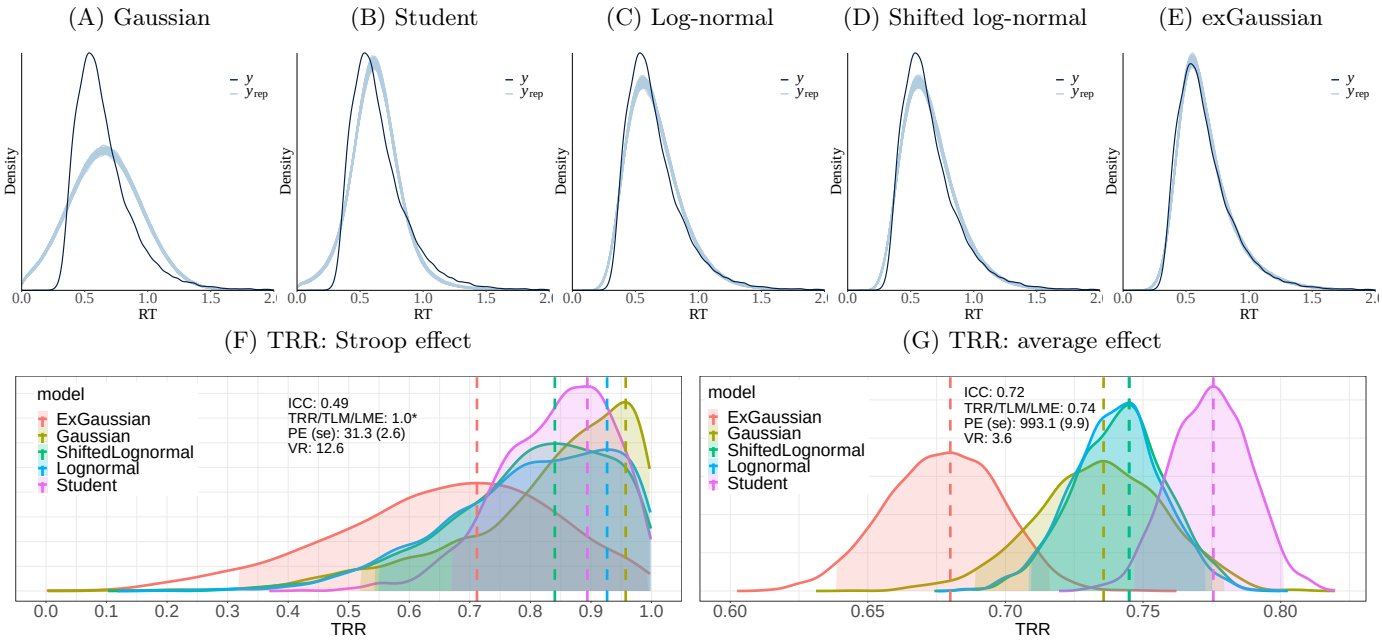


Figure 4: Comparisons among five distribution candidates for RT datat of Stroop effect. (A-E) Visual comparisons are illustrated with each of the five panels showing the posterior predictive density (light blue cloud) that is composed of 200 sub-curves each of which corresponds to one draw from the posterior distribution. Overlaid with the predictive simulations is the raw data (solid black curves with linear interpolation). Differences between the solid black curve and the light blue cloud indicate how well the respective model fits the raw data. (F) TRR distributions for Stroop effect in Study 1 of Hedge et al. (2018) are shown for the five likelihood distributions. The TRR estimates were based on 2000 draws from MCMC simulations of 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 fitted the data the worst. The exGaussian distribution provided the lowest TRR with a mode of 0.71 while achieving the best fit among the five. (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. Their uncertainty is much narrower than the Stroop effect (F).

through leave-one-out cross-validation on the posterior likelihood. The predictive accuracy of the exGaussian model was further confirmed by its best posterior predictive check among the five candidates (Fig. 4).

The modeling capability of the Bayesian framework can be illustrated through model comparisons. The success of the conventional framework heavily relies on assuming a Gaussian distribution as a prior because its nice properties such as ordinary least squares and convenient inferences through standard statistics such as Student's t. However, the convenience may come with a cost: when the prior Gaussian is ill-adapted, one may sacrifice in accuracy. Such a cost can be demonstrated through the Stroop dataset. The TRR estimation based on Gaussian prior had the worst fit among the five priors (Fig. 4A-E), similar to the report by Haines et al. (2020). In contrast, exGaussian was the clear winner thanks to its well-known capability in handling skewed and outlying data such as reaction time that is lower-bounded; such an adaptivity is important considering the arbitrariness involved in the common practice of data cleansing regarding outlying values. The outperformance of exGaussia over the 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 distribution as opposed to specifying a more complex variance-covariance structure as in Haines et al. (2020) than a single dispersion parameter σ_0 for the likelihood.

The rich information from Baysian modeling is worth noting. The TRR estimation based on the BML model (17) with ExGaussian is presented in Fig. 4. Rather than a simple point estimate under the conventional statistical 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, the underestimation of the conventional ICC with the misleading aspect of a point estimate

(ICC = 0.49) was clearly revealed through the Bayesian framework. The adaptivity 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. The BML models for a single 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 an experimental dataset

3.1 Data description

Modified Erikson Flanker Task Data were previously published as part of a larger sample to examine associations between anxiety and neural responses to errors in youth (CITE). Forty-two subjects participated the study and provided usable data: 24 adults (>18 years; age: 26.81 ± 6.36; XX female) and 18 youth (<18 years; age: 14.01 ± 2.48; XX female). They performed a modified Eriksen Flanker task (CITE) with 432 experimental trials during fMRI scanning in each of two separate sessions approximately 53 (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 milliseconds. The trial ended with a blank response screen of 1.700 ms. The task was completed in four runs with three blocks per run to provide intermittent performance feedback to maximize commission errors (see CITE for details). Stimulus presentation and jitter orders were optimized and pseudorandomized using the make_random_timing.py program in AFNI.

Imaging 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/afn Cox, 1996) with standard preprocessing including despiking, slice-timing correction, distortion correction, alignment of all volumes to a base volume with minimum outliers, nonlinear registration to the MNI template, spatial smoothing with a 6.5mm FWHM kernel, masking, and intensity scaling. 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 also were excluded. Participants 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.

Region-of-interest (ROI) Selection

Independent ROIs were selected (to be added)

Table 2: Abbreviations for brain regions

Abbr.	Region	Abbr.	Region	Abbr.	Region
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

Subject-level Analysis At the individual subject level, we analyzed brain activity with a linear regression 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 hemodynamic response. The effects of interest at the condition level were two main contrasts: Cognitive Conflict (Incongruent Correct Responses > Congruent Correct Responses) and Error (Incongruent Commission Errors > Incongruent Correct Responses). All participants contributed to the conflict contrast, only participants with sufficient commission errors in the incongruent condition contributed to the error contrast (27 what is this 27?). List here how many trials for each conflict condition and for error condition: mean ± SD. We analyzed the subject-level data at the whole-brain level as well as within each of the 12 ROIs (Table 2) with two distinct approaches: conventional CLM with regressors created at the condition level and TLM with trial-level regressors.

3.2 TRR estimation for behavioral data

The RT data from the Flanker task are structured as follows. There were two subsets of data: one for the conflict between the two conditions and one for the error responses. For the conflict subset, there were 32005 observations from 42 subjects who performed two sessions of the Flanker task with the congruent and incongruent conditions, which corresponds to about 190 trials per condition per session per subject. The RT values for the conflict data ranged within [4, 1669] ms. For the error subset, there were 11366 observations from 42 subjects who performed two sessions of the Flanker task with the congruent and incongruent conditions, resulting in approximately 68 trials per condition per session per subject. The RT values for the error responses were within [9, 1686] ms.

(A) Contrast between incongruent and congruent

(B) Average between incongruent and congruent

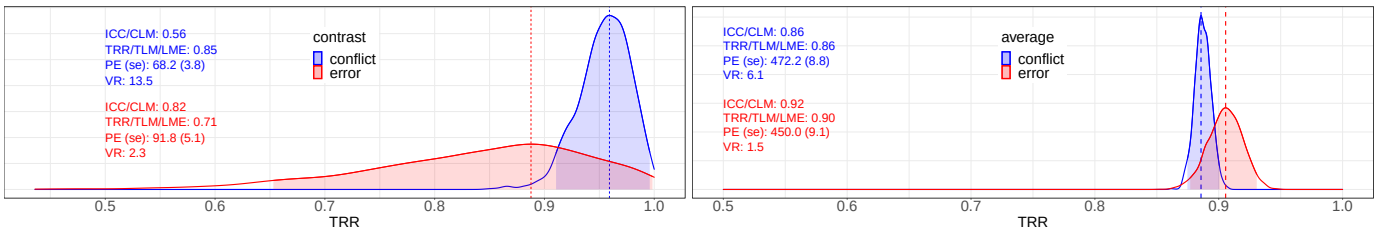


Figure 5: TRR distributions of behavioral data (RT) from Flanker task. The TRR estimates for the contrast (A) and average (B) between the two conditions were based on 2000 draws from MCMC simulations of the BML model with an exGaussian likelihood and are shown here as a kernel density estimate which smooths the posterior samples. The dashed vertical lines indicate the mode (peak) of the 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 strong indicator for the degree of ICC underestimation. Population effects (PE) and standard errors (se) are shown for reference.

The TRR estimates for the RT data of the Flanker task were relatively high. Five distributional candidates were considered under the BML frmework (17): Gaussian, Student's t, log-normal, shifted log-normal and exGaussian. Model comparisons and validations for each of the two RT subsets (conflict and error) indicated

that, similar to the situation with the RT data of the Stroop task (Fig. 4), exGaussian was appropriate although Student's t was also equally fitting. The TRR was largely around 0.9 for both the contrast and the average between the two conditions of the conflict and error data; however, the TRR estimates for the contrast were more concentrated than the average effect and it was the opposite for the error responses (Fig. 5). The individual differences were slightly more reliable for the contrast of the conflicts than that of he errors and were less so for the average effect. The underestimation of the conventional ICC based on point estimation was largely negligible excluding the contrast of conflict, as indicated by the corresponding variability ratio.

3.3 Whole-brain TRR estimation for neuroimaging data

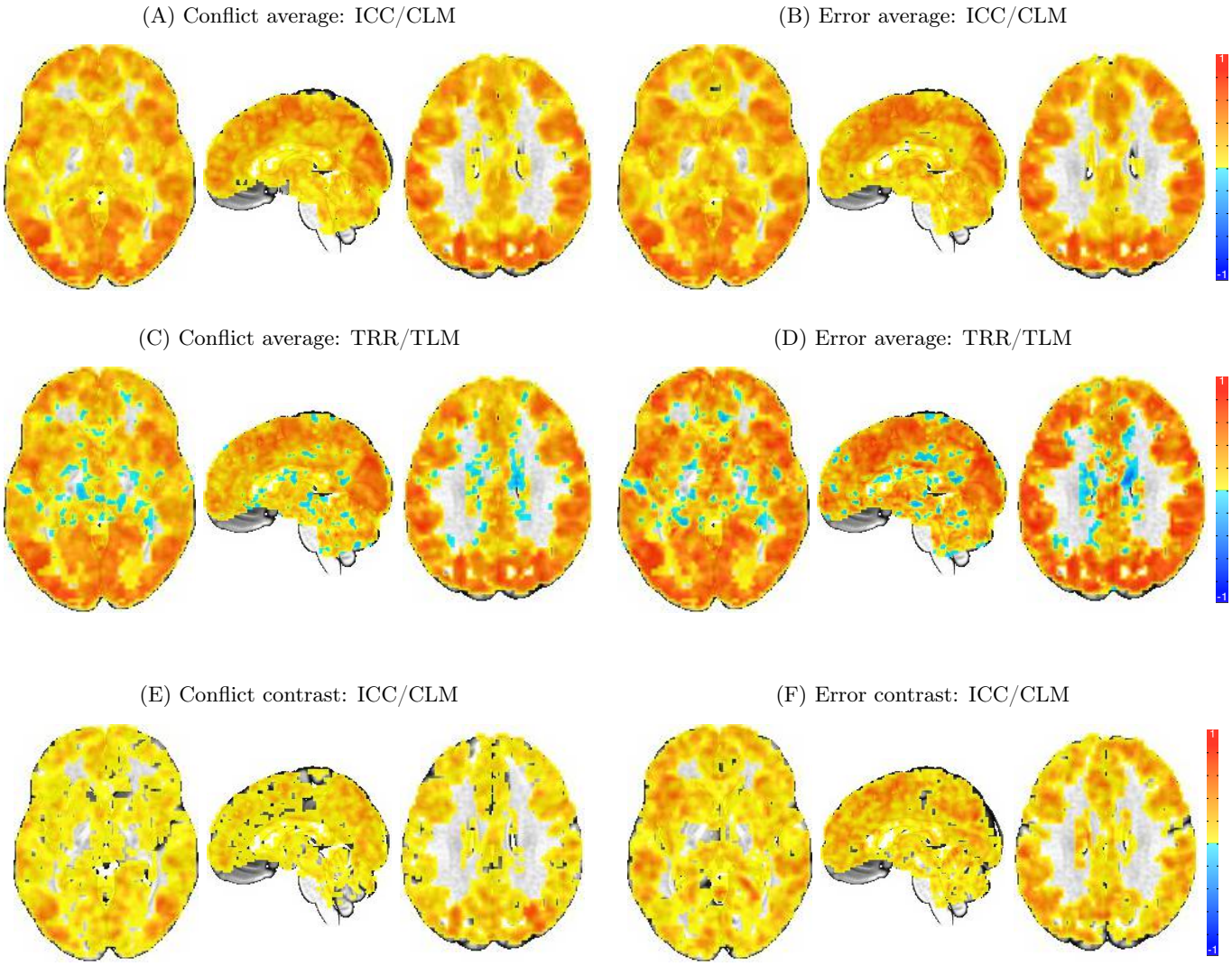


Figure 6: Whole-brain voxel-level test-retest reliability estimates for the FMRI dataset of Flanker task. The conventional ICC values for the average conflict effect (A) showed negligible underestimation compared to the TRR (C) estimated based on TLM. In contrast, the conventional ICC values for the average error effect (B) showed moderate or negligible amount of underestimation compared to the TRR estimation (D) based on TLM. The ICC values for the contrast of conflict (E) and error (F) were much smaller than their average counterparts while the TLM-based approach numerically failed at most voxels in the brain. The three slices of axial ($Z = 0$), sagittal ($Y = 14$) and axial ($Z = 28$) planes are oriented in the radiological convention (left 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-convariance 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.

The TRR analysis for the neuroimaging data was performed as follows at the whole-brain voxel level. As the BML model is not computationally feasible, two modeling frameworks were adopted: conventional ICC with the program 3dICC through condition-level modeling (2) and the LME model (12) with the program

3dLMEr through trial-level modeling. There were totally 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 responses. The input for ICC computation was composed of a condition-level contrast per session from each of the 24 subjects. The trial-level input for TRR estimation had the same structure as the RT data: 32006 and 11367 three-dimensional volumes of trial-level effects for conflict and error, respectively, from two conditions, two sessions and 42 subjects.

The individual differences were largely reliable among most brain regions for the average effect while moot for the contrast. High or moderate amount of TRR for the average effect of both conflict and error were revealed through the point estimates based on the LME model through trial-level modeling (6C vs D). The ICC underestimation (Fig. 6A vs C) was mostly insignificant for the average of conflict responses while small but noticeably for the average of error responses (Fig. 6B vs D). The difference in terms of ICC underestimation between conflict and error was likely due to a much larger number of trials with conflict than error (190 vs 68). Based on the ICC results (Fig. 6E,F), the reliability of individual differences for the contrast of both conflict and error between the two conditions was deemed mostly lackluster across the brain. Unlike the successful execution for the average effect (Fig. 6C,D), numerical failures occurred for most voxels in the brain with the contrast for both conflict and error (results not shown). Thus, the extent of ICC underestimation for the contrast would be explored at the region level.

3.4 Region-level TRR estimation for neuroimaging data

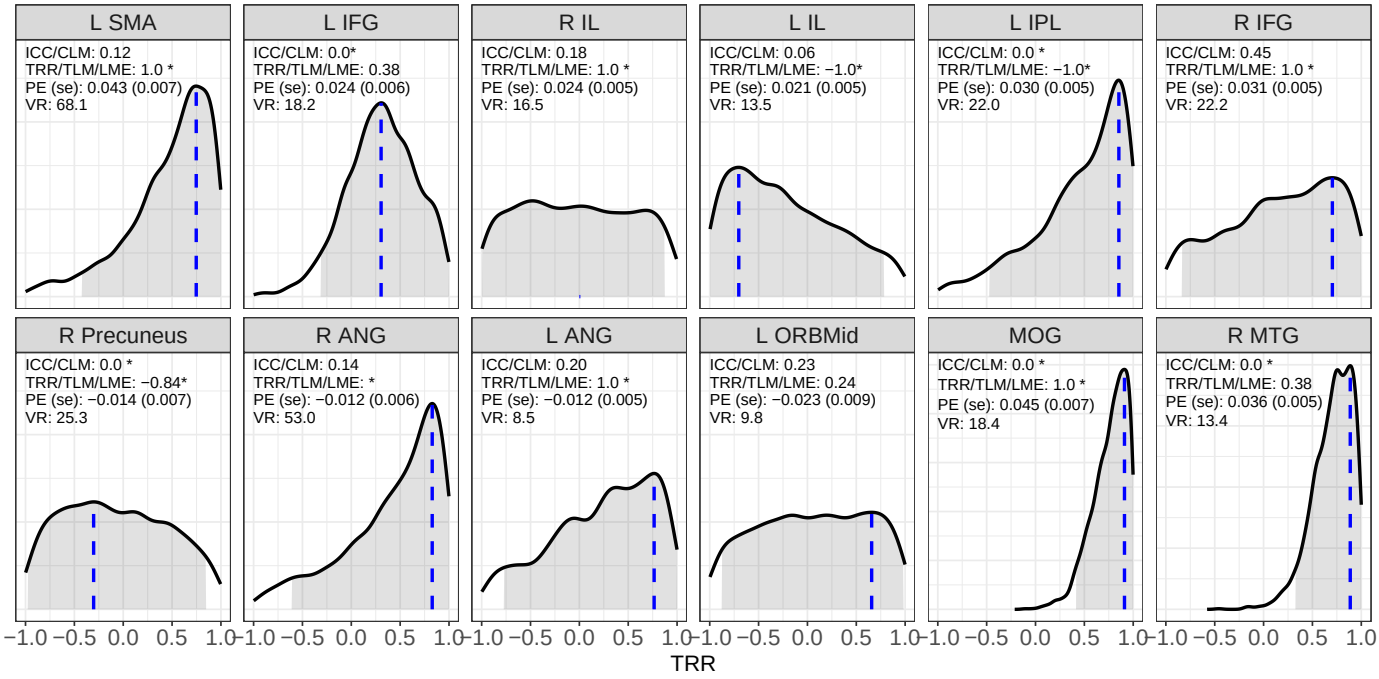
TRR estimation at the region level was performed through BML. Similar to the whole-brain data, for each of the two data subsets, there were totally 32005 trial-level effects for each of the 12 ROIs from 42 subjects for the two conditions of congruent and incongruent during two sessions. The BML model (19) was adopted using the program TRR with the trial-level effect estimates from each subject as input. Student's t-distribution was utilized for cross-trial variability so that any potential outlying values could be accommodated (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.

Large variations existed across ROIs in terms of TRR magnitude and precision (Figs. 7,8). Some regions showed high or moderate TRR with relatively low uncertainty; some regions exhibited high TRR with wide range of uncertainty; others were difficult to assess as their TRR distributions were quite diffusive. For example, the following regions demonstrated high reliability with relatively high precision: MOG and right MTG for both the conflict contrast (Fig. 7A) and the conflict average effect (8A); left SMA and PreCG for the error contrast; left SMA and PreCG the error average effect; left SMA, left IL, right IL, PreCG and MOG for error average effect. Some regions had reasonably high reliability but with moderate or poor precision (e.g., left SMA, left IPL, right ANG 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 the conventional point estimation would be simply out of place.

Simone, could you add something here from the science perspective about the test-retest reliability among those regions? Also, which ROIs were selected based on their cluster survival and which ones were not?

Variability ratio is an approximate indicator for ICC underestimation. The linear underestimation rate (7 or (13) is derived under the assumption of Gaussian distribution for cross-trial variations. Such assumption is largely violated for both behavioral data (e.g., RT for Stroop and Flanker tasks) and BOLD response. Nevertheless, VR offers an index that illustrates why a large heterogeneity exists across various scenarios (e.g., contrast as opposed to single effect, differences across brain regions). For instance, a large VR is usually associated with substantial ICC underestimation in cases such as the Stroop contrast dataset (Fig. 4F), Flanker contrast dataset (Fig. 5), brain regions of left SMA, left IPL. right ANG, MOG and right MTG for

(A) Conflict contrast



(B) Error contrast

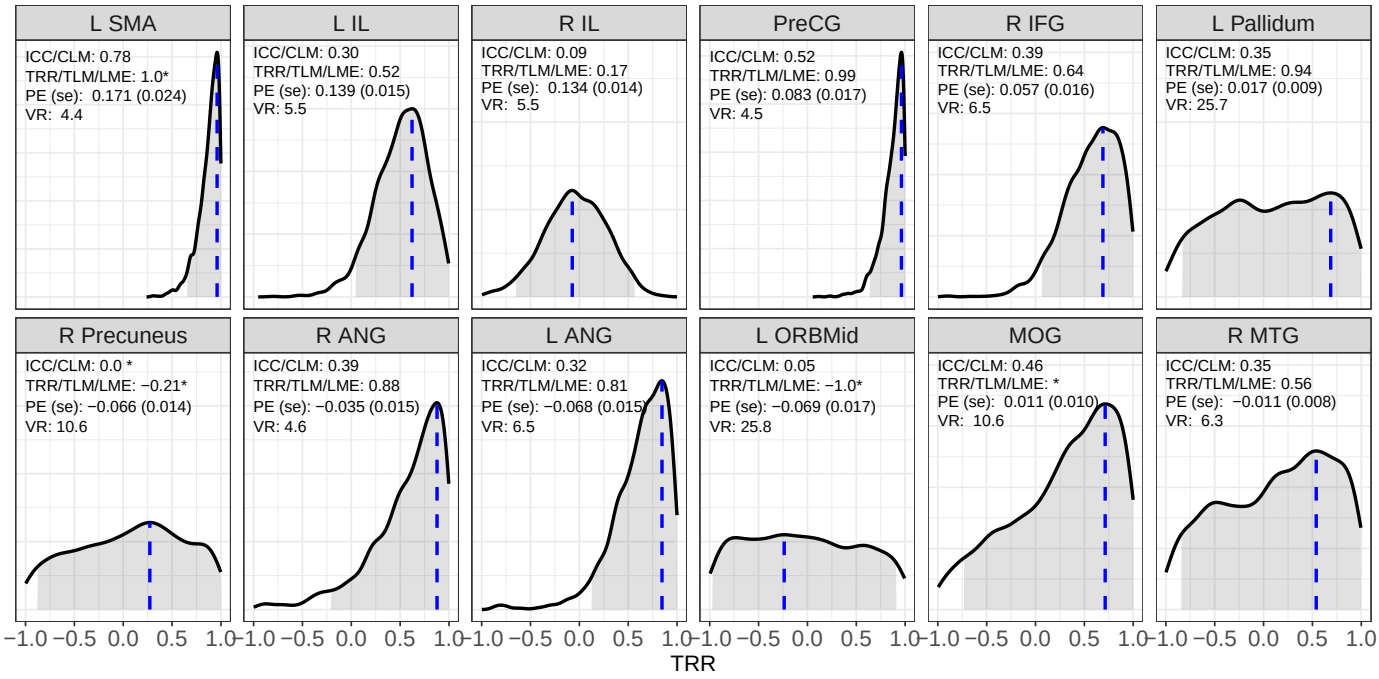


Figure 7: TRR distributions for the contrast of Flanker task at 12 brain regions. The TRR estimates for the contrast of the conflict (A) and error (B) contrast were obtained using the program TRR based on 2000 draws from MCMC simulations of the BML model with an exGaussian likelihood. Each TRR posterior distributions is shown here as a kernel density estimate that smooths the posterior samples. The blue vertical lines indicate 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.

conflict contrast, etc.

Variability ratio is also an approximate indicator for the amount of TRR estimation uncertainty and explains the higher difficulty of TRR estimation for a contrast relative to a single effect. For example, a large VR signified a large extent of TRR uncertainty at all of the regions for conflict contrast except for MOG and right MTG (Fig. 7). The TRR estimation is much more difficult for a contrast than a single effect, and point

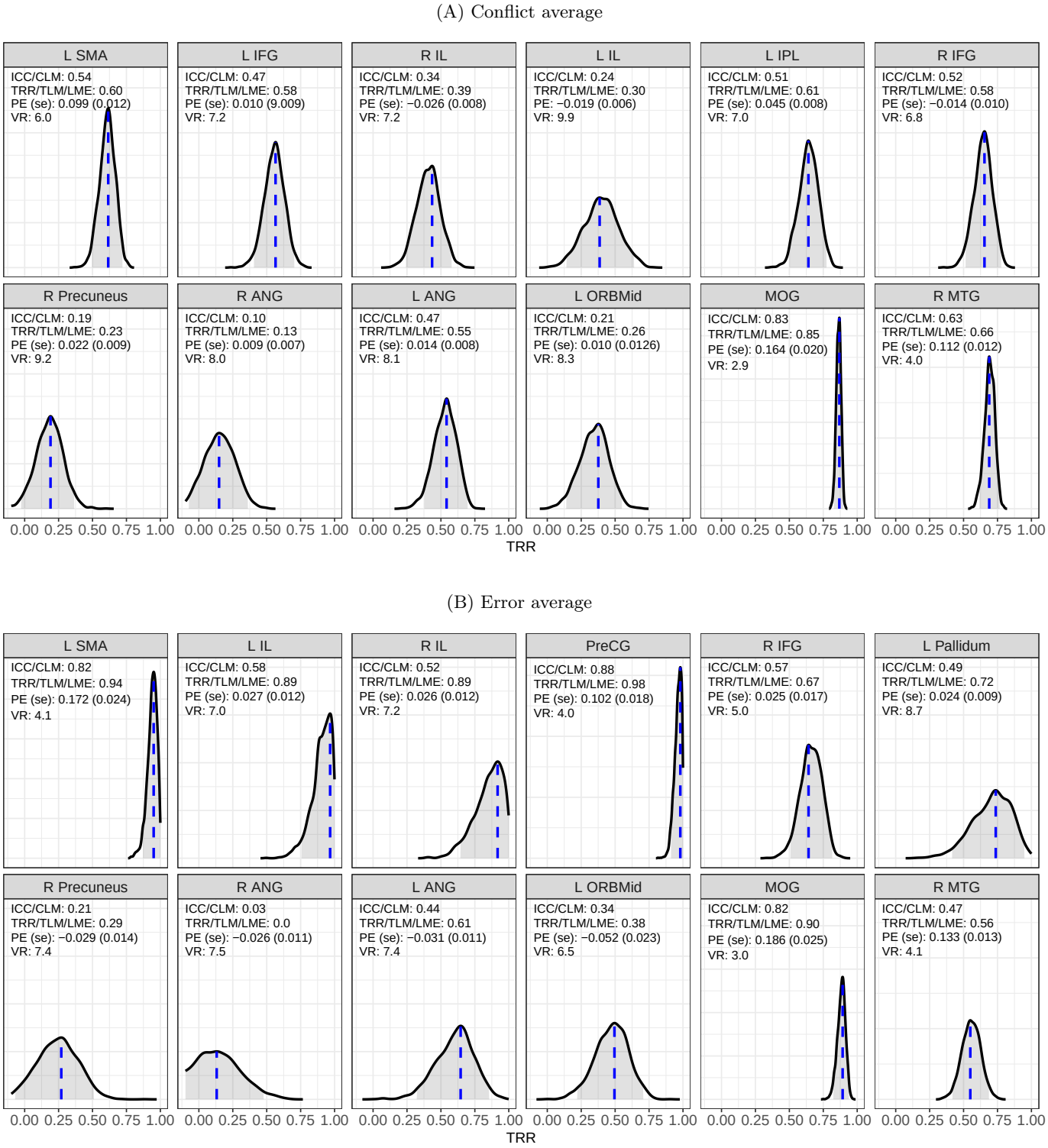


Figure 8: TRR distributions for the average effect of Flanker task at 12 brain regions. The TRR estimates for the average effect of conflict (A) and error (B) contrast were obtained using the program TRR based on 2000 draws from MCMC simulations of the BML model with an exGaussian likelihood. Each TRR posterior distributions is shown here as a kernel density estimate that smooths the posterior samples. The blue vertical lines indicate 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.

estimates are simply not good representations for TRR. For instance, TRR estimation for single effects (e.g. average effect in Fig. 8) tend to be more concentrated with moderate or even negligible ICC underestimation (Fig. 4G, Fig. 6, Fig. 8). As a comparison, ten out of the twelve regions had a substantial range of uncertainty for the TRR of both conflict and error contrast (Fig. 7) that were usually associated with a large variability ratio. It is for this reason that TRR cannot simply expressed as point estimates through the conventional

ICC.

4 Discussion

Test-retest reliability is of great interest in exploring the consistent pattern of individual differences. A high TRR signifies the internal replicability of a test and ensures that the effect estimates assessed under one sitting or one time point are stable as well as representative over different circumstances or time. To avoid potential confounds of time or age, the experiment is typically designed with two time points that span a short period (e.g., a few weeks). A low TRR may indicate inconsistent and inexplicable variations at the subject level; more importantly, it would discourage or even abandon the adoption of the effect of interest as untrustworthy measurement for device (e.g., MRI scanner) reliability and clinical applications such as biomarkers, personalized interventions and individualized diagnoses. For example, when neuroimaging data from a group of subjects are acquired before and after a hardware/software upgrade of an MRI scanner or across two scanners, low TRR may cause grave concern regarding the reliability as well as reproducibility across scanners.

Proper modeling is pivotal in accurately assessing TRR. The aim of statistical modelling is data abstraction through an idealistic formulation about the potential data generating process. Our investigation here illustrates that data should not be forced to adapt to a Procrustean bed of model formulation at hand through information reduction; rather, one should construct an adaptive model that (a) reasonably characterizes the data hierarchy, (b) handles outliers and skewness in a principled way, and (c) properly quantifies uncertainty. Through TRR modeling, we intend to demonstrate these three dimensions along which the conventional statistical framework struggles to handle. When substantiating a theoretical hypothesis, one usually employs experimental trials as the lowest level of data hierarchy whose variability is typically ignored and shrouded during preprocessing such as averaging or complete pooling in neuroimaging. In addition, Gaussianity is largely taken for granted as a default prior but may wreak havoc on estimation accuracy. Lastly, point estimation such as ICC under the conventional statistical framework, offer little room for uncertainty quantification.

4.1 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 instances. From the 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.

4.2 BML as an adaptive framework to address the inadequacy of ICC

In recent years, TRR has been found much lower than anticipated when assessed by the conventional ICC. Such low ICC values have occurred with behavior data (e.g., RT) for inhibition effects and implicit attitude tasks for measuring implicit bias as well as many neuroimaging studies including both task-related and resting state experiments. The seemingly baffling situation has led to fruitful explorations in modeling improvements (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 with many trials. 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) lack of reliability when modeling assumptions are violated or when numerical failures occur. As shown in the case of inter-rater reliability, ICC was originally designed for data structure without the notion of many trials. Nowadays almost all experiments involve a large number of trials due to subtle effect sizes. Nevertheless, just because trial-level effects are of no interest does not necessarily mean they could be fully ignored. Similarly, averaging trial-level effects, despite its long history, as commonly practiced in condition-level modeling in psychometrics and neuroimaging does not properly account for them and equates to hideously carrying them over to subsequent analyses without general awareness.

The serious problem of underestimation foils the generalizability of ICC as a reliability metric. With trial-level effects not explicitly characterized in the associated ANOVA/LME model, the conventional ICC is linearly related to the hypothetical TRR with an underestimation rate that depends on two factors: the trial sample size and the relative magnitude of cross-trial variability to cross-subject variability. Specifically, the larger the number of trials (or cross-trial variability), the less (or more) severe the ICC underestimation. As ICC is implicitly contaminated by cross-trial variability, and is thus sensitive to trial sample size. As a result, ICC values reported in the literature may have embedded a different degree of underestimation due to different trial sample sizes; therefore, results may not necessarily be comparable across experiments, leading to a portability issue (e.g., Rounder and Haaf, 2019). Conceptually, such underbiased ICC estimates present a challenge as the generalizability about the reliability of individual differences is lost.

Hierarchical models such as LME and BML through TLM allows one to disentangle trial-level effects from the rest, providing more accurate embodiment of TRR for the underlying data generating process.

By definition, ICC as a TRR measure concerns about subject-level effects; thus, trial-level effects should be appropriately untangled in effects partitioning. In other words, even though trial-level effects are of no interest, their accountability in the modeling scaffold is crucial because the disentanglement of the cross-trial variability from the cross-subject variability allows the accurate revelation 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 generating process and simultaneously incorporates all the information available through regularization and shrinkage. As a result, one effectively gains a high predictive accuracy for TRR estimation.

The BML framework is well-suited for TRR estimation. The adoption of the Bayesian formulation is not intended to inject prior information, but to overcome several limitations of the LME approach through Monte Carlo simulations. Despite its accommodation of multilevel data structure, LME models may encounter numerical hiccups, provide point estimates with no easy access to estimation uncertainty and have a limited flexibility of handling deviations from a Gaussian prior such as data skewness and outliers. BML can further overcome theses LME issues. For example, rather than censoring data through brute force and 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 V, the modeling investment is worthwhile in revealing TRR values above 0.9 in a real experimental data (Fig. 4).

4.3 Important role of trial sample size

Two types of sample size, trials and subjects, are involved in typical psychometric and neuroimaging studies. Per central limit theory, a reasonably large sample size of an experiment unit is amenable to conventional properties such as Gaussian distribution that 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 whose reasonable range cannot necessarily be met nor easily predetermined in real practice. In fact, sample size 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). Between the two types of samples, the number of trials tends to have an anonymous status. On one hand, they are adopted with the intention to achieve robust effect estimates at the condition level; on the other hand, they do not occupy an equivalent position in modeling as its subject counterpart. Not only are cross-trial variations usually not captured in modeling, but also is the realization largely lacking as to its pivotal role in achieving accurate effect estimation including both overall effects at the population level (Chen et al., 2020) as well as measures such as TRR at the subject level.

The number of trials plays a much more crucial role than that of subjects in the amount of ICC underestimation under the conventional formulation. Likely based on an extended perception from the population-level effect estimation, 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 applicability of that extension is quite limited, and the trial sample size has not gained much attention until recently (Rouder et al., 2019). In fact, the trial number maintains 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.

Trial sample size is also more pivotal in determining the precision of TRR estimation and a substantially large number of trials may be required to achieve reasonable precision. Even though underestimation is no longer an issue under BML, the uncertainty of TRR estimation as represented by, for example, standard

derivation 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, 7 and 8), the precision of TRR estimation varies substantially across brain regions or between simple effects and contrasts. To dissolve those diffusive posterior distribution of TRR, a few hundred or even more trial samples may have to be adopted. In reality, such experiment designs may not be sustainable due to the extraordinarily length and financial burden.

4.4 High cross-trial variability

Experimental data indicate that cross-trial variability is much larger than cross-subject variability. If the former is roughly in the same order of magnitude or smaller, the conventional ICC might be considered a shortcut to approximate TRR as shown in the formulations (7) and (13). However, real experiments including both behavioral and neuroimaging data point to a much larger cross-trial 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 two negative impacts: the underestimation of TRR when the conventional ICC formulation is adopted and a poor precision of TRR estimation under the BML framework.

The substantially larger cross-trial variability remains to be 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 (Table 1). 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. However, they are not purely random fluctuations in the sense that contralateral regions illustrate a high degree of trial-level synchronization (Chen et al., 2020). In addition, the cross-trial fluctuations are to some extent associated with behavioral measures such as reaction time and stimulus rating that are typically modeled through trial-level modulation analysis at the subject level. One may hypothesize that the random appearance of effects across trials may be caused by momentary lapses in attention. Or, brain regions may constantly undergo some intrinsic fluctuations while external stimuli or tasks are simply perturbations that ride above the large wave of internal neuronal activities.

4.5 Difficulty of obtaining high TRR precision on a contrast

The precision of TRR estimation under BML depends on the relative magnitude of cross-trial variability. Precision information is not readily accessible under the conventional ICC framework, but it may evince through a different channel such as estimation bias and numerical issues. Even though the magnitude of the variability ratio V would not on average cause any bias on TRR estimation, it could have a substantial impact on the uncertainty of TRR estimates. If cross-trial variability σ_0 is relatively large, it may dwarf cross-subject variability σ_{τ_r} and result in imprecise TRR estimation with a diffusive or even close to uniform posterior distribution (Fig. 7). A close examination of such undesirable situations is definitely necessary.

Simple effects are relatively easy to achieve a reasonable extent of 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 in the range of a few to 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 8). 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 hard to attain under some circumstances. The reason is the following. Cross-trial variability σ_0 measures the fluctuations surrounding the trial-level 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 the latter case measures the fluctuations regarding the contrast, not per the individual condition effects nor at trial level. The contrast between two conditions is usually a few times or more smaller in scale in terms of effect magnitude. Therefore, the variability ratio V is usually larger for a contrast than for a single effect, resulting in a higher severity of the impact on the TRR precision for a contrast. Another aggravating factor in neuroimaging 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. 7). In general, we recommend the adoption of the BML framework for its flexibility and adaptivity in closely characterizing the data information. Despite the potential difficulty of achieving a high precision for TRR estimation, the resulting posteriors from a BML model would likely encapsulate the distribution shape regardless of its centrality or diffusivity.

4.6 The relationship between population-level effects and TRR

Population effects are not necessarily tied to the reliability of individual differences, and TRR of individual difference should be open to broader explorations regardless the strength of population-level effects (magnitude or statistics). As clearly illustrated in Fig. 1, various scenarios may occur without contradictions: large population effects may correspond to a strong or weak TRR, and low population effects could be associated with a strong or weak TRR. The disconnection between the two is not commonly recognized in the field. In general practice, the investigator usually chooses regions based on their involvement in strong population effects. In other words, it is the existence of strong evidence of population-level effects that usually piques the interest of exploring the reliability of individual differences. However, one should not be limited to such evidence.

Another complexity in neuroimaging involves the issue of multiplicity. As massively univariate analysis is widely adopted, multiplicity is an intrinsic issue that results in substantial information waste (Chen et al., 2019) and leads to a heavy penalty in the process of multiple testing adjustment. The detrimental impact of the currently common practice of dichotomization through multiple testing adjustment further exacerbates the difficulty in region detection purely based on the commonly accepted rigidity of statistical evidence in the form of surviving clusters, leading to missing opportunities in TRR estimation. For example, the statistical evidence associated with the four regions (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 their selection had been based on the statistical strength at the population level.

4.7 The awareness of modeling assumptions

The construction, comparison and validation of models are a complex process. It might be a cliche to state that all models are wrong (Box, 1976), but we all should be aware of the pitfalls in the unending pursuit of potential modeling improvements. The critical assessment of various underlying assumptions of each model is always a necessity when we learn from the data in the presence of uncertainty, and TRR estimation illustrates this point from multiple angles.

The importance of characterizing data hierarchy cannot be overlooked. Cross-trial variability parallels to its cross-subject counterpart. The typical averaging step in behavioral data or the split between the subject and population steps through condition-level modeling in neuroimaging implicitly means a complete pooling strategy for cross-trial variability with an underlying assumption that all trials share exactly the same effect. Such a focus on condition-level effect has been broadly explored recently in neuroimaging (e.g., Westfall et al., 2017; Chen et al., 2020) and may result in distortions of both effect estimation and statistical evidence. If cross-subject variability is currently required to be properly accounted for through partial pooling as commonly practiced in population-level modeling in neuroimaging as well as in TRR estimation, there is no legitimate reason not to properly handle cross-trial variability.

Distributional assumption is another aspect that is usually unrecognized. With only two parameters of location and scaling, Gaussian distribution has many desirable properties such as guaranteed asymptotic convergence with large enough samples and numerical frugality. Its dominant adoption convincingly demonstrates the power and adaptivity of the Gaussian distribution. However, sample size could be an expensive commodity in real practice including TRR estimation for both behavioral and neuroimaging data collection. When not large enough, a sample size may render data with a skewed or diffusive distribution. A typical approach is to censor out those oddities. In addition to the arbitrariness involved (e.g., how to select a particular threshold?), the performance of the adopted model is rarely justified nor compared to potential candidates. In contrast, the accommodation of the Bayesian framework illustrates the distributional flexibility of data fitting including the incorporation of measurement errors.

Model comparison and validation are also important in improving the accuracy of TRR estimation. When multiple model candidates are available, an effective and robust approach is necessary. For reaction time data, in addition to the conventional Gaussian model, one could adopt a principled approach in handling data skewness and outliers. For example, through model comparisons and validation (Fig. 4), our investigation indicated that exponentially modified Gaussian distribution might be better suited for the reaction time data than its Gaussian and log-normal counterparts, demonstrating its strength in accommodating data skewness and outliers and confirming the adaptivity of the exGaussian distribution for RT data in the literature (Ratcliff, 1979).

Point estimation can be misleading and may cause numerical failures. As a first-order moment, an effect estimator for the mean of data samples is widely adopted in conventional statistics, and its uncertainty can be reasonably characterized through the associated standard deviation. The popularity of point estimation is ascribed to its information extraction 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 could misrepresent the whole picture 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. The conventional ICC may have given a familiar face of point estimation for TRR in common practice. The BML framework clearly illustrates that such a perception should be abandoned. Furthermore, the statistical evidence for ICC based on Fisher-transformation or F-statistic (Chen et al., 2018) would not do justice to presenting the TRR precision. A large degree of uncertainty may show up as numerical failures in the LME model, necessitating more trials (and more subjects to a lesser extent) and better experimental designs.

4.8 Limitations

Currently the BML framework is practically not feasible TRR analysis at the whole brain level, and is limited to behavioral and region-based neuroimaging data. Because of 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. For behavioral or region-level neuroimaging data, TRR analysis can be performed through, for example, the program TRR. For whole-brain voxel-level analysis, the LME approach may work for a single condition or a contrast with large effects using the program 3dLMEr; 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 (Rounder et al., 2019), we surmise that hundreds of trials might be required for neuroimaging studies to narrow down TRR estimation uncertainty.

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.

5 Conclusion

The conventional intraclass correlation could underestimate the test-retest reliability to various extent for datasets with multiple trials. Its loss of accuracy is two-fold: first, the underlying model 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 estimation uncertainty. We recommend that TRR be estimated as the subject-level correlation across repetitions through a Bayesian multilevel model that maps the data structure as accurate as possible, and 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. Our work was inspired by the modeling platforms of Haines et al. (2020) and Rounder 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 package brms (Bürkner, 2018) and with lmer the R package 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,

$$\xi_s = \frac{1}{2}(\tau_{1s} + \tau_{2s}),\ \eta_s = \frac{1}{2}(\tau_{1s} - \tau_{2s}), \tag{20}$$

we have,

$$\begin{aligned}
&\overline{y}_{1s\cdot} \sim \mathcal{N}(a_1 + \xi_s + \eta_s,\ \frac{1}{m}\sigma_0^2),\ \overline{y}_{2s\cdot} \sim \mathcal{N}(a_2 + \xi_s - \eta_s,\ \frac{1}{m}\sigma_0^2),\\
&\mathrm{Var}(\xi_s) = \frac{1}{2}(1+\rho)\sigma_\tau^2,\ \mathrm{Var}(\eta_s) = \frac{1}{2}(1-\rho)\sigma_\tau^2,\ \mathrm{Cov}(\xi_s,\ \eta_s) = \frac{1}{4}(\mathrm{Var}(\xi_s) - \mathrm{Var}(\eta_s)) = 0,\\
&\mathrm{Var}(\overline{y}_{1s\cdot}) = \mathrm{Var}(\xi_s) + \mathrm{Var}(\eta_s) + \mathrm{Cov}(\xi_s,\ \eta_s) + \frac{1}{m}\sigma_0^2 = \sigma_\tau^2 + \frac{1}{m}\sigma_0^2,\\
&\mathrm{Var}(\overline{y}_{2s\cdot}) = \mathrm{Var}(\xi_s) + \mathrm{Var}(\eta_s) - \mathrm{Cov}(\xi_s,\ \eta_s) + \frac{1}{m}\sigma_0^2 = \sigma_\tau^2 + \frac{1}{m}\sigma_0^2,\\
&\mathrm{Cov}(\overline{y}_{1s\cdot},\ \overline{y}_{2s\cdot}) = \mathrm{Var}(\xi_s) - \mathrm{Var}(\eta_s) = \rho\sigma_\tau^2.
\end{aligned} \tag{21}$$

Through the notations,

$$V = \frac{\sigma_0}{\sigma_\tau},\ U = \frac{1}{1 + \frac{1}{m}V^2}, \tag{22}$$

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

$$\mathrm{ICC(3,1)} = \frac{\mathrm{Cov}(\overline{y}_{1s\cdot},\ \overline{y}_{2s\cdot})}{\sqrt{\mathrm{Var}(\overline{y}_{1s\cdot})\ \mathrm{Var}(\overline{y}_{2s\cdot})}} = \frac{\rho\sigma_\tau^2}{\sigma_\tau^2 + \frac{1}{m}\sigma_0^2} = \frac{1}{1 + \frac{1}{m}V^2}\rho = U\rho, \tag{23}$$

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 (21) indicate that $\mathrm{Var}(\overline{y}_{1s\cdot}) = \mathrm{Var}(\overline{y}_{2s\cdot}) = \sigma_\tau^2 + \frac{1}{m}\sigma_0^2$. That is, the variability of cross-trial averages is composed of two components, one associated with cross-subject variability 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),

$$\mathrm{ICCa} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2 - \frac{1}{m}\sigma_0^2}. \tag{24}$$

B Simulations for a single 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) Fixed standard deviations (scaling parameters): $\sigma_{\tau_1} = \sigma_{\tau_2} = \sigma_\tau = 1$

2) 4 different TRR values were chosen: $\rho = 0.3$, 0.5, 0.7 and 0.9.
3) 4 different sample sizes of subjects: $n = 20$, 40, 70 and 100.
4) 4 different sample sizes of trials per repetition: $m = 20$, 40, 70 and 100.
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 based on CLM through ANOVA/LME (2) through aggregation across trials;
   (b) TRR ρ based on TLM through LME (5);
   (c) conventional ICC adjusted by removing the cross-trial variability $\frac{\sigma_0^2}{m}$.

Each of these $4 \times 4 \times 4 \times 2 \times 3 = 384$ combinations was simulated 1000 times , using the function *lmer* in the *R* package lme4 (Bates et al., 2015) with the following iterative steps.

i) For each subject s, obtain subject-level effects during the two repetitions through random sampling: $\begin{bmatrix}\tau_{1s}\\ \tau_{2s}\end{bmatrix} \sim \mathcal{N}(\mathbf{0}_{2\times 1},\ \boldsymbol{R}_{2\times 2})$, where $\boldsymbol{R} = \begin{bmatrix}\sigma_{\tau_1} & 0\\ 0 & \sigma_{\tau_2}\end{bmatrix}\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\begin{bmatrix}\sigma_{\tau_1} & 0\\ 0 & \sigma_{\tau_2}\end{bmatrix} = \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix},\ s = 1, 2, .., n.$
ii) Construct the simulated data per TLM under LME (5):
$y_{rst} \sim \mathcal{N}(a_r + \tau_{rs},\ \sigma_0^2),\ r = 1, 2;\ s = 1, 2, .., n;\ t = 1, 2, .., m.$
iii) Solve the two LME models of CLM (2) and TLM (5).
iv) Retrieve the simulated parameters including the three TRR estimates. Specifically, the conventional ICC is obtained through CLM (3) while the TRR ρ is estimated through TLM (5). In addition, we adjust the conventional ICC by removing the cross-trial variability $\frac{\sigma_0^2}{m}$ through the formula (**??**).

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

Simulations for reliability with an effect contrast under the LME model (5) are similar to the situation with a single effect but with a slightly higher complexity. 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,

$$\begin{aligned}&(\mu_{11s}, \mu_{12s}, \mu_{21s}, \mu_{22s})^T \sim \mathcal{N}(\mathbf{0}_{4\times 1},\ \boldsymbol{P}_{4\times 4}),\\ &\boldsymbol{P} = \text{diag}(q_{11}, q_{12}, q_{21}, q_{22})\ \boldsymbol{C}\ \text{diag}(q_{11}, q_{12}, q_{21}, q_{22}),\\ &s = 1, 2, ..., n,\end{aligned} \tag{25}$$

where $\boldsymbol{P}$ and $\boldsymbol{C}$ are the variance-covariance and correlation matrix for the four effects, respectively. 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 C is of the following structure,

$$\begin{bmatrix}1 & \pi & \theta & \eta\\ \pi & 1 & \eta & \theta\\ \theta & \eta & 1 & \pi\\ \eta & \theta & \pi & 1\end{bmatrix}, \tag{26}$$

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, their correlation matrix 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),

$$\boldsymbol{C} = \begin{bmatrix} 1 & \rho_0 & 0 & 0 \\ \rho_0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho_1 \\ 0 & 0 & \rho_1 & 1 \end{bmatrix}, \tag{27}$$

where the correlation between the two varying intercept components

$$\rho_0 = \frac{\text{cov}(\tau_{1s},\ \tau_{2s})}{\sqrt{\text{var}(\tau_{1s})\ \text{var}(\tau_{2s})}} = \frac{\text{cov}(\frac{1}{2}(\mu_{11s}+\mu_{21s}),\ \frac{1}{2}(\mu_{12s}+\mu_{22s}))}{\sqrt{\text{var}(\frac{1}{2}(\mu_{11s}+\mu_{21s}))\ \text{var}(\frac{1}{2}(\mu_{12s}+\mu_{22s}))}} = \frac{\pi+\eta}{1+\theta}., \tag{28}$$

and the two varying slope components

$$\rho_1 = \frac{\text{cov}(\lambda_{1s},\ \lambda_{2s})}{\sqrt{\text{var}(\lambda_{1s})\ var(\lambda_{2s})}} = \frac{\text{cov}(\mu_{11s}-\mu_{21s},\ \mu_{12s}-\mu_{22s})}{\sqrt{\text{var}(\mu_{11s}-\mu_{21s})\ \text{var}(\mu_{12s}-\mu_{22s})}} = \frac{\pi-\eta}{1-\theta}. \tag{29}$$

The correlation of 0s in (27) can be similarly derived as in (28) and (29).

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

The underestimation for a contrast is worse than the case of a single effect. Similar to the situation with a single effect, the extent of ICC underestimation for a contrast depends on two factors, the trial sample size m and the relative magnitude of cross-trial variability ratio V. One difference is that, because a contrast involves two single effects, the underestimation is more severe as shown by the presence of 2 in the denominator of the underestimation rate U.

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 condition contrast

Simulations for reliability with a condition-level contrast were performed through the TCM with the LME model (12). With the assumption of homoscedasticity across the four condition-level effects in (25): $q_{11} = q_{21} = q_{12} = q_{22} = 1$, the cross-session reliability or TRR in the TLM with the LME model (12) is determined as ρ_0 in (28) and ρ_1 in (28).

$$\rho = \frac{cov(\frac{1}{2}(\tau_{1p1} + \tau_{2p1}),\ \frac{1}{2}(\tau_{1p2} + \tau_{2p2}))}{\sqrt{var(\frac{1}{2}(\tau_{1p1} + \tau_{2p1}))\ var(\frac{1}{2}(\tau_{1p2} + \tau_{2p2}))}} = \frac{\theta_1 + \theta_3 + \theta_4 + \theta_6}{2\sqrt{(1+\theta_2)(1+\theta_5)}}. \tag{30}$$

The following four varying parameters were considered

1) 4 different reliability values were chosen with ρ = 0.3, 0.5, 0.7 and 0.9 that, respectively, correspond to four sets of correlations in $\boldsymbol{C}$ per (30):

$$\begin{aligned}
&(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = \\
&(0.7, 0.5, 0.55, 0.55, 0.5, 0.7), \\
&(0.7, 0.5, 0.45, 0.45, 0.5, 0.7), \\
&(0.7, 0.5, 0.35, 0.35, 0.5, 0.7), \\
&(0.7, 0.5, 0.25, 0.25, 0.5, 0.7).
\end{aligned}$$

2) 4 different numbers of subjects: P = 20, 40, 70 and 90.
3) 4 different numbers of trials per session: T = 20, 40, 70 and 90.
4) 4 different ratios of cross-trial variability σ_0 relative to cross-subject variability s: $\frac{\sigma_0}{s}$ = 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 on CLM under ANOVA/LME (2) through aggregation across trials;
   (b) cross-session reliability ρ based on trial-level modeling (TLM) through the LME formulation (12);
   (c) conventional ICC based on CLM adjusted by removing the cross-trial variability $\frac{2\sigma_0^2}{m}$.

Each of these $4 \times 4 \times 4 \times 2 \times 3 = 384$ 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:
$$\begin{bmatrix} \tau_{1p1} \\ \tau_{1p2} \\ \tau_{2p1} \\ \tau_{2p2} \end{bmatrix} \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{R}^{(1)}), \text{where } \boldsymbol{R}^{(1)} = \begin{bmatrix} 1 & \theta_1 & \theta_2 & \theta_3 \\ \theta_1 & 1 & \theta_4 & \theta_5 \\ \theta_2 & \theta_4 & 1 & \theta_6 \\ \theta_3 & \theta_5 & \theta_6 & 1 \end{bmatrix},\ p = 1, 2, .., P.$$
2. Construct the simulated data per the LME formulation (10):
$y_{crst} \sim (N)(a_{cs} + \tau_{cps},\ \sigma_0^2),\ r = 1, 2;\ s = 1, 2, .., n;\ t = 1, 2, .., m.$
3. Solve the two LME models, CLM (2) and TLM (12), using the function *lmer* in the *R* package lme4 (Bates et al., 2015).
4. Retrieve the simulated parameters including the three reliability estimates. Specifically, the conventional ICC is obtained through the CLM formula (3) while the cross-session reliability ρ is estimated through the LME model (12) with TLM. In addition, we adjust the conventional ICC by removing the cross-trial

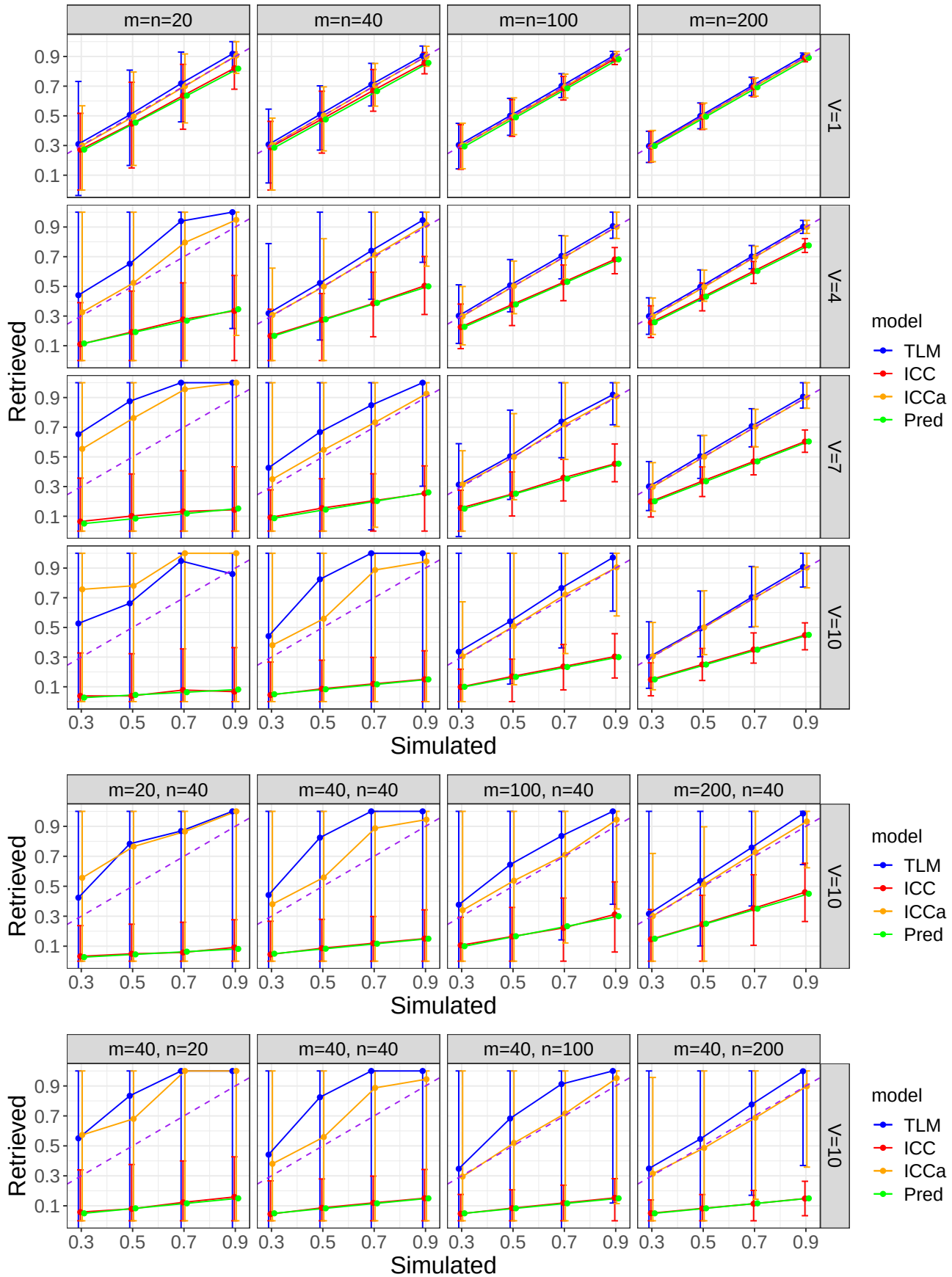


Figure 9: Simulation results for a condition contrast. The four columns correspond to the sample size of subjects and trials while the four rows are the varying standard deviation ratios of $\frac{\sigma_0}{\sigma_\tau}$. The x- and y-axis are the simulated and retrieved TRR, respectively. Each data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

variability $\frac{2\sigma_0^2}{m}$, per the same logic in (24), from the denominator of the CLM formula (3),

$$\text{ICCa} = \frac{\widetilde{\sigma}_\tau^2}{\widetilde{\sigma}_\tau^2 + \sigma_e^2 - \frac{2\sigma_0^2}{m}}. \tag{31}$$

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., trial, subject) effects considered here is Gaussian, as specified in the respective model; for example, see the distribution assumptions in the BML models (16, 18, 17, 19). If justified, one could adopt other priors like Student's t for the effects across trials and subjects, just as for the likelihood (or the prior for the response variable y in the BML models. In addition, prior distributions (usually called hyperpriors) are needed for three types of model parameters in each model: (a) population effects or location parameters ("fixed effects" under LME, such as intercept and slopes), (b) standard deviations or scaling parameters for lower-level effects ("random effects" under LME), 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 effects (e.g., population-level intercept and slopes). In contrast, weakly-informative priors are utilized for standard deviations of lower-level parameters such as varying slope, subject-, trial- and region-level effects, 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.

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

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

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. https://doi.org/10.1177/0956797620916786

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. https://doi.org/10.1016/j.jmva.2009.04.0

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

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