Intraclass correlation: improved modeling approaches and applications for neuroimaging

Gang Chen*[a], Paul A. Taylor[a], Simone P. Haller[b], Katharina Kircanski[b], Joel Stoddard[c], Daniel S. Pine[d], Ellen Leibenluft[b], Melissa A. Brotman[b], and Robert W. Cox[a]

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

[b]Section on Mood Dysregulation and Neuroscience, Emotion and Development Branch, National Institute of Mental Health, USA

[c]Division of Child and Adolescent Psychiatry, Department of Psychiatry, University of Colorado School of Medicine, USA

[d]Section on Development and Affective Neuroscience, Emotion and Development Branch, National Institute of Mental Health, USA

Abstract

Intraclass correlation (ICC) is a reliability metric that gauges similarity when, for example, entities are measured under similar, or even the same, well-controlled conditions, which in MRI applications include runs/sessions, twins, parent/child, scanners, sites, etc. The popular definition and interpretation of ICC are usually framed statistically under the conventional ANOVA platform. Here, we show that the ANOVA framework is often limited, rigid, and inflexible in modeling capabilities. Additionally we provide a comprehensive overview of ICC analysis in its prior usage in neuroimaging. These intrinsic limitations motivate several novel improvements. Specifically, we start with the conventional ICC model under the ANOVA platform, and extend it along two dimensions: first, fixing the failure in ICC estimation when negative values occur under degenerative circumstance, and second, incorporating reliability information of effect estimates into the ICC model. These endeavors lead to four modeling strategies: linear mixed-effects (LME), Bayesian mixed-effects (BME), multilevel mixed-effects (MME), and Bayesian multilevel mixed-effects (BMME). Compared to ANOVA, each of these four models directly provides estimates for fixed effects as well as their statistical significances, in addition to the ICC estimate. These new modeling approaches can also accommodate missing data as well as

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

fixed effects for confounding variables. More importantly, we show that the novel MME and BMME approaches offer more accurate characterization and decomposition among the variance components, leading to more robust ICC computation. Based on these theoretical considerations and model performance comparisons with a real experimental dataset, we offer the following general-purpose recommendations. First, ICC estimation through MME or BMME is preferable when precision information is available for the effect estimate; precision information provides weights that more accurately allocate the variances in the data. When precision information is unavailable, ICC estimation through LME or the novel BME is the preferred option. Second, even though the absolute agreement version, ICC(2,1), is presently more popular in the field, the consistency version, ICC(3,1), is a practical and informative choice for whole-brain ICC analysis that achieves a well-balanced compromise when all potential fixed effects are accounted for. Third, approaches for clear, meaningful, and useful result reporting in ICC analysis are discussed. All models, ICC formulations, and related statistical testing methods have been implemented in an open source program `3dICC`, which is publicly available as part of the AFNI suite. Even though our work here focuses on the whole brain level, the modeling strategy and recommendations can be equivalently applied to other situations such as voxel, region, and network levels.

Introduction

Recently, reliability and reproducibility have been hot topics in science in general and in the neuroimaging community in particular. It is known that neuroimaging data are very noisy and large proportion of the data variability cannot be properly accounted for: It is fair to assess that less than half of the data variability can currently be explained in the typical data analysis. For example, the typical analytical approach is to make a strong and unrealistic assumption that a hemodynamic response is the same across brain regions, subjects, groups, different tasks or conditions. In addition, even though large amount of physiological confounding effects are embedded in the data, it remains a daunting task to incorporate the physiological noise in the model.

Recent surveys indicate that about 60% of published experiments failed to survive replication in psychology (Baker, 2015) and about 40% in economics (Bohannon, 2016), and the situation with neuroimaging is equally, if not more, pessimistic (Griffanti et al., 2016). While in neuroimaging, reproducibility typically is a qualitative description as to whether an activation cluster is reported across many studies, reliability of an effect estimate can be defined as the variation in repeated measurements performed on the same measuring entities (e.g., subjects in neuroimaging) under the identical or approximately the same experimental conditions (NIST, 2007). Specifically, reliability can be quantitatively defined as the agreement or consistency across two or more measurements, and intra-class correlation (ICC) has been specifically developed for this purpose (Shrout and Fleiss, 1979; McGraw and Wong, 1996).

Generally speaking, the conventional ICC metric indicates agreement, similarity, stability, consistency, or reliability among multiple measurements of a quantity. For instance, such a quantity can be the ratings of n targets assessed by k raters or judges in a classical example (Shrout and Fleiss, 1979). In the neuroimaging

context, when the same set of measuring entities (e.g., subjects or families) go through the same experiment protocol under the same conditions, ICC can be utilized to assess the data quality. Those multiple measurements can be the effect estimates from n subjects (or families) under k different replications (e.g., runs, sessions, scanners, sites, twins, siblings, parent-child pairs, studies, assessments, diagnoses, or analytical methods). The same quantity (rating or effect estimate) across those k replications in the ICC definition is reflected in the word *intra*class, as opposed to the Pearson (or *inter*class) correlation coefficient that reveals the linear relationship between two quantities that can be of different nature (e.g., brain response and cortex thickness).

Here we first review the various versions of ICC definition and their computation formulations under their classic ANOVA platform, and then discuss their limitations and drawbacks as our motivations to develop more extended models. Even though our work mainly focuses on whole brain data analysis in neuroimaging, the methodologies can be applied to other contexts or fields when the underlying assumptions are met.

Various types of ICC

In this section, we introduce the three classic types of ICC, motivating each from their basic statistical model and describing their interpretation, applicability, and generalization. Throughout this article, regular italic letters in lower (e.g., α) stand for scalars and random variables; boldfaced italic letters in lower ($\boldsymbol{a}$) and upper ($\boldsymbol{X}$) cases for column vectors and matrices, respectively; Roman and Greek letters for fixed and random effects on the righthand side of a model equation. In the neuroimaging context, let y_{ij} be the effect estimate (BOLD response in percent signal change or connectivity measurement) from the ith level of factor A and jth level of factor B (e.g., the ith member of family j) ($i = 1, 2, ..., k; j = 1, 2, ..., n$). A one-way random-effects ANOVA can be formulated to decompose the response variable or effect estimate y_{ij} as

$$y_{ij} = b_0 + \lambda_j + \epsilon_{ij}, \tag{1}$$

where b_0 is a fixed effect or constant representing the overall average, λ_j codes the random effect of the jth level of factor B (e.g., family j), which is assumed to be independent and identically distributed (i.i.d.) with Gaussian distribution $\mathscr{G}(0, \sigma_\lambda^2)$, and the residuals ϵ_{ij} are also i.i.d. with $\mathscr{G}(0, \sigma_\epsilon^2)$. In addition, λ_j and ϵ_{ij} are assumed to be independent with respect to each other. A prototypical application for the model (1) is the scenario with the effect estimates of k siblings or twins ($k = 2$) from each family. The reason that the effects of factor A are not explicitly modeled is that there is no meaningful way to assign an order or sequence among the levels of factor A consistently among the levels of factor B (e.g., ordering twins within each family).

The ICC for the model in (1) is defined as

$$\rho_1 = \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_\epsilon^2}, \tag{2}$$

which can be estimated by

$$\hat{\rho}_1 = \frac{MS_\lambda - MS_\epsilon}{MS_\lambda + (k-1)MS_\epsilon}, \tag{3}$$

where MS_λ and MS_ϵ are the mean squares (MSs) associated with the family effects λ_j and the residuals, respectively, in the ANOVA framework of model (1). The definition (2) is usually referred to as ICC(1,1) in the literature (e.g., Shrout and Fleiss, 1979; McGraw and Wong, 1996).

To make statistical inference, Fisher's transformation for ICC value ρ (McGraw and Wong, 1996),

$$z = \frac{1}{2}\sqrt{\frac{2(n-2)(k-1)}{k}} \ln \frac{1+(k-1)\rho}{1-\rho} \tag{4}$$

approximately follows a standard Gaussian $\mathscr{G}(0,1)$ under the null hypothesis H_0: $\rho = 0$, and offers a solution for significance testing. However, a better technique is to formulate an F-statistic (McGraw and Wong, 1996),

$$F_1(n-1, n(k-1)) = \frac{MS_\lambda}{MS_\epsilon}, \tag{5}$$

whose distribution is exact under H_0: $\rho_1 = 0$, unlike the Fisher transformation (4).

The meaning of the ICC can be interpreted by four common perspectives:

i. As the definition (2) itself indicates, the ICC is the proportion of total variance that is attributed to a random factor (or accounted for by the association across the levels of the random factor). For instance, if the variance associated with some random factor B increases, the levels of factor B (e.g., families in the model (1)) would be less similar while the levels of some factor A (e.g., siblings or twins) tend to be relatively more similar, leading a higher ICC value. This interpretation leads to the non-negativity of ICC in the range of $[0, 1]$.

ii. The ICC is the expected correlation between two effect estimates that are randomly drawn among the levels of factor A. For instance, ICC(1,1) for a family study shows the genetic relatedness among monozygotic or dizygotic twins, full or half siblings, cousins, and so on[1]. As the ordering among the siblings or between twins within each family is difficult to define meaningfully, neither random nor fixed effects are considered for the individual family members in the model (1), and the corresponding ICC(1,1) is used to assess the familial resemblance as a whole. Specifically, with the assumptions in the model (1), ICC(1,1) is essentially the Pearson correlation of the effect estimates between two siblings or twins, i_1 and i_2 ($i_1 \neq i_2$), within a family,

$$Corr(y_{i_1 j}, y_{i_2 j}) = \frac{Cov(\lambda_j + \epsilon_{i_1 j}, \lambda_j + \epsilon_{i_2 j})}{\sqrt{Var(\lambda_j + \epsilon_{i_1 j}) Var(\lambda_j + \epsilon_{i_2 j})}} = \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_\epsilon^2}.$$

[1]The interpretation of genetic relatedness can also be applied to the case of ICC(3,1) between parent and child.

However, it is worth emphasizing that this equivalence between ICC and Pearson correlation holds because of the following two facts: 1) ICC is a relationship gauge between any two measuring entities in light of the same physical measure (e.g., BOLD response for all family members in ICC(1,1) as opposed to Pearson correlation between weight and height), and 2) it is assumed that any two levels of factor A share the same mean and variance, b_0 and σ_λ^2, with the model (1). When some measuring entities have generally higher (or lower) effect estimates relative to the group mean effect, or when there is some extent of consistency among the factor B levels within each measuring entity, then those effect estimates are correlated, and the ICC formulation (2) basically captures that correlation or consistency.

iii. ICC is an indicator of homogeneity of the responses among the levels (e.g., families) of the random factor B: a higher ICC means more similar or homogeneous responses. On the other hand, when an ICC value is close to zero, the effect estimates associated with a measuring entity are no more similar than those from different families, and the random effect components could be removed from the model (1).

iv. ICC reflects the extent of common conditions (e.g., same task and scanning parameters) or common genetic information (e.g., siblings, twins) that the effect estimates share. The ICC would be higher if effect estimates associated with the same factor level (e.g., family) were under more similar environments and genetic influence.

The determining choice of fixed versus random effects for an explanatory variable in a model can be subtle, and the distinction is usually determined in light of the nature of the factor: interchangeability of factor levels, or whether there exists some systematic difference, across the factor levels. For example, families or subjects can be often considered as the levels of a random factor because: 1) they are each recruited through a random sampling process as representatives (or samples) of a potential or conceptual population (as embodied in the assumption of Gaussian distribution $\mathscr{G}(0, \sigma_\lambda^2)$ for the random effects λ_j in the model (1)), achieving the goal of generalization in statistical inferences; 2) their order does not matter in the model; and 3) a particular set of families or subjects can in practice be replaced by another set. In contrast, parents and children are typically handled as fixed effects because of the lack of interchangeability (due to significant difference in age, experience, etc.). In neuroimaging scanners or sites can be thought of as the levels of either a random- or fixed-effects factor, depending on whether the scanning parameters are similar or different across scanners or sites. Similarly, runs or sessions should be treated as random effects if no significant systematic difference exists across runs or sessions; otherwise, they should be modeled as fixed effects when habituation or familiarity effect of the task is substantial.

Because of the distinction between fixed versus random effects, there are two other ICC definitions that are more popular due to their wider applicability. Let y_{ij} be the effect estimate from subject j at the ith level of within-subject (or repeated-measures) factor A ($i = 1, 2, ..., k; j = 1, 2, ..., n$). When the factor A (e.g., runs,

sessions, scanners, sites) is modeled as random effects, we have a two-way random-effects ANOVA system,

$$y_{ij} = b_0 + \pi_i + \lambda_j + \epsilon_{ij}, \tag{6}$$

where π_i is the level-specific random effect associated with the ith level of factor A, and λ_j represents the subject-specific random effect[2]. The associated ICC is defined as

$$\rho_2 = \frac{\sigma_\lambda^2}{\sigma_\pi^2 + \sigma_\lambda^2 + \sigma_\epsilon^2}, \tag{7}$$

with an extra variance term in the denominator due to the presence of random effects for factor A in the model (6). The ICC in this case can be computed by

$$\hat{\rho}_2 = \frac{MS_\lambda - MS_\epsilon}{\frac{k}{n}(MS_\pi - MS_\epsilon) + MS_\lambda + (k-1)MS_\epsilon}, \tag{8}$$

Both Fisher's transformation (4) and the F-statistic,

$$F_2(n-1, (n-1)(k-1)) = \frac{MS_\lambda}{MS_\epsilon}, \tag{9}$$

offer approximate significance testing for the null hypothesis H_0: $\rho_2 = 0$.

When the factor A (e.g., runs, sessions, scanners, sites, parent-child pairs) is modeled as fixed effects, we have a two-way mixed-effects ANOVA structure,

$$y_{ij} = b_0 + b_i + \lambda_j + \epsilon_{ij}, \tag{10}$$

where b_i and λ_j represent the fixed effects of factor A and random effects of subjects (or families in the case of parent versus child), respectively. The associated ICC is defined as

$$\rho_3 = \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_\epsilon^2}, \tag{11}$$

which is the same formula as (2), but we note that the presence of b_i here means that the two models would have different estimates of σ_ϵ^2. The ICC in this case can still be computed similarly as (3),

$$\hat{\rho}_3 = \frac{MS_\lambda - MS_\epsilon}{MS_\lambda + (k-1)MS_\epsilon}, \tag{12}$$

[2]For simplicity, the notations for the model terms and for the corresponding variance and MS terms in the ICC formulas are undifferentiated across models. To avoid confusions, we emphasize that the values for the same notation term may change across models. For example, residuals ϵ_{ij} and its variance estimate $\hat{\sigma}_\epsilon^2$ may differ, depending on how the effects associated with factor A are modeled.

Table 1: Hypothetical effect estimates (e.g., BOLD response in percent signal change) with a relationship from five subjects during two sessions ($y_{2j} = y_{1j} + 0.2, j = 1, 2, ..., 5$; ICC(1,1) = 0.43, ICC(2,1) = 0.56, ICC(3,1) = 1, Pearson correlation $r = 1$).

subject	s1	s2	s3	s4	s5
session 1, y_{1j}	0.1	0.2	0.3	0.4	0.5
session 2, y_{2j}	0.3	0.4	0.5	0.6	0.7

with an exact F-statistic,

$$F_3(n-1,(n-1)(k-1)) = \frac{MS_\lambda}{MS_\epsilon}, \tag{13}$$

for significance testing under the null hypothesis H_0: $\rho_3 = 0$.

The two ICC definitions, (7) and (11), are popularly notated as ICC(2,1) and ICC(3,1), respectively, and are sometimes referred to as "intertest ICCs," extensions of the classic inter-rater reliability (Shrout and Fleiss, 1979). When there are only two levels for factor A ($k = 2$), these two versions are usually referred to as test-retest reliability measures. The four interpretation perspectives for ICC(1,1) also apply directly to these two popular versions. For example, with the assumptions in the model (1), ICC(2,1) is a special case of Pearson correlation of the effect estimates between any two levels of factor A, i_1 and i_2 ($i_1 \neq i_2$),

$$Corr(y_{i_1j}, y_{i_2j}) = \frac{Cov(\pi_{i_1} + \lambda_j + \epsilon_{i_1j}, \pi_{i_2} + \lambda_j + \epsilon_{i_2j})}{\sqrt{Var(\pi_{i_1} + \lambda_j + \epsilon_{i_1j})Var(\pi_{i_2} + \lambda_j + \epsilon_{i_2j})}} = \frac{\sigma_\lambda^2}{\sigma_\pi^2 + \sigma_\lambda^2 + \sigma_\epsilon^2}.$$

ICC(3,1) is more similar to Pearson correlation than ICC(1,1) and ICC(2,1) in the sense that each level of factor A is assumed to have a different mean, but still unlike Pearson correlation, the assumption of same variance holds across the levels of factor A for all the three ICC definitions.

In addition, ICC(2,1) and ICC(3,1) are sometimes called *absolute* and *consistent* agreement, respectively. The distinction between these two ICC types can be hypothetically illustrated by paired effect estimates (e.g., in percent signal change for an FMRI experiment) from five subjects during two sessions (Table 1). The three ICC values are all nonnegative because they represent a proportion of total variance embedded in the data, and they generally follow a sequential order (cite...): $ICC(1,1) \leq ICC(2,1) \leq ICC(3,1)$.

The first index in the ICC(...,...) notation specifies the ICC type, while the second indicates the relationship between two *single* measurements (e.g., between two twins for ICC(1,1) or two levels of factor A for ICC(2,1) and ICC(3,1)). For each of the three *single measurement* ICC types, there is another version, called *average measurement* ICC, which shows the relationship between two sets of average measurements among the k levels of factor A, and with notations ICC(1,k), ICC(2,k), and ICC(3,k), they are similarly defined as the single measurements version except that the terms in the denominator, σ_ϵ^2, $\sigma_\pi^2 + \sigma_\epsilon^2$, and σ_ϵ^2, for ICC(1,1), ICC(2,1) and ICC(3,1), respectively, are each scaled by a factor of k^{-1}. By definition, the average measurements ICC is larger than its single measurements counterpart. In addition, a similar correlation interpretation about the average measurement ICCs can be seen with, for example, ICC(2,k),

$$Corr(\frac{1}{k}\sum_{i=1}^{k} y_{ij}^{(1)}, \frac{1}{k}\sum_{i=1}^{k} y_{ij}^{(2)}) = \frac{Cov(\frac{1}{k}\sum_{i=1}^{k}\pi_i^{(1)} + \lambda_j + \frac{1}{k}\sum_{i=1}^{k}\epsilon_{ij}^{(1)}, \frac{1}{k}\sum_{i=1}^{k}\pi_i^{(2)} + \lambda_j + \frac{1}{k}\sum_{i=1}^{k}\epsilon_{ij}^{(2)})}{\sqrt{Var(\frac{1}{k}\sum_{i=1}^{k}\pi_i^{(1)} + \lambda_j + \frac{1}{k}\sum_{i=1}^{k}\epsilon_{ij}^{(1)})Var(\frac{1}{k}\sum_{i=1}^{k}\pi_i^{(2)} + \lambda_j + \frac{1}{k}\sum_{i=1}^{k}\epsilon_{ij}^{(2)})}} = \frac{\sigma_\lambda^2}{\frac{1}{k}\sigma_\pi^2 + \sigma_\lambda^2 + \frac{1}{k}\sigma_\epsilon^2},$$

where the superscript such as those in $y_{ij}^{(1)}$ and $y_{ij}^{(2)}$ indicates a particular set of data substantiation, and thus $\pi_i^{(m)} \overset{iid}{\sim} \mathscr{G}(0, \sigma_\pi^2)$, $\lambda_j^{(m)} \overset{iid}{\sim} \mathscr{G}(0, \sigma_\lambda^2)$, $\epsilon_{ij}^{(m)} \overset{iid}{\sim} \mathscr{G}(0, \sigma_\epsilon^2)$. As the average measurement ICC is less popular in practice, we hereafter focus on the their single measurement counterpart, but our modeling work below can be directly extended to the average measurements ICC.

Motivations for further modeling work

The ANOVA framework has been adopted in computing ICC for historical reasons because ANOVA was developed and became widely adopted in early 20th century: the framework is widely introduced in basic statistics textbooks, and the MS terms are computationally efficient. Various computational tools are widely available through ANOVA (e.g., packages *irr* and *psych* in R). However, the ANOVA approach does have both limitations for interpretation and practical drawbacks for calculation:

i. *Negative value.* Although the ICC value should be theoretically nonnegative per its definition as the proportionality of the total variance (as in (7)), its estimation (as in (8)), may become negative due to the fact that the numerator in the computational formula is the difference between two MS terms. Such cases are uninterpretable. Importantly, in neuroimaging they are not rare occurrences, with a large number of such voxels appearing in the brain (and in any tissue type).

ii. *Missing data.* In common practical situations, missing data may occur. As data balance is essential in partitioning the MS terms, ANOVA cannot properly handle missing data due to the breakdown of the underlying rigid variance-covariance structure. For example, the data for the six subjects who missed scanning for one session as shown in Table 2 would have to be abandoned due to the rigidity of the ANOVA structure.

iii. *Confounding effects.* Under some circumstances it might be desirable to incorporate explanatory variables into a model, so that the variability due to those confounding effects can be properly accounted for. For example, subject groupings (e.g., sex, handedness, genotypes), age, and reaction time are typical candidates in neuroimaging that often warrant their inclusion in the model. However, the rigid ANOVA structure usually does not easily allow for such inclusions[3].

[3]One exception is that, when a subject-grouping factor (e.g., males versus females) is considered, it is possible to construct the involved MS terms in the special case of having exactly equal number of subjects across all groups. However, even for such a balanced scenario, specific MS terms would have to be derived for each ICC computation formula.

Table 2: A hypothetical study of 17 subjects with missing data (1: available, 0: missing)

subject	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16	s17
session 1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
session 2	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1
session 3	1	0	1	1	1	0	1	1	1	1	1	0	1	1	0	1	1

iv. *Sampling errors.* Conceptually, the residual term ϵ_{ij} in the ICC model can be considered to represent measurement or sampling errors. In other words, the underlying assumption is that all the effect estimates y_{ij} share the same sampling variance for the measurement errors. However, unlike the typical situation in other fields where the effect estimates are usually direct measurements (e.g., scores assessed by raters), the effect estimates in neuroimaging are obtained through a data reduction process with a regression model for each subject. That is, each effect estimate is associated with a standard error that indicates the precision[4] information of the estimation, and heterogeneity or (possibly unfounded) heteroscedasticity is expected because the standard error varies across subjects as well as between twins or parents/children, across runs/sessions or scanners/sites. When the standard error for the effect estimate is ignored in favor of a homoscedasticity assumption in the ANOVA formulation (as widely practiced in neuroimaging when computing the ICCs, as well as in group analysis), it begs the question: what is the impact for the ICC estimate when heterogeneity of standard error is substantially present across the measuring entities?

v. *Type selection.* The applicability of ICC(1,1) is limited to situations where the levels of the repeated-measures factor are measuring entities that are difficult to assign meaningful orders or sequences such as twins. However, between ICC(2,1) and ICC(3,1), the choice becomes challenging for the investigator, with nontrivial impact on the ICC results: other than some prior knowledge about the potential existences of confounding effects or systematic difference across the levels of the repeated-measures factor, there is no solid statistical tool under ANOVA to leverage one choice over the other. For example, is there any statistical evidence that could allow us to decide unequivocally on ICC(2,1) or ICC(3,1) by treating runs or sessions as random or fixed effects? In addition, the typical whole brain analysis through a massively univariate approach at the voxel level may further aggravate the choice.

vi. *Discrete responses.* The conventional ICC definition hinges around the Gaussianity assumption about the response variable (e.g., BOLD response in neuroimaging). However, in real practice, an extension may be desirable in cases where the response variable is discrete instead of being continuous. For instance, a diagnosis is usually binary (e.g., schizophreniac/autistic/bipolar or not), while events are countable (e.g., number of epileptic episodes during a day). Such scenarios are beyond the scope of ANOVA.

Our ICC modeling work here hinges around these five limitations about the ANOVA solution. We first discuss four modeling approaches as extensions to the ANOVA framework, and then use an experimental dataset

[4]Precision is defined as the reciprocal of the variance.

Table 3: Issues in ICC computations under ANOVA, and summary of which mixed-effects models can (✓) or cannot (✗) address each.

Issues	ANOVA	LME	BME	MME	BMME
negative ICC	✗	✓	✓	✓	✓
zero ICC	✗	✗	✓	✗	✓
missing data	✗	✓	✓	✓	✓
confounding effects	✗	✓	✓	✓	✓
sampling error	✗	✗	✗	✓	✓
type selection	✗	✓	✓	✓	✓
discrete responses	✗	✓	✓	✓	✓

to examine the performance of the various modeling methods. A literature survey and further discussion are presented at the end. The implementations of our modeling work are publicly available for voxel-wise computation through program `3dICC` as part of the AFNI suite (Cox, 1996).

Theory: three extended ICC models

Here we propose four mixed-effects models: linear mixed-effects (LME) model, Bayesian mixed-effects (BME) model, and meta-analytic mixed-effects (MME) model. Each of the three models addresses the limitations and drawbacks we discussed earlier, and provides incremental improvements (Table 3).

Linear mixed-effects (LME) modeling

Whenever multiple values (e.g., two effect estimates from two scanning sessions) from each measuring entity (e.g., subject or family) are correlated (e.g., the levels of a within-subject or repeated-measures factor), the data can be formulated with an LME model, sometimes also referred to as a multilevel or hierarchical model. One natural extension to the ANOVA modeling in (1), (6), and (10) is to simply treat the model conceptually as LME, reformulating neither the equations nor their ICC definitions. This LME approach for ICC has perviously been implemented in the program 3dLME (Chen et al., 2013) for voxel-wise data analysis in neuroimaging. For example, the ICC(2,1) model is an LME with two crossed random-effects terms, whose applications can be seen under other circumstances such as inter-subject correlation analysis (Chen et al., 2016) and psycholinguistic studies (e.g., Baayen et al., 2008).

However, the application of LME methodology for ICC does not stop at the conceptual level, and in fact it has several advantages in some aspects of computation where limitations are present under the ANOVA framework. Specifically, as the variances for the random effects components and the residuals are directly estimated through optimizing the restricted maximum likelihood (REML) function, and thus the ICC value is computed with variance estimates $\hat{\sigma}^2_\pi, \hat{\sigma}^2_\lambda$, and $\hat{\sigma}^2_\epsilon$, through the definitions in (2), (7), or (11), instead of with their counterpart with MS terms, (3), (8), or (12), under ANOVA. Therefore, just as the theoretical quantities, the estimated ICC is nonnegative by definition, avoiding the interpretability difficulty with ANOVA when negative ICC estimate occurs. Similarly, the three F-statistic formulas, (5), (9), and (13), can be expressed in

terms of variance estimates as well,

$$F_i = \frac{k\hat{\sigma}_\lambda^2}{\hat{\sigma}_\epsilon^2} + 1, i = 1, 2, 3. \tag{14}$$

Another convenient byproduct from the LME model interpretation of (6) for ICC(2,1) is that the ICC for the absolute agreement between any two measuring entities (e.g., subjects) can be easily obtained through ratio among the variances,

$$\hat{\rho}_2 = \frac{\hat{\sigma}_\pi^2}{\hat{\sigma}_\pi^2 + \hat{\sigma}_\lambda^2 + \hat{\sigma}_\epsilon^2},$$

which is usually not discussed under the ANOVA framework.

In regard to the type selection choice between ICC(2,1) and ICC(3,1), which is an ambiguous one to some extent in the ANOVA framework, in the LME model (10) the fixed effects associated with the levels of the repeated-measures factor A can now be directly examined through statistical assessment: If the fixed effects b_i can be deemed negligible (i.e., no confounding effects nor systematic differences across the factor A levels), the ICC(2,1) offers a better metric because the model (6) is more parsimonious. We will elaborate this point later through an experimental dataset.

Furthermore, missing data can be naturally handled in LME because parameters are estimated through the optimization of the (restricted) maximum likelihood function where a balanced structure is not required. As long as the missingness can be considered random, not structural, (i.e., no systematic pattern exists among the missing data) and there are enough measuring entities present (e.g., at least 10) for each level of the repeated-measures factor, then the computation can still be performed and essentially no information loss is involved.

In addition, the extension to incorporate confounding effects is readily available through adding more fixed-effects terms into the model. For instance, the ICC(2,1) model (6) can be expanded to an LME model with two crossed random-effects components,

$$y_{ij} = b_0 + \sum_{l=1}^{m} b_l x_j^{(l)} + \pi_i + \lambda_j + \epsilon_{ij}, \tag{15}$$

where $x^{(1)}, x^{(2)}$, ..., $x^{(m)}$ are m explanatory variables (e.g., sex, age) that can be either categorical or quantitative, and b_1, b_2, ..., b_m are their corresponding fixed effects.

Last, generalized linear models can be incorporated into the mixed-effects framework, and such an extension would allow us to handle discrete responses such as binary or countable data. For example, if the investigator is interested in assessing the reliability between computer and human diagnosis of a disease (schizophrenia, autism, or bipolar disorder), logistic modeling can be applied easily under the mixed-effects platform through logit link function. Similarly, countable variables (e.g., number of seizures during a fixed period of time) can

be modeled through a Poisson distribution with log link function.

For the convenience of further discussion, we adopt the conventional LME platform for ICC with a more generic and inclusive formulation (Pinheiro and Bates, 2004; Chen et al., 2016) than (1), (6), or (10). The following formulation contains all all the aforementioned models as special cases, and will be further discussed hereafter:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{16}$$

where the known vector[5] $\boldsymbol{y}_{kn\times 1} = vec(y_{ij})$ is the vectorization or column-stacking of effect estimates y_{ij}; $\boldsymbol{b}_{m+1}$ contains the unknown fixed-effects to be estimated; the known model (or design) matrix $\boldsymbol{X}_{kn\times(m+1)}$ codes the explanatory variables for the fixed effects; the known model (or design) matrix $\boldsymbol{Z}$ are usually a column-wise subset of $\boldsymbol{X}$; $\boldsymbol{\theta}$ contains the unknown random effects to be estimated; and $\boldsymbol{\epsilon}_{kn\times 1}$ contains the unkown residuals. The distributional assumptions are that $\boldsymbol{\theta} \sim \mathscr{G}(\mathbf{0}, \boldsymbol{V}), \boldsymbol{\epsilon} \sim \mathscr{G}(\mathbf{0}, \boldsymbol{R})$, $\boldsymbol{V}$ and $\boldsymbol{R}$ are the unknown variance-covariance matrices for the random effects $\boldsymbol{\theta}$ and residuals $\boldsymbol{\epsilon}$, respectively, $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ are independent of each other (i.e., $\boldsymbol{\theta}\perp\boldsymbol{\epsilon}$, or $Cov(\boldsymbol{\theta}, \boldsymbol{\epsilon}) = \mathbf{0}$). For the ICC context with no missing data, $\boldsymbol{R} = \sigma_\epsilon^2 \boldsymbol{I}_{kn}$; for ICC(1,1) and ICC(3,1), $\boldsymbol{V} = \sigma_\lambda^2 \boldsymbol{I}_{kn}$, and for ICC(2,1), $\boldsymbol{V} = \boldsymbol{I}_n \otimes diag(\sigma_\pi^2, \sigma_\lambda^2)$.

For those ANOVA cases where a negative ICC value would occur due to the subtraction between two MS terms, LME avoids negativity by having a boundary at 0, via a positive definiteness for the variance-covariance matrix when estimating the variance components through optimizing the likelihood function within the nonnegative domain of the variance components. With these considerations in mind, the question for those ambiguous values in the LME framework becomes: are those regions in the brain fully uncorrelated across the levels of the repeated-measures factor, or is the zero variance estimate simply some artifact providing a zero; or is it a consequence of convergence failure in the optimization algorithms when solving LME? This is addressed by introducing a novel improvement to the conventional LME, and all the advantages of LME over ANOVA will carry over to the other variants of mixed-effects models herein.

Bayesian Mixed Effects (BME) model: LME assisted by a Bayesian nudge

When a zero variance estimate occurs, typically the corresponding likelihood profile is decreasing or close to a flat line at zero, leading to the vulnerability of numerically falling to the boundary of zero. Such a scenario may occur when the sample size is small or when the noise is substantially larger than the signal, derailing the LME capability to provide a meaningful variance estimate in this near-zero boundary value of zero. It is unfortunately not rare to have a small number of subjects or to have a neuroimaging signal submerged or drowned out with noise. Compared to a negative ICC value (or negative variance estimate σ_λ^2), having a floor for σ_λ^2 at 0 avoids an uninterpretable situation; however, even the zero estimate for ICC may be questionable

[5] Without loss of generality, the dimensions shown here for vectors and matrices are assumed to have no missing data, by default. When missing data occur, the dimensions can be adjusted accordingly.

to some extent: do we truly believe that all the subjects in a study have *exactly* the same average effect across the levels of factor A?

Variances in LME are estimated by optimizing REML, which is equivalent to the posterior marginal mode, averaging over a uniform prior on the fixed effects. To pull out of the trapping area surrounding the likelihood boundary, one possibility is to apply a weakly informative prior distribution for the variance parameters (Chung et al., 2013). This Bayesian approach can be conceptualized as forming a compromise between two pulling directions: the anchoring influence of the prior information and the wriggling power of the data themselves. With a proper prior distribution, one may prevent a numerical boundary estimate from occurring by "nudging" the initial variance estimate by no more than one standard error, leading to negligible influence from the prior when the information directly from the data should be fully relied upon.

Here we adopt a weakly informative prior distribution, gamma density function (Chung et al., 2013), for a variance component v in the LME model,

$$h(v;\eta,\kappa) = \frac{\kappa^{\eta}}{\Gamma(\eta)}\sigma^{\eta-1}e^{-\kappa v}, \qquad \eta,\kappa > 0, \tag{17}$$

where η and κ are the shape and rate (inverse scale) parameters, respectively. In practice, the gamma density function has two desirable properties for our nudging purpose here: 1) a positive and constant derivative at zero when $\eta = 2$, guaranteeing a nudge toward the positive direction, and 2) $h(0;\eta,\kappa) = 0$ when $\eta > 1$, allowing the variance to hit the boundary value of zero when the true value is zero. With $\eta = 2$, the prior is uninformed (and improper) within $(0,\infty)$ when κ approaches 0, and becomes gradually informative (but still weak) when κ is away from 0. Therefore, a parameter set of $\eta = 2$ with a small rate parameter κ produces a positive estimate for the variance but does not overrule what the data reveals.

Typical Bayesian methodology involves estimating the posterior distribution through sampling with simulations (e.g., Markov chain Monte Carlo). However, here the prior can be directly incorporated into the likelihood function for LME, hence no simulations are required in order to nudge the variance estimate out of the boundary value for the ICC computation, and usually there is little extra computational cost added to the classic LME computations. In fact, an overall higher efficiency can be achieved under some circumstances because the prior may speed up the convergence process otherwise stuck or slowed in the trapping area close to the boundary.

With both classic LME and its nudging version BME, there remains one last limitation mentioned in the Introduction to be overcome for ICC estimation: How can the LME model utilize the precision information (i.e., standard error) associated with the effect estimates from the individual subject analysis? Would the precision information provide a more accurate partitioning among the variance components including situations when ANOVA renders negative ICC or when LME forces σ_{λ}^2 to be 0?

Multilevel mixed-effects (MME) modeling

The effect estimates of interest in neuroimaging are usually not raw data or direct measures, but instead regression coefficients as output from a time series regression model; as such, these effect estimates have their own measurement uncertainties or confidence intervals. A close inspection of the LME model for ICC, (1), (6), or (10), reveals that the residual term ϵ_{ij} represents the uncertainty embedded in the effect estimates. An underlying assumption in the classic LME model holds that the measurement errors share the same uncertainty: $\epsilon_{ij} \overset{iid}{\sim} \mathscr{G}(0, \sigma_\epsilon^2)$. Since the effect estimates come from a time series regression model separately for each subject, their precision is not necessarily the same and may vary significantly across subjects for various reasons, including variations in trial sequence, data points, and outliers. Therefore, in practice, each effect estimate is associated with a different variance for its measurement error: $\epsilon_{ij} \sim \mathscr{G}(0, \sigma_{ij}^2)$. The true value of σ_{ij}^2 is usually unknown, but its estimation is readily available, conveniently embedded in the denominator of the t-statistic for each effect estimate out of the individual subject analysis. Such an approach has previously been developed for simple group analysis, with the standard error for the effect estimate incorporated into the group model in neuroimaging (e.g., FLAME in FSL (Woolrich et al., 2004); 3dMEMA (Chen et al., 2012)).

Here we extend the LME model, (1), (6), or (10), through replacing the assumption $\epsilon_{ij} \sim \mathscr{G}(0, \sigma_\epsilon^2)$ with an unknown σ_ϵ^2 in the model by $\epsilon_{ij} \sim \mathscr{G}(0, \sigma_{ij}^2)$, where the variance estimate $\hat{\sigma}_{ij}^2$ for σ_{ij}^2 is assumed to be available, which is usually the case for task response. One notable exception would be for the correlation coefficient from seed-based analysis with, for example, resting state data unless unequal number of data points or confounding effects exists across subjects. We call this approach mlultilevel mixed-effects (MME) modeling, with the phrase *multilevel* reflecting the fact that the modeling approach borrows part of a methodology typically adopted in meta analysis when a summarization analysis is performed across previous studies, each of which provided both effect estimate and its standard error. The MME counterpart for the standard LME formulation (16) is extended to have the assumption $\boldsymbol{R} = diag(vec(\hat{\sigma}_{ij}^2))$.

Although computationally more sophisticated due to the involvement of more than one variance component in the case of the model (6) for ICC(2,1), the basic numerical scheme remains similar to our previous work for group analysis (Chen et al., 2012). That is, the variance components for the random effects are iteratively solved through optimizing REML with the estimates $\hat{\sigma}_{ij}^2$ for measurement precision playing a weighting role: an effect estimate y_{ij} with a smaller (or larger) $\hat{\sigma}_{ij}^2$ has a larger (or smaller) impact on the estimation of the variance components and fixed effects. It is this strategy of differential weighting that separates MME from the three previous models in which each of the effect estimates is treated equally.

One adjustment for MME specific to the ICC case is the following. The variance for the residuals, σ_ϵ^2, in the ICC definitions under all other models is no longer available, due to the replacement of the residual term with an unknown variance by the measurement error with a known variance. In its place, we substitute $\hat{\sigma}_\epsilon^2$ with the weighted (or "typical") - instead of arithmetic - average of the sample variances $\hat{\sigma}_{ij}^2$ in light of the generic

model (16) (Higgins and Thompson, 2002; Viechtbauer, 2010),

$$(\hat{\sigma}_\epsilon^2)_{\boldsymbol{W}} = \frac{T-p}{tr(\boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W})}, \tag{18}$$

where T is the total number of data points in $\boldsymbol{y}$ (where $T = kn$ if no missing data occur); p is the column rank of $\boldsymbol{X}$; tr denotes the trace operator; $\boldsymbol{X}$ and $\boldsymbol{W} = \boldsymbol{R}^{-1} = diag(\frac{1}{\hat{\sigma}_{11}^2}, ..., \frac{1}{\hat{\sigma}_{kn}^2})$ are the model matrix for the fixed effects in the model (16) and the weighting matrix, respectively. The differential weighting is reflected in the heterogeneous diagonals of the variance-covariance matrix $\boldsymbol{R}$ for the measurement errors, which becomes more transparent with a reduced form of $(\hat{\sigma}_\epsilon^2)_{\boldsymbol{W}}$ for the case of no missing data without any explanatory variables, i.e., $\boldsymbol{X} = \boldsymbol{1}_{kn\times 1}$, in the model (16),

$$(\hat{\sigma}_\epsilon^2)_{\boldsymbol{W}} = \frac{(kn-1)\sum_{i=1}^k \sum_{j=1}^n w_{ij}}{(\sum_{i=1}^k \sum_{j=1}^n w_{ij})^2 - \sum_{i=1}^k \sum_{j=1}^n w_{ij}^2}, \tag{19}$$

where the weights $w_{ij} = \frac{1}{\hat{\sigma}_{ij}^2}$.

Bayesian MMM (BMME): MME assisted by a Bayesian nudge

A zero variance estimate may still occur under MME for the same reason as in LME, namely that the corresponding likelihood profile peaks at zero. Similarly, the same gamma density function (17) can be adopted as a weakly informative prior distribution (Chung et al., 2013), for a variance component v in the MME model (e.g., see the example in (17) here), parallel to the extension from LME to BME. Here, as well, as no posterior samplings are involved in the process, the prior sometimes can even speed up the convergence process compared to MME. We note that BMME solves all the issues raised with the current ANOVA framework (Table 3).

Performance comparisons among the models

Model implementations

Here we have discussed three types of ICC that can be estimated through five modeling strategies (ANOVA, LME, BME, MME, and BMME), leading to total of 15 scenarios, each of which is accompanied by an F-statistic defined by either (5) or (9). The five models (each with three ICC types and corresponding F-statistics) are implemented in a publicly available program `3dICC` in AFNI, making use of the R packages *psych* (Revelle, 2016), *lme4* (Bates et al., 2015), *blme* (Chung et al., 2013), and *metafor* (Viechtbauer, 2010), respectively. The program `3dICC` takes a three-dimensional dataset as input for whole-brain voxel-wise analysis, but it can also be utilized to analyze a single voxel, region, or network. Additionally, parallel computing is available with multiple CPUs through the R package *snow* (Tierney et al., 2016). For each ICC model except ANOVA, the fixed effects (intercept or group average effect for each of the three models, as well as additional comparisons among factor A levels for ICC(3,1)) and their corresponding t-statistics are provided in the output from `3dICC`,

in addition to ICC and the associated F-statistic value.

Experimental testing dataset

To demonstrate the performances of our four proposed modeling approaches relative to ANOVA, we utilize experimental data from a previous FMRI study (Haller et al., 2017). Briefly, 25 healthy volunteers (mean age = 13.97 years, SD = 2.22 years, range = 10.04-17.51 years; 60% female) were asked to judge the gender of happy, fearful and angry face emotions. Each emotion was displayed at three intensities (50%,100%, 150%). A neutral condition, representing 0% intensity, was included for each face emotion (i.e., three neutral subsets were created, one for each face emotion). MRI images were acquired in a General Electric 3T scanner (Waukesha, WI, USA), and the participants completed two MRI scanning sessions approximately two-and-a-half months apart (mean = 75.12 days, SD = 15.12 days, range: 47-109 days). The FMRI echoplanar images (EPIs) were collected with the following scan parameters: flip angle = 50 degrees, echo time = 25 ms, repetition time = 2300 s, 47 slices, planar field of view = 240 $\times$ 240 mm^2, acquisition voxel size = 2.5$\times$2.5$\times$3 mm^3, and three runs with 182 volumes for each in a total acquisition time of 21 minutes. The parameters for the anatomical MPRAGE images were: flip angle = 7°, inversion time = 425 ms, and acquisition voxel size = 1 mm isotropic.

The EPI time series went through the following preprocessing steps in AFNI: de-spiking, slice timing and head motion corrections, affine alignment with anatomy, nonlinear alignment to a Talairach template TT_N27, spatial smoothing with a 5 mm full-width half-maximum kernel and scaling by the voxel-wise mean. Individual TRs and the immediately preceding volume were censored if: 1) the motion shift (defined as Euclidean norm of the derivative of the translation and rotation parameters) exceeded 1 mm between TRs; or 2) more than 10% of voxels were outliers. Only trials with accurate gender identification were included in the final analysis, but incorrect trials were also modeled as effects of no interest. Separate regressors were created for each of 13 event types (i.e., 30 trials for each face emotion at each intensity, neutral trials represented by three regressors of 30 trials each, and incorrect trials). Six head motion parameters and baseline drift using third order Legendre polynomials were included as additional regressors. The two sessions were analyzed separately, but the three runs were concatenated and then entered into a time series regression model with a serial correlation model of ARMA(1,1) for the residuals.

The effect estimate in percent signal change, combined with the variance for the corresponding measurement errors, for the neutral condition associated with angry face-emotion from the individual subject analysis, was adopted for comparing the five ICC models: ANOVA, LME, BME, MME, and BMME. ICC(1,1) is not applicable in this case, but both ICC(2,1) and ICC(3,1) and their F-statistics were computed for each of the four models; the session effect and the corresponding t-statistic were also examined. The runtime was about one hour for each of the analyses with 16 CPUs on a Linux system (Fedora 14) with Intel® Xeon® X5650 at 2.67 GHz.

Model comparisons

ANOVA rendered a substantial number of voxels with negative ICC values (first column in Panel A, Fig. 1; first row and first column in Fig. 3). LME typically provided virtually the same ICC estimates as ANOVA, except that those negative ICC estimates were replaced with 0 (uncolored voxels in the second column in Panel B, Fig. 1; scatterplot cells (1, 2) and (2, 1) in Fig. 3). It is worth noting that a significant proportion of voxels with negative or zero ICC from ANOVA or LME appeared in grey matter. For BME, we tested four different priors by varying the rate parameter κ at values of 0, 0.1, 0.3, and 0.5 (with η fixed at 2); their differences in ICC estimates across the four κ values were negligible. Furthermore, the computation cost is a decreasing function of κ (substantially highest at $\kappa = 0$). In light of these results, we set an empirical prior of gamma density (17) at $\eta = 2, \kappa = 0.5$ for neuroimaging data.

For the voxels with negative ICC values from ANOVA or with zero ICC values from LME, BME offered positive but generally small ICC estimates; the nudging effect of BME is relatively small when the ICC value from ANOVA/LME is positive but small (less than 0.3), and it is negligible when the ICC value from ANOVA/LME is moderate or large (larger than 0.3) (third column in Panel A, Fig. 1; scatterplot cells (1, 3) and (3,1) in Fig. 3). Some of the ICC estimates from MME were larger to varying extent than ANOVA/LME/BME, while some were smaller (fourth column in Panel A in Fig. 1; "fat blobs" in the scatterplot cells (1, 4), (2, 4), (3, 4), and their symmetric counterparts in Fig. 3); there were a higher number of voxels with larger ICC from MME relate to ANOVA/LME/BME than those with smaller ICC values (slightly redder voxels in fourth column than the first three columns, Panel A in Fig. 1; more dots with ICC greater than 0.75 on the MME side of the green diagonal line in the scatterplot cells (1, 4), (2, 4), (3, 4), and their symmetric counterparts in Fig. 3). However, there were still some number of voxels had zero ICC estimates from MME (uncolored voxels in the fourth column, Panel A, Fig. 1), although less than LME. Last, BMME shared similar comparisons with MME relative to ANOVA/LME/BME (fifth column, Panel A, Fig. 1; scatterplot cells (1, 5), (2, 5), (3, 5), and their symmetric counterparts Fig. 3). The differences between BMME and MME were parallel to those between BME and LME (fifth column, Panel A, Fig. 1; scatterplot cells (4, 5) and (5, 4), Fig. 3); that is, BMME provided small but positive ICC estimates for those voxels with zero ICC under MME, and the nudging effect was negligible when the effect estimate was relatively reliable. The F-values across the five models followed roughly the same patterns as the ICC values (Panel B in Fig. 1) because the F formulation is closely related to its ICC counterpart, and the high reliability measure in the brain shared the same regions with group average effect revealed from the four mixed-effects models (first two rows in Fig. 2).

The differences between ICC(2,1) and ICC(3,1) (Panel A in Fig. 1; diagonal cells in Fig. 3) were mostly small for all the five models except for those regions where session effect were substantial (third row, Fig. 2). Because any systematic differences between the two sessions were accounted in the ICC(3,1), but not ICC(2,1), model, the former rendered slightly higher ICC estimates. This is demonstrated in Fig. 3, by the fact that most voxels were above the green diagonal line in the diagonal cells of scatterplots. In addition, one noteworthy

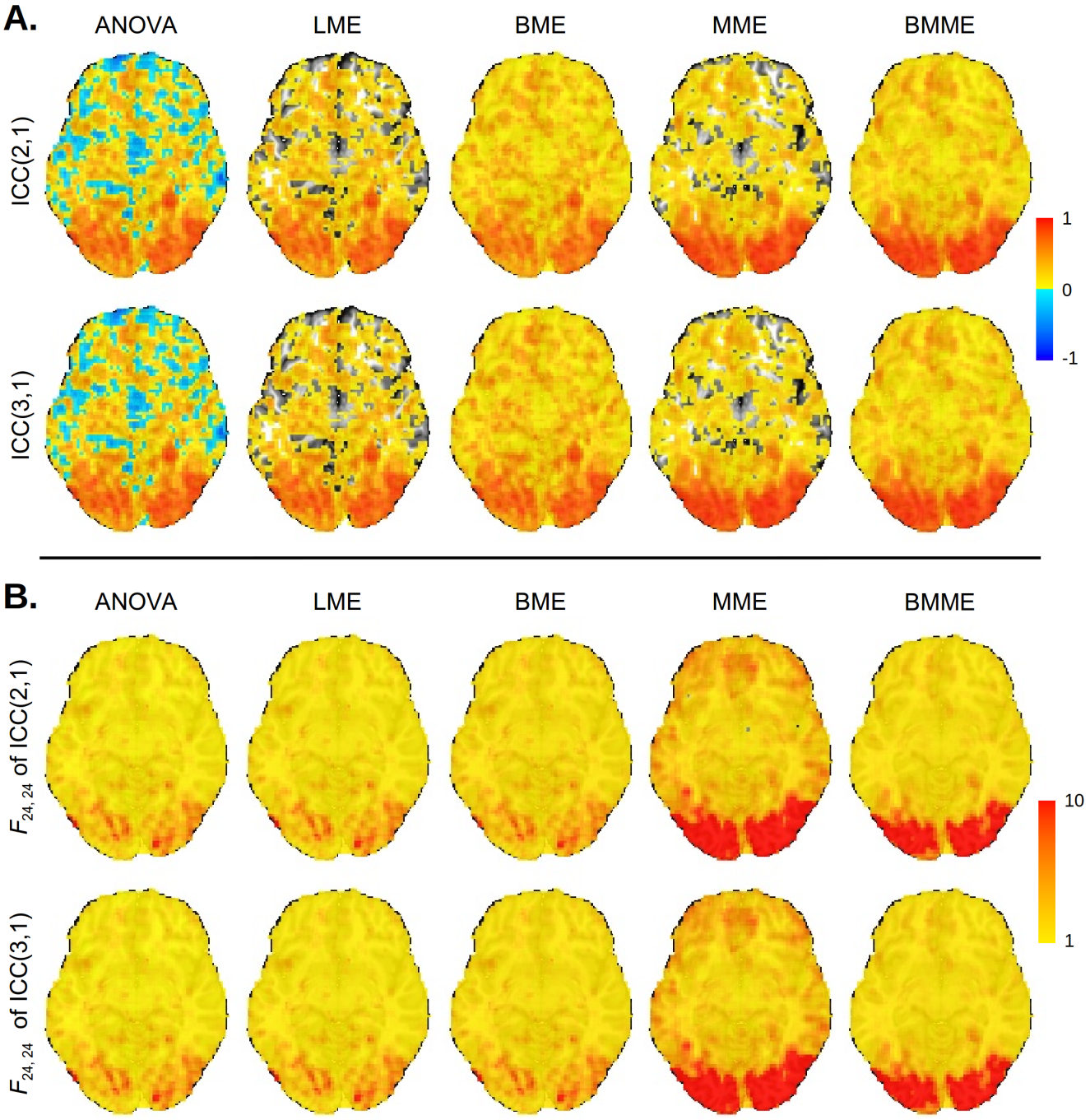

Figure 1: **Panel A**. ICC maps on an axial slice ($Z = -6$ mm, TT standard space; radiological convention: left is right) with a whole brain dataset through four models. The conventional ANOVA (first column) and LME (second column) rendered substantial number of voxels with negative (blue) and zero (not colored) ICC values, respectively. **Panel B**. The F-statistic maps corresponding to each of the ICC maps in Panel **A** are shown with colors in the range of [1, 10], with the lower bound of 1 defined by the formulation of F-statistic in (9). In each case, the degrees of freedom were (24, 24).

phenomenon is that BMME narrowed the differences between the two ICC types, as represented in Fig. 3 by the thinner band in scatterplot cell (5, 5) relative to the other diagonal cells.

To gain better insights into the relative performances of the five models, we demonstrate three scenarios with three representative voxels (Table 4 and Fig. 4; effect estimates and their variances in Appendix A). At voxel V_1 (from the right middle occipital gyrus), the ICC estimates from the ANOVA, LME and BME approaches were similar across all the five models. Both MME and BMME estimates were not much different because those effect estimates were roughly equally precise (first two columns in the right panel). The ICC estimate for the two types did not differ much either as the difference between the two sessions was small (0.01% signal change, Table 4).

At voxel V_2 (from the right lingual gyrus) ANOVA produced negative ICC values because MS_λ is less than MS_ϵ in the corresponding ICC formulation (7) or (11). LME avoided negativity by taking the lower boundary value of zero that is allowed for variance, while BME rendered a positive but small ICC estimate. In contrast, both MME and BMME achieved more accurate variance estimates in the sense that the availability of variances for measurement errors provided a way to reallocate or redistribute the components in the total variance. In other words, the substantial amount of heteroscedasticity as shown in Fig. 4 (third and fourth column in the right panel), allowed differential weightings among the effect estimates with a higher emphasis on the more precise ones and downgrading for the less reliable ones. The dramatically different ICC values at voxel V_2 from MME and BMME, relative to the three other ICC models, can be understood by examining the wide range of effect estimates as well as their heterogenous variances, as shown in the wide color spectrum of Fig. 4. It is worth noting that the presence of substantial session effect (close to 0.1%, $p = 5.7 \times 10^{-5}$) led to a moderate difference between ICC(2,1) and ICC(3,1) for MME, but the analogous difference for BMME was negligible; that is, BMME tended to render similar ICC value between the two types regardless of the presence of session effect, just as shown in the whole brain data (scatterplot cell (5, 5) in Fig. 3).

Last, at voxel V_3 (from the right lingual gyrus) ANOVA, LME and BME revealed moderately reliable effect estimates with similar ICC estimates. However, both MME and BMME rendered higher ICC values, due to due to the presence of moderate amount of heteroscedasticity, similar to the situation with voxel V_2 even though less dramatic here (last two columns in both panels, Fig. 4). Also similar to voxel V_2, the session effect (0.08%, $p = 1.1 \times 10^{-5}$) caused lower estimate from MME for ICC(2,1) than ICC(3,1), but BMME estimated virtually the same reliability between the two ICC types. It is also interesting to note that, at both voxels V_2 and V_3, the t-statistic for the fixed effect of session was much higher when the precision information was considered than in the other two models, LME and BME (Table 4). This phenomenon demonstrates the potential impact and importance of modeling precision in neuroimaging group analysis (Worsley et al., 2002; Woolrich et al., 2004; Chen et al., 2012).

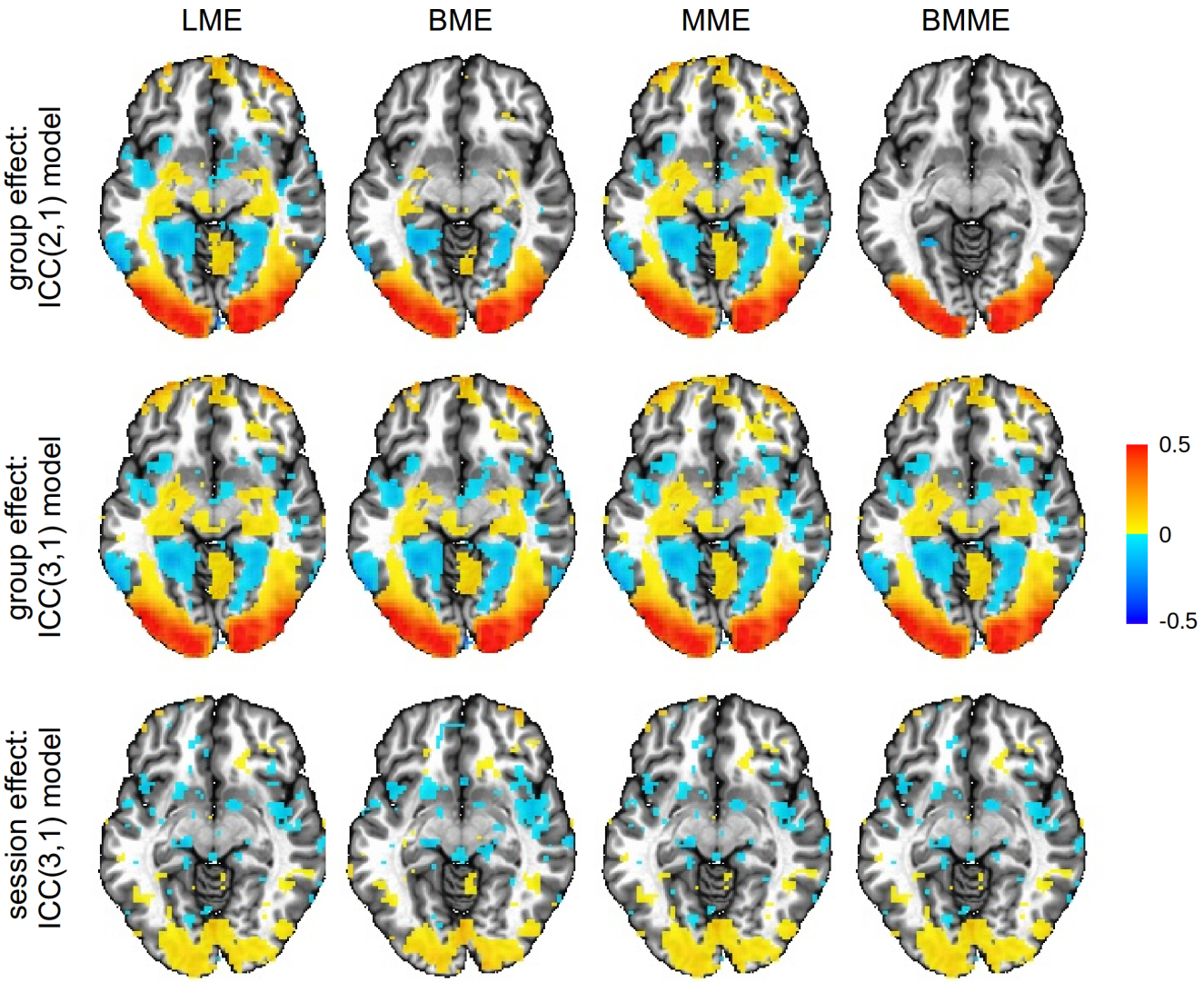

Figure 2: The first two rows show group average effect on an axial slice ($Z = -6$ mm, TT standard space; radiological convention: left is right) with a whole brain dataset estimated through the four mixed-effects models. The third row is the session effect in the ICC(3,1) model. While the maps display the session effect, they were thresholded at a liberal significance level of 0.1 with 24 degrees of freedom for the corresponding t-statistic. The estimates for the fixed effects are not directly available from ANOVA, and therefore are not displayed. As the two Bayesian approaches estimated slightly higher variances under the ICC(2,1) model, some regions in the frontal area failed to survive the liberal thresholding for BME and BMME as shown in the first row. In contrast, all the four models demonstrated similar results under the ICC(3,1) model as shown in the second row.

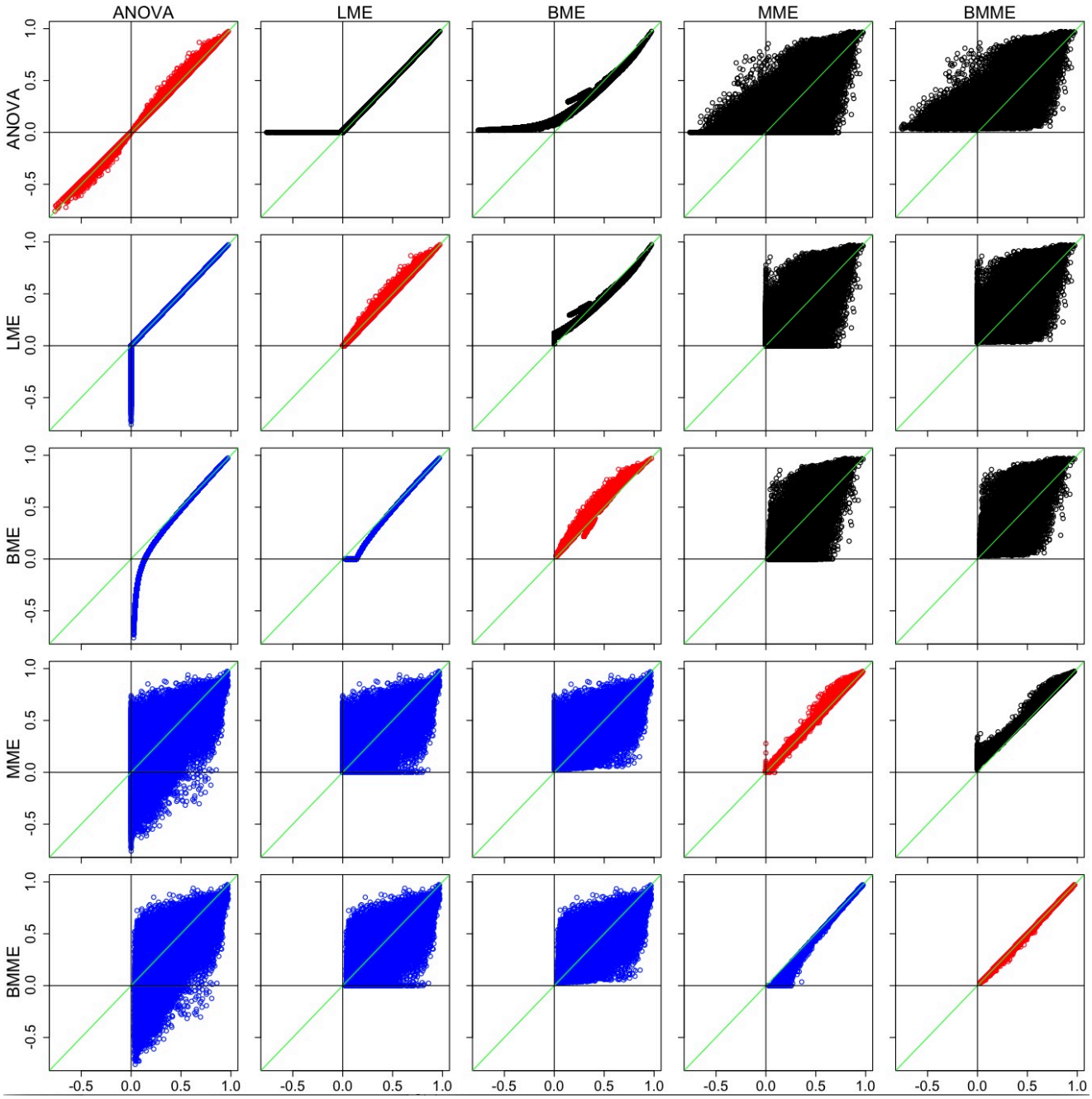

Figure 3: Comparisons across the five models and between ICC(2,1) and ICC(3,1) through scatterplots of ICC with the voxels in the brain. The x and y axes are a pair of ICC values between two of the five models with their labels shown at the left and top, respectively. The 10 combinatorial comparisons for ICC(2,1) are illustrated in the upper triangular cells in black; the 10 combinatorial comparisons for ICC(3,1) are listed in the lower triangular cells in blue; and the diagonals compare the two ICC types in red for each of the five models with ICC(2,1) and ICC(3,1) as x and y axes, respectively. In each plot the green diagonal line marks the situation when ICC(2,1) = ICC(3,1).

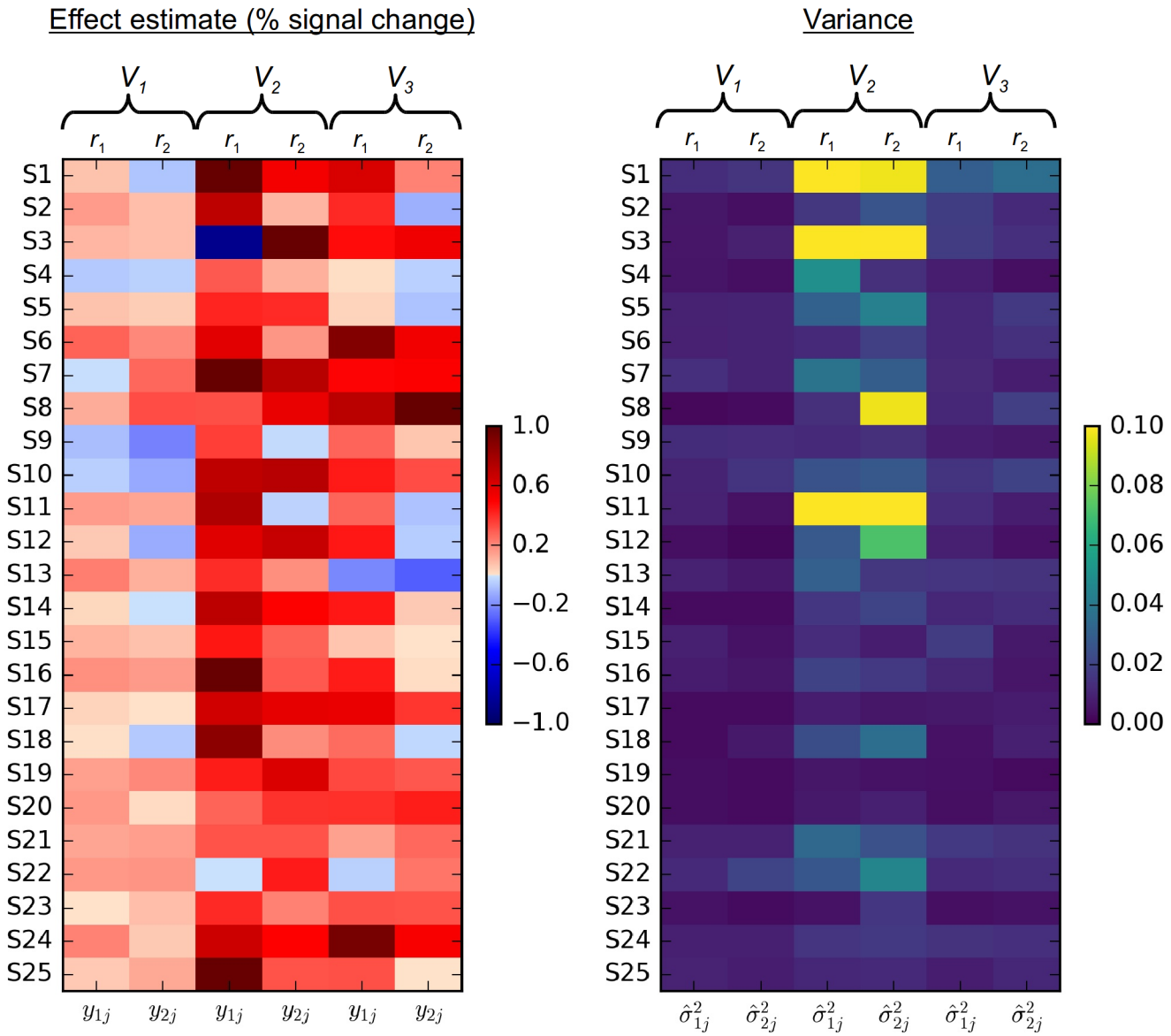

Figure 4: Example illustrations of heterogeneity and heteroscedasticity at three voxels with results shown in Table 4. Effect estimates (left) during the two sessions (r_1 and r_2) and the corresponding variances (right) for the measurement errors for the 25 subjects at three grey matter voxels (V_1, V_2, and V_3) from the axial slice shown in Fig. 1. The real data are listed in Appendix A. Substantial session effect can be seen in voxel V_2 (left matrix, middle two columns) while the session effect is negligible at voxels V_1 and V_3. A large amount of variability existed for the effect estimates at voxel V_1 (left matrix, first two columns), leading to negative ICC estimate by ANOVA and zero by LME, while BME managed to deal with the degenerative situation with a small, but positive, ICC estimate. MME provided a positive and relatively large ICC estimate through weighting from the precision information (right matrix, first two columns). The variability for the precision of the effect estimate is moderate at voxel V_2 (right matrix, middle two columns), but minimal at voxel V_3 (right matrix, right two columns).

Table 4: Results at three example voxels[a] shown in Fig. 4

Voxel	Model	ICC(2,1)			ICC(3,1)			session effect		
		value	$F_{24,24}$	p-value	value	$F_{24,24}$	p-value	value	t_{24}	p-value
V_1*	ANOVA	0.53	3.3	0.0024	0.53	3.3	0.0024	-	-	-
	LME	0.531	3.292	0.0025	0.534	3.292	0.0025	0.012	1.144	0.26
	BME	0.500	3.578	0.0014	0.552	3.468	0.0017	0.012	1.159	0.26
	MME	0.504	3.033	0.0043	0.504	3.030	0.0043	0.008	0.786	0.44
	BMME	0.529	3.246	0.0027	0.527	3.231	0.0028	0.008	0.789	0.44
V_2	ANOVA	-0.27	0.56	0.92	-0.28	0.56	0.92	-	-	-
	LME	0	1	0.5	0	1	0.5	0.073	1.469	0.15
	BME	0.044	1.126	0.39	0.058	1.123	0.39	0.073	1.499	0.15
	MME	0.470	4.464	2.5e-4	0.631	4.422	2.7e-4	0.091	4.876	5.7e-5
	BMME	0.652	4.744	1.5e-4	0.649	4.693	1.7e-4	0.091	4.878	5.7e-5
V_3	ANOVA	0.57	4.8	1.4e-4	0.66	4.8	1.4e-4	-	-	-
	LME	0.574	4.819	1.4e-4	0.656	4.819	1.4e-4	0.080	3.372	0.0025
	BME	0.509	5.072	8.9e-5	0.665	4.970	1.1e-4	0.080	3.398	0.0023
	MME	0.743	13.851	5.8e-9	0.864	13.748	6.3e-9	0.076	5.52	1.1e-5
	BMME	0.870	14.435	3.8e-9	0.870	14.344	4.0e-9	0.076	5.517	1.1e-5

[a]The locations for the three voxels are the following (coordinates are in mm in TT space). V_1: (39, 64, -6) in right middle occipital gyrus; V_2: (11, 88, -11) in right lingual gyrus; V_3: (11, 86, -6) in right lingual gyrus. Any fixed effects such as session difference here can be easily estimated through any of the four mixed-effects models for ICC(3,1), but they are usually not provided through ANOVA.

Discussion

Reliability is a crucial foundation for scientific investigation in general, and it has been a challenging issue for neuroimaging in particular as a hot topic over the years (e.g., Bennett and Miller, 2010). ICC offers a metric that can measure reliability under relatively strict circumstances. If the same scanning parameters are applied to the same cohort of subjects or families that undergo the same set of tasks, ICC shows the reliability across the levels of a categorical variable such as twins, sessions, or scanners. Nevertheless, it is difficult to keep scanning circumstances perfectly constant across sessions in neuroimaging, thus systematic differences may slip in one way or another. On the other hand, ICC can be applied to neuroimaging in the classic sense (i.e., inter-rater reliability or concordance), assessing the reliability between two diagnostic approaches (e.g., human versus automatic method) on a disease (e.g., schizophrenia, autism, or depression) or two different analytical methods applied to the same collection of data.

As shown here, there are three types of ICC, each of which can be estimated through various models such as ANOVA, LME, BME, MME, and BMME. On one hand, the ICC metric offers a unique approach to measuring the reliability of neuroimaging data under some well-controlled conditions; on the other hand, the investigator may still face a daunting job in deciding which ICC type, as well as which model, is most appropriate to apply in a given experiment or setup.

Literature survey of ICC for neuroimaging

ICC has been applied to neuroimaging data for over 10 years, mainly to examine reliability under various scenarios, and ICC(2,1) is largely adopted in the field using the ANOVA approach. ICC(2,1) has been used to

show reliability under various scenarios, for example, at the regional level (Fiecas et al., 2013), at the network level for resting-state (Guo et al., 2012; Cao et al., 2014), at the whole brain level with ANOVA (Kristo et al., 2014; Zanto et al., 2014; Quiton et al., 2014) and with LME using a precursor of `3dICC` (Fiecas et al., 2013; White et al., 2016; Haller et al., 2017). ICC(3,1) has been applied at the regional level for task-related FMRI data (Cáceres et al., 2009; Jaeger et al., 2015) and for MEG data (Recasens and Uhlhaas, 2017), at the network level for resting-state data (Braun et al., 2012), and at the whole brain level for task-related FMRI data, using ANOVA in PASW and/or Matlab (Cáceres et al., 2009; Plichta et al., 2012; Brandt et al., 2013) and in SAS (Fournier et al., 2014). It should be noted that in many cases the ICC type adopted in a study was not clearly stated, often due to the ambiguity in terms of the model involved (Zuo et al., 2010b; Lin et al., 2015; Shah et al., 2016).

When described, ICC(1,1) and ICC(3,1) have been employed at times in the literature. For example, ICC(1,1) has been applied to brain networks based on resting-state FMRI data (Wang et al., 2011), to functional near-infrared spectroscopy (fNIRS) data (e.g., Bhambhani et al., 2006; Pichta et al., 2006; Pichta et al., 2007; Zhang et al., 2011; Tian et al., 2012), and to resting-state data at the whole brain level (Zuo et al., 2010a), as well as to task-related data through LME at the regional level (Töger et al., 2017). However, we note that in a large number of publications, the adoption of the ICC type was neither explicitly explained nor justified, which makes precise interpretation difficult.

On the whole brain level, in addition to the ICC computation through LME as implemented in the open source AFNI program `3dLME` (Chen et al., 2013), there have been a few Matlab toolboxes publicly available: three using ANOVA (Cáceres et al., 2009; Fiecas et al., 2013; Molloy and Birn, 2014), and one using both ANOVA and LME in Matlab (Zuo et al., 2010b). The concept of ICC has also been extended to characterize cross-subject heterogeneity when the effect estimate precision is incorporated into FMRI group analysis (Chen et al., 2012), as well as to reveal the relatedness among subjects in inter-subject correlation (ISC) analysis with data from naturalistic scanning (Chen et al., 2017a).

Effect estimates as input for ICC analysis

In practice, several factors can contribute to having poor data quality and accuracy in neuroimaging. Having relatively low signal-to-noise ratio is a major issue, and suboptimal modeling due to poor understanding about the major components in the signal is another. By some estimates, less than half (and in some cases down to 20-30%) of data variability can be accounted for in a typical FMRI data analysis at the individual level (Gonzalez-Castillo et al., 2017). Some rigor and standardization steps are required to achieve more accurate reproducibility and reliability. At present, the cumulative impact of altering preprocessing and modeling strategy in an analysis pipeline is largely unknown.

For example, it is well known that the absolute values of the FMRI-BOLD signal from the scanner have arbitrary units with scaling fluctuations among subjects, and therefore some kind of calibration is needed during

preprocessing if the effect estimates are to be used in further analyses at the group level, for both typical group analysis as well as ICC estimation. Such a calibration should take into consideration the fact that the baseline varies across brain regions, subjects, groups, scanners and sites. However, radically different scaling strategies are in use, and their effects at the group level remain unclear. For instance, global or grand mean scaling, as currently typically practiced, can only account for part of the cross-subject variability, but it does not address the cross-region variability within an individual brain as a practical reality in neuroimaging scanning, and therefore it may lead to difficulty in interpreting and comparing the effect estimates. Analyses on networks including ICC and causal modeling would be affected as well. In contrast, voxel-wise scaling as a calibrator, even though imperfect due to the ambiguity in the baseline definition (Chen et al., 2017), provides a more accurate characterization than the alternatives. Because of these considerations, in general we recommend voxel-wise mean scaling during preprocessing, so that a directly interpretable and compatible measure in percent signal change can be adopted for the effect estimates that are to be taken for group analysis including ICC.

Lastly, each effect estimate in neuroimaging is typically derived from a model coefficient, and is intrinsically associated with some extent of uncertainty that may vary across subjects. As the precision information of the single subject effect estimate (embedded in the denominator of t-statistic) is required for the preferred MME and BMME, it is important to model the temporal correlation in the residuals at the individual level to avoid inflating the precision (or underestimating the uncertainty) for the effect estimate to be used in the ICC model.

Which ICC model to use?

Standard ICC in most literature to date contains several aspects of ambiguity, greatly hindering meaningful interpretation. Among the thousands of voxels in a typical whole-brain neuroimaging dataset, negative ICC values unavoidably, and even frequently, show up, though they are usually not reported in the literature. In addition, the selection of a given ICC type, especially when choosing between ICC(2,1) and ICC(3,1), is usually not clearly explained. In some cases, the chosen method is ill-suited to the given analysis scenario (e.g., if ICC(1,1) was used between two sessions).

When precision information for the measurement errors of the effect estimate is available, we recommend using BMME for the following three reasons: 1) the precision information offers a more robust estimate for ICC, as well as for the fixed effects in the model; 2) the Bayesian aspect of the approach leads to avoidance of the uninterpretable situation of a negative ICC from ANOVA or an unrealistic zero ICC estimate from plain LME; 3) as demonstrated here (scatterplot cell (5, 5) in Fig. 3 and Table 4), BMME tends to render roughly the same ICC estimate regardless of the type the investigator adopts. When the precision information is unavailable (e.g., when one has the correlation value as the effect of interest from seed-based analysis or psycho-physiological interaction analysis), we recommend using BME because of its capability to provide a reasonable ICC estimate even when the model becomes degenerate (i.e., the numerical model is singular or unidentifiable).

One may argue that those voxels with negative ICC estimates through ANOVA or zero values through

LME do not warrant any special attention. However, there are practical reasons that make their realistic ICC estimation still desirable. First of all, as previously mentioned, negative or zero ICC values are neither interpretable nor realistic, especially for grey matter regions in the brain. Second, realistic ICC estimates provide a starting point to achieve model tuning or improvement through potentially incorporating confounding effects or varying the modeling strategy (e.g., fixed- vs random-effects). Finally, having a low but more realistic ICC estimate at a region for a particular study is useful when comparing reliability across different studies. For example, with a region showing relatively high ICC in a previous study, a low ICC value in a subsequent investigation would prompt further exploration as to why this occurs as one cannot rely on interpreting negative or zero values for clear comparisons.

The use of a Bayesian approach may raise questions regarding its insertion of arbitrariness into the estimation process. Each variance component in the ICC formulation is estimated as a point value; however, it is worth noting that the numerical value of a variance component does not exactly fall at its numerical estimate, but varies within some range (e.g., 95% confidence interval). When a zero estimate occurs for a variance component, it usually tends to be imprecise and underestimated, and such estimation should not be taken at face value (Chung et al., 2013). A zero estimate for variance and ICC can also cause inflated inferences on the fixed effects such as group average as well as systematic differences across the factor A levels in the ICC(3,1) model. On the other hand, the Bayesian approach can be conceptualized as a tug of war between the prior and data. As the results from our experimental dataset demonstrate here (e.g., Fig. 1), a weakly informative prior for ICC estimation can pull a degenerate situation (e.g., zero variance estimate) out of the boundary and render a more reasonable estimate, while little impact will incur when the data contain strong information, which would overrule the prior.

Four aspects of MME and BMME are worth highlighting here. First, these two multilevel approaches estimate ICC differently from the other three methods to some extent, as indicated in those "fat blobs" among cells (1,4), (2,4), (3,4), (1,5), (2,5), and (3,5) for ICC(2,1), and (4,1), (4,2), (4,3), (5,1), (5,2), and (5,3) for ICC(3,1). It should be stressed that the focus is not on which method leads to a higher or lower estimate, but on which model provides a more accurate characterization about the reliability measure. Second, just as the levels of a within-subject factor are treated as simultaneous variables of a multivariate model (as in the AFNI group analysis program `3dMVM`) for repeated-measures ANOVA in neuroimaging group analysis[6] (Chen et al., 2014), so are the factor A levels in MME and BMME for ICC(3,1) modeled here in a multivariate fashion, due to the mothod's modeling advantages and flexibility in statistic formulations. Third, the measurement errors associated with the factor A levels in a generic model of the form in (6) or (10) can be correlated, thus the variance-covariance matrix $\boldsymbol{R}$ should be semi-definite in general. It is because of the presence of covariances in $\boldsymbol{R}$ that the modeling approach adopted in 3dMEMA of AFNI and FLAME of FSL cannot instead be utilized

[6]Due to the complexity of handling within-subject factors, some group analyses in published Neuroimaging literature are still not performed correctly, leading to substantial number of publications with inflated statistical inferences (McLaren, 2010; Chen et al., 2014).

to perform a paired test in cases where the two effect estimates are entered separately as input (even if the program permits such an option), as the measurement errors corresponding to the two effect estimates are usually correlated[7]. However, $\boldsymbol{R}$ is actually a diagonal matrix in the neuroimaging ICC context, because the measurement errors can be reasonably assumed to be independent with each other (e.g., across runs or sessions). Finally, nonparametric methods (e.g., bootstrapping, permutations) may offer a robust venue for controlling family-wise error for relatively simple models at the group level; however, under some scenarios parametric approaches provide more specific and accurate characterization about the effect of interest through some quantifying parameters (e.g., variance components in the ICC context) in the model, which are currently both valid and irreplaceable, as shown here with ICC computations through ANOVA, LME, BME, MME, and BMME, as well as the group analysis approach through incorporating the precision information (Worsley et al., 2002; Woolrich et al., 2004; Chen et al., 2012).

Which ICC type to adopt?

As there is usually one single effect estimate for each subject per scanning situation, our discussion here focuses on single-measurement ICC. Among the three ICC types, ICC(1,1) is likely the easiest to recognize which scenarios can be applied to. It can be used when there is no apparent distinction for a sequence of the levels for the factor that is associated with the multiple measurements in the ICC model, and it typically applies to the situation of studying twins or siblings.

In contrast, the other two types are utilized for scenarios such as having two or more runs, sessions, scanners, sites, or between parent and child. The reliability from ICC(2,1) demonstrates absolute agreement to the extent that the values exactly match between any two levels of the factor A, while ICC(3,1) shows the consistency or the extent that the effect estimates match across the factor A levels *after* accounting for potential systematic differences or other fixed effects. In other words, if the systematic differences across the levels of factor A are negligible, the two ICC estimates would be similar. On the other hand, if the systematic differences or confounding effects are substantial, both ICC values tend to differ to some extent, and they lead to different interpretations. However, the existence of systematic differences itself warrant further exploration about the source or nature of those fixed effects. For example, what is the association between the ICC map and the activation map (i.e., intercept in the model (15) or (16))?

Due to the simultaneity of analyzing all the voxels, it is unrealistic to choose one ICC type for some regions while selecting another for the rest. Per the discussion here between ICC(2,1) and ICC(3,1), we generally recommend the latter for whole brain analysis. In doing so, potential fixed effects are properly accounted for. More importantly, it is not the ICC interpretation in the sense of absolute agreement that is of utmost importance, but the extent of agreement after potential fixed effects are all explained. Furthermore, with ICC(3,1), the investigator can address the following questions: 1) which regions in brain show systematic

[7]Instead, such a paired test can be validly performed by taking the contrast and its standard error as input.

effects across the levels of factor A? 2) how do those systematic effects correspond to the ICC maps in these regions? and 3) are the systematic effects caused by habituation or attenuation?

Result reporting

Clear and accurate scientific reporting is important for result reproducibility, reliability and validity as well as for meta-analysis. The present reporting convention in neuroimaging is especially discouraging, with a lopsided focus on statistics alone, e.g., due to oversight or limitations in software implementations (Chen et al., 2017b), leading to widespread incomplete reporting in the literature. It cannot be emphasized enough that the effect estimates involved in a study should be reported in addition to the statistical significance, and the same emphasis should be applied to ICC analysis. Specifically, the investigator should explicitly state the ICC type and the model adopted, as well as the justification for such choices. One may notice that the ICC formulation for ICC(1,1) in (2) and (3) is exactly the same as ICC(3,1) in (11) and (12), which means that reporting the ICC formula would not be enough to reveal the whole picture because the two underlying models, (1) and (10), are dramatically different (and so are the two resulting ICC estimates).

With regards to the criteria for reliability, a loose rule of thumb has been adopted for ICC values as the following (Cicchetti, 1994): [0, 0.4), poor; [0.4, 0.6), fair; [0.6, 0.75), good; and [0.75, 1], strong. However, the statistical significance based on the Fisher transformation (4), or preferably, the F-statistic (5), (9) and (13), is a more accurate option. Nevertheless, the F-statistic is not necessarily a basis for clusterization, but together with the ICC value, it serves as some auxiliary information to gauge the reliability about the effect of interest from the conventional analytical pipeline. Finally, all the fixed effects including the intercept (group average) are crucial part of the model and should be discussed in the paper as well, as demonstrated in Fig. 2 here.

We note also that a popular modeling practice in neuroimaging is to include an explanatory variable of no interest, after which the investigator then reports that such an inclusion somehow automatically takes care of all effects associated with that variable without further discussing its impact (i.e., "covarying out age or sex effect"). Simply labeling an explanatory variable as a covariate, nuisance variable, or regressor of no interest (e.g, sex, age, reaction time) does not de facto validate its incorporation, nor does it change the way the variable is treated in the model. In fact, as typically seen in the literature, the term of covariate, nuisance variable, or variable of no interest may open a loophole to sloppy or slippery modeling by simply adding the variable as an additive term in the model. In other words, the terminology tends to hide the fact or even excuse that any potential interactions of the covariate with other explanatory variables are not considered in the model; furthermore, the terminology does not warrant an immunity for no examination, reporting, or discussion for the effect of the variable. We strongly believe that the investigator should explore and report the effect of each covariate[8] used in neuroimaging group analysis. Along the same line, the fixed effects should be also discussed

[8]There are two popular usages for the words of covariate in the field, one focuses on the nature of the variable (quantitative) while the other emphasizes the fact that the investigator is not interested in the variable. The latter usage should be discouraged for two reasons. First, the labeling may cause suboptimal modeling, as we argue here; second, treating a categorical variable (e.g., sex) as a variable of no interest may invite some unintended issues. For example, coding sex as a dummy variable, if not done

and explained together with ICC(3,1) under the reliability setting, as exemplified here in Fig. 1 and Fig. 2.

Conclusion

One potential problem with the classic definition of ICC is that negative ICC values may appear, a scenario which is almost certain to show up in a whole brain neuroimaging analysis. Here we extend the conventional ICC definition and its computations to the frameworks of LME, BME, MME, and BMME modeling to address this issue and other difficulties. Such an extension not only offers wider modeling flexibility such as the inclusion of various fixed effects, but also avoids the interpretability problem of negative ICC values under ANOVA. We offer our recommendations in model adoption and ICC type selection as well as in result reporting. All modeling strategies and ICC types as well as the estimation and statistic testing for the fixed effects are currently available in the AFNI program `3dICC`.

Acknowledgments

The research and writing of the paper were supported by the NIMH and NINDS Intramural Research Programs (ZICMH002888) of the NIH/HHS, USA. We are indebted to Wolfgang Viechtbauer and Vincent Dorie for their precious help and for the *R* packages *metafor* and *brms* they build and maintain, respectively.

Appendix A. FMRI processing

The general sequence of FMRI data preprocessing steps was described in the subsection **Experimental testing dataset** under the section **Performance comparisons among the models**. However, for greater specificity and reproducibility, in this appendix we also include the exact command that was implemented for the processing. While there were several processing steps specified, each with many user-chosen options, it is possible to provide the exact pipeline in a succinct manner because the processing steps and options were created and specified using *afni_proc.py* in AFNI. This tool permits the user full freedom to tailor a desired pipeline that may be reliably duplicated for the entire group, stored for future reference and published with a study for unambiguous description.

properly, may lead to problematic interpretations about other effects in the model due to the complications involved in centering.

```
#!/bin/tcsh

# Set top level directory structure
set subjID = $1
set currDir = 'pwd'
set anatDir = ../freesurfer.anat/${subjID}/mri
set epiDir = afni
set stimDir = stim.files

# miscellaneous parameters or options
set motion_max = 1.0 ; set delete_nfirst = 4 ; set costfunc = lpc+zz

# run afni_proc.py to create a single subject processing script
afni_proc.py -subj_id ${subjID}                                          \
    -script proc.script.FINAL.1.20.17.${subjID}  -scr_overwrite          \
    -blocks despike tshift align tlrc volreg blur mask scale regress     \
    -copy_anat $anatDir/brainmask.nii                                    \
    -tcat_remove_first_trs $delete_nfirst                                \
    -dsets     $epiDir/r01+orig   $epiDir/r02+orig  $epiDir/r03+orig     \
    -blur_size 5       -out_dir NL.results                               \
    -anat_unif_GM no   -anat_has_skull no            -tlrc_NL_warp       \
    -volreg_align_e2a  -volreg_align_to MIN_OUTLIER  -volreg_tlrc_warp   \
    -align_opts_aea -cost $costfunc -giant_move -AddEdge                 \
    -regress_stim_times                                                  \
        $stimDir/${subjID}n_A_timing.1D                                  \
        $stimDir/${subjID}a50_timing.1D                                  \
        $stimDir/${subjID}a100_timing.1D                                 \
        $stimDir/${subjID}a150_timing.1D                                 \
        $stimDir/${subjID}n_F_timing.1D                                  \
        $stimDir/${subjID}f50_timing.1D                                  \
        $stimDir/${subjID}f100_timing.1D                                 \
        $stimDir/${subjID}f150_timing.1D                                 \
        $stimDir/${subjID}n_H_timing.1D                                  \
        $stimDir/${subjID}h50_timing.1D                                  \
        $stimDir/${subjID}h100_timing.1D                                 \
        $stimDir/${subjID}h150_timing.1D                                 \
        $stimDir/${subjID}w_timing.1D                                    \
    -regress_stim_labels                                                 \
        n_A a50 a100 a150  n_F f50 f100 f150  n_H h50 h100 h150 w        \
    -regress_local_times          -regress_censor_outliers 0.1           \
    -regress_basis 'BLOCK(2,1)'   -regress_censor_motion $motion_max     \
    -regress_est_blur_epits       -regress_est_blur_errts                \
    -regress_reml_exec    -regress_compute_fitts    -regress_opts_3dD    \
    -allzero_OK           -regress_opts_reml        -GOFORIT             \
    -regress_make_ideal_sum sum_ideal.1D                                 \
    -gltsym 'SYM: +n_A +a50 +a100 +a150 +n_F +f50 +f100 +f150 +n_H +h50 +h100 +h150 '  \
    -glt_label 1 Positive_Control

tcsh -xef proc.script.${subjID} |& tee proc.script.${subjID}.output
```

Table 5: A compact `tcsh` script that contains the succinct, tailored *afni_proc.py* command used to generate the full processing pipeline (> 500 lines) in AFNI for this study. To implement across the group, one simply loops through a list of subjects, entering the given file name as the sole command line argument, which is passed to the variable `$subjID`. Here, stimulus variables are encoded as: h = happy, n = neutral, f = fearful, w = wrong; and each is followed by the duration (50, 100, 150 s).

Appendix B. Data at three voxels from 25 subjects with two sessions that are illustrated in Fig. 4

Voxel	V_1				V_2				V_3			
Session	1		2		1		2		1		2	
Data	y_{1j}	$\hat{\sigma}^2_{1j}$	y_{2j}	$\hat{\sigma}^2_{2j}$	y_{1j}	$\hat{\sigma}^2_{1j}$	y_{2j}	$\hat{\sigma}^2_{2j}$	y_{1j}	$\hat{\sigma}^2_{1j}$	y_{2j}	$\hat{\sigma}^2_{2j}$
S1	0.075	0.013	-0.067	0.015	1.164	0.147	0.529	0.098	0.621	0.030	0.217	0.036
S2	0.160	0.006	0.081	0.004	0.705	0.016	0.100	0.026	0.407	0.018	-0.111	0.012
S3	0.101	0.006	0.084	0.009	-0.862	0.262	1.121	0.354	0.473	0.018	0.541	0.014
S4	-0.055	0.006	-0.041	0.004	0.297	0.050	0.113	0.014	0.012	0.008	-0.042	0.004
S5	0.075	0.010	0.053	0.010	0.416	0.031	0.412	0.044	0.036	0.011	-0.074	0.017
S6	0.282	0.009	0.199	0.010	0.590	0.012	0.166	0.019	0.906	0.011	0.539	0.014
S7	-0.012	0.014	0.273	0.010	1.042	0.038	0.727	0.030	0.486	0.012	0.506	0.008
S8	0.123	0.002	0.334	0.003	0.321	0.014	0.563	0.098	0.715	0.011	1.056	0.019
S9	-0.079	0.013	-0.213	0.013	0.366	0.012	-0.023	0.014	0.276	0.008	0.067	0.007
S10	-0.040	0.010	-0.127	0.015	0.705	0.026	0.734	0.028	0.439	0.015	0.331	0.020
S11	0.157	0.010	0.135	0.005	0.758	0.141	-0.033	0.159	0.280	0.013	-0.072	0.008
S12	0.056	0.004	-0.120	0.002	0.588	0.028	0.683	0.072	0.449	0.009	-0.052	0.005
S13	0.224	0.010	0.114	0.007	0.403	0.031	0.185	0.016	0.201	0.016	-0.294	0.015
S14	0.024	0.003	-0.007	0.003	0.713	0.016	0.501	0.021	0.453	0.011	0.061	0.013
S15	0.107	0.009	0.077	0.005	0.453	0.013	0.283	0.008	0.066	0.018	0.001	0.007
S16	0.182	0.008	0.160	0.006	1.120	0.020	0.303	0.017	0.438	0.011	0.019	0.006
S17	0.035	0.003	0.000	0.003	0.648	0.008	0.577	0.006	0.569	0.007	0.382	0.008
S18	0.014	0.003	-0.057	0.007	0.872	0.024	0.189	0.036	0.264	0.005	-0.025	0.009
S19	0.151	0.004	0.200	0.003	0.436	0.006	0.616	0.005	0.343	0.005	0.306	0.003
S20	0.165	0.004	0.020	0.003	0.278	0.007	0.386	0.009	0.398	0.004	0.434	0.007
S21	0.136	0.010	0.151	0.009	0.318	0.034	0.318	0.025	0.141	0.017	0.267	0.015
S22	0.165	0.012	0.177	0.021	-0.004	0.027	0.444	0.046	0.043	0.012	0.242	0.013
S23	0.005	0.005	0.085	0.003	0.410	0.005	0.217	0.016	0.322	0.004	0.313	0.005
S24	0.213	0.009	0.061	0.009	0.663	0.016	0.493	0.017	0.938	0.015	0.517	0.014
S25	0.059	0.010	0.132	0.008	1.191	0.011	0.318	0.012	0.324	0.008	0.001	0.011

References

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

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci. 1191:133-155.

Bhambhani, Y., Maikala, R., Farag, M., Rowland, G., 2006. Reliability of near-infrared spectroscopy measures of cerebral oxygenation and blood volume during handgrip exercise in nondisabled and traumatic

brain-injured subjects. J Rehabil Res Dev. 43(7):845-56.

Braun, U., Plichta, M.M., Esslinger, C., Sauer, C., Haddad, L., Grimm, O., Mier, D., Mohnke, S., Heinz, A., Erk, S., Walter, H., Seiferth, N., Kirsch, P., Meyer-Lindenberg, A., 2012. Test-retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. Neuroimage 59(2):1404-1412.

Brandt, D.J., Sommer, J., Krach, S., Bedenbender, J., Kircher, T., Paulus, F.M., Jansen, A., 2013. Test-Retest Reliability of fMRI Brain Activity during Memory Encoding. Frontiers in Psychiatry 4:163.

Cao, H., Plichta, M.M., SchŁfer, A., Haddad, L., Grimm, O., Schneider, M., Esslinger, C., Kirsch, P., Meyer-Lindenberg, A., Tost, H., 2014. Test-retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state. Neuroimage 84:888-900.

Cáceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. NeuroImage 45(3):758-768.

Chen et al., 2012. FMRI Group Analysis Combining Effect Estimates and Their Variances. NeuroImage. NeuroImage 60:747-765.

Chen, G., Saad, Z.S., Britton, J.C., Pine, D.S., Cox, R.W., 2013. Linear Mixed-Effects Modeling Approach to FMRI Group Analysis. NeuroImage 73:176-190.

Chen, G., Taylor, P.A., Shin, Y.W., Reynolds, R.C., Cox, R.W., 2017a. Untangling the Relatedness among Correlations, Part II: Inter-Subject Correlation Group Analysis through Linear Mixed-Effects Modeling. Neuroimage 147:825-840.

Chen, G., Taylor, P.A., Cox, R.W., 2017b. Is the statistic value all we should care about in neuroimaging? Neuroimage 147:952-959.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., Liu, J. 2013. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. Psychometrika 78(4):685-709.

Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment 6 (4):284-290.

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29:162-173. http://afni.nimh.nih.gov.

Fiecas, M., Ombao, H., van Lunen, D., Baumgartner, R., Coimbra, Feng D., 2013. Quantifying Temporal Correlations: A Test-Retest Evaluation of Functional Connectivity in Resting-State fMRI. NeuroImage 65: 231-241.

Fournier, J.C., Chase, H.W., Almeida, J., Phillips, M.L., 2014. Model Specification and the Reliability of fMRI Results: Implications for Longitudinal Neuroimaging Studies in Psychiatry. PLoS ONE 9(8):e105169.

Gelman, A., Hennig, C., 2016. Beyond subjective and objective in statistics. http://www.stat.columbia.edu/ gelman/research/unpublished/objectivityr3.pdf

Gonzalez-Castillo, J., Chen, G., Nichols, T., Cox, R.W., 2017. Bandettini, P.A., Variance Decomposition for Single-Subject task-based fMRI activity estimates across many sessions. NeuroImage 154:206-218.

Guo, C.C., Kurth, F., Zhou, J., Mayer, E.A., Eickhoff, S.B., Kramer, J.H., Seeley, W.W., 2012. One-year test-retest reliability of intrinsic connectivity network fMRI in older adults. Neuroimage 61(4):1471-1483.

Haller, S.P., Kircanski, K., Stoddard, J., White, L., Chen, G., Sharif-Askary, B., Zhang, S., Towbin, K.E., Pine, D., Leibenluft, E., Brotman, M.A., 2017. Reliability of Neural Activation and Connectivity during Implicit Face Emotion Processing in Youth. In preparation.

Higgins, J.P.T., Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine 21(11), 1539-1558.

Jaeger, L., Marchal-Crespo, L., Wolf, P., Riener, R., Kollias S., Michels, L., 2015. Test-retest reliability of fMRI experiments during robot-assisted active and passive stepping. Journal of NeuroEngineering and Rehabilitation 12:102

Kristo, G., Rutten, G.-J., Raemaekers, M., de Gelder, B., Rombouts, S.A.R.B., Ramsey, N.F., 2014. Task and Task-Free FMRI Reproducibility Comparison for Motor Network Identification. Human Brain Mapping 35:340-352.

Lin, Q., Dai, Z., Xia, M., Han, Z., Huang, R., Gong, G., Liu, C., Bi, Y., He, Y., 2015. A connectivity-based test-retest dataset of multi-modal magnetic resonance imaging in young healthy adults. Scientific Data 2.

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

Molloy, E.K., Birn, R.M., 2014. Tools for test-retest FMRI studies. ZENODO. DOI:10.5281/zenodo.49326.

Plichta, M.M., Herrmann, M.J., Baehne, C.G., Ehlis, A.C., Richter, M.M., Pauli, P., Fallgatter, A.J., 2006. Event-related functional near-infrared spectroscopy (fNIRS): are the measurements reliable? Neuroimage 31(1):116-124.

Shrout, P., Fleiss, J., 1979. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86(2):420-428.

Luke Tierney, A. J. Rossini, Na Li and H. Sevcikova (2016). snow: Simple Network of Workstations. R package version 0.4-2. https://CRAN.R-project.org/package=snow

Plichta, M.M., Herrmann, M.J., Baehne, C.G., Ehlis, A.C., Richter, M.M., Pauli, P., Fallgatter, A.J., 2007. Event-related functional near-infrared spectroscopy (fNIRS) based on craniocerebral correlations: reproducibility of activation? Hum Brain Mapp 28(8):733-741.

Quiton, R.L., Keaser, M.L., Zhuo, J., Gullapalli, R.P., Greenspan, J.D., 2014. Intersession reliability of fMRI activation for heat pain and motor tasks. NeuroImage: Clinical 5:309-321.

Recasens, M., Peter J. Uhlhaas, 2017. Test-retest reliability of the magnetic mismatch negativity response to sound duration and omission deviants. NeuroImage. In press.

Revelle, W. (2016) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.6.9.

Shah, L.M., Cramer, J.A., Ferguson, M.A., Birn, R.M., Anderson, J.S., 2016. Reliability and reproducibil-

ity of individual differences in functional connectivity acquired during task and resting state. Brain Behav 6(5):e00456.

Tian, F., Kozel, F.A., Yennu, A., Croarkin, P.E., McClintock, S.M., Mapes, K.S., Husain, M.M., Liu, H., 2012. Test-retest assessment of cortical activation induced by repetitive transcranial magnetic stimulation with brain atlas-guided optical topography. J Biomed Opt 17(11):116020.

Töger, J., Sorensen, T., Somandepalli, K., Toutios, A., Lingala, S.G., Narayanan, S., Nayak, K., 2017. Test-retest repeatability of human speech biomarkers from static and real-time dynamic magnetic resonance imaging. Journal of the Acoustical Society of America 141:3323

Viechtbauer, W., 2010. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1-48.

Wang, J.H., Zuo, X.N., Gohel, S., Milham, M.P., Biswal, B.B., He, Y., 2011. Graph Theoretical Analysis of Functional Brain Networks: Test-Retest Evaluation on Short- and Long-Term Resting-State Functional MRI Data. PLoS One 6(7), e21976.

White, L.K., Britton, J.C., Sequeira, S., Ronkin, E.G., Chen, G., Bar-Haim, Y., Shechner, T., Ernst, M., Fox, N.A., Leibenluft, E., Pine, D.S., 2016. Behavioral and neural stability of attention bias to threat in healthy adolescents. Neuroimage 136:84-93.

Woolrich, M.W., Behrens, T.E.J., Beckmann, C.F., Jenkinson, M., Smith, S.M., 2004. Multilevel linear modeling for FMRI group analysis using Bayesian inference. NeuroImage 21(4), 1732-1747.

Worsley, K.J., Liao, C., Aston, J., Petre, V., Duncan, G.H., Morales, F., Evans, A.C., 2002. A general statistical analysis for fMRI data. NeuroImage 15, 1-15.

Zanto, T.P., Pa, J., Gazzaley, A., 2014. Reliability measures of functional magnetic resonance imaging in a longitudinal evaluation of mild cognitive impairment. NeuroImage 84:443-452.

Zhang, H., Duan, L., Zhang, Y.J., Lu, C.M., Liu, H., Zhu, C.Z., 2011. Test-retest assessment of independent component analysis-derived resting-state functional connectivity based on functional near-infrared spectroscopy. Neuroimage 55(2):607-615.

Zuo, X.N., Martino, A.D., Kelly, C., Shehzad, Z.E., Gee, D.G.,Klein, D.F., Castellanos, F.X., Biswal, B.B., Milham, M.P., 2010a. The oscillating brain: Complex and reliable. NeuroImage 49(2):1432-1445.

Zuo, X.N., Kelly, C., Adelstein, J.S., Klein, D.F., Castellanos, F.X., Milham, M.P., 2010. Reliable intrinsic connectivity networks: TestÐretest evaluation using ICA and dual regression approach. NeuroImage 49(3):2163-2177.