To pool or not to pool: Can we ignore cross-trial variability in FMRI?

Gang Chen*[a], Srikanth Padmala[b], Yi Chen[c,d], Paul A Taylor[a], Robert W Cox[a], and Luiz Pessoa[e]

[a]Scientific and Statistical Computing Core, NIMH, National Institutes of Health, USA
[b]Centre for Neuroscience, Indian Institute of Science, Bangalore, India
[c]German Center for Neurodegenerative Diseases, Magdeburg, Germany
[d]IKND, Universität Magdeburg, Germany
[e]Department of Psychology, Department of Electrical and Computer Engineering, Maryland Neuroimaging Center, University of Maryland, College Park, USA

Abstract

In this work, we investigate the importance of explicitly accounting for cross-trial variability in neuroimaging data analysis. To attempt to obtain reliable estimates in a task-based experiment, each condition is usually repeated across many trials. The investigator may be interested in (a) condition-level effects, (b) trial-level effects, or (c) the association of trial-level effects with the corresponding behavior data. The typical strategy for condition-level modeling is to create one regressor per condition at the subject level with the underlying assumption that responses do not change across trials. In this methodology of *complete pooling*, all cross-trial variability is ignored and dismissed as random noise that is swept under the rug of model residuals. Unfortunately, this framework invalidates the generalizability from the confine of specific trials (e.g., particular faces) to the associated stimulus category ("face"). Here we propose an adaptive and computationally tractable framework that meshes well with the current two-level pipeline and explicitly accounts for trial-by-trial variability. The trial-level effects are first estimated per subject through *no pooling*. To allow generalizing beyond the particular stimulus set employed, the cross-trial variability is modeled at the population level through *partial pooling* in a multilevel model, which permits accurate effect estimation and characterization. Alternatively, trial-level estimates can be used to investigate, for example, brain-behavior associations or correlations between brain regions. Furthermore, our approach allows appropriate accounting for serial correlation, handling outliers, adapting to data skew, and capturing nonlinear brain-behavior relationships. By applying a Bayesian multilevel model framework at the level of regions of interest to an experimental dataset, we show how multiple testing can be addressed and full results reported without arbitrary dichotomization. Our approach revealed important differences compared to the conventional method at the condition level, including how the latter can distort effect magnitude and precision. Notably, in some cases our approach led to increased statistical sensitivity. In summary, our proposed framework provides an effective strategy to capture trial-by-trial responses that should be of interest to a wide community of experimentalists.

Introduction

The workhorse of functional magnetic resonance imaging (FMRI) studies is the *task design*, where it is possible to experimentally manipulate conditions to investigate the brain basis of perception, cognition, emotion,

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

and so on. The reliability of a task-based experiment hinges on having a reasonably large number of *repetitions* associated with a condition. Such repetitions are usually termed "trials", and each trial is considered to be an instantiation of an idealized condition. For example, in an emotion study with three conditions (positive, neutral and negative), the investigator may show 20 different human faces of each emotional valence to the subject in the scanner. From the statistical perspective, the number of trials serves as the sample size for each condition and, per the law of large numbers in probability theory, the average effect estimate for a specific condition should approximate the (idealized) expected effect with increased certainty as the number of trials grows.

Statistics lives by and flourishes in the rich variability of the data. The ultimate goal of most neuroimaging studies lies in generalizing results at the population level: the objective is to make statements that go beyond the particular samples studied. Thus, variability across samples serves as a key yardstick to gauge the evidence for the impact of experimental manipulations. More generally, in a neuroimaging study, at least four interrelated levels of variability are woven into the data tapestry – subjects, brain regions, trials and consecutive time points – all of which deserve proper statistical treatment.[1] (a) Among these four levels, cross-subject variability is the easiest and most straightforward to handle. As the experimental subjects usually can be considered as independent and identically distributed, cross-subject variability is typically captured through a Gaussian distribution at the population level. In other words, each participant's effect is considered to be drawn from a hypothetical population that follows a Gaussian distribution. (b) Because FMRI data inherently form a time series, strategies must be developed to handle the sequential dependency in the data. As the underlying mechanisms of BOLD response are not fully understood, the current models cannot exhaustively account for various effects and confounds; thus, temporal structure remains in the model residuals. The awareness of this issue has indeed generated various strategies of autoregressive (AR) modeling to tackle it. (c) Multiplicity is an intrinsic issue of the massively univariate approach adopted in neuroimaging with voxels or regions treated as independent units. Various strategies have been developed, including cluster-based inferences, Gaussian random field theory and permutation-based methods. At the level of regions, we recently proposed an integrative approach that handles cross-region variability with a Bayesian multilevel (BML) model that dissolves the conventional multiplicity issue (Chen et al., 2019a, 2019b). (d) Lastly, until recently, trial-by-trial response variability had received little attention (Westfall et al., 2017; Yarkoni, 2019). The central objective of the present study is to develop a multilevel framework to effectively handle this source of variability.

What is trial-by-trial variability? Clearly, multiple sources contribute to cross-trial fluctuations, although these are rather poorly understood. When the fluctuations are of no research interest, they are often treated as *random noise* under the assumption that the "true" response to, say, a fearful face in the amygdala exists, and deviations from that response constitute random variability originating from the measurement itself or from neuronal/hemodynamic sources. Consider a segment of a simple experiment presenting five faces (Fig.1). In the standard approach, the time series is modeled with a single regressor that takes into account all face instances (Fig.1a,b). The fit, which tries to capture the mean response, does a reasonable job at explaining signal fluctuations. However, the fit is clearly poor in several places.

It is important to properly account for cross-trial variability. Under the condition-level modeling utilized in standard data analysis, trial-by-trial fluctuations are flatly swept under the rug of the model residuals, creating at least three problems: (a) an unrealistic assumption of "fixed" responses across trials, (b) the loss of hierarchical structure across the two different levels – trial and TR – of data variability, and (c) the inability to legitimately generalize from the confine of specific trials (e.g., 5 neutral faces from a given stimulus dataset, Fig. 1f) to the condition category (e.g., neutral face, Fig. 1g). This last point means that, strictly speaking, the domain of generalizability of experiment is the set of trials, which is clearly not the way experimentalists interpret their findings. If one adopts a principled trial-level modeling as developed here (see also Westfall et al., 2017), trial-based regressors can be utilized to capture trial-by-trial fluctuations, thereby potentially capturing overall signal

[1] The variabilities across these four levels are usually considered random effects under the conventional statistical framework. In contrast, variations associated with, for example, conditions (e.g. positive, neutral and negative), subject groups (e.g., patients and controls) or quantitative variables (e.g., age, RT), are treated as fixed effects at the population level.

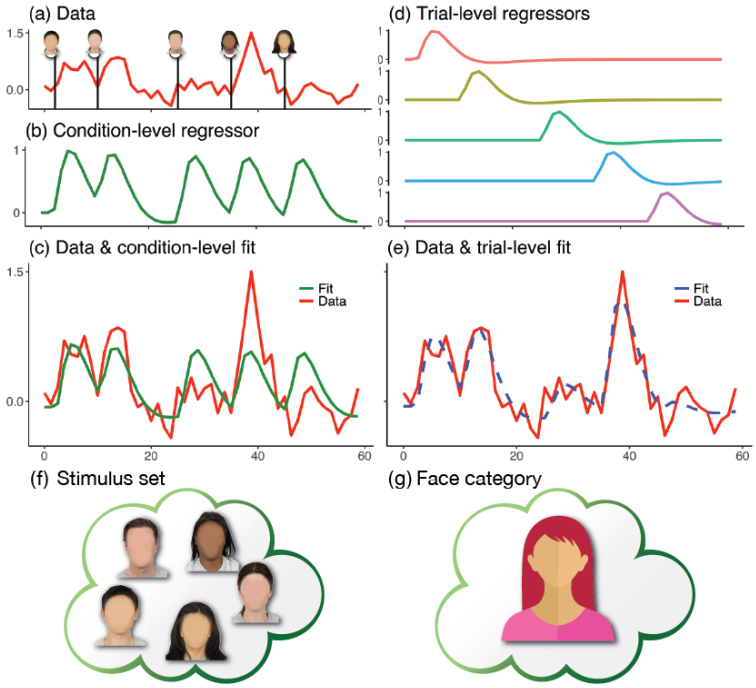


Figure 1: Time series modeling in neuroimaging. Consider an experiment with five face stimuli. (a) Hypothetical times series (scaled by mean value) is shown at a brain region associated with five stimuli. (b) The conventional modeling approach assumes that all stimuli produce the same response with one regressor. (c) An effect estimate (in percent signal change or scaling factor for the regressor (b)) is associated with the fit (green) at the condition level. (d) An alternative approach models each stimulus separately with one regressor per stimulus. (e) Trial-level modeling provides an improved fit (dashed blue). (f) The set of five stimuli (specific faces, blurred for privacy only) serves as a representation of and potential generalization to a condition category (face). (g) As described in the paper, trial-level estimates can be integrated via partial pooling such that inferences can be made at the general category level.

fluctuations better (Fig. 1d,e). Importantly, as the varying trial response is explicitly accounted for, inferences can be made at the desired level (e.g., neutral face, Fig. 1g).

There are at least three instances where the investigator is actually interested in trial-by-trial variability. An important application is when relating cross-trial fluctuations to behavior. For example, trial-level effects can be associated with success/failure in task performance (Ress et al., 2000; Pessoa et al., 2002; Sapir et al., 2005; Lim et al., 2009). Another use is in correlation analysis being established in a trial-by-trial fashion, at times called "beta series correlation" (Rissman et al., 2004). Trial-level responses are also used for prediction purposes, including multivoxel pattern analysis (MVPA), support vector machines (SVM), and neural networks more generally. Our research goal is to develop an adaptive methodology that can capture cross-trial fluctuations effectively, thus allowing it to be applied to the cases above as well as typical population-level analysis. Indeed, whereas we illustrate the approach with behavioral and FMRI data, it can equally be applied to M/EEG and calcium imaging data.

Conventional time series modeling strategies

The conventional whole-brain voxel-wise analysis adopts a massively univariate approach with a two-level procedure: the first is at the subject level, and the second at the population level.[2] The split between these two levels is usually due to two reasons. One is model complexity: because of idiosyncrasies across subjects (e.g., different trial effects and confounds, varying AR structures), it is generally unwieldy to integrate all subjects into

[2]Due to the difficulty and varying strategies of handling the discontinuities cross runs and sessions, the analytical pipeline of FMRI analysis can be described in the literature with a two-, three- or even four-level procedure depending on the specific pipeline or software. For example, the analysis for each run may be labeled as the first level, followed by a second level that summarizes the effect estimates across runs (and a third level for across-session summary) through simple averaging or a fixed-effects model; the analysis for generalization at the population level is thus termed as third (or fourth) level. As the data across runs and sessions can be integrated into one model at the subject level through a numerical scheme (e.g., Chen et al., 2012), here we stick to a two-level description for simplification. To avoid the messy terminology in the field, we directly describe the two levels as subject and population instead of their ordinal sequence.

a single model. The second consideration is practicality. In particular, it is computationally impractical to solve one "giant", integrative model even if one could build it.

The statistical model at the subject level is time series regression[3] solved through generalized least squares (GLS). The preprocessed EPI data y_k is fed into a time series regression model as the response variable on the left-hand side,

$$\text{GLS}: y_k = \alpha_0 + \alpha_1 z_{1k} + ... + \alpha_m z_{mk} + \beta_1 x_{1k} + \beta_2 x_{2k} + ... + \beta_n x_{nk} + \epsilon_k,\ k = 1, 2, ..., K, \tag{1}$$

where k indexes discrete time points, and the residuals ϵ_k are assumed to follow a Gaussian distribution. Between the two sets of regressores, the first set z_{ik} $(i = 1, 2, .., m)$ contains various covariates including slow drifts (e.g., polynomial or sine/cosine terms associated with low-frequency signals), head-motion variables, outlier censoring and physiological confounds such as cardiac and breathing effects, while the second set x_{jk} $(j = 1, 2, .., n)$ is associated with the experimental conditions. Correspondingly, there are two groups of effect parameters: the first set α_i $(i = 1, 2, .., m)$ is usually of no interest to the investigator while the second set $\beta_j,\ j = 1, 2, ..., n$ is the focus of specific research questions.

The construction of condition regressors x_{jk} $(j = 1, 2, .., n)$ in the GLS model (1) largely depends on the research focus. For most investigations, the interest is placed on the effects at the condition level, and the trials of each condition are treated as multiple instantiations of the event of interest. While various approaches are adopted to construct the condition regressors x_{jk}, they are typically treated with the assumption that the response magnitude remains the same across all trials of a condition (Fig. 1b,c). Specifically, one regressor per condition is constructed through the convolution of the individual trial duration with a fixed-shape hemodynamic response function.[4] Note that the fixed-response-magnitude approach can be relaxed in certain scenarios. For example, one may modulate the trial-level response by creating another regressor through auxiliary information (e.g., RT). At present, we will focus on an alternative approach: to capture the trial-level effects, one feeds one regressor per trial to the GLS model (1); trial-level modulation, if desired, will be performed at the population level.

Another complexity involves the residuals ϵ_k in the GLS model (1). If the residuals are white (i.e., no autocorrelation), time series regression can be numerically solved through ordinary least squares (OLS) or maximum likelihood. However, it has been long recognized that temporal correlation structure exists in the residuals (e.g., Bullmore et al., 1996) because some components in the data either are unknown or cannot be properly accounted for. Failure to model the autocorrelation may lead to inflated reliability (or underestimated uncertainty) of the effect estimates. Three strategies that utilize GLS have been proposed to improve the model by characterizing the temporal correlations in the residuals ϵ_k. First, an early approach was to characterize the autocorrelation with a uniform first-order autoregressive model for the whole brain (Friston et al., 2002). Second, a localized AR(1) model was developed later so as to consider neighboring voxels within each tissue type (Woolrich et al., 2001). Third, an even more flexible approach was created using an autoregressive moving average ARMA(1,1) structure that accommodates the model at the voxel level through the program 3dREMLfit in AFNI (Chen et al., 2012). A recent comparison study has shown that the performances of the three methods matche their respective modeling flexibility, complexity and adaptivity (Olszowy et al., 2019).

The conventional modeling of condition regressors in the GLS model (1) can be further extended. For one, we can take inspiration from typical population-level analysis, which includes a term for each subject so that subject-specific effects are properly accounted for. The same approach can be adopted at the trial level to account for trial-by-trial variability. In particular, the assumption that all the trials of a given condition share the same brain response magnitude should be viewed skeptically (Fig. 1d,e). Critically, from a modeling perspective, treating

[3]The popular term for the subject-level analysis is general linear model (GLM) in neuroimaging. However, a more accurate description of the modeling approach is *time series regression*, especially considering the nature of input data and the complex issue of delicately handling the temporal structure embedded in the residuals through generalized least squares (GLS) (cf., ordinary least squares (OLS) for GLM).

[4]The alternative approaches with multiple basis functions share the common assumption of same response magnitude across trials.

trials as "fixed effects" is tantamount to limiting the focus of the study to the trial instantiations employed, foiling the validity of the experimenter's goal to generalize from the particular samples used (e.g., specific faces utilized in the experiment) to the generic level (e.g., human faces in general). Needless to say, the latter generalization is taken for granted in neuroimaging studies. Here, we argue that the modeling strategy adopted should address this issue head on, and we demonstrate a trial-level modeling approach to achieving this goal.

Methods

Perspectives on trial-level modeling

Our motivation is to directly model trial-level effects at the subject level and to account for across-trial variability at the population level. In doing so, the conventional assumption of constant response across trials is abandoned in light of the following two perspectives.

1) **Research focus**. Depending on the specific research hypothesis, one may be interested in: (a) trial-level effect estimates for each subject, so that those effects can be utilized for predictions or correlativity among regions; (b) association of trial-level effects with behavioral data; or (c) condition-level effects. We will focus on the latter two which involve population-level analysis.

2) **Modeling perspective**. The BOLD response magnitude varies across trials, but what is the nature of the trial-to-trial fluctuations? There are three modeling strategies depending on the ultimate research goal, mapping to three different data pooling methods (Chen et al., 2019b). The first, commonly adopted approach assumes that the underlying BOLD response does not change from trial to trial and that the observed fluctuations are noise or random sampling variability (Fig. 1b,c). Thus, the average response is estimated across trials to represent condition-level effects. This approach can be considered to be *complete pooling*, where all the "individuality" of trials is ignored in the model. Technically, the approach precludes generalization to the trial category in question, and does not allow extending one's conclusions from the specific trials used in the experiment to situations beyond the trials employed (Yarkoni, 2019). In contrast, it is possible to adopt a *no pooling* strategy at the subject level, and estimate each trial's response separately (Fig. 1d,e); in other words, each trial is fully unique and assumed to be unrelated to other trials. Between the two extremes, a middle ground can be taken at the population level such that the cross-trial variations are considered as random samples of the condition-level effect (cf. subjects as samples of an idealized population). This characterization of randomness allows the investigator to make the generalization from the specific trials instantiated in the experiment to the *concept* of a condition category, the idealized population from which trials are envisioned to be random samples. With such *partial pooling* approach, information can be loosely but meaningfully shared across trials.

Neuroimaging is no stranger to dealing with the three pooling methods. In fact, the issue about cross-trial variability basically runs parallel to its cross-subject counterpart. The typical split between the subject- and population-level analysis means that a no-pooling strategy is adopted at the individual subject level in the sense that each subject is assumed to have unique response effects; then partial-pooling is typically followed up at the population level with a Gaussian distribution for cross-subject variability. In the early days, there were even choices between fixed- versus random-effects analysis at the population level; such a comparison is just another way to elaborate the differences between complete and partial pooling. Today, complete pooling for cross-subject variability (or fixed-effects analysis) is typically considered unacceptable (leading to paper rejection!), and the adoption of partial pooling (or random-effects analysis) at the population level is routine practice. It is exactly the same underlying rationale that we wish to address in the context of cross-trial variability, thus we believe there are no legitimate reasons preventing the analyst from the adoption of a more general pooling methodology.

Population analysis through trial-level modeling

We start with a linear mixed-effects (LME) platform for population analysis. The model incorporates trial-level effect estimates y_{st} under one condition from individual subjects based on the GLS model (1),

$$\begin{aligned} & y_{st} = \alpha_0 + \xi_s + \eta_t + \epsilon_{st}; \\ & \xi_s \sim \mathcal{N}(0,\ \lambda^2),\ \eta_t \sim \mathcal{N}(0,\ \omega^2),\ \epsilon_{st} \sim \mathcal{N}(0,\ \sigma^2); \\ & s = 1, 2, ..., S;\ t = 1, 2, .., T. \end{aligned} \tag{2}$$

The indices s and t code subjects and trials, respectively; α_0 is the intercept that embodies the overall effect at the population level; ξ_s and η_t represent cross-subject and cross-trial effects (random effects); ϵ_{st} is the residual term and usually assumed to follow a Gaussian distribution. When explanatory variables are involved (e.g., between- and/or within-subject variables), the model can be naturally extended by augmenting the intercept term α_0. The LME framework (2) with a crossed or factorial random-effects structure can be numerically analyzed by, for example, the program 3dLMEr (Chen et al., 2013) in AFNI (Cox, 1996) at the whole-brain voxel-wise level.

How does the conventional approach compare to the LME formulation for trial-level modeling? In the conventional approach trial effects are obviously not modeled at the subject level. To a first approximation, the condition-level effect can be conceptualized as the arithmetic mean of the trial-level effect estimates $\overline{y}_{s\cdot} = \frac{1}{T}\sum_{t=1}^{T} y_{st}$. When no trial-level information is available, the LME model (2) simply reduces to the conventional Student's t-test for the condition-level effects $\overline{y}_{s\cdot}$:

$$\begin{aligned} & \overline{y}_{s\cdot} = \alpha_0 + \epsilon_s; \\ & s = 1, 2, ..., S. \end{aligned} \tag{3}$$

With assumption of an identical and independent distribution of cross-trial effects $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(0,\ \omega^2)$, the missing component of the cross-trial effects η_t in the GLM (3) relative to the LME counterpart (2) means that the variability,

$$\mathrm{Var}(\frac{1}{T}\sum_{t=1}^{T}\eta_t) = \frac{\omega^2}{T}, \tag{4}$$

is not accounted for in the condition-level approach.[5] Notice that this variability depends on and is sensitive to T, the number of trials. Therefore, if T is large enough, the variability may become inconsequential, in which case complete pooling could be justified as a reasonable approximation. However, given that the number of trial repetitions is relatively small, such scenario appears unrealistic. Accurately mapping the data hierarchy and explicitly characterizing cross-trial variability, as represented by the trial-specific terms η_t and its distribution $\mathcal{N}(0, \omega^2)$ in the LME model (2), legitimizes the generalizability from the specific trials to a general category.

We reiterate that it is through the explicit capture of cross-trial variability that provides a solid foundation for generalization. As a routine practice, nowadays cross-subject variability is properly accounted for at the population level, and such accountability is evidenced in conventional models as simple as Student's t-tests, GLM and AN(C)OVA, or as the subject-specific terms ξ_s and their distribution $\mathcal{N}(0, \lambda^2)$ in the above LME platform (2). However, the same rationale has not been adopted and applied to the cross-trial variability, even though the adoption of many exemplars of a condition in experimental designs is intended for generalization.

One potential improvement of the LME model (2) is the incorporation of effect precision. The subject-level effect estimates (e.g., y_{st} in the LME model (2)) from the GLS model (1) are estimated, naturally, with some degree of uncertainty (embodied by the standard error, $\hat{\sigma}_{st}$). As the whole analysis pipeline is broken into the two levels of subject and population, theoretically it is desirable to explicitly incorporate the reliability

[5] The i.i.d assumption about cross-trial effects η_t in the derivation is likely violated for several reasons including potential serial correlation. However, the overall logic remains applicable.

information of the effect estimates into the population-level model so that the information hierarchy would be largely maintained. In standard practices, subject-level standard error is usually ignored at the population level; such practice assumes that the uncertainty is either exactly the same across subjects or negligible relative to the cross-subject variability (Chen et al., 2012). To address this shortcoming, some population-level methods have been developed to incorporate both effect estimates from the subject level and their standard errors (Worsley et al., 2002; Woolrich et al., 2004; Chen et al., 2012). However, this integration approach has not gained much traction in practice due to its small potential gain (Mumford et al., 2009; Chen et al., 2012; Olszowy et al., 2019). Within the LME framework, unfortunately there is no easy solution to consider these standard errors. In contrast, taking them into account is a natural component of BML modeling, and we will explore the role of precision information in the current context of trial-level modeling.

Another possible improvement of the LME model (2) is outlier handling. Due to the substantial expansion in the number of regressors involved in trial-level modeling, the chance of having outlying effect estimates cannot be ignored. However, it is a challenge to handle outliers and data skew within the LME framework. A typical approach is to set hard bounds, thus constraining data to a predetermined interval in order to exclude outliers. In contrast, by adopting a BML framework, outliers can be accommodated in a principled manner with the utilization of non-Gaussian distributions for data variability.

Four different BML models are considered here. Following our recent Bayesian approach (Chen et al., 2019a), we formulate the models within a single integrative platform at the level of region of interest (ROI) to capture the hierarchical structure among three intersecting levels: subjects, trials and regions. Specifically, the trial-level effect estimates y_{str} are modeled as follows:

$$\begin{aligned}
\text{base model - } \mathsf{M0}: \quad & y_{str} \sim \mathcal{N}(\mu,\ \sigma^2); \\
\text{model with standard error - } \mathsf{Me}: \quad & y_{str} \sim \mathcal{N}(\mu,\ \hat{\sigma}^2_{str}); \\
\text{model with } t\text{-distribution - } \mathsf{Mt}: \quad & y_{str} \sim \mathcal{T}(\nu, \mu,\ \sigma^2); \\
\text{hybrid model - } \mathsf{Mh}: \quad & y_{str} \sim \mathcal{T}(\nu, \mu, \hat{\sigma}^2_{str}); \\
& s = 1, 2, ..., S;\ t = 1, 2, .., T;\ r = 1, 2, ..., R;
\end{aligned} \tag{5}$$

where r, s and t index the ROIs, subjects and trials, and the parameter ν is the number of degrees of freedom for the t-distribution. The four BML models differ along two dimensions in a crossed manner: (a) whether the uncertainty information $\hat{\sigma}^2_{str}$ (effect variance from the subject level) is incorporated and propagated from the subject to population level, and (b) whether Gaussian $\mathcal{N}$ or Student's t-distribution $\mathcal{T}$ is assumed for the response variable. Under the Bayesian framework, the BML models (5) and the ones hereafter are expressed as a distribution or likelihood function, rather than as an equation (like the LME model in (2)). Hence, the parameter μ in the four models (5) can be further specified as follows:

$$\begin{aligned}
& \mu = \alpha_0 + (\xi_s + \eta_t + \zeta_r) + (\iota_{st} + \kappa_{sr} + \upsilon_{tr}); \\
& \xi_s \sim \mathcal{N}(0,\ \lambda^2),\ \eta_t \sim \mathcal{N}(0,\ \omega^2),\ \zeta_r \sim \mathcal{N}(0,\ \theta^2),\ \iota_{st} \sim \mathcal{N}(0, \pi^2),\ \kappa_{sr} \sim \mathcal{N}(0, \phi^2),\ \upsilon_{tr} \sim \mathcal{N}(0, \psi^2); \\
& s = 1, 2, ..., S;\ t = 1, 2, .., T;\ r = 1, 2, ..., R.
\end{aligned} \tag{6}$$

As previously, the indices s and t code for subjects and trials, respectively; α_0 is the intercept; ξ_s and η_t represent cross-subject and cross-trial effects; ζ_r accounts for the effect of the rth region. Compared to its LME counterpart (2), the four BML models incorporate brain regions (indexed by r), augmenting the LME model to a platform with three crisscross levels. Due to the addition of this cross-region dimension, it is possible to further include the three two-way interaction terms, ι_{st}, κ_{sr} and υ_{tr} among the three effects of subjects, trials and regions (with parentheses grouping the three single levels and their interactions).

The relationships among the four BML models are as follows. Relative to the base model M0, Me takes into account the precision of the effect estimate y_{str} by utilizing the standard error $\hat{\sigma}_{str}$ from the subject level.

To handle outliers and skew in the data y_{str}, we replace the Gaussian distribution with Student's t-distribution and formulate the third BML model, Mt. Thanks to its leptokurtic property (i.e., having "heavy tails"), the t-distribution (with the Gaussian distribution as its asymptote) has increasingly more mass in the tails as the degrees of freedom decrease, effectively counteracting the potential impact of outlying values. Lastly, the models Me and Mt can be combined to incorporate uncertainty and to handle outliers simultaneously, leading to the hybrid Mh. In all model versions, the intercept α_0 can be expanded with terms to accommodate terms such as subject-grouping and quantitative covariates.

Handling behavioral covariates and nonlinearity

Trial-level modeling can be extended to incorporate behavioral variables. In conventional approaches, the association between trial-level effects and behavior can be modeled by creating a modulatory variable at the subject level. Accordingly, instead of one, two regressors are constructed per condition. The first is the typical regressor for the average condition effect (here, the behavioral measure is considered at a center value, such as the subject's mean). The second regressor codes, for example, for the linear relationship between BOLD response and the behavioral measure. When trial-level effects are directly estimated at the subject level, the following LME can be adopted at the population level:

$$\begin{aligned}
&y_{st} = \alpha_0 + \alpha_1 x_{st} + \xi_{0s} + \xi_{1s} x_{st} + \epsilon_{st};\\
&(\xi_{0s}, \xi_{1s})' \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{\Lambda}),\ \epsilon_{st} \sim \mathcal{N}(0, \sigma^2);\\
&s = 1, 2, ..., S;\ t = 1, 2, .., T;
\end{aligned} \tag{7}$$

where x_{st} is the behavioral measure of the sth subject at the tth trial. The parameters α_0 and ξ_{0s} are population- and subject-level intercepts, respectively. The effect of the behavioral variable x_{st} on the response variable y_{st} is captured through the slope parameter α_1 at the population level, while its subject-level counterpart is characterized by the slope parameters ξ_{1s}. The 2×2 variance-covariance matrix $\boldsymbol{\Lambda}$ reflects the relationship between the subject-level intercept ξ_{0s} and slope ξ_{1s}.

The modeling of behavioral covariates can be altered to relax the linearity assumption. Polynomials (e.g., quadratic terms) can be used but still require some extent of prior knowledge and assumption about the relationship. Alternatively, we can adopt smoothing splines with a set of basis functions defined by a modest sized set of knots. For example, we can use penalized cubic smoothing splines $s(\cdot)$ to achieve a counterbalance between the goodness of fit and the curvature or wiggliness measured by the integrated square of second derivative (Wood, 2017):

$$\begin{aligned}
&y_{st} = s(x_{st}) + \xi_s + \epsilon_{st};\\
&\xi_s \sim \mathcal{N}(0, \lambda^2),\ \epsilon_{st} \sim \mathcal{N}(0, \sigma^2);\\
&s = 1, 2, ..., S;\ t = 1, 2, .., T.
\end{aligned} \tag{8}$$

Alternative trial-level modeling

Westfall et al. (2017) proposed an integrative approach to address the trial-level generalization problem. Accordingly, they relaxed the assumption of fixed BOLD response across trials, and directly modeled trial-to-trial fluctuations. In addition, both the subject and population levels were merged into one model. In the following description, we have slightly modified and generalized their original notation with data of I experimental

conditions from S subjects,

$$
\begin{aligned}
& y_{sk} = \alpha_0 + \alpha_1 y_{s,k-1} + \alpha_2 y_{s,k-2} + \sum_{i=1}^{I} (\beta_i + \xi_{si}) \sum_{t=1}^{T_i} x_{sitk} + \sum_{i=1}^{I} \sum_{t=1}^{T_i} \eta_{it} x_{sitk} + \epsilon_{sitk}; \\
& \xi_{si} \overset{i.i.d.}{\sim} \mathcal{N}(0, \lambda_i^2),\ \eta_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \omega_i^2),\ \epsilon_{sk} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2); \\
& i = 1, 2, ..., I;\ s = 1, 2, ..., S;\ k = 3, 4, ..., K; t = 1, 2, ..., T_i.
\end{aligned} \tag{9}
$$

Indices s, i, t and k code subject, condition, trial and time, respectively; T_i is the number of trials for the ith condition; α_0 is the overall intercept; λ_i^2 and ω_i^2 characterize the cross-subject and cross-trial variability, respectively, for the ith condition; x_{sitk} is the trial-level regressor; the effect associated with the regressor x_{sitk} of the ith condition is partitioned into two components, β_i for the average (fixed) component across all trials and ξ_{si} for the subject-specific (random) component; η_{it} represents the cross-trial (random) effect shared by all subjects; ϵ_{sitk} is the residual term with the assumption of white noise (no serial correlation) and variance σ^2. Note that an autoregressive AR(2) structure with two parameters α_1 and α_2 is explicitly modeled with lagged effects as regressors, instead of being embedded in the residuals as typically practiced in the field (Worsley et al., 2002; Woolrich et al., 2001; Chen et al., 2012). All random effects and residuals are assumed Gaussian. In addition, likely for computational simplifications, the intercept α_0, AR effects α_1 and α_2, cross-trial effect η_{it} are assumed to be the same across subjects. Due to the unavailability of numerical implementations and the intractable computational cost, the LME model was solved at the region level in Westfall et al. (2017) through the NiPyMC Python package. Finally, they focused solely on conventional statistical evidence and its dichotomization, whereas we wish to consider both effect magnitude and the associated statistical evidence through a more continuous view of statistical support.

Trial-level modeling and study goals

We will use an FMRI dataset to demonstrate our trial-level modeling framework that blends in well with the current analytical pipeline. At the subject level, the effect estimate at each trial is obtained with no pooling through the GLS model (1), with the temporal correlation in the residuals captured via an ARMA(1,1) structure. At the population level, in parallel to cross-subject variability, the trial-level effects are modeled through partial pooling to address the following question: What are the differences and consequences compared to the conventional approach of complete pooling? The common practice in neuroimaging is largely limited on statistical evidence followed by artificial dichotomization; thus, relatively little attention is paid to effect magnitude. For example, Westfall et al. (2017) reported substantially inflated statistical values (1.5-3.0 times) when complete pooling was adopted. Here, we wish to explore whether we could replicate the inflation of statistical evidence while emphasizing the impact of trial-level modeling on both effect estimate and its uncertainty. Overall, the issues that we want to address include:

1) extent of cross-trial variability;
2) variability of autocorrelation structure in the subject-level residuals;
3) impact of directly modeling the autocorrelation with lagged effects;
4) cross-trial fluctuations as an indication of synchrony among regions;
5) importance of incorporating precision information in model formulation;
6) handling of data skew and outliers;
7) reporting full results in a comprehensive fashion.

Most of the models in this paper are under the Bayesian framework using Stan (Carpenter et al., 2017) through the R package brms (Bürkner, 2018). The choice of Bayesian modeling was made for multiple reasons, most notably its ability to incorporate multiplicity and to provide a straightforward interpretation of effect estimates through posterior distributions, instead of point estimates and significance testing thresholding. Each Bayesian model here is specified with a likelihood function, followed by priors for lower-level effects (e.g., trial,

region, subject). The hyperpriors employed for model parameters (e.g., population-level effects, variances in prior distributions) are discussed in Appendix A. Note, however, that if the ROI-related components in our models are excluded, the models can be applied at the whole-brain voxel level under the conventional LME framework.

Trial-level modeling of FMRI data

Experimental data

We adopted a dataset from a previous experiment (Padmala et al., 2017). A cohort of 57 subjects was investigated in a 3T scanner. Each subject performed 4 task types,[6] each of which was repeated across 48 trials. Each task started with a 1-s cue phase indicating the prospect of either reward (`Rew`) or no-reward (`NoRew`) for performing the subsequent task correctly. The cue was followed by a 2-6-s variable delay period. The task stimulus itself was displayed for 0.2 s. Participants had to perform a challenging perceptual task when confronted with either a negative (`Neg`) or neutral (`Neu`) distractor. The subject was then expected to respond within 1.5 s. The total $4 \times 48 = 192$ trials were randomly arranged and evenly divided across 6 runs with 32 trials in each run. The TR was 2.5 s. In the analyses that follow, only correct trials were employed.

We sought to investigate the interaction between motivation (reward) and emotion (distraction) in both behavior and brain data. The experiment manipulated two factors: one was the prospect (Pro) of being either rewarded or not while the other was the distractor (Dis) displayed, which was either negative or neutral. In terms of brain responses, there were six effects of interest: two cue types (`Rew`, `NoRew`) and four prospect-by-distractor task types (`NoRew_Neg`, `NoRew_Neu`, `Rew_Neg`, `Rew_Neu`) following a 2×2 factorial structure. In terms of behavior, the focus was the recorded trial-level RT on the same four task types. In addition, the relationship between brain response and behavioral RT was of interest. Here, variable names with a first capital letter (e.g., Pro and Dis for the manipulation factors) symbolize population effects (or fixed effects under the conventional framework), whereas those with a first lowercase letter (e.g. subj, trial and roi for subject, trial and ROI) indicate lower-level (or random) effects.

Behavioral data analysis

We first analyzed behavioral performance at the condition level. The success rate, rate_{ijs}, measures the proportion of correct responses (out of 48 trials) of sth subject under the task of ith prospect and jth distractor. The data could be analyzed with a binomial distribution through a logistic model. However, to aid interpretability, as the number of trials of each task was reasonably large, the binomial distribution was approximated as Gaussian; thus, we opted to model the success rate data in a manner that follows the base model $\mathsf{M0}$ in (5):

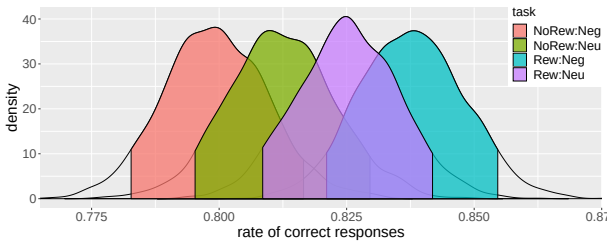
$$\begin{aligned}
&\text{rate}_{ijs} \sim \mathcal{N}(\mu,\ \sigma^2);\\
&\mu = \text{Pro}_i * \text{Dis}_j + \text{subj}_s;\\
&\text{subj}_s \sim \mathcal{N}(0,\ \lambda^2);\\
&i = 1, 2;\ j = 1, 2;\ s = 1, 2, ..., 57.
\end{aligned} \tag{10}$$

The term subj_s captures the subject-specific effect, $\text{Pro}_i * \text{Dis}_j$ represents $\alpha_0 + \text{Pro}_i + \text{Dis}_j + \text{Pro}_i : \text{Dis}_j$, where α_0 is the intercept (0th-order interaction), Pro_i the effect associated with the $i$the level of prospect, Dis_j the effect associated with the jth level of distractor, and $\text{Pro}_i : \text{Dis}_j$ the second-order interaction between the two variables.[7] We also fitted the accuracy data using a t-distribution, after the model Mt; however, the latter did not improve model fit considerably.

[6] The terms "condition" and "task" are interchangeable in the literature in describing a stimulus type. Here we use "condition" to describe a general category of trials to avoid any potential confusion since the experiment involves both cues and tasks.

[7] In general, two factors of m and n levels, respectively, have totally an intercept, $m-1$ and $n-1$ terms for the individual effects of the two factors, and $(m-1)(n-1)$ interactions. How these total mn terms are formulated depends on the specific parameterization method such as dummy and deviation coding.

(a) Posterior distributions of correct response rates



(b) Main effects and interactions

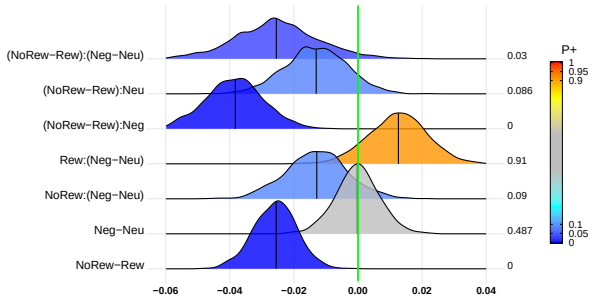


Figure 2: Summary of accuracy based on the BML model (10). (a) Response accuracy and the associated 95% quantile interval were estimated for each of the 4 tasks. (b) Among the posterior distributions of accuracy, the bottom two rows are the main effects while the top five rows show the interactions. At the right side of each distribution lists the posterior probability of each effect being positive, P+ (area under the curve to the right of the green line indicating zero effect), also color-coded in the distribution shading. The vertical black line under each distribution is the median (or 50% quantile). This figure corresponds to Fig. 3B in Padmala et al. (2017).

The response accuracy data rate_{ijs} is consistent with the following conclusions. Accuracy varied to some extent across the four conditions (Fig. 2a). For the main effects of the two factors (prospect and distractor, bottom two rows, Fig. 2b), the subjects had a lower response accuracy for the `NoRew` condition than `Rew`, while the accuracy for the two distractors types `Neg` and `Neu` was comparable. The overall interaction between prospect and distractor, (`NoRew`−`Rew`):(`Neg`−`Neu`), was fairly robust (top row, Fig. 2b). Specifically, the prospect effect (`NoRew`−`Rew`) was larger under the `Neg` distractor than `Neu` (second and third row, Fig. 2b) while the distractor effect (`Neg`−`Neu`) was largely in the opposite direction between the two prospects of `NoRew` and `Rew` (fourth and fifth row, Fig. 2b).

Now we focus on the RT data at the trial level. The RT analyses had to deal with the issue of trial- versus condition-level dichotomy, illustrating the differentiation between complete and partial pooling. Across participants, the number of correct responses ranged from 28 to 47 out of 48. We constructed a M0-type model that directly accounts for cross-trial variability,

$$\begin{aligned}
&\text{RT}_{ijst} \sim \mathcal{N}(\mu,\ \sigma^2);\\
&\mu = \text{Pro}_i * \text{Dis}_j + \text{subj}_s + \text{trial}_t + \text{subj}_s : \text{trial}_t;\\
&\text{subj}_s \overset{i.i.d.}{\sim} \mathcal{N}(0,\ \lambda^2);\ \text{trial}_t \overset{i.i.d.}{\sim} \mathcal{N}(0,\ \omega^2);\ \text{subj:trial}_{st} \overset{i.i.d.}{\sim} \mathcal{N}(0,\ \pi^2);\\
&i = 1, 2;\ j = 1, 2;\ s = 1, 2, ..., 57;\ t = 1, 2, ..., T_{ijs}\ (28 \le T_{ijs} \le 47).
\end{aligned} \tag{11}$$

The terms Pro_i and Dis_j are the effects associated with the prospect and distractor level, respectively; subj_s, trial_t, and subj:trial_{st} are the varying effects associated with the sth subject, tth trial and their interaction, respectively; T_{ijs} is the number of correct responses of the sth subject during the task of ith prospect and jth distractor. Examination of the RT data indicated that the overall distribution was skewed to some extent (Fig. 3a). Thus, we explored two modified models using either a Student t-student or an exponentially modified Gaussian distribution (Palmer et al., 2011) to handle the skew, simply by replacing $\mathcal{N}(\mu, \sigma^2)$ in the model (11) with $\mathcal{T}(\nu, \mu, \sigma^2)$ or $\mathcal{EMG}(\mu, \sigma^2, \beta)$, respectively, where ν is the parameter that codes the degrees of freedom for the t-distribution, and β is an exponential decay parameter for the exGaussian distribution. These two models produced similar effect estimates and statistical evidence. However, they provided improved fitting: the estimated degrees of freedom, ν, had a mean of 3.4 with 95% quantile interval of [3.1, 3.7], consistent the skewness of the data; skewness was also accommodated by the exponential decay parameter estimate $\beta = 106.19 \pm 1.81$ ms.

The RT data supports the following conclusions. The posterior distribution was different among the four tasks (Fig. 3b). For the main effects of the two factors (prospect and distractor, bottom two rows, Fig. 3c), RTs were substantially shorter during `Rew` trials, and `Neg` distractors robustly slowed down behavior. The overall interaction between prospect and distractor had strong support (top row, Fig. 3c).

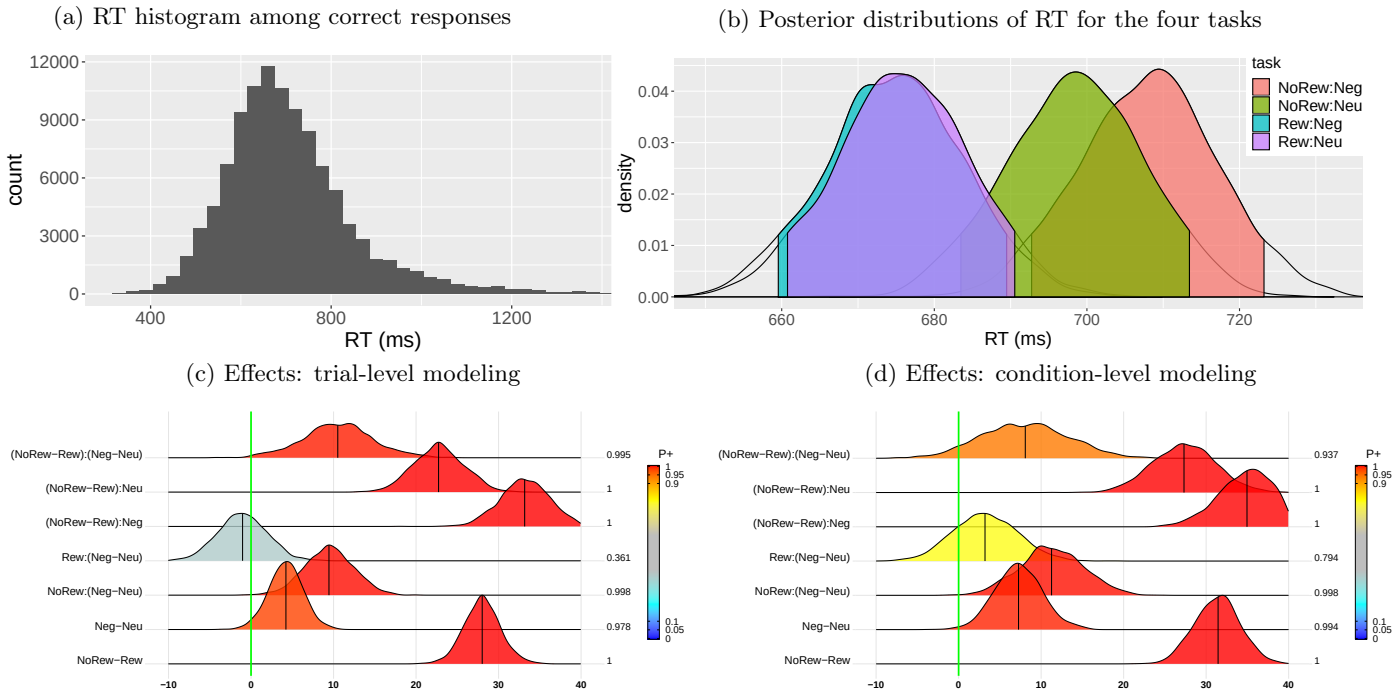


Figure 3: Summary of RT data based on the BML model (11) with t-distribution. (a) The histogram of RT among correct response trials shows the aggregated information across the trials (within [28, 47]), 4 tasks and 57 subjects (bin width: 30 ms). (b) RT and the associated 95% quantile intervals were shown for each of the 4 tasks with an overall mean of 689.3ms and s.d. of 8.8ms. (c) Among the posterior distributions, the bottom two rows are the main effects while the top five rows show the the interactions. At the right side of each distribution lists the posterior probability of each effect being positive, P+ (area under the curve to the right of the green line indicating zero effect), also color-coded in the distribution shading. The black vertical segment under each distribution shows the median. (d) The counterpart result of (c) based on the condition-level RT effects aggregated cross trials (corresponding to Fig. 3A in Padmala et al. (2017)).

To gauge the effectiveness of trial-level modeling, we also analyzed the RT data at the condition level. As typically practiced for condition-level effects in neuroimaging, we aggregated the RT data across trials within each condition through averaging (i.e., complete pooling), thereby assuming the same RT across all trials under each task:

$$\begin{aligned}
&\overline{\mathrm{RT}}_{ijs\cdot} \sim \mathcal{T}(\mu,\ \sigma^2);\\
&\mu = \mathrm{Pro}_i * \mathrm{Dis}_j + \mathrm{subj}_s;\\
&\mathrm{subj}_s \overset{i.i.d.}{\sim} \mathcal{N}(0,\ \lambda^2);\\
&i = 1, 2;\ j = 1, 2;\ s = 1, 2, ..., 57;
\end{aligned} \tag{12}$$

where $\overline{\mathrm{RT}}_{ijs\cdot} = \frac{1}{T_{ijs}} \sum_{t=1}^{T_{ijs}} \mathrm{RT}_{ijst}$. The major difference of the model (12) relative to the trial-level model (11) lies in the omission of terms related to the trial-level effects, trial_t. On the surface, the statistical evidence (Fig. 3d) based on the condition-level model (12) was similar to its trial-level counterpart (Fig. 3c). This is not surprising given the massive evidence for most effects. However, the results also illustrate the higher sensitivity and efficiency of partial pooling relative to complete pooling in that the interaction effect, (`NoRew−Rew`):(`Neg−Neu`), received only modest support under the condition-based model while being convincingly affirmed by the trial-level model. We note that the interaction was the chief concern in the original study (Padmala et al., 2017), which would not be deemed "statistically significant" under the traditional dichotomous framework. Overall, this example illustrates how data variability is more accurately decomposed and characterized through the trial-level model than the aggregation approach.

Neuroimaging data analysis

Time series data were preprocessed using AFNI at each voxel. Steps included cross-slice alignment, cross-TR alignment (mitigation of head motion), cross-subject alignment (normalization to standard space), spatial smoothing (FWHM: 6 mm) and voxel-wise scaling to 100 through dividing the data by the mean signal. To illustrate our modeling framework, we analyzed the data at the ROI level to highlight effective ways to visualize the full results without thresholding, and to demonstrate how BML aids in handling multiple testing. Among the 11 selected ROIs, seven were based on their involvement in attention and executive function more generally: left/right frontal eye fields (FEF), left/right anterior insula (Ins), left/right intraparietal sulcus (IPS), supplementary/pre-supplementary motor area (SMA). We included four additional ROIs, the left/right ventral striatum (VS) and left/right amygdala (Amyg), which are known for their involvement in reward and affective processing, respectively. The ROIs were defined as follows: insula masks were from Faillenot et al. (2017); ventral striatum masks were based on Pauli et al. (2016); amygdala ROIs were defined from Nacewicz et al. (2014); for the remaining regions the peak coordinates of the analysis by Toro et al. (2008) were used to create spherical ROIs.

Trial-level effect estimation at the subject level

Trial-level effects were estimated for each subject as follows. For each ROI, time series data were extracted and averaged across all voxels. The resulting representative time series was analyzed by applying the GLS model (1) with 3dREMLfit in AFNI. Six effects of interest were considered at the condition level: two cue types (`Rew` and `NoRew`) and four tasks (`Rew_Neg`, `Rew_Neu`, `NoRew_Neg` and `NoRew_Neu` factorially combined in terms of the factors Pro and Dis). We compared two approaches: the conventional condition-level method of creating one regressor per condition, and our approach of modeling each trial with a separate regressor. Each regressor was created by convolving a 1-s rectangular wave with an assumed HRF filter (Gamma variate). Multiple regressors of no interest were also included in the model: separate third-order Legendre polynomials for each run; regressors associated with 6 head-motion effects and their first-order derivatives; and regressors for trials with incorrect responses. In addition, we censored time points for which head motion was deemed substantial. The 6 runs of data were concatenated with the cross-run gaps properly handled (Chen et al., 2012). With 48 originally planned trials per task, each of the four tasks were modeled at the trial level, resulting in T_{ijs} ($28 \leq T_{ijs} \leq 47$) regressors associated with the ith prospect and jth distractor; each of the two cues were modeled with $T_{i\cdot s} = \sum_{j=1}^{2} T_{ijs}$ regressors. Each of the error trials and the corresponding cues were modeled separately. For comparison, condition-level effects were also estimated directly for each subject through two approaches. First, each condition was modeled with a regressor that is associated with the T_{ijs} trials. Second, each condition was modeled with two regressors, one was associated with the average RT across the T_{ijs} trials while the other captured the modulation effect of RT.

Four approaches were adopted in handling the correlation structure in the residuals: (a) OLS with the assumption of white noise, (b) AR(1), (c) AR(2) and (d) ARMA(1,1), with the latter three models numerically solved through GLS. In addition, we compared the model with AR(2) for the residuals to an AR(2) model with lagged effects of the BOLD signal, as suggested in Westfall et al. (2017). Our comparisons (Appendix B) indicated that AR(2) and ARMA(1,1) for the model residuals rendered similar effect estimates and both slightly outperformed AR(1); thus, all the effect estimates for further analyses were from ARMA(1,1).

Trial-level modeling is vulnerable to the multicollinearity problem. The original experiment was neither intended nor optimally designed for trial-level modeling. Indeed, a few subjects had highly correlated regressors at the trial level between a cue and its subsequent task (correlations among the regressors were below 0.6 for most subjects except for seven who had correlation values above 0.9 among a few regressors). Close inspection revealed that the high correlations were mostly caused by motion effects and the associated data censoring, or by a short separation between a cue and the following task.

(a) Cross-trial synchrony between the two amygdala regions

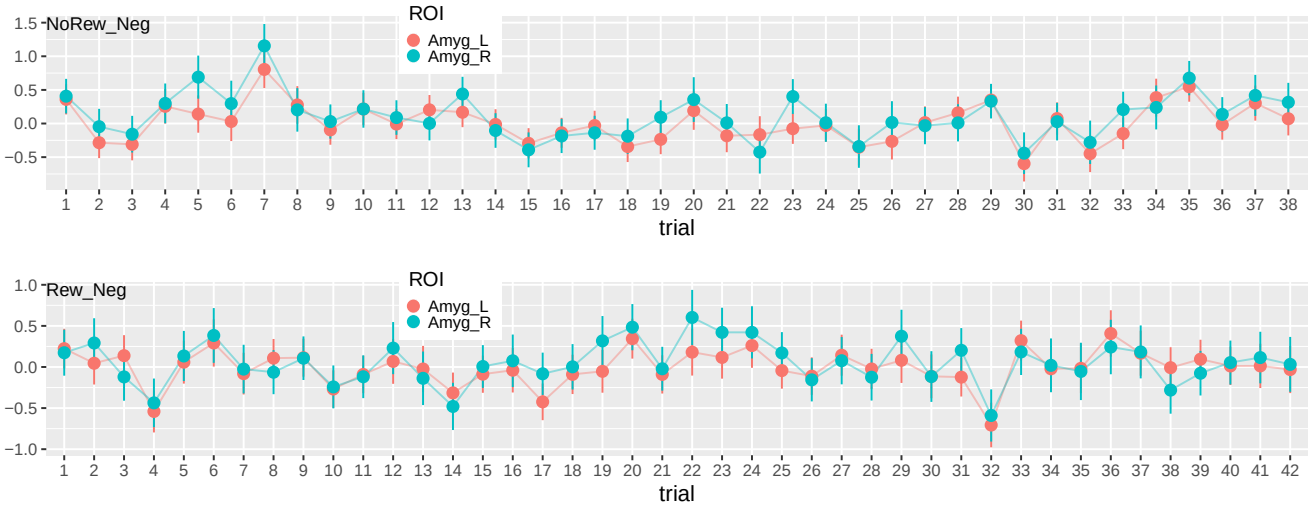


Figure 4: Synchronization among brain regions. The effect estimates (dots) with their standard errors (line segments) were obtained through the GLS model with ARMA(1,1). Some extent of synchrony existed across trials between the left and right amygdalas of a subject under two different tasks of `NoRew_Neg` (upper panel) and `Rew_Neg` (lower panel).

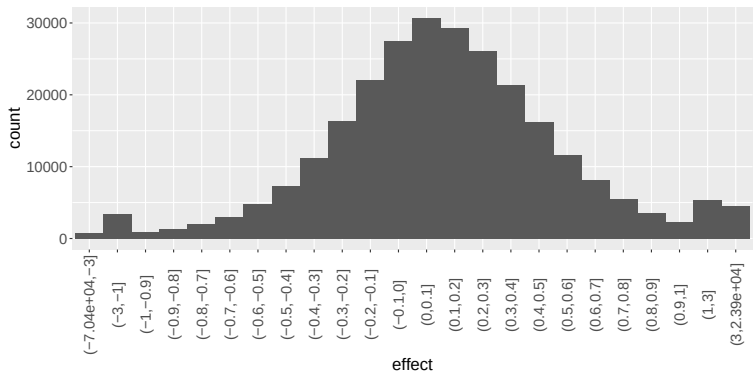


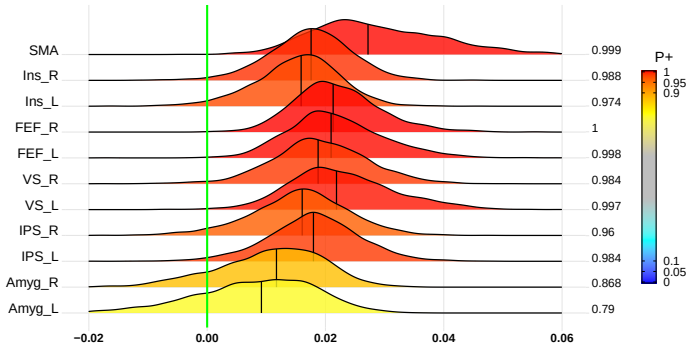
Figure 5: Distribution of the effect estimates from the GLS model with ARMA(1,1). With 11 ROIs and 57 subjects, there were $11 \times \sum_{s=1}^{57} \sum_{i=1}^{2} \sum_{j=1}^{2} T_{ijs} = 98461$ trial-level effect estimates ($28 \leq T_{ijs} \leq 47$) among the 4 tasks. A small portion (450, 0.42%) were outlying values beyond the range of [-2, 2] with the most extremes reaching -70000 and 23900. To effectively accommodate outliers, the x-axis was shrunk beyond (-1, 1).

Trial-level effects varied substantially without a clear pattern (Figs. 4,12). Across trials, the estimated BOLD response changed substantially, and occasionally showed negative estimates. Such seemingly random fluctuations appeared across all conditions, regions and subjects. Possible factors influencing trial responses include fluctuations in attention, familiarity, habituation effects, poor modeling and pure noise. Despite the absence of a clear pattern, it is quite revealing to observe some degree of synchronization between the five contralateral region pairs: an association analysis rendered a regression coefficient of 0.73 ± 0.04 between the region pairs, indicating that a 1% signal change at a right brain region was associated with about 0.73% signal change at its left counterpart (Figs. 4,12).

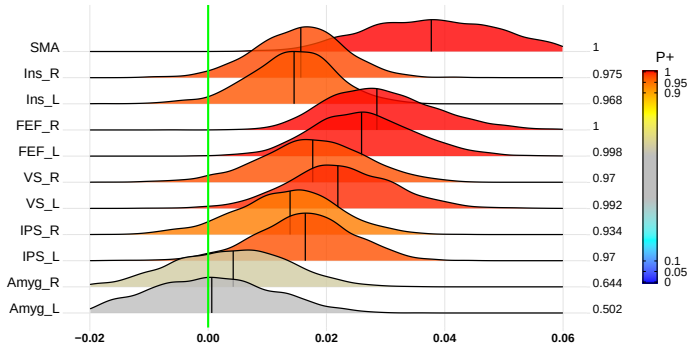
Condition effect estimation at the population level

We started with population-level analyses for the four tasks. Inspection of the histogram of effect estimates from the GLS model with ARMA(1,1) in (1) revealed a fraction of outliers that were beyond [-2, 2] (Fig. 5), which were traced mostly to the censoring of time points due to head motion. If not handled properly, extreme values would likely distort the analysis. In our investigation of task-phase effects, we utilized four approaches to handling outliers: (a) M0: brute force removal of values outside [-2, 2]; (b) Me: incorporation of uncertainty for effect estimates; (c) Mt: adoption of t-distribution; and (d) Mh: hybrid of Me and Mt. Specifically, the four models in (5) were applied to the effect estimates y_{ijrst} and their variances $\hat{\sigma}^2_{ijrst}$ with the indices i and j coding

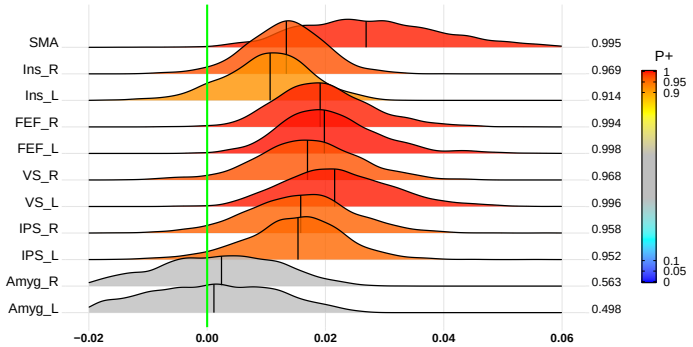
(a) Model M0: outlier removal



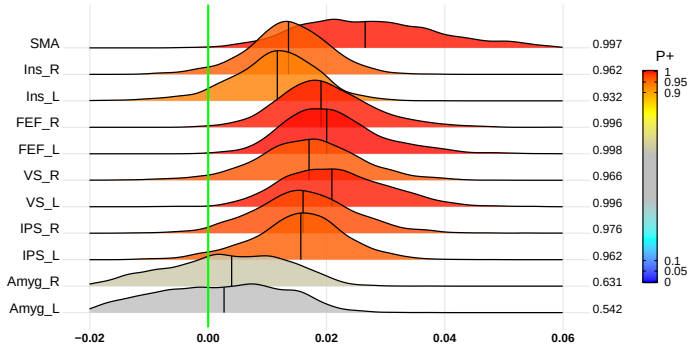
(b) Model Mt: Student's t



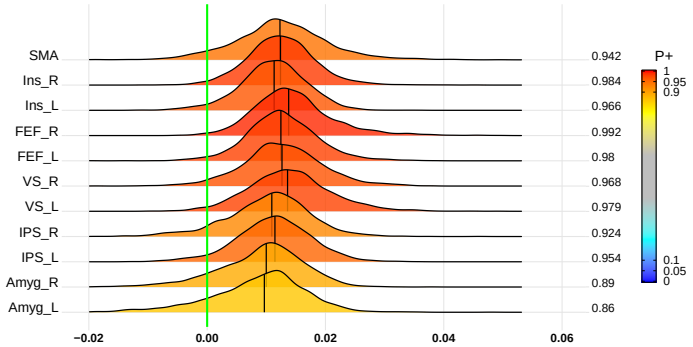
(c) Model Me: uncertainty incorporated



(d) Model Mh: hybrid of Mt and Me



(e) Conventional approach: condition-level effects as input



(f) Model Mh: with RT effects controlled

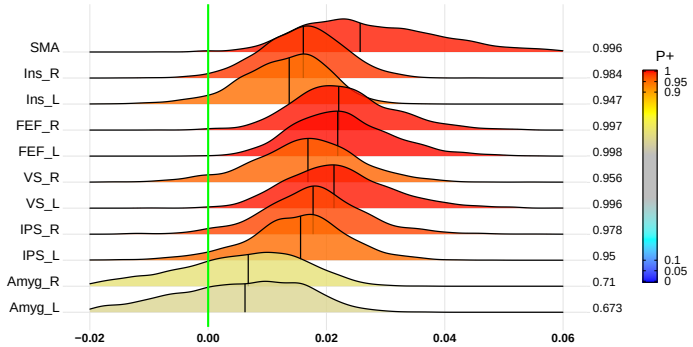


Figure 6: Interaction (`NoRew`−`Rew`):(`Neg`−`Neu`) at the population level. The value at the right end of each line indicates the posterior probability of the effect being great than 0 (vertical green line), color-coded in the area under each posterior density. (a) M0: values beyond [-2, 2] were considered as outliers and removed. (b) Mt: outliers and skewness were accommodated with t-distribution. (c) Me: uncertainty of effect estimates was used to deal with outliers. (d) Mh: both uncertainty of effect estimates and t-distribution were adopted. (e) Conventional approach: condition-level effects from each subject were fitted in the model (14). (f) Covariate modeling: trial-level effects were modeled with RT as a covariate in the model (17).

for the two factors of prospect and distractor:

$$\begin{aligned}
&\mu = \text{Pro}_i * \text{Dis}_j + \text{Pro}_i * \text{Dis}_j : \text{roi}_r + \text{subj}_s + \text{trial}_t + \text{roi}_r\text{:subj}_s + \text{roi}_r\text{:trial}_t + \text{subj}_s\text{:trial}_t;\\
&(\text{Pro}_i * \text{Dis}_j : \text{roi}_r) \sim \boldsymbol{\mathcal{N}}(\mathbf{0},\ \boldsymbol{\Theta}),\ \text{subj}_s \sim \mathcal{N}(0,\ \lambda^2),\ \text{trial}_t \sim \mathcal{N}(0,\ \omega^2),\\
&\text{roi}_r\text{:subj}_s \sim \mathcal{N}(0,\ \phi^2),\ \text{roi}_r\text{:trial}_t \sim \mathcal{N}(0,\ \psi^2),\ \text{subj}_s\text{:trial}_t \sim \mathcal{N}(0,\ \pi^2);\\
&i = 1,2;\ j = 1,2;\ s = 1,2,...,S;\ t = 1,2,..,T_{ijs};\ r = 1,2,...,R;
\end{aligned} \tag{13}$$

where the 4×4 variance-covariance matrix $\boldsymbol{\Theta}$ captures the cross-region variability among the four tasks.

All the four models generated relatively similar posterior distributions for the interaction effect (Fig. 6a-d). First, the base model M0 overestimated the effect magnitude for the two amygdala ROIs. To appreciate the differences between the models, consider the number of degrees of freedom for the t-distribution, ν, in Mt which is adaptively estimated from the shape of the data distribution. The strength of t-distribution in handling heavier tails and potential outliers is demonstrated by the small degrees of freedom estimated as $\nu = 3.24 \pm 0.03$ in Mt (cf. the Gaussian distribution of M0 with $\nu = \infty$). The inclusion of uncertainty in Me also allowed effective handling of extreme values (similar estimates as Mt, in particular for the left/right amygdala). This can be appreciated by noting the decreased role of the t-distribution in the hybrid model Mh, where $\nu = 61.8 \pm 5.1$. Overall, instead of relying on a predetermined threshold value to handle outliers in M0, models Mt, Me and Mh offer principled approaches to adjusting to the shape of the data distribution and the presence of potential outliers. Although the practical differences in the present dataset were small, we believe they have broader usefulness. In the remainder of the paper, we utilized the model Mh given its more general formulation. It is worth noting that the important role of standard errors in the models Me and Mh necessitates the accurate accountability of the serial correlation in the GLS model (1).

How do the above results compare to the conventional approach of condition-level modeling through complete pooling? To perform this evaluation, we defined a GLS model (1) with ARMA(1,1) at the subject level that contained six condition-level effects with the assumption that the BOLD response was the same across the trials under each condition. At the population level, task effects y_{ijrs} and their standard errors $\hat{\sigma}_{ijrs}$ were fitted with the BML model Mh in (5):

$$\begin{aligned}
&\mu = \text{Pro}_i * \text{Dis}_j + \text{Pro}_i * \text{Dis}_j : \text{roi}_r + \text{subj}_s + \text{roi}_r\text{:subj}_s;\\
&\text{Pro} : \text{Dis} : \text{roi}_r \sim \boldsymbol{\mathcal{N}}(\mathbf{0},\ \boldsymbol{\Theta}),\ \text{subj}_s \sim \mathcal{N}(0,\ \lambda^2),\ \text{roi}_r\text{:subj}_s \sim \mathcal{N}(0,\ \phi^2);\\
&i = 1,2;\ j = 1,2; s = 1,2,...,S;\ r = 1,2,...,R;
\end{aligned} \tag{14}$$

with definitions as before, and where $\boldsymbol{\Theta}$ is a 4×4 variance-covariance matrix for the cross-region variability among the four tasks. Compared to the trial-level modeling approaches (Fig. 6a-d), the condition-level modeling approach produced similar statistical evidence (Fig. 6e), but exhibited inflated reliability (i.e., narrower posteriors), as well as underestimated effect magnitude (densities closer to 0) at most ROIs. In other words, complete pooling tended to homogenize effect estimates and inflate their certainty.

How about the conventional modulation analysis? Under this approach, cross-trial variability is accounted for via a linear modulation of the RT data at the subject level. At the population level we applied model Mh in (14) to the condition-level estimates that associated with RT modulation. The resulting interaction effects at the population level (not shown here) were very similar to the ones without the RT modulation (Fig. 6e). Thus, in this case, a modulation regressor did not substantially alter the posterior densities of the interaction effects.

Now we switch to investigate the cue-phase responses. The Mh model (5) was applied with the trial-level

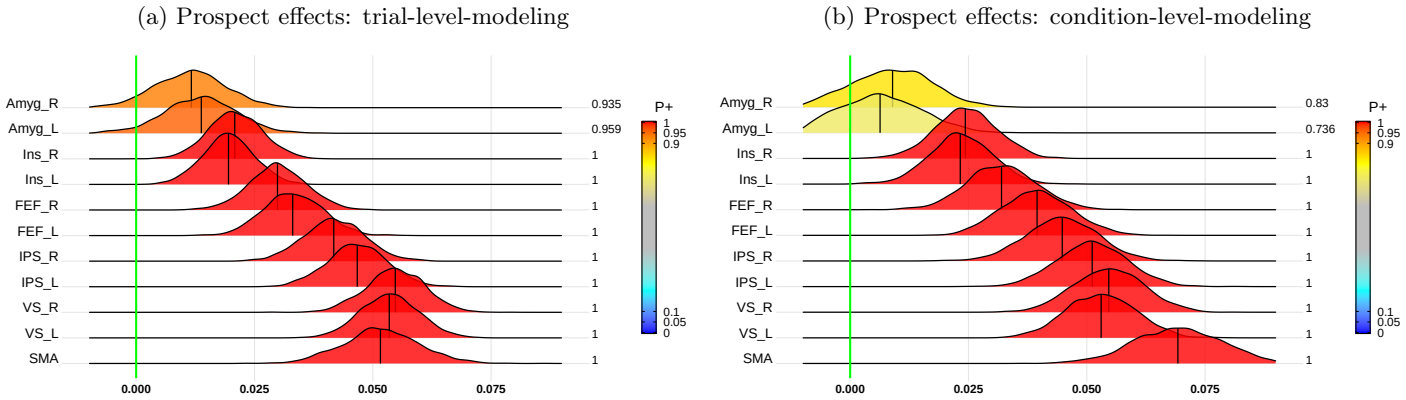


Figure 7: Prospect effect (`Rew`−`NoRew`) during cue phase at the population level. Even though the two approaches of trial- and condition-level modeling agreed with each other to some extent in terms of statistical evidence for the contrast between `Rew` and `NoRew`, trial-level modeling (a) showed stronger evidence for both left and right amygdala than its condition-level counterpart (b).

effect estimates y_{irst} and standard errors $\hat{\sigma}_{irst}$ for the cue types `NoRew` and `Rew`:

$$\begin{aligned}
&\mu = \text{Pro}_i + \text{Pro}_i : \text{roi}_r + \text{subj}_s + \text{trial}_t + \text{roi}_r\text{:subj}_s + \text{roi}_r\text{:trial}_t + \text{subj}_s\text{:trial}_t;\\
&\text{Pro} : \text{roi}_r \sim \boldsymbol{\mathcal{N}}(\mathbf{0},\ \boldsymbol{\Theta}),\ \text{subj}_s \sim \mathcal{N}(0,\ \lambda^2),\ \text{trial}_t \sim \mathcal{N}(0,\ \omega^2),\\
&\text{roi}_r\text{:subj}_s \sim \mathcal{N}(0,\ \phi^2),\ \text{roi}_r\text{:trial}_t \sim \mathcal{N}(0,\ \psi^2),\ \text{subj}_s\text{:trial}_t \sim \mathcal{N}(0,\ \pi^2);\\
&i = 1, 2;\ s = 1, 2, ..., S;\ t = 1, 2, .., T_{ijs};\ r = 1, 2, ..., R;
\end{aligned} \tag{15}$$

where $\boldsymbol{\Theta}$ is a 2×2 variance-covariance matrix for the cross-region variablity between the two cue types. Most ROIs showed extremely strong evidence for a prospect effect with greater responses during `Rew` relative to `NoRew` (Fig. 7a), although the righ/left amygdala showed weaker support.

Again, to compare the trial-level approach to the conventional condition-level strategy, the Mh model (5) was fit with the condition-level effect y_{irs} and standard errors $\hat{\sigma}_{irs}$ for the cue types `NoRew` and `Rew`:

$$\begin{aligned}
&\mu = \text{Pro}_i + \text{Pro}_i : \text{roi}_r + \text{subj}_s + \text{roi}_r\text{:subj}_s;\\
&\text{Pro} : \text{roi}_r \sim \boldsymbol{\mathcal{N}}(\mathbf{0},\ \boldsymbol{\Theta}),\ \text{subj}_s \sim \mathcal{N}(0,\ \lambda^2),\ \text{roi}_r\text{:subj}_s \sim \mathcal{N}(0,\ \phi^2);\\
&i = 1, 2;\ s = 1, 2, ..., S;\ r = 1, 2, ..., R.
\end{aligned} \tag{16}$$

Most of the results from the condition-level approach (Fig. 7b) were similar to the trial-based analysis (Fig. 7a), because of the large effects sizes. However, the condition-level approach did not capture the amygdala effects well, where evidence in their favor was rather weak. In contrast, the trial-based analysis garnered much stronger evidence, and at least the left amygdala would cross a typical one-sided 0.05 statistical threshold (although we believe this dichotomous procedure is detrimental to progress).

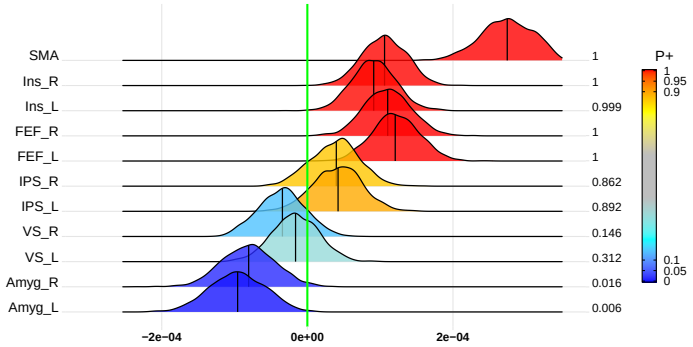
Association analysis with behavioral data

To probe the association between the BOLD response during the task phase and the RT data, we adopted the Mh model (5) for the trial-level effect estimates y_{ijrst} and their standard errors $\hat{\sigma}_{ijrst}$:

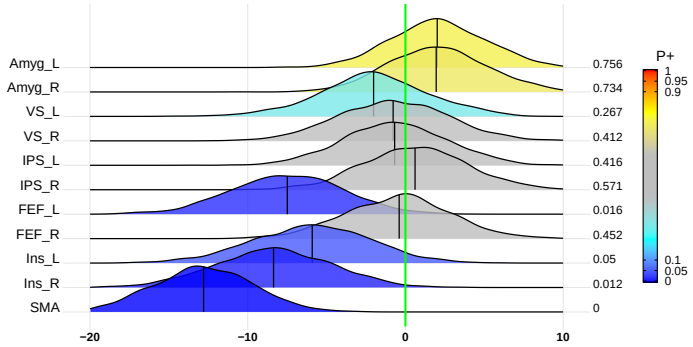
$$\begin{aligned}
&\mu = \text{Pro}_i * \text{Dis}_j * \text{RT}_{ijst} + \text{Pro}_i * \text{Dis}_j * \text{RT}_{ijst} : \text{roi}_r + \text{RT}_{ijst} : \text{subj}_s + \text{RT}_{ijst} : \text{roi}_r\text{:subj}_s;\\
&\text{Pro}_i * \text{Dis}_j * \text{RT}_t : \text{roi}_r \sim \mathcal{N}(0,\ \boldsymbol{\Theta}),\ \text{RT}_{ijst} : \text{subj}_s \sim \mathcal{N}(0,\ \boldsymbol{\Lambda}),\ \text{RT}_{ijst} : \text{roi}_r\text{:subj}_s \sim \mathcal{N}(0,\ \boldsymbol{\Phi}),\\
&t = 1, 2, .., T_{ijs};\ r = 1, 2, ..., R;\ i = 1, 2;\ j = 1, 2; s = 1, 2, ..., S;
\end{aligned} \tag{17}$$

where $\boldsymbol{\Theta}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ are 8×8, 2×2 and 2×2 variance-covariance matrices for the respective effects across regions, subjects and the interactions between regions and subjects. As each quantitative variable requires two parameters

(a) Linearity of task effects with RT: TLM



(b) Linearity of task phase RT with cue effects: TLM



(c) Linearity of task effects with RT modulation

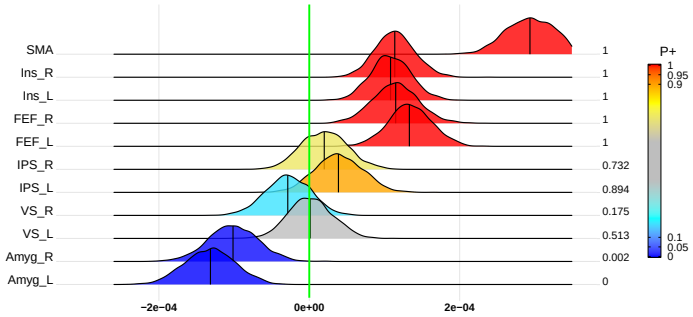


Figure 8: Linear association of task and cue effects with task phase RT at the population level. (a) Linear dependence of trial-level effects during the task phase on RT was assessed in the model (17). (b) Linear dependence of RT during the task phase on the trial-level effects during the cue phase was assessed in the model (18). (c) RT modulation effect during the task phase from the subject level was evaluated in the model (14).

(intercept and slope) in the model, $\text{Pro}_i * \text{Dis}_j * \text{RT}_{ijst}$ expands to 8 effects, leading to a 8×8 variance-covariance matrix $\boldsymbol{\Theta}$ for the cross-region variability. There was a strong indication of linearity between the overall task effects and RT in all the ROIs, except in the left and right ventral striatum (Fig. 8a). In addition, when RT was considered as a confounding variable, the interaction (`NoRew-Rew`):(`Neg-Neu`) showed compatible result (Fig. 6f) as its counterpart (Fig. 6d) from the model (13) without RT modulation.

The linear association between cue effects and subsequent task phase RT was also explored. In other words, how were trial-by-trial fluctuations during the cue phase related to task execution? To do so, the model `M0` in (5) was applied to the behavior data RT_{ist} as response variable and the cue effects x_{irst} as explanatory variable,

$$\begin{aligned}
&\mu = x_{irst} + x_{irst} : \text{roi}_r + x_{irst} : \text{subj}_s + x_{irst} : \text{roi}_r{:}\text{subj}_s; \\
&x_{irst} : \text{roi}_r \sim \mathcal{N}(0,\ \boldsymbol{\Theta}^2),\ x_{irst} : \text{subj}_s \sim \mathcal{N}(0,\ \boldsymbol{\Lambda}^2),\ x_{irst} : \text{roi}_r{:}\text{subj}_s \sim \mathcal{N}(0,\ \boldsymbol{\Phi}^2), \\
&s = 1, 2, ..., S;\ t = 1, 2, .., T_{i\cdot s};\ r = 1, 2, ..., R;\ i = 1, 2.
\end{aligned} \tag{18}$$

where $\boldsymbol{\Theta}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ are 2×2 variance-covariance matrices for the respective effects across regions, subjects and the interactions between regions and subjects. Evidence for linear association between the cue phase responses and subsequent behavior was very robust in the SMA, left FEF, and left/right insula (Fig. 8b).

How does the linearity assessed above compare to the conventional modulation nethod? To evaluate such scenario, we applied the `Mh` model (5) with the formulation (14) to the RT effects y_{ijs} for the four tasks and their standard error $\hat{\sigma}_{ijs}$ from the modulation analysis at the subject level. Compared to trial-level modeling (Fig. 8a), the RT effects based on modulation (Fig. 8c) showed very similar results.

Is linearity too strong an assumption even tbough frequently assumed in investigating the relationship between FMRI signals and covariates of interest (e.g., RT)? To address this question, we focused on one of the cue/task combinations, namelly `Rew_Neg`, involving reward cues and negative distractors. The trial-level effects y_{rst} and

(a) Linear fitting: `Rew_Neg` effects with corresponding RT (b) Smooth splines: `Rew_Neg` effects with corresponding RT

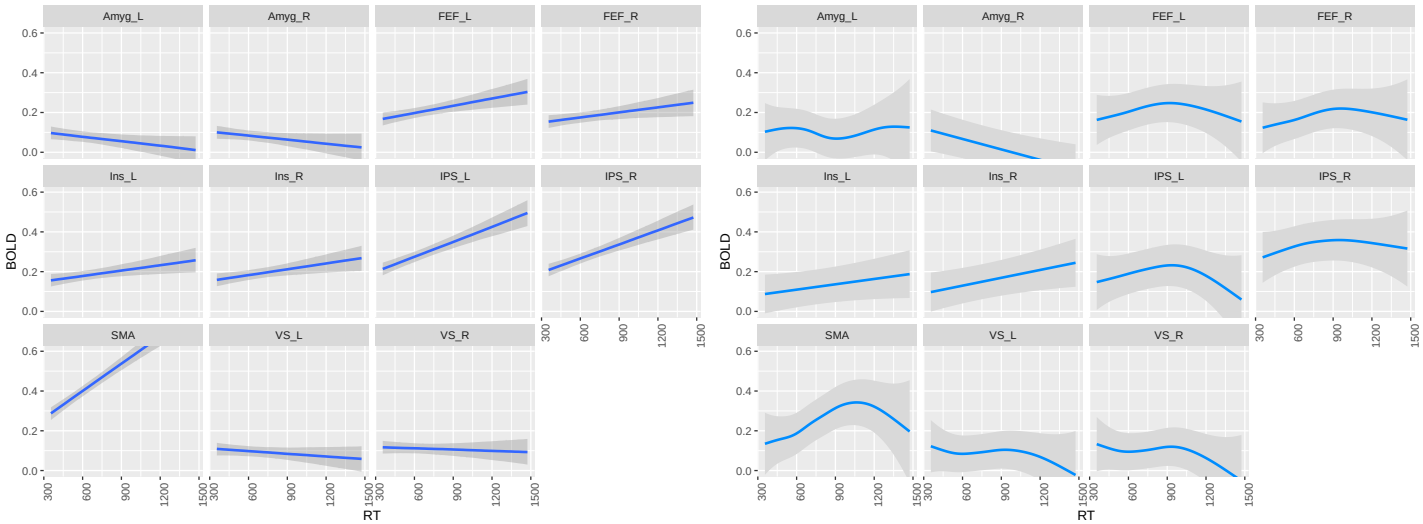


Figure 9: Comparisons of association analysis under the task `Rew_Neg` between linear fitting and smoothing splines. For better visualization on the dependence of trial-level effects on RT, the trends are shown with their 95% uncertainty bands. (a) Linear fitting was assessed in the model (19). (b) Association analysis was evaluated through smoothing splines in the model (20).

its variance $\hat{\sigma}^2_{rst}$ under the task `Rew_Neg` were fitted as follows,

$$\begin{aligned}
&y_{rst} \sim \mathcal{T}(\nu,\ \mu,\ \hat{\sigma}^2_{rst});\\
&\mu = \mathrm{RT}_{st} : \mathrm{ROI}_r + \mathrm{RT}_{st} : \mathrm{subj}_s + \mathrm{roi}_r{:}\mathrm{subj}_s;\\
&\mathrm{RT}_{st} : \mathrm{subj}_s \sim \mathcal{N}(0,\ \mathbf{\Lambda}),\ \mathrm{roi}_r{:}\mathrm{subj}_s \sim \mathcal{N}(0,\ \mathbf{\Phi});\\
&t = 1, 2, .., T_s;\ r = 1, 2, ..., R.\ s = 1, 2, ..., S.
\end{aligned} \tag{19}$$

Separately, a nonlinear function was applied to the RT via smoothing splines:

$$\begin{aligned}
&y_{rst} \sim \mathcal{T}(\nu,\ \mu,\ \hat{\sigma}^2_{rst});\\
&\mu = s(\mathrm{RT}_{st}) : \mathrm{ROI}_r + \mathrm{subj}_s + \mathrm{roi}_r{:}\mathrm{subj}_s;\\
&\mathrm{subj}_s \sim \mathcal{N}(0,\ \mathbf{\Lambda}),\ \mathrm{roi}_r{:}\mathrm{subj}_s \sim \mathcal{N}(0,\ \mathbf{\Phi});\\
&t = 1, 2, .., T_s;\ r = 1, 2, ..., R;\ s = 1, 2, ..., S;
\end{aligned} \tag{20}$$

where the smoothing function $s(\cdot)$ adopts a cubic spline basis defined by a set of knots spread evenly across the RT range and penalized by the conventional integrated square second-derivative cubic-spline term. The dimensionality of the basis expansion (i.e., number of knots) was automatically chosen through generalized cross-validation so that simplicity was balanced against explanatory power (Wood, 2017).

Fitting the data with linear and nonlinear models yielded some similarities, but multiple differences were also observed (Fig. 9). For example, linear fitting revealed positive trends at the SMA and the right/left IPS between task responses and the corresponding RT (Fig. 9a), whereas the spline fittings uncovered more complex relationships (Fig. 9b). The largely parallel trends observed across the contralateral region pairs provide some validation for both the linear and nonlinear fittings. Nevertheless, the nonlinear results suggest that linearity is likely too strong an assumption across the whole RT range; thus, the statistical evidence for linearity might have been inflated. For example, linearity might be applicable for certain RT ranges, but support for it might be limited at lower and higher values of RT with fewer data points. In addition, the uncertainty under the model assuming linearity (19) appears to have been considerably underestimated, especially away from the central RT values.

Discussion

Experimental sciences aim for generalizability. In doing so, they draw inferences about populations – idealized, theoretical constructs – from samples of, for example, trials and subjects. The ability to generalize is typically achieved through framing the cross-sample variability with an appropriate statistical distribution (e.g., Gaussian, Student-t). At the same time, when carefully choosing a stimulus set and subjects, an experimentalist obviously aims to draw conclusions that reach beyond the particular instances utilized. As recognized by Westfall et al. (2017), the standard analysis framework in neuroimaging does not lend itself to such a goal in the case of stimuli, and the same logic employed in bridging subjects to "population effects" is required. The present paper develops a two-level approach that is well adapted to the current analytical streamline; in addition, partial pooling is applied to trial-level effects at the population level to tackle the generalizability problem. We illustrated the effectiveness of the approach via a series of analysss of an FMRI dataset from a rich experimental paradigm with multi-phase trials, including cue stimuli and subsequent task execution.

Why should we more accurately account for trial-level variability?

The issue of cross-sample variability has been recognized for several decades across multiple research areas. For example, Clark (1973) pointed out that neglecting the problem of the "language-as-fixed-effect fallacy ... can lead to serious error", and even alluded to earlier warnings that were largely ignored in the literature, including the report by Coleman (1964). Given the common practice of aggregation across trials, even classical experiments such as Stroop and flanker tasks, which one would anticipate to show high reliability, yielded lackluster results (Rouder and Haaf, 2019).

In neuroimaging, trial-level variability is typically bundled together and flattened with the residuals of the GLS model at the subject level when condition-level effects are of interest. Such practice assumes that all trials have exactly the same BOLD response. In our analysis of an FMRI dataset, considerable variability was observed across trials (Figs. 4,12). Although the attribution of this variability to "pure noise" cannot be excluded, our results collectively point in a different direction: the variability is meaningful. Indeed, we interpret our results as suggesting that, without directly capturing trial-level effects, population-level estimates can be compromised. In particular, our results suggest that it is possible to underestimate effect magnitudes, while in some cases overestimating certainty (Fig. 6d,e; Fig. 7).

Ignoring cross-trial variability also leads to the loss of legitimately being able to generalize beyond the stimulus set used. As emphasized by Clark (1973), a serious implication is that studies will be "particularly vulnerable to lack of replicability". Generalization from a stimulus set to a category requires proper model construction (Coleman, 1964; Clark, 1973; Westfall et al., 2017; Yarkoni, 2019). When condition-level effects are inferred without accounting for cross-trial variability, technically speaking, the conclusions are applicable only to the particular trials in the experimental design, not even to similar cases from the same category. Trials are typically conceptualized as originating from a population that follows specific distributional assumptions (e.g., Gaussian), thus supporting the generalization from specific trials to the associated category. In addition, the explicit accountability of cross-trial variability provides more accurate characterization of effect uncertainty. Currently, modeling cross-subject variability is considered standard in the field as a way to draw population-level inferences. We believe the same should be considered for cross-trial variability.

Modeling cross-trial variability is important even if the difference in statistical evidence is small practically. First, although the fact that cross-trial variability diminishes as the sample size increases (see expression (4)), it is not practically possible to realistically determine the "required" number of trials. As the sample size in our experimental data was reasonably large (57 subjects and 28-47 trials per task), the differences in statistical evidence between trial- and condition-level modeling were not large (e.g., Fig. 6d,e). Nevertheless, meaningful differences were observed, such as the strength of the evidence for cue effects in the amygdala (Fig. 7). More importantly, condition-level modeling showed distortions in both the magnitude and uncertainty of effect estimates

(e.g., Fig. 6d,e). The importance of the increased sensitivity of the approach developed here will be particularly important when focusing on interaction effects, for which sample sizes need to be relatively large. Given that an important goal of neuroimaging studies is indeed to study interactions, we believe our approach offers a valuable gain to investigators. Furthermore, whereas investigators are generally cognizant of the need to employ enough subjects, awareness about requirements about trial sample size remains limited. We believe that considerations about trial sample size should be on a comparable footing as those of the number of participants.

Modeling trial-level responses is also of great value when studying brain-behavior relationships. A rich literature has investigated how trial-by-trial fluctuations in behavior are associated with intertrial variability (Ress et al., 2000; Pessoa et al., 2002; Pessoa and Padmala, 2005; Sapir et al., 2005; Lim et al., 2009). Many of these studies have proposed that the most likely source of the association is related to trial-by-trial changes in attention. Another potential source (Fox et al., 2006; Fox et al., 2007) is that instrinsic signal fluctuations account for much of intertrial variability in human behavior. Specifically, spontaneous fluctuations of the BOLD signal in resting-state studies also contributed to fluctuations during behavioral tasks. Evidences showed that ongoing intrinsic activity accounted for 60% of the variability in brain responses during a simple button-pressing task (Fox et al., 2007). Overall, although cross-trial variability has been framed in terms of issues of "fixed" versus "random" effects (Westfall et al., 2017), we believe it is of value to conceptualize the problem from a much broader perspective.

We also investigated the integrative LME modeling approach proposed by Westfall et al. (2017) from the following perspectives. (a) *AR structure.* Their model explicitly accounts for the serial correlation of the times series with lagged effects as explanatory variables, instead of capturing the AR structure in the residuals as typically practiced in the field. Such an approach remains controversial (e.g., Achen, 2000; Keele and Kelly, 2006; Bellemare et al., 2017; Wilkins, 2018), and had a dramatically large impact on the results with our dataset (Figs. 12b,13, Appendix B). In addition, it is likely inaccurate to assume that all the subjects share the same AR structure, as well as same baseline and cross-trial effects. (b) *Assumption of white noise.* Due to the violation of endogeneity (e.g., omitted variables, measurement error), the residuals in the integrative LME model (9) would be correlated with the lagged response variables. Accordingly, it would still be important to model the temporal structure in the residuals. (c) *Focus on statistical evidence.* Westfall et al. (2017) reported statistic values without accompanying effect estimates. We believe this practice, which is common in neuroimaging, is problematic because it follows a binary logic of "real" and "not true" effects; in addition, study reports without revealing effect magnitudes contribute to the reproducibility problem.

At the population level, what are the practical consequences of ignoring cross-trial variability? Westfall et al. (2017) made the strong cautionary warning that it could produce inflation 1.5 to 3 times of the values of the relevant statistic employed. The changes that we observed, important as they were, were less dramatic, and in some cases involved *deflation* of statistical evidence (e.g., cue effects, Fig. 7). Nevertheless, some substantial differences in effect magnitude and reliability were observed too (e.g., interactions between prospect and distractor, Fig. 6d,e). It is possible that differences between the present approach and that by Westfall et al. (2017) were partly influenced by some modeling choices including AR handling and other modeling assumptions. In this context, it is informative that we also observed a small amount of deflation in statistical evidence when data were aggregated across trials for the behavioral data (Fig. 3). In fact, our observations are more aligned with a similar assessment of test-retest reliability in psychometrics (Rouder and Haaf, 2019).

Benefits of two-level modeling approach

We propose a unified statistical platform that addresses the generalizeability issue through a two-level modeling approach (Table 1). Instead of adopting the conventional *complete pooling* (all trials essentially averaged), we directly estimate the trial-level effects at the subject level through *no pooling* (parameter estimates obtained for each trial separately). At the population level *partial pooling* is adopted via a hierarchical model to achieve generalizability from specific trials to condition category. Note that although we adopted a Bayesian platform here,

Table 1: Parallelism between cross-subject and cross-trial variability

Strategy	No Pooling	Partial Pooling	Complete Pooling	Association
Info. Sharing	none (individual)	some (adaptive)	full (uniform)	some (associative)
Subject Effect	unique for sth subject: $\xi_s \sim \mathcal{U}(-\infty, \infty)$ (`current practice`)	regularized: $\xi_s \sim \mathcal{N}(0, \lambda^2)$ (`current practice`)	same across subjects: $\xi_s = 0$ (`'fixed effects'`)	subject-level covariate (e.g., age)
Trial Effect	unique for tth trial: $\eta_t \sim \mathcal{U}(-\infty, \infty)$ (***proposed practice***)	regularized: $\eta_t \sim \mathcal{N}(0, \omega^2)$ (***proposed practice***)	same across trials: $\eta_t = 0$ (`current practice`)	trial-level behavior (e.g., RT)
Variance Prior	∞	finite (λ^2, ω^2)	0	-
Properties	unbiased but unreliable	biased but more accurate	homogenized	-
Applicability	prep. for population analysis; predictions; cross-region correlations	subjects to population, trials to category	-	controllability, correlation

the framework can also be implemented with an LME model, if one adopts a whole-brain voxel-wise approach and is not interested in partial pooling across ROIs. We now discuss a few strengths of our approach.

1) **Computational feasibility**. Our two-level approach attains computational tractability by segregating subject and population analyses. Importantly, at the same time, with subject-level uncertainty (standard errors) carried to the population level, any potential information loss is likely minimal relative to a "single-step" integrative method. The BML models can be implemented through the R package brms, which builds on top of the Stan language. At the same time, equivalent LME models can be performed at the voxel level through the program 3dLMEr publicly available in the AFNI suite.

2) **Flexibility and adaptivity**. The two-level approach can be adopted for several research objectives, including (a) condition-level effect estimation at the population level; (b) classification and machine learning applications utilizing trial-level estimates; (c) correlativity analysis based on trial-level effects; and (d) brain-behavior association. At a basic level, close examination and visualization of the trial-level effects become possible (e.g., synchrony between contralateral regions or among regions in a network).

3) **Outlier handling**. The possibility of outliers and data skew needs to be carefully considered in trial-by-trial analyses. The present framework flexibility accounts for these possibilities by (a) incorporating reliability information and (b) regularizing the estimation via, for example, Student's t-distribution.

4) **Modeling options**. Standard modulation analysis is able to investigate associations between behavioral variables and BOLD responses at the trial level. Our approach allows the evaluation of brain-behavior associations by assuming a linear relationship at the population level. In addition, the approach offers the investigator the opportunity to flexibly explore nonlinear relationships with smoothing splines.

Additional trial-level modeling issues

Our study also sheds insights about other aspects of FMRI data analysis.

1) **The importance of modeling AR structure in the residuals**. Our investigation on AR effects (Appendix B) confirmed a previous study (Olszowy et al., 2019) and indicated that the AR structure in the residuals varies substantially across regions, tasks and subjects. Therefore, we recommend that, to obtain reasonably accurate standard errors for effect estimates, a GLS model with the temporal structure in the residuals be accounted for with preferably AR(2) or ARMA(1, 1) for a TR around 2 s. With shorter TRs, a higher-order AR structure would be likely needed (Olszowy et al., 2019; Luo et al., 2020).

2) **Incorporating effect uncertainty in population analysis**. Should standard errors of effect estimates be modeled at the population level? Previous studies suggested that the benefit was minimal (Mumford et al., 2009; Chen et al., 2012; Olszowy et al., 2019). However, since trial-level modeling is more prone to multicollinearity and may result in unreliable effect estimates, uncertainty information provides a robust mechanism to counter the impact of outliers at the population level. As the accountability of the serial correlation in the residuals of the time series regression model is influential on the accuracy of the standard error for each effect estimate, it

becomes important to more accurately model the AR structure at the subject level.

3) **Multicollinearity**. As each trial is estimated as a separate regressor, careful experimental design and trial-order optimization must be carefully considered. For analyses based on behavioral performance which cannot be optimized in advance, particular attention should be paid to multicollinearty. However, given the overall two-stage estimation procedure, mutlicollnearity may pose less of a problem than typically assumed, although it will likely affect the precision of the estimated effects.

Conclusions

In the present study, we investigated the extent and impact of trial-by-trial responses in FMRI data. Several modeling strategies were developed and evaluated. At the trial level, responses were estimated through a GLS model with serial correlations accounted for, whereas population-level analysis was carried via a hierarchical model that effectively characterized effect structure, allowing generalizability from the specific stimuli employed to the generic category. Additional applications of the approach employed here include the analysis of brain-behavior associations, trial-based correlation analysis, as well as trial-level classification and machine learning.

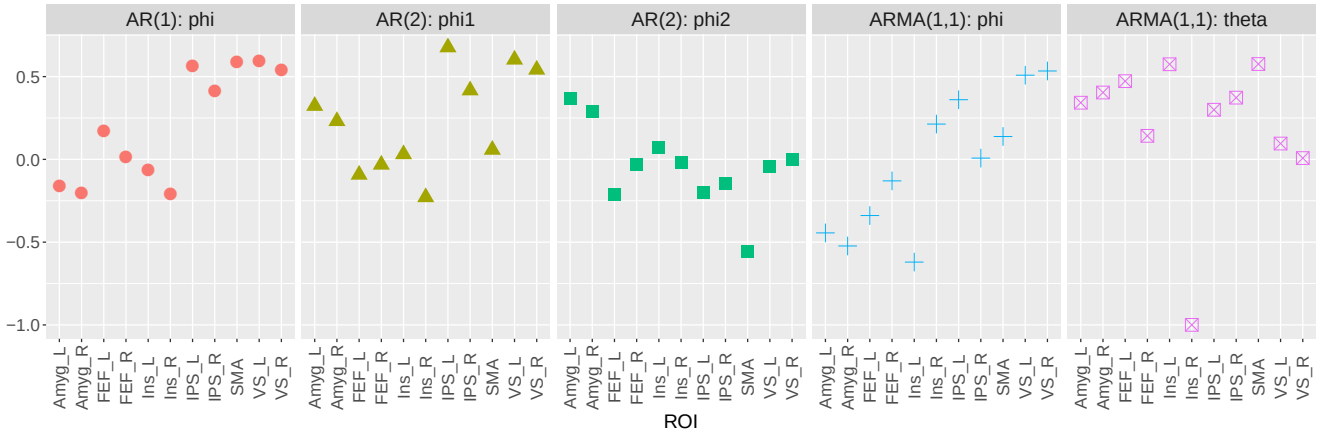
Acknowledgments

The research and writing of the paper were supported (GC, PAT, and RWC) by the NIMH and NINDS Intramural Research Programs (ZICMH002888) of the NIH/HHS, USA. LP's research was supported in part by NIMH (R01 MH071589 and R01 MH112517). We are appreciative of the technical support from the Stan (Carpenter et al., 2017) and R (R Core Team, 2019) communities. Most of the modeling work was performed in Stan through the R package brms (Bürkner, 2018), and the figures were generated with the R package ggplot2 (Wickham, 2009). We thank Kelly Morrow, Chirag Limbachia and Anastasiia Khibovska for assistance in generating some of the figures, and Kelly Morrow for help in creating the regions of interest.

Appendix A. Hyperpriors adopted for BML modeling

The prior distribution for all the lower-level (e.g., trial, ROI, subject) effects considered here is Gaussian, as specified in the respective model; for example, see the distribution assumptions in the BML model (6). If justified, one could adopt other priors like Student's t for the effects across trials, regions and subjects, just as for the likelihood (or the prior for the response variable y in the BML model (5)). In addition, prior distributions (usually called hyperpriors) are needed for three types of model parameters in each model: (a) population effects or location parameters ("fixed effects" under LME, such as intercept and slopes), (b) standard deviations or scaling parameters for lower-level effects ("random effects" under LME), and (c) various parameters such as the covariances in a variance-covariance matrix and the degrees of freedom in Student's t-distribution. Noninformative hyperpriors are adopted for population effects (e.g., population-level intercept and slopes). In contrast, weakly-informative priors are utilized for standard deviations of lower-level parameters such as varying slope, subject-, trial- and region-level effects, and such hyperpriors include a Student's half-$t(3, 0, 1)$ or a half-Gaussian $\mathcal{N}_+(0, 1)$ (a Gaussian distribution with restriction to the positive side of the respective distribution). For variance-covariance matrices, the LKJ correlation prior (Lewandowski, Kurowicka, and Joe, 2009) is used with the shape parameter taking the value of 1 (i.e., jointly uniform over all correlation matrices of the respective dimension). Lastly, the standard deviation σ for the residuals utilizes a half Cauchy prior with a scale parameter depending on the standard deviation of the input data. The hyperprior for the degrees of freedom, ν, of the Student's t-distribution is $\Gamma(2, 0.1)$. The consistency and full convergence of the Markov chains were confirmed through the split statistic $\hat{R}$ being less than 1.1 (Gelman et al., 2013). The effective sample size (or the number of independent draws)

(a) Serial correlation estimation across 11 ROIs of one subject



(b) Serial correlation estimation across 57 subjects at VS_L

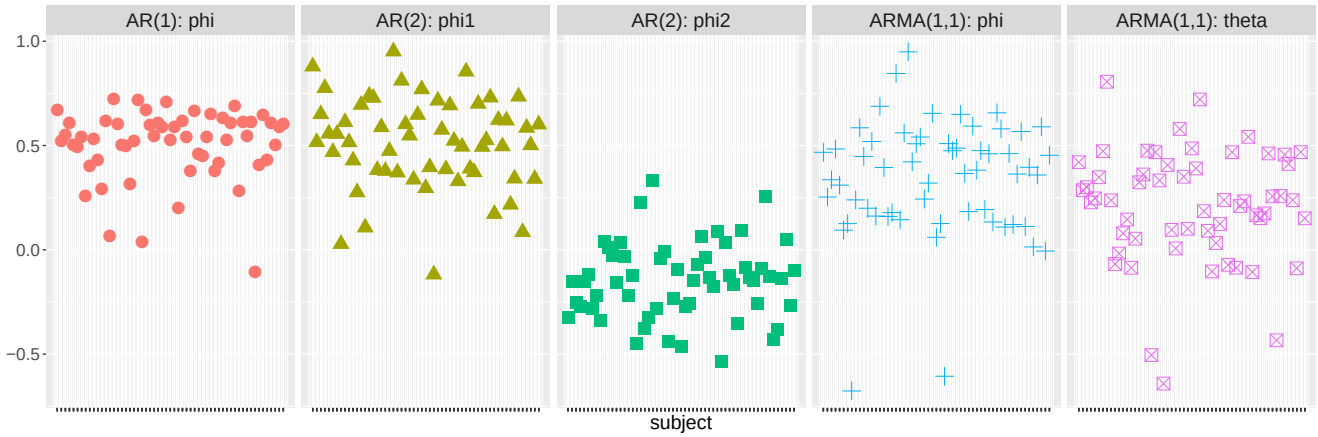


Figure 10: Variations of temporal correlation across regions and subjects. The overall average first-order AR parameter of trial-level modeling across all the 11 ROIs and 57 subjects was 0.50 $\pm$0.20, 0.47 $\pm$0.28 and 0.33 $\pm$0.38 for AR(1), AR(2) and ARMA(1,1), respectively; the second-order parameter for AR(2) and moving average parameter for ARMA(1,1) were -0.13 ∓ -0.17 and 0.18 $\pm$0.34, respectively. The relative magnitude of these AR parameters indicated that the first AR parameter captured substantially large proportion of the serial correlation while the second parameter in AR(2) and ARMA(1,1) remained helpful.

from the posterior distributions based on Markov chain Monte Carlo simulations was more than 200 so that the quantile (or compatibility) intervals of the posterior distributions could be estimated with reasonable accuracy.

Appendix B. Handling autocorrelation in FMRI data

The amount of temporal correlation embedded in the residuals of the time series regression with trial-level modeling was substantial with large variations across regions, tasks and subjects (Fig. 10). Specifically, the overall serial correlation across the 11 ROIs and 57 subjects was 0.50 $\pm$0.20, 0.47 $\pm$0.28 and 0.33 $\pm$0.38 assessed from the AR(1), AR(2) and ARMA(1,1) models, respectively), indicating that some large amount of effects were not properly accounted for through the explanatory variables. With condition-level modeling, cross-trial fluctuations would become part of the residuals; thus, the AR effects would be different and likely stronger,

The performances of the OLS approach were compromised due to the presence of persistent temporal correlation in the model residuals. Based on the Gauss-Markov theorem, the OLS method would still provide consistently unbiased estimates, with the caveat that the precision for the effect estimates tends to be inflated. However, the asymptotic property of the unbiasedness heavily relies on a large sample size, which cannot necessarily be met nor easily predetermined in real practice. With the current dataset, the OLS solutions showed some extent of over- and under-estimation compared to the three AR models (Fig. 11). In addition, a slight amount of underestimated uncertainty (or inflated precision) about the OLS effect estimates is evident compared to their AR counterparts (Fig. 11).

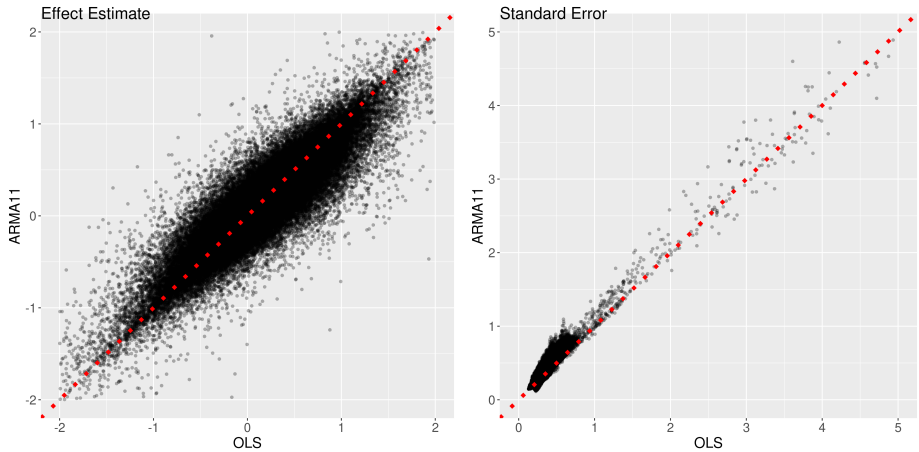
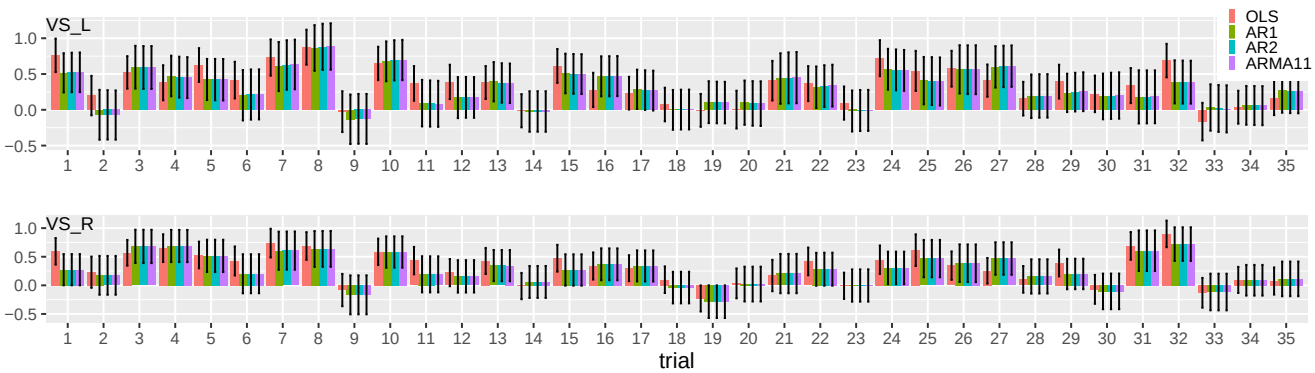


Figure 11: Comparisons of OLS and ARMA(1,1) in effect estimate and uncertainty. The effect estimates (left) and their standard errors (right) are shown for the total $2 \times 11 \times \sum_{s=1}^{57} \sum_{i=1}^{2} \sum_{j=1}^{2} T_{ijs} = 200640$ trial-level effects among the two cues and four tasks. The theoretical unbiasedness of OLS estimates can be verified by the roughly equal number of data points on the two sides of the diagonal line (dotted red). However, the instability of OLS estimation is shown by the fat cloud surrounding the diagonal line: slightly overestimation (or underestimation) of OLS was shown by 52.7% (or 45.5%) of data points above (or below) the x-axis. The precision inflation of OLS can be assessed by the proportion of data points (97.5%) above the dotted red line.

(a) Comparisons of effect estimates for `NoRew_Neg` among 4 AR models



(b) Comparison of effect estimates for `NoRew_Neg` between residual AR(2) and response AR(2)

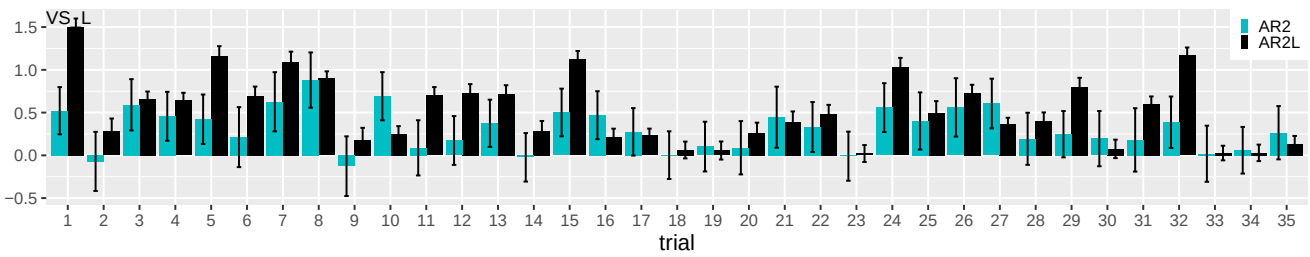


Figure 12: Trial-level effects under the task `NoRew_Neg` from one subject. (a) Effect estimates are shown at two contralateral regions, left (upper row) and right (lower row) ventral striatum. Black segments indicate one standard error, and the colors code the four different AR models (OLS, AR(1), AR(2) and ARMA(1,1)) for the residuals in the GLS model (1). Only 35 trials (out of 48) were successfully completed by the subject. Despite substantial amount of cross-trial variability, some consistent extent of synchronization was revealed across all the four models and all the five contralateral region pairs (only one pair shown here). (b) Effect estimates (AR2L, black) at left ventral striatum were obtained with AR effects modeled as second-order lagged effects of the EPI time series in the model (21) as implemented in Westfall et al. (2017). The same AR(2) results from (a) are shown (AR2, iris blue) as a comparison. The impact of incorporating lagged effects in the model was quite evident with both effect estimates and their precision substantially higher at some trials.

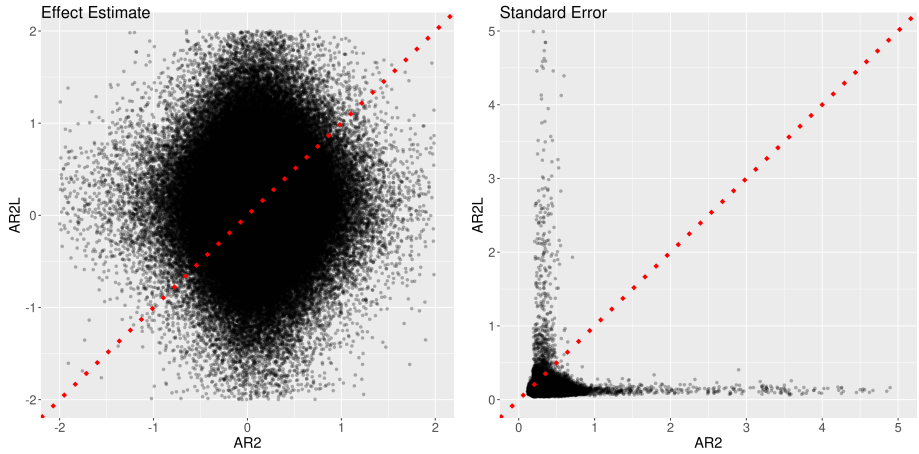


Figure 13: Comparisons of two approaches in AR handling. Two models were adopted to fit the data at the 11 ROIs, one (x-axis: AR2) with the GLS model (1) plus an AR(2) structure and the other (y-axis: AR2L) with the model (21) that mimicked the approach by Westfall et al. (2017). The effect estimates (left) and their standard errors (right) are shown for the total $2 \times 11 \times \sum_{s=1}^{57} \sum_{i=1}^{2} \sum_{j=1}^{2} T_{ijs} = 200640$ trial-level effects among the two cues and four tasks. The substantial amount of deviation of the effect estimates from the diagonal line (dotted red) indicates the dramatic differences between the two models. The precision underestimation of the model with lagged effects (AR2L) can be assessed by the proportion of data points (98.3%) below the dotted red line.

Among the three AR models, both the AR(2) and ARMA(1,1) models slightly edged out AR(1) due to the extra accountability from the second AR parameter. While a large amount of autocorrelation was explained through the first-order parameters among the three AR models (first, second and fourth columns, Fig. 12), the second parameter for AR(2) and ARMA(1,1) provided less but still sizeable amount of autocorrelation accountability (third and fifth columns, Fig. 12). In light of the observations that both the AR(2) and ARMA(1,1) results were hardly differentiable (Fig. 12), we opted to adopt the ARMA(1,1) model in the current study.

What if the serial correlation is directly modeled as delayed effects of the EPI time series, as adopted in Westfall et al. (2017)? To explore the impact of such approach, we analyzed the EPI data of the 11 ROIs at the subject level with the following model,

$$\begin{aligned} & y_k = \phi_1 y_{k-1} + \phi_2 y_{k-2} + \alpha_0 + \alpha_1 z_{1k} + ... + \alpha_m z_{mk} + \beta_1 x_{1k} + \beta_2 x_{2k} + ... + \beta_n x_{nk} + \epsilon_k; \\ & \epsilon_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2); k = 1, 2, ..., K. \end{aligned} \tag{21}$$

where k indexes the time points, ϕ_1 and ϕ_2 are the first- and second-order AR parameters for the lagged effects of the EPI signal y_k. As both y_{k-1} and y_{k-2} are largely correlated with all the regressors, the impact on effect estimates was substantially evident across subjects, regions, conditions and trials (Figs. 12b,13). In addition to a varying amount of increase on some effect estimates, the uncertainty (standard error) was quite smaller for most effect estimates.

References

Achen, C. H. (2001). Why lagged dependent variables can suppress the explanatory power of other independent variables. Annual Meeting of the Political Methodology Section of the American Political Science Association, UCLA, July 20-22, 2000.

Amrhein, V., Greenland, S., McShane, B., 2019. Scientists rise up against statistical significance. Nature 567, 305-307.

Bellemare, M.F., Masaki, T., Pepinsky, T.B., (2017). Lagged Explanatory Variables and the Estimation of Causal Effect. The Journal of Politics 79(3):949-963.

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal 10(1):395-411.

Bullmore, E., Brammer, M., Williams, S.C., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., Sham, P. (1996). Statistical methods of estimation and inference for functional MR image analysis. Magn. Reson. Med. 35:261-277.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A. (2017). Stan: A probabilistic programming language. Journal of Statistical Software 76(1).

Chen, G., Saad, Z.S., Nath, A.R., Beauchamp, M.S., Cox, R.W. (2012). FMRI Group analysis combining effect estimates and their variances. NeuroImage 60:747-765.

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

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

Chen, G., Xiao, Y., Taylor, P.A., Riggins, T., Geng, F., Redcay, E., 2019a. Handling Multiplicity in Neuroimaging through Bayesian Lenses with Multilevel Modeling. Neuroinformatics 17(4):515-545.

Chen, G., Taylor, P.A., Cox, R.W., Pessoa, L., 2019b. Fighting or embracing multiplicity in neuroimaging? neighborhood leverage versus global calibration. NeuroImage (in press).
doi: 10.1016/j.neuroimage.2019.116320

Clark, H.H., 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior 12(4):335-359.

Coleman, E.B. (1964). Generalizing to a language population. Psychological Reports 14(1):219-226.

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

Faillenot, I., Heckemann, R.A., Frot, M., Hammers, A. (2017). Macroanatomy and 3D probabilistic atlas of the human insula. NeuroImage 150:88-98.

Fox, M.D., Snyder, A.Z., Zacks, J.M., Raichle, M.E. (2006). Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. Nature neuroscience 9(1):23-25.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Raichle, M. E. (2007). Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior. Neuron 56(1):171-184.

Friston, K. J., Glaser, D.E., Henson, R.N., Kiebel, S., Phillips, C., Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: applications. Neuroimage 16:484-512.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). Bayesian Data Analysis. Chapman & HallCRC Press, London, third edition.

Keele, L., Kelly, L.J. (2006). Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables. Political Analysis Volume 14(2):186-205.

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

Lim, S.L., Padmala, S., Pessoa, L. (2009). Segregating the significant from the mundane on a moment-to-moment basis via direct and indirect amygdala contributions. Proceedings of the National Academy of Sciences 106(39):16841-16846.

Luo, Q., Misaki, M., Mulyana, B., Wong, C.-K., Bodurka, J. (2020). Improved autoregressive model for correction of noise serial correlation in fast fMRI. Magn Reson Med. doi: 10.1002/mrm.28203

McElreath R., 2016, Statistical Rethinking: A Bayesian Course with Examples in R and Stan. Chapman & Hall/CRC Press.

McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L., 2019. Abandon Statistical Significance. The American Statistician 73:sup1, 235-245.

Mumford, J.A., Nichols, T.E. (2009). Simple group fMRI modeling and inference. NeuroImage 47(4):1469-1475.

Nacewicz, B.M., Alexander, A.L., Kalin, N.H., and Davidson, R.J. (2014). The neurochemical underpinnings

of human amygdala volume including subregional contributions. In: Annual meeting of the Society of Biological Psychiatry (New York, NY).

Olszowy, W., Aston, J., Rua, C., Williams, G.B. (2019). Accurate autocorrelation modeling substantially improves fMRI reliability. Nat Commun 10, 1220.

Padmala, S., Sirbu, M., Pessoa, L. (2017). Potential reward reduces the adverse impact of negative distractor stimuli. Soc Cogn Affect Neurosci. 12(9):1402-1413.

Palmer, E.M., Horowitz, T.S., Torralba, A., Wolfe, J.M. (2011). What are the shapes of response time distributions in visual search? J Exp Psychol 37(1):58-71.

Pauli, W.M., O'Reilly, R.C., Yarkoni, T., Wager, T.D. (2016). Regional specialization within the human striatum for diverse psychological functions. Proceedings of the National Academy of Sciences 113(7):1907-1912.

Pessoa, L., Gutierrez, E., Bandettini, P., Ungerleider, L. (2002). Neural correlates of visual working memory: fMRI amplitude predicts task performance. Neuron 35(5):975-987.

Pessoa, L., Padmala, S. (2007). Decoding near-threshold perception of fear from distributed single-trial brain activation. Cerebral cortex 17(3):691-701.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ress, D., Backus, B.T., and Heeger, D.J. (2000). Activity in primary visual cortex predicts performance in a visual detection task. Nat. Neurosci. 3:940-945.

Rissman, J., Gazzaley, A., D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage 23(2):752-63.

Rouder, J.N., Haaf, J.M. (2019). A psychometrics of individual differences in experimental tasks. Psychonomic Bulletin & Review 26:452-467.

Sapir, A., D'Avossa, G., McAvoy, M., Shulman, G.I., and Corbetta, M. (2005). BOLD signals for spatial attention predict performance in a motion discrimination task. Proc. Natl. Acad. Sci. USA 102:17810-17815.

Toro, R., Fox, P. T., Paus, T. (2008). Functional coactivation map of the human brain. Cerebral cortex 18(11):2553-2559.

Vehtari, A., Gelman, A., Gabry J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. In Statistics and Computing 27:1413-1432.

Westfall, J., Nichols, T.E., Yarkoni, T. (2017). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. Wellcome Open Research 1:23.

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

Wilkins, A.S., 2018. To Lag or Not to Lag?: Re-Evaluating the Use of Lagged Dependent Variables in Regression Analysis. Political Science Research and Methods 6(2):393-411.

Wood, S. N. (2017). Generalized Additive Models: An Introduction with R (2nd ed). Chapman & HallCRC.

Woolrich, M. W., Ripley, B. D., Brady, M., Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14:1370-1386.

Woolrich, M.W., Behrens, T.E.J., Beckmann, C.F., Jenkinson, M., Smith, S.M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. NeuroImage 21(4):17320-1747.

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.

Yarkoni, Tal. (2019). The Generalizability Crisis. PsyArXiv. November 22. doi:10.31234/osf.io/jqw35.