METHOD ARTICLE

Fixing the stimulus-as-fixed-effect fallacy in task fMRI [version 1; peer review: 1 not approved]

Jacob Westfall[1]*, Thomas E. Nichols[2], Tal Yarkoni[1]*

[1]Department of Psychology, University of Texas, Austin, USA
[2]Department of Statistics & WMG, University of Warwick, Coventry, UK

* Equal contributors

v1 **First published:** 09 Dec 2016, **1**:23
https://doi.org/10.12688/wellcomeopenres.10298.1
Latest published: 17 Mar 2017, **1**:23
https://doi.org/10.12688/wellcomeopenres.10298.2

Abstract

Most functional magnetic resonance imaging (fMRI) experiments record the brain's responses to samples of stimulus materials (e.g., faces or words). Yet the statistical modeling approaches used in fMRI research universally fail to model stimulus variability in a manner that affords population generalization, meaning that researchers' conclusions technically apply only to the precise stimuli used in each study, and cannot be generalized to new stimuli. A direct consequence of this *stimulus-as-fixed-effect fallacy* is that the majority of published fMRI studies have likely overstated the strength of the statistical evidence they report. Here we develop a Bayesian mixed model (the random stimulus model; RSM) that addresses this problem, and apply it to a range of fMRI datasets. Results demonstrate considerable inflation (50-200% in most of the studied datasets) of test statistics obtained from standard "summary statistics"-based approaches relative to the corresponding RSM models. We demonstrate how RSMs can be used to improve parameter estimates, properly control false positive rates, and test novel research hypotheses about stimulus-level variability in human brain responses.

Keywords

experimental design, statistical modeling, Bayesian modeling, functional magnetic resonance imaging, mixed-effect modeling, stimulus-as-fixed-effect fallacy

Open Peer Review

Approval Status ✓ ? ?

	1	2	3
version 2 (revision) 17 Mar 2017	✓ view	? view	? view
version 1 09 Dec 2016	✗ view		

1. **Sophie Donnet**, University of Paris-Saclay, Paris, France
2. **Robert Leech**, Imperial College London, London, UK
   Romy Lorenz, Imperial College London, London, UK
3. **Michael B. Miller**, University of California Santa Barbara, Santa Barbara, USA
   Benjamin O. Turner, University of California Santa Barbara, Santa Barbara, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Tal Yarkoni (tyarkoni@utexas.edu)

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Wellcome Trust [100309]; and the National Institutes of Health [R01MH096906].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Westfall J, Nichols TE and Yarkoni T. **Fixing the stimulus-as-fixed-effect fallacy in task fMRI [version 1; peer review: 1 not approved]** Wellcome Open Research 2016, **1**:23 https://doi.org/10.12688/wellcomeopenres.10298.1

First published: 09 Dec 2016, **1**:23 https://doi.org/10.12688/wellcomeopenres.10298.1

Introduction

Consider two potential titles of a hypothetical neuroimaging paper: (1) "Inferior frontal gyrus responds more strongly to the words 'chair,' 'house,' and 'tree' than to 'run,' 'pay,' and 'speak'"; and (2) "Inferior frontal gyrus responds more strongly to nouns than to verbs". These two titles may superficially appear to describe exactly the same set of findings, since categories like Noun and Verb are necessarily comprised entirely of individual exemplars like 'chair' and 'speak'. Yet there can be little doubt that most neuroimaging researchers forced to choose between the two titles above would opt for the latter, which makes a far more interesting scientific statement. After all, we typically do not care about individual words like 'chair', except insofar as they exemplify broader populations of items that share similar properties. As the psycholinguist Edmund Coleman observed over 50 years ago, "many studies of verbal behavior have little scientific point if their conclusions have to be restricted to the specific language materials that were used in the experiment" (Coleman, 1964; cf. Clark, 1973). The same is no doubt true of the stimuli used in modern neuroimaging studies.

Choosing between hypothetical titles like those above may seem purely a matter of preference--a researcher simply decides that she cares more about the underlying population than about the individual stimuli, and can then proceed to describe her results as such. But the conceptual move from stimulus-level to population-level inference is not automatically justified. It must be explicitly supported by appropriate statistical inference. In studies where a sample of participants respond to a sample of stimuli, as is the case in a vast number of fMRI studies, the correct analysis that allows generalization to both participant and stimulus populations involves fitting a mixed-effects model with crossed random effects of both participants and stimuli (Baayen *et al.*, 2008; Judd *et al.*, 2012). This is not a hypothetical concern: in a review of 100 random task-based fMRI articles extracted from the Neurosynth database (see Online Methods for details), we found that 63/100 (95% Jeffreys interval = [53%, 72%]) used multiple stimuli in a context where generalization over stimuli was clearly indicated. Yet while virtually every fMRI study conducted over the past 15 years has modeled subject as a random factor (Penny *et al.*, 2003), we are aware of only two published fMRI studies that have discussed the problem of unmodeled stimulus-related variance from a methodological perspective (Bedny *et al.*, 2007; Donnet *et al.*, 2006), and know of no primary fMRI studies that have modeled participants and stimuli as crossed random factors.

The consequences of this stimulus-as-fixed-effect fallacy, as it is called in psycholinguistics (Clark, 1973), are potentially devastating. Strictly speaking, the *p*-values (or other inferential statistics) reported in the entire fMRI literature to date are valid only for the exact stimuli used in each study. The conclusions cannot be generalized to a broader population of stimuli without risking inflated Type I error (cf. Donnet *et al.*, 2006). Previous work in other domains (and replicated here) demonstrates that this inflation can be dramatic, with the Type 1 error rate frequently exceeding 50% under realistic conditions (Judd *et al.*, 2012; Wickens & Keppel, 1983).

Here we develop a Bayesian mixed model (the random stimulus model; RSM) that directly estimates the degree of stimulus variability in fMRI data and properly adjusts the key parameter estimates to account for uncertainty due to stimulus sampling. We then apply this model to a variety of real fMRI datasets with diverse stimulus samples and experimental designs, comparing the results from the standard statistical model that ignores stimulus variability to the results from the corresponding RSM. Our findings suggest that an unknown but possibly large fraction of published fMRI findings are likely to be false positives driven by unmodeled stimulus-level variability. We demonstrate that the magnitude of the problem can be considerably ameliorated by employing large stimulus samples and/or presenting different subjects with different stimuli -- in fact, in the limiting case where every participant receives a completely unique stimulus set, the standard model is the statistically appropriate model. Finally, we show how the stimulus-level parameter estimates produced by RSMs can be used to generate and test novel research hypotheses, opening up a powerful new method for studying the neural substrates of cognition.

Methods

fMRI datasets

All analyses used publicly available data obtained from one of three sources. The HCP analyses used task fMRI data from the Human Connectome Project's "100 unrelated subject" release, accessible via the online Connectome Workbench (http://www.humanconnectome.org). These data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. All HCP analyses used the preprocessed data release (https://db.humanconnectome.org/data/projects/HCP_900). No further processing of the time series was performed prior to region-based averaging and mixed-effects modeling (see below). Methods have been previously described in detail in (Glasser *et al.*, 2013).

The emotion regulation dataset was obtained from OpenfMRI, and is publicly available (accession number, ds000009; https://openfmri.org/dataset/ds000009/). Note that although the full dataset includes n = 24 subjects, only 11 subjects had preprocessed data available. We therefore conducted analyses with the convenience n = 11 sample, since subject sample size was irrelevant for our purposes. Experimental design and preprocessing procedures for this dataset have been previously described in Cohen (2009; http://gradworks.umi.com/34/01/3401764.html).

The IAPS dataset previously used in Chang *et al.* (2015) was obtained from the NeuroVault whole-brain image repository (Gorgolewski *et al.*, 2015). Images were downloaded via the NeuroVault API from the corresponding image collection (http://neurovault.org/collections/503/). The dataset contains 30 trial-level estimates for each of 172 participants (30 images in total). On each trial, participants passively viewed either a negative or a neutral IAPS image (15 of each). All methods have been previously described in Chang *et al.* (2015).

Statistical modeling

The standard model. Consider a hypothetical fMRI experiment in which participants view 20 stimuli, half belonging to one stimulus category and half to another (as in Figure 1), and we are interested in the difference in neural response between these two categories. Let Y_{it} be the neural response of the ith subject at the tth time point in a particular voxel or region of interest (ROI), with all preprocessing already carried out and with each participant's data separately standardized. The standard statistical model used to analyze such data is:

$$Y_{it} = \beta_0 + (\beta_1 + p_{1i})X_{1it} + (\beta_2 + p_{2i})X_{2it} + e_{it},$$

where X_1 and X_2 are the regressors representing the idealized neural responses toward the two categories of stimuli; β_0 is the fixed intercept; β_1 and β_2 are the fixed effects of the two neural regressors; p_1 and p_2 are normally distributed participant effects of the neural regressors, representing stable subject-to-subject variability in the degree of neural response toward both stimulus types; and the e terms are normally distributed, observation-level errors with an AR(2) covariance structure. The regressors X_1 and X_2 are formed by summing over the idealized neural responses toward the individual stimuli comprising each stimulus category,

$$X_{1it} = \sum_{j=1}^{10} x_{ijt} \text{ and } X_{2it} = \sum_{j=11}^{20} x_{ijt}.$$

where x_{ijt} is the idealized neural response of the ith participant for the jth stimulus at time t. These idealized neural responses

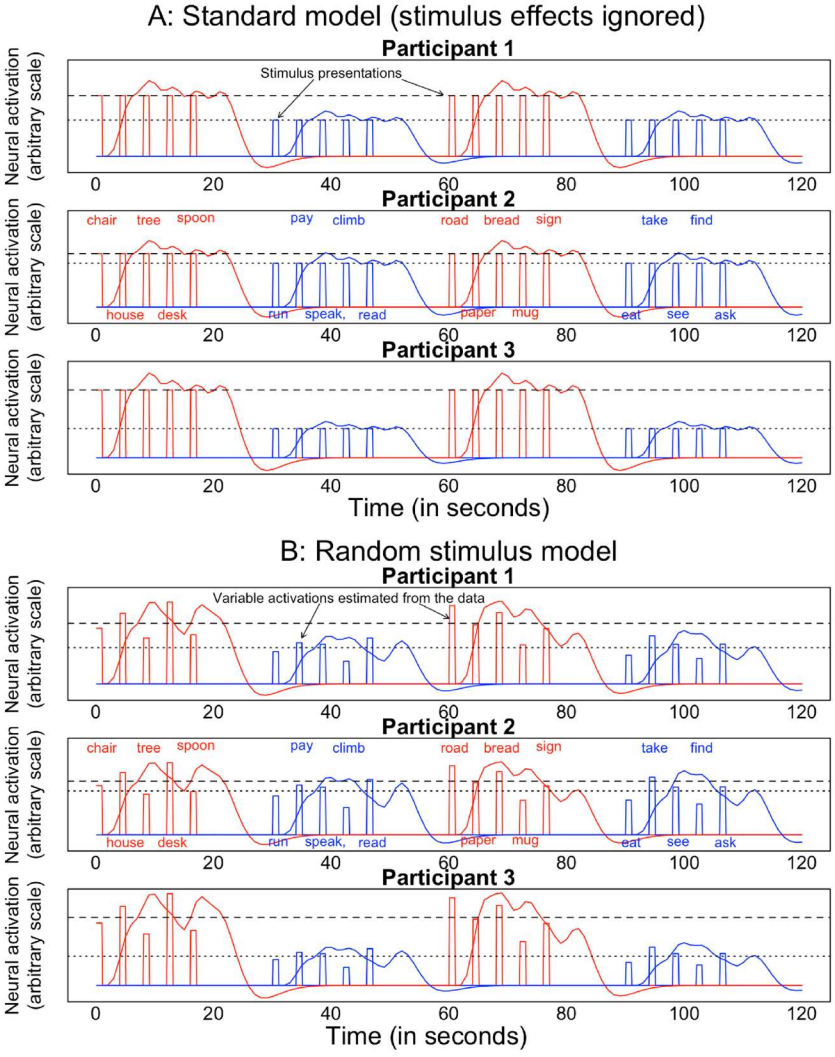

Figure 1. Idealized data from the standard model and the random stimulus model (RSM) for a hypothetical experiment. Stimuli belonging to one of two conditions (nouns in red, verbs in blue) are presented in alternating blocks, with the same stimuli presented in the same order to each participant. Both models incorporate subject-level variability in the magnitude of the category difference. For example, Participant 2 shows a small category difference while Participant 3 shows a large category difference. In the standard model (**A**), neural responses are assumed to be equal in magnitude for all stimuli in a category. The RSM (**B**) relaxes this assumption and estimates a separate response for each stimulus. For example, the noun "desk" elicits consistently high responses for all participants, while the noun "spoon" elicits consistently low responses for all participants.

are based on convolving a stimulus presentation sequence with a hemodynamic response function (Poldrack *et al.*, 2011).

Random stimulus model (RSM). The standard model posits that the stimulus-level regressors are all identical in amplitude, differing only in their presentation times (Figure 1A), a dubious assumption in most fMRI studies. In the RSM, we relax this assumption and allow the stimulus-level regressors to have distributions of amplitudes that are to be estimated from the data (Figure 1B); these amplitudes are common over subjects, but vary randomly per stimuli. To achieve this, we add a set of terms $s_j x_{ijt}$ to the model, where the s_j are normally distributed stimulus effects, representing stable stimulus-to-stimulus variability in the strength of the neural response. The resulting model is:

$$Y_{it} = \beta_0 + (\beta_1 + p_{1i})X_{1it} + (\beta_2 + p_{2i})X_{2it} + \sum_{j=1}^{20} s_j x_{ijt} + e_{it}.$$

This model cannot be fit using standard mixed modeling statistical packages, such as lme4 in R or SAS PROC MIXED, because these packages assume that each row of the dataset is associated with one and only one level of each random factor, i.e., a single participant and a single stimulus. (Though standard software can fit a slightly simplified, approximate version of this model, quite similar to what (Gelman & Hill, 2007) refer to as a "no-pooling model;" see Supplementary File 1 for an application of this approximate model to the (Chang *et al.*, 2015)). However, for the RSM, the measurements at each time point are influenced not by a single random stimulus effect (s_j), but rather by *all* of the random stimulus effects ($\Sigma_j s_j$). Despite this complication, it is relatively straightforward to fit the RSM as a Bayesian model using a probabilistic programming framework, such as BUGS, JAGS, or Stan. For the models in this paper, we used the PyMC3 Python package (Patil *et al.*, 2010; Salvatier *et al.*, 2015), which is built on the Theano deep learning package (Bastien *et al.*, 2012; Bergstra *et al.*, 2010; version 0.9.0.dev2) and implements the state-of-the-art No U-Turn MCMC Sampler (Hoffman & Gelman, 2014). An alpha version of our *NiPyMC* analysis package is available online (https://github.com/PsychoinformaticsLab/nipymc; DOI, 10.5281/zenodo.168087; Yarkoni & Westfall, 2016). In Supplementary File 1 we give the full statistical details of the specific models we estimated in our reanalyses, including the precise distributional assumptions and the variations of the basic model that we applied to each individual dataset.

We note that there are various specification options that could be applied to the standard model and RSM described here and in Supplementary File 1, for example, different choices of HRF, autocorrelation parameters, motion correction, image realignment, and so on. While such options can certainly impact overall data quality and test statistics (cf. Carp, 2012) they are extremely unlikely to affect the central conclusions supported by the present results. To exert a non-negligible impact on our results, these specification options would need to have very different impacts on the standard model and RSM (otherwise the extensions would simply lead to the test statistics from both models increasing or decreasing more or less in unison, leaving their relative differences essentially unchanged). We are aware of no a priori reasons to expect this to be the case for any of the methodological procedures employed with any frequency in the literature, and reiterate that comparably large decreases in test statistics have been repeatedly observed in other domains of psychology when including random stimulus effects (Judd *et al.*, 2012; Wolsiefer *et al.*, 2016).

Simulations

We conducted an extensive series of simulations in order to validate and to better understand the properties of our proposed RSM. Our first goal was to verify that the RSM could adequately recover true parameter values. Our second goal was to identify the conditions under which using a RSM produces the greatest attenuation of test statistics compared to the standard model. In orthogonal, ANOVA-like designs where the appropriate RSM can be fit in standard mixed modeling software, it can be shown that the test statistic for the standard model that ignores stimulus variability will be inflated by a factor of roughly $\sqrt{\frac{E+mP+nS}{E+mP+S}}$, where n is the number of participants, m is the number of stimuli, and E, P, and S are, respectively, the error variance, participant variance, and stimulus variance (the exact expression depends on the experimental design). While we cannot safely assume that the more complicated fMRI RSM will follow a similar inflation factor, this does give us several hypotheses about the qualitative conditions under which we should expect the worst inflation in fMRI data. Specifically, the degree of inflation should increase with participant sample size, decrease with stimulus sample size, and increase with stimulus variability. In Appendix 1 we describe the results of our simulations in detail. Here we summarize the basic structure of the simulations and their results.

In each run of the simulation, we generated data according to the RSM for a block-design experiment involving participants responding to stimuli nested in two stimulus categories. The test of interest in these simulated experiments is the difference in the fixed regression coefficients for the two stimulus categories (i.e., whether there is greater activation for one stimulus category than for the other). We varied three primary factors in our simulations: the participant sample size (n = 16, 32, or 64), the stimulus sample size (m = 16, 32, or 64), and the degree of random stimulus variability (zero, moderate, or high). Note that when the random stimulus effects have zero variance, the RSM is statistically equivalent to the standard model. We included this condition in order to investigate the performance of the RSM when the standard model is the correct model. For each simulated experiment, we fit four statistical models: the standard model, the RSM, the standard SPM-style "summary statistics" model, and a fourth model that we call the Fixed Stimulus Model, which we describe in Supplementary File 1. Here we focus on comparing the performance of the standard model and RSM (though, in practice, the three non-RSM models all display essentially indistinguishable behavior across all simulations).

Literature review

For our survey of the literature using task-based fMRI, we randomly selected task fMRI papers, without replacement and with uniform sampling probability, until 100 experiments were obtained (each paper can contain 1 or more experiments). Four of the papers we sampled described two experiments, one paper described three experiments, and the rest described a single experiment, so that

the 100 experiments we sampled came from 94 unique papers. We coded each experiment for (a) whether the stimuli were crossed with participants or nested in participants, (b) the type of stimuli used, (c) the total number of stimuli, (d) the number of stimulus categories, and (e) whether the study was eligible to have applied a RSM to the data obtained from the study. Generally speaking, experiments were deemed eligible to have applied a RSM if (i) the experiment used more stimuli than stimulus categories (so that the individual stimulus effects are statistically identifiable), and (ii) the sampled stimuli could not be considered to fully exhaust the population of stimuli over which generalization was intended. In the handful of cases where it was not totally clear from the text whether a RSM could have been applied, we decided to err on the conservative side and deem the study ineligible. A spreadsheet with the detailed study-level results of our survey can be found at https://github.com/PsychoinformaticsLab/nipymc (Yarkoni & Westfall, 2016). We ultimately found that 63/100 of the experiments (95% Jeffreys interval = [53%, 72%]) were eligible to have applied a RSM, and thus the published test statistics for these experiments are likely inflated relative to the more appropriate RSM test statistics.

Results

Simulations

When the true data generating process contained zero stimulus variability, the standard model and RSM yielded very similar test statistics for the stimulus category difference, for all participant and stimulus sample sizes. The exception was that when the stimulus sample size was small (m = 16), the test statistics from the RSM were slightly attenuated (by about 18%) compared to the standard model test statistics. However, this attenuation disappeared with increasing stimulus sample size, so that at m = 32 and m = 64, the test statistics from the two models were essentially identical, as should be the case given the lack of true stimulus variability. This provides evidence that the reduced test statistics observed in our reanalyses of real datasets (most of which were based on a sample of 100 participants) are not simply the result of the RSM always yielding lower test statistics. Instead, the RSM tends to yield lower test statistics when they should in fact be lower, namely, when there is random stimulus variability in the data that is ignored by the standard model.

When the true data generating process contained moderate stimulus variability, the RSM yielded consistently lower test statistics than the standard model. This reduction was exacerbated when the stimulus sample size was small (m = 16) and attenuated somewhat when the stimulus sample size was large (m = 64). The opposite pattern held for the participant sample size: the reduction in the RSM test statistic was largest when the participant sample size was large (n = 64) and smallest when the participant sample size was small (n = 16). These patterns are consistent with what has been observed in previous behavioral work (Judd *et al.*, 2012). In the best case (n = 16 and m = 64) the RSM test statistics were still reduced by an average of 17% compared to the standard model test statistics. In the worst case (n = 64 and m = 16), the RSM test statistics were reduced by an average of 67%. Finally, when the true data generating process contained high stimulus variability, the same qualitative patterns held, but the reduction in test statistics was even greater, ranging from 41% in the best case to 81% in the worst case. Importantly, only the RSM correctly estimated the variability in the average condition difference across simulated datasets; the standard errors from both the standard model and other simpler (but incorrect) approximations consistently underestimated the true variance of the condition differences across simulated datasets.

Does the amygdala preferentially respond to emotional faces?

To illustrate the scope and magnitude of the stimulus-as-fixed-effects fallacy in fMRI, we first focus our attention on one of the most well-established neuroimaging findings: the role of the amygdala in affective processing. The amygdala has been shown in hundreds of fMRI studies to increase activation in response to biologically salient stimuli - most notably faces - and to show a particular strong response to negative affect-provoking stimuli, such as fearful or angry face expressions (Breiter *et al.*, 1996; Morris *et al.*, 1996). However, the number of stimuli used in studies demonstrating this effect is often small - in many cases, fewer than 10 stimuli per experimental condition. As we note above, this is precisely the situation in which inflated statistical significance is expected.

To quantify the effects of modeling stimulus as a random factor on the amygdala response to emotionally salient stimuli, we used data (n = 111 subjects) from the Human Connectome Project (HCP; Barch *et al.*, 2013; Van Essen *et al.*, 2013). In the HCP Emotion Processing Task, adapted from an earlier task used by Hariri and colleagues (Hariri *et al.*, 2002), participants view blocks of faces (20 in total) or geometric shapes (3 in total), and make an unrelated perceptual matching judgment. The face stimuli have either fearful or angry facial expressions (10 per condition). We first analyzed the data using the standard model, where subjects (but not stimuli) were modeled as random effects. For each contrast of interest, we define a test statistic $z = \mu/\sigma$, where μ and σ are the mean and standard deviation, respectively, of the posterior samples for the associated parameter estimate (cf. Kruschke, 2013). Consistent with previous reports on this dataset (Barch *et al.*, 2013) and the broader literature, when analyzing the data under the standard model, we found a robust increase in amygdala activation for face stimuli relative to shape stimuli (z = 26), and a smaller but still notable increase for angry faces relative to fearful faces (z = 3.3). These results are illustrated in Figure 2 (top).

As noted above, the analysis under the standard model fails to account for the uncertainty inherent in the fact that the effect of stimulus category is based on a small and highly variable stimulus sample. When we modeled the data using the RSM, we found that the test statistic for the face vs. shape contrast was reduced by 89% (from z = 26 to z = 2.8) compared to the standard model, and the test statistic for the anger vs. fear contrast was reduced by 78% (from z = 3.3 to z = 0.7). In the former case, the effect remained intact (but would have failed to do so in a smaller sample more typical of the modal fMRI study, and would not have survived multiple comparisons correction even at the current sample size). In the latter case, the remaining effect was negligible, providing essentially no basis for concluding that the amygdala responds differentially to angry vs. fearful faces. These results are illustrated

in Figure 2 (bottom). Thus, simply accounting for natural variability in the sampled stimuli was sufficient to turn a seemingly robust and possibly scientifically intriguing finding into an unremarkable result that probably does not merit further consideration. The striking effect of explicitly modeling stimulus as a random effect is further illustrated in Figure 3.

Whole-brain analysis reveals differential stimulus sensitivity

Next, we extended the analysis to the rest of the brain, fitting the same RSM in 100 different regions-of-interest (ROIs; we used an ROI rather than voxel-wise approach, due to the computational demands of the RSM). As Figure 4 illustrates, the impact of modeling stimulus as a random factor were no less dramatic than in the amygdala for most brain regions. For the two test statistics (i.e., face vs. shape contrast and anger vs. fear contrast), the median ratios of the standard model z-statistic over the RSM z-statistic were 3.3 and 5.96, respectively, indicating reductions of 70% and 83% when random stimulus effects were added. When thresholding brain activity at even a relatively liberal threshold of z = 3, only 2 out of 100 regions (compared to 59 out of 100 in the classical analysis) remained statistically significant, and *no* region showed a significant difference between angry and fearful faces (as compared to 27 regions in the classical analysis).

Intriguingly, the fusiform face area (FFA; Kanwisher *et al.*, 1997), which showed the most robust face-related increase in the standard model (z = 31), failed to show a significant difference between faces and shapes in the RSM. This counterintuitive result can be understood by considering the large amount of stimulus-level variability in FFA responses to faces (Figure 5). Since the face vs. shape test statistic in the RSM depends on the ratio between the between-condition and within-condition (i.e., stimulus-level) variances, a brain region that is extremely sensitive to different stimuli of the same modality may counterintuitively fail to show a consistent difference between faces and shapes precisely *because* it is extremely sensitive (but differentially so) to individual faces.

Although the resulting reduction in the test statistic may come as an unpleasant surprise, there is an important silver lining: the ability to quantify the variance in brain activity specifically related to individual stimuli provides a powerful tool for identifying brain regions sensitive to different classes of stimuli. To illustrate, Figure 5 displays the stimulus-level variability captured by the model in each brain region. Not surprisingly, stimulus-related variability was greatest in visual cortical regions; however, a number of other brain regions also showed considerable stimulus sensitivity in response to faces, including motor cortex, anterior insula

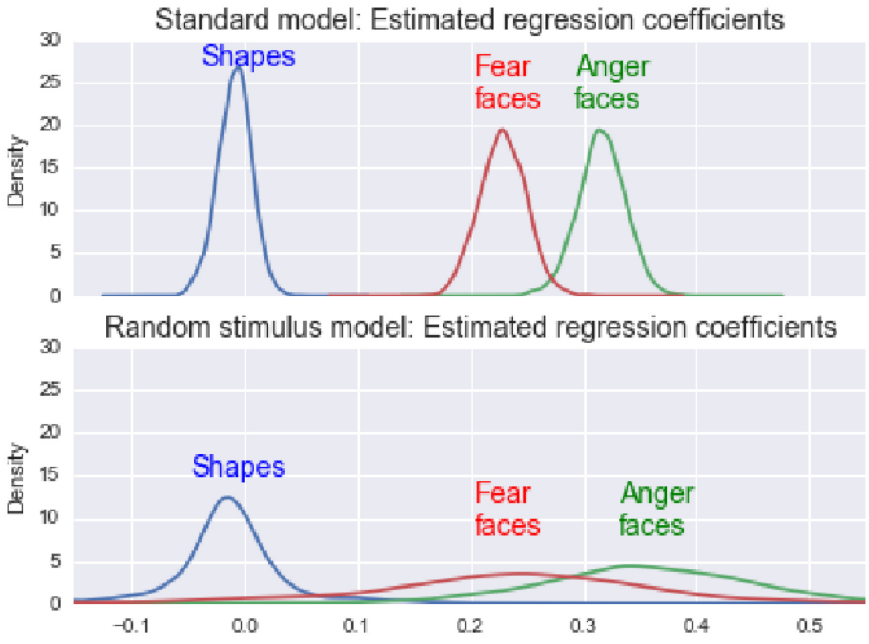

Figure 2. Posterior samples of the regression coefficients associated with each stimulus category. The results indicate the predicted magnitude of average amygdala response in response to each category, under both the standard model and the random stimulus model (RSM). Under the standard model, there is clear separation of the estimated amygdala responses toward all three stimulus categories, with anger faces evoking a somewhat stronger response than fear faces, and both face stimuli evoking a much stronger response than the simple shape stimuli. Under the RSM, the means of the regression coefficients are about the same, but they are estimated with far more uncertainty. The result is that while the face vs. shape contrast is still clearly discernible, the anger vs. fear contrast is no longer distinguishable from sampling/measurement error.

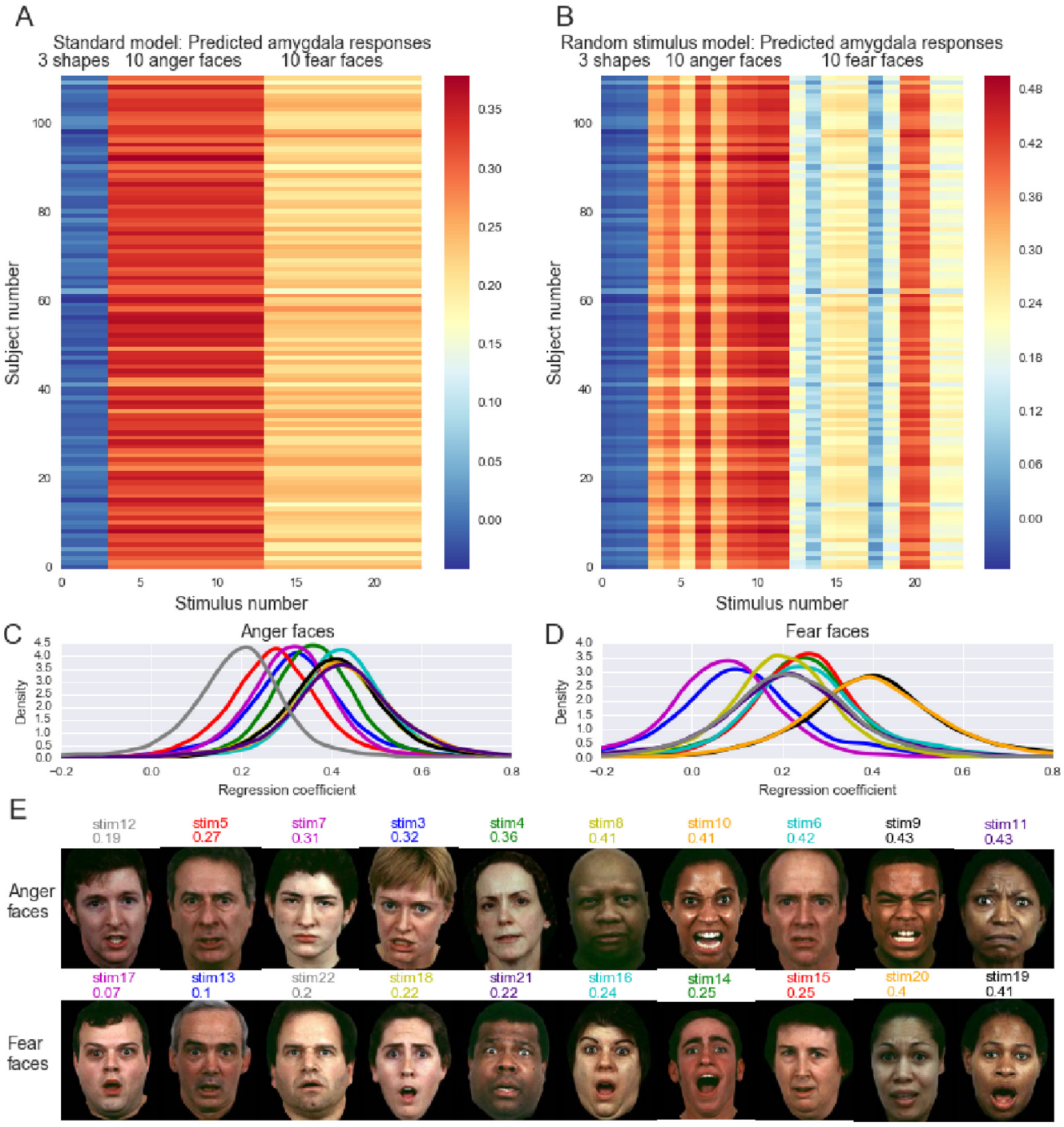

Figure 3. Magnitude of amygdala response predicted by the standard model (panel **A**) and the random stimulus model (RSM) (panel **B**). This is represented as subject×stimulus matrices, where each row (111 in total) represents a unique subject and each column (23 in total) represents a unique stimulus. Each entry of the matrix gives a (posterior mean) regression coefficient corresponding to the model's prediction for that subject's amygdala response toward that stimulus. Notice that the standard model assumes that a subject has the same amygdala response toward all stimuli in a given category - in other words, it assumes no random stimulus variability. While this may not be an entirely unreasonable assumption for the three relatively impoverished shape stimuli (a circle and two ovals), it is a patently absurd assumption for the faces, as a cursory visual inspection of the stimuli makes clear (panel **E**). When we add random stimulus effects to the standard model, resulting in the RSM, we find that random stimulus variability (evident in the within-category variance of the column means) is in fact one of the chief sources of variation in the data. The images in panel **E** are sorted within each emotion condition by their posterior sample means (panels **C** and **D**), which are printed below the stimulus labels. (Faces images from Human Connectome Task fMRI battery, used with permission. Copyright 2012.)

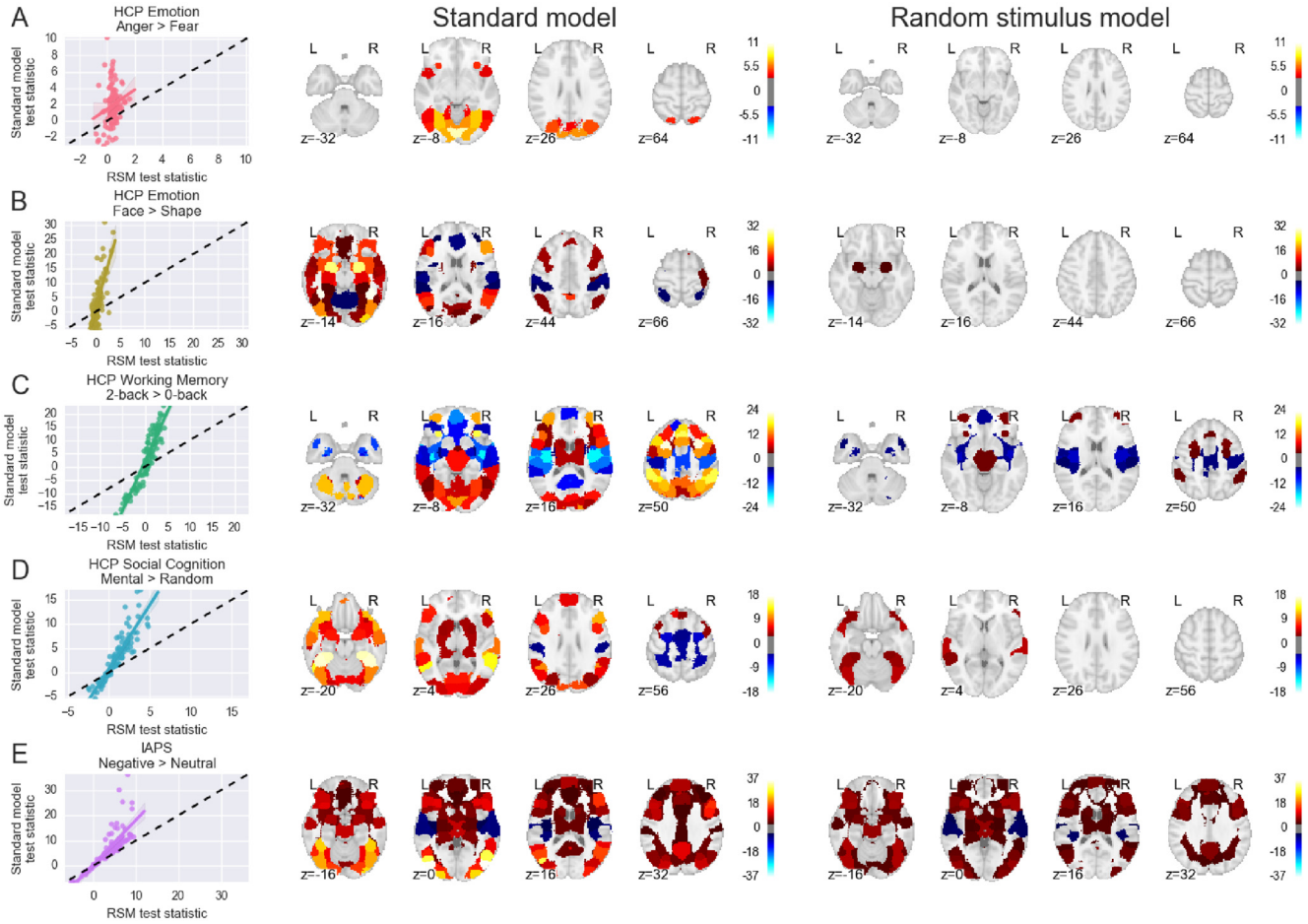

Figure 4. Whole-brain results for five contrasts from four different datasets when modeled with either a standard approach or a random stimulus model (RSM). Each row displays results for a different dataset and/or contrast (see main text for details). Left column: scatter plot displaying the relationship between region-of-interest (ROI)-level test statistics from the standard model (y-axis) and RSM (x-axis). Each point represents a single ROI from a 100-region whole-brain clustering. Middle and right columns: axial slices displaying ROI-level test statistics (z statistics, defined as in the main text) from the standard and RSM analyses, respectively. Maps are thresholded at $|z| > 3.3$ - comparable to using $p < .001$, uncorrected, in a traditional frequentist analysis - in order to illustrate the significant drop in test statistics in most datasets when including random stimulus effects.

(particularly for anger faces), and portions of anterior PFC. As one might intuitively expect, the variability in responses to the 3 simple geometric shapes was muted in comparison to the response to faces.

Generalization to other datasets

To ensure that the HCP Emotion Task was not an outlier, and that our conclusions apply more generally, we repeated our random stimulus analyses on several other datasets. We fit RSM models to two other HCP functional tasks - the Social Cognition Task and the Working Memory Task - as well as a non-HCP emotion processing dataset (Chang *et al.*, 2015). These tasks differed widely in experimental design, stimulus modality (video clips, images, and audio narratives), number of stimuli (10 in the social cognition task, 96 in the WM task), and putative psychological processes. Nevertheless, when contrasting the RSM with the classical model, test statistics for critical comparisons were reduced considerably in all datasets (the median reductions across all 100 ROIs ranged from 12% to 83% in the datasets we examined; Figure 4 and Figure 6). In general, the rank-order stability of regions was high across the two analyses in terms of the test statistics (mean $r = 0.77$). Thus, the global pattern of activity across the whole brain was, at least in the tested datasets, relatively conserved in the RSM, and the drop in test statistics largely reflected the increased variance of the fixed-effect estimates of the experimental conditions (cf. Figure 2).

The critical role of stimulus sample size

Why were the critical test statistics from the RSM model so small compared to the standard analysis, despite the relatively large participant samples used in these analyses? The likely culprits here are the relatively small stimulus sets used in these experiments. As far as the RSM is concerned, the stimuli used in a study are just as important as the human subjects---both ultimately represent sources of random variation in the data that we would like to generalize

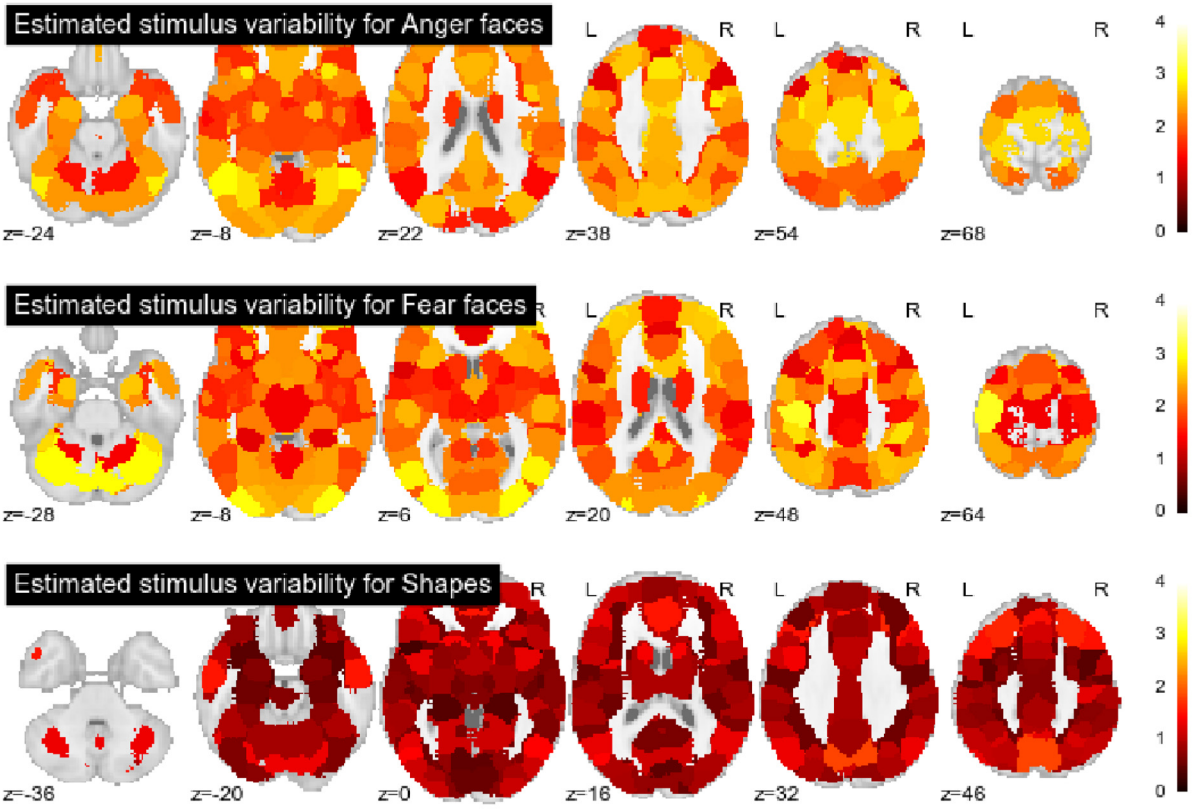

Figure 5. Model-estimated stimulus variability (standard deviation of the random stimulus effects) for the three stimulus categories in the HCP Emotion Processing Task, separately for each region-of-interest (ROI). ROIs with greater estimated stimulus variability showed greater sensitivity to idiosyncratic stimulus differences within each stimulus category.

over when estimating brain responses to different experimental conditions. Most researchers would question the wisdom of conducting an fMRI study that compared, say, 5 highly variable subjects in one condition to 5 highly variable subjects in another condition, yet many researchers routinely make essentially the same mistake when sampling stimuli. The problem is exacerbated in many cases, including in many of the present datasets, when stimuli are presented in exactly the same order to all participants - an approach that conflates order effects with stimulus effects, necessarily inflating the variance seemingly accounted for by the latter.

The important silver lining to this otherwise grim analysis is that test statistic inflation in the classical model is related in predictable ways to stimulus sample size (Judd *et al*., 2012; Wickens & Keppel, 1983). Thus, it should be possible to minimize the gap between the standard model and the RSM by increasing the number of stimuli in one's experiment. To test this prediction empirically, we used two additional datasets (Figure 6). First, we applied the RSM to the HCP language task, which included a Math condition in which participants provided forced-choice answers to auditorily presented mental arithmetic problems. In contrast to the other HCP tasks, stimuli in the Math condition are adaptively chosen from a large set of over 7,000 candidate mental arithmetic problems based on each participant's in-task performance. We consequently predicted that RSM estimates should be very close to standard model estimates for this experimental condition. This prediction was confirmed (Figure 6B): test statistics from the two models were very similar across the brain when comparing the Math condition to the implicit resting baseline (mean $|z|$ = 8.47 vs. 8.12; 4% reduction). This consistency across models contrasted sharply with the large reduction observed for the Language condition (mean $|z|$ = 4.83 vs. 2.67; 45% reduction), which presented the same 6 stimuli to nearly all subjects. As a consequence of the loss of precision in the Language condition, test statistics for the Math vs. Language contrast also showed a considerable decline (mean $|z|$ = 13.75 vs. 5.96; 57% reduction). Figure 6C displays estimates from the standard model and RSM for a sample region (V5/MT in visual cortex), clearly illustrating the selective increase in uncertainty in the story condition.

Second, we analyzed an unpublished emotion regulation dataset (Cohen, 2009; http://gradworks.umi.com/34/01/3401764.html), publicly available from the OpenfMRI.org repository (Poldrack & Gorgolewski, 2015), in which 11 participants passively viewed either negative or neutral pictures. Importantly, each participant viewed 60 different stimuli (from a total set of 120). Theoretically, this "partially-crossed" design should considerably reduce the discrepancy between the RSM and classical model

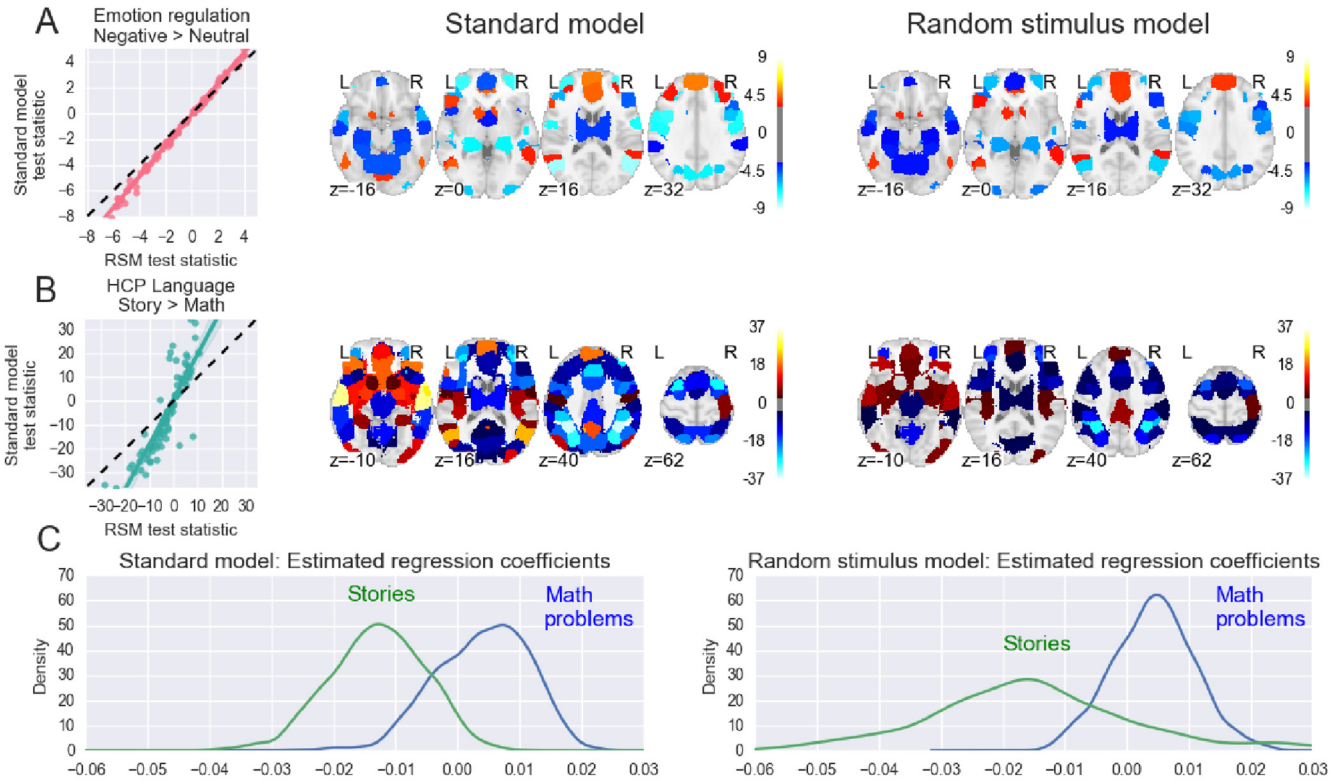

Figure 6. Whole-brain results for three datasets modeled with either a standard approach or a random stimulus model. The interpretation is the same as in Figure 4.

(Westfall *et al.*, 2014), and this is precisely what we found: test statistics from the two models were very similar across the brain when comparing passive viewing of neutral vs. negative images (Figure 6A; mean $|z| = 2.77$ vs. 2.30 for standard model vs. RSM, respectively; median ratio = 1.18). These results confirm that there is indeed a predictable and robust relationship between the stimulus sampling scheme used in an fMRI experiment and the degree of test statistic inflation one can expect to observe when using the standard (incorrect) model.

Stimulus-level parameter estimates as a tool for exploration

While the primary reason to include random stimulus effects in one's model is to ensure that statistical inferences can be safely generalized to new stimuli, an important secondary benefit is that this approach facilitates data exploration and hypothesis generation. The inclusion of a separate parameter for each stimulus allows one to estimate the unique pattern of whole-brain activation associated with each stimulus. Inspection of these estimates may help identify novel features of the data or design that can be subsequently tested formally.

To illustrate, consider the parameter estimates displayed for individual faces in Figure 3E. Qualitatively, there appears to be a potential trend for black faces to elicit larger amygdala responses than white faces. To formally test this hypothesis, we obtained race judgments of the 20 faces from 3 lab members blind to our hypothesis and with no knowledge of the parameter estimates. When the RSM model was recomputed with an additional fixed effect coding stimulus race, the resulting posterior estimate was suggestive of a weak race effect ($z = 1.55$; the other parameter estimates were all virtually unchanged). Of course, this particular analysis is circular, since the hypothesis was generated and tested using the same data. The important point, however, is that even this cursory visual inspection of the stimulus-level estimates was sufficient to suggest a scientifically interesting hypothesis that could be readily tested using independent data. Indeed, a number of previous studies (Lieberman *et al.*, 2005) have reported race-related amygdala activation patterns consistent with our conjecture (though none included random stimulus effects, and hence the existing evidence for race effects in the amygdala is itself very likely overstated).

Discussion

We have shown that the universal failure of fMRI studies to include random stimulus effects in statistical models can have a substantial, and almost invariably deleterious, effect on reported results. At root, the problem lies in a mismatch between researcher intent and statistical implementation: neuroimaging researchers intend for their statistical conclusions to generalize across populations of stimuli similar, but not identical to, the ones they tested; however, conventional statistical procedures only allow conclusions to be drawn with respect to the exact stimuli used in each study. Our literature survey suggests that the ramifications of this discrepancy

between intent and praxis are likely to be very large: in a survey of 100 articles, we found that use of a RSM was clearly indicated in over 60% of cases. This is a conservative lower bound of the true extent of the problem, as many of the remaining studies could not have used a RSM due to otherwise avoidable limitations of their experimental design (e.g., using only a single stimulus per condition).

Given that the RSM test statistics we obtained were frequently reduced by 50% or more relative to those obtained using a standard analysis, one implication of our findings is that a large fraction of the results reported in the fMRI literature are likely to be severely inflated. Moreover, when the true mean stimulus effect is zero, variation in stimuli will generally exert some non-zero influence on brain activity (as was evident in all of the datasets we tested), which in turn will inflate Type I error. In simulations, which we describe in detail in Supplementary File 1, we found that the Type 1 error rate for the standard model in the presence of unmodeled stimulus variability, using a nominal decision threshold of $\alpha = .001$, ranged from about .01 to about .4, depending on the sample sizes and the level of stimulus variability. At a threshold of $\alpha = .05$ (as would be common in many hypothesis-driven ROI-level tests) the Type 1 error rate was as high as .65, still under relatively conservative assumptions (e.g., a maximum participant sample size of 64 and a minimum stimulus sample size of 16).

Although its consequences are severe, the statistical argument that we've presented - based on the idea that the experimental stimuli are typically a random factor and not a fixed factor - is conceptually subtle. While the fixed vs. random distinction is not always defined consistently (Gelman & Hill, 2007, p. 245), the classical definition of a random factor from the literature on analysis of variance is that the levels of the factor that appeared in the experiment (i.e., the stimuli that were used) do not fully exhaust the theoretical population of levels that might have been used. Importantly, this definition does *not* imply that the stimuli were selected haphazardly. Indeed, typically experimenters take great care in selecting an appropriate stimulus set to be used in the study. Rather, to say that the stimuli are "random" is simply to say that there are, in principle, other possible stimuli that could have served the experimenter's purposes just as well as those that were in fact used. Using this definition, our literature survey suggests that a RSM is the most statistically appropriate model to use in a majority of task fMRI studies. But exceptions do exist for certain special cases. Below we list a few necessary conditions for fitting a RSM, some more conceptual and some purely statistical.

First, the stimuli in question must be inherently discrete entities, such as distinct words or photographs. An example of stimuli that, on these grounds, would *not* be modeled as random would be the varying doses of some drug administered to the subjects. While the doses are nominally discrete in that only a finite number of the range of possible doses are administered, these doses represent points on a well-defined continuum, and would simply be treated as a fixed predictor or covariate in the analysis. Second, and as mentioned above, a RSM is only indicated when the stimuli used in the study do not fully exhaust the theoretical population of stimuli that might have been used. An example of an experiment that does not satisfy this condition would be a study of brain responses toward single-digit numbers, in which the study employs all possible single-digit numbers. Third, there must be at least some degree of overlap in the stimulus sets that each subject receives. In what is overwhelmingly the most common case, every subject receives the same stimulus set, and a RSM can be estimated relatively easily. Less frequent, but not uncommon, are experiments where subjects receive different subsets of stimuli from a larger stimulus pool, but there is some overlap in the stimulus sets, such that at least some of the stimuli receive responses from more than one subject. An example would be the Math stimuli in the HCP Language task. A RSM would generally be appropriate here as well. In the least common case (6/100 of the experiments in our survey), every subject receives a completely unique stimulus set, with no overlap between sets, such that stimuli are strictly nested in participants. In this case the standard model would be a statistically appropriate model, and in fact subjects would need to respond multiple times to each stimulus for a RSM to be statistically identifiable at all.

We are aware of only two previous papers that have discussed the issue of random stimulus variability in fMRI (Bedny *et al*., 2007; Donnet *et al*., 2006). While conceptually similar, these two papers take different approaches than the one described here, and it is worth noting the differences. Donnet *et al*. (2006) describe a model with random stimulus effects that are different for every subject, so that the corresponding variance component is more akin to a subject-by-stimulus interaction variance component than to the stimulus variance components incorporated in our models. They also do not discuss inference on the fixed effects of activation magnitude. Bedny *et al*. (2007) do discuss inference on the fixed effects, but they focus primarily on conducting a separate subject-wise analysis (i.e., what we have called the standard model, which ignores stimulus variance) and stimulus-wise analysis (i.e., the conceptual complement of the subject-wise analysis, which includes stimulus variance but ignores participant variance). This approach is common in psycholinguistics, and it is certainly a step up from running only the standard or subject-wise model, but it is not equivalent to the full, correct model with crossed random effects of participants and stimuli. In particular, it does not succeed in maintaining the nominal Type 1 error rate (Raaijmakers, 2003; Raaijmakers *et al*., 1999).

What can researchers do to address the stimulus-as-fixed-effect fallacy? Broadly speaking, there are two possible strategies. The best practice approach to address the issue is to explicitly include random effects in one's model for every factor a researcher intends to generalize over. In the present study, we used MCMC sampling to fit our mixed-effects models; however, other approaches (based on maximum likelihood estimation, variational inference, etc.) are also available (Bates *et al*., 2015). The primary downside of an estimation-based solution is that it is computationally intensive and may be technically demanding. At present, no major fMRI analysis package supports RSMs of the kind we employ here, limiting the ability of most researchers to produce correct inferences. While we have open-sourced the *NiPyMC* Python package we used to fit the models reported here (http://github.com/PsychoinformaticsLab/nipymc; Yarkoni & Westfall, 2016), this should be viewed as a provisional (and not particularly scalable) solution

until more robust and widely-used packages such as FSL, SPM, or AFNI introduce support for random stimulus effects.

Alternatively, a less effective, but much simpler approach to the problem, is to use as large a stimulus sample as is practically feasible. In previous work, we have shown that the number of stimuli can impose a hard cap on statistical power in the RSM: when stimulus samples are very small, it may be impossible to obtain statistically significant estimates of the fixed effects, no matter how many thousands of subjects one samples (Westfall *et al.*, 2014). This is evident in the present findings, where test statistics from fMRI experiments with only a few stimuli (such as the HCP Emotion and Social Cognition tasks) showed precipitous drops in the RSM, whereas those generated by designs with many stimuli are often negligible (e.g., Figure 6). The primary (and significant) downside of a stimulus-maximizing heuristic is that it is only a heuristic--there is no guarantee that the resulting test statistic will closely approximate the one that would have been obtained through the explicit inclusion of random stimulus effects. In particular, if the degree of random stimulus variability is large, then a huge number of stimuli may be required before the two sets of test statistics closely converge. Nevertheless, in the absence of analysis tools capable of correctly modeling multi-stimulus designs, we strongly encourage researchers to always include as many stimuli as possible in their designs. Importantly, unlike increases in participant sample size, adding stimuli rarely incurs any additional cost. Researchers can usually easily increase the number of stimuli by either (a) eschewing repeated presentation of a few stimuli in favor of single presentation of many different stimuli, or (b) using a "partially crossed" design where each participant responds to a different subset of stimuli (Westfall *et al.*, 2014). These approaches allow one to enjoy the statistical power benefits of a large stimulus sample without increasing data collection requirements.

Software availability

Source code for *NiPyMC* analysis package: http://github.com/PsychoinformaticsLab/nipymc/

Archived source code at time of publication: DOI, 10.5281/zenodo.168087 (Yarkoni & Westfall, 2016)

License: NiPyMC is distributed under The MIT License.

Author contributions

JW & TY were responsible for the conception & design of the work, and the analysis of the data; all authors contributed to the interpretation of the results, drafting and revising the work.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the Wellcome Trust [100309]; and the National Institutes of Health [R01MH096906].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Supplementary File 1: Further information regarding statistical models.

Click here to access the data.

References

Baayen RH, Davidson DJ, Bates DM: **Mixed-effects modeling with crossed random effects for subjects and items.** *J Mem Lang.* 2008; **59**(4): 390–412.
Publisher Full Text

Barch DM, Burgess GC, Harms MP, *et al.*: **Function in the human connectome: task-fMRI and individual differences in behavior.** *Neuroimage.* 2013; **80**: 169–189.
PubMed Abstract | Publisher Full Text | Free Full Text

Bastien F, Lamblin P, Pascanu R, *et al.*: **Theano: new features and speed improvements.** *arXiv: 1211.5590 [cs.SC],* 2012.
Reference Source

Bates D, Douglas B, Martin M, *et al.*: **Fitting Linear Mixed-Effects Models Using lme4.** *J Stat Softw.* 2015; **67(1)**: 1–48.
Publisher Full Text

Bedny M, Aguirre GK, Thompson-Schill SL: **Item analysis in functional magnetic resonance imaging.** *Neuroimage.* 2007; **35**(3): 1093–1102.
PubMed Abstract | Publisher Full Text

Bergstra J, Breuleux O, Bastien F, *et al.*: **Theano: a CPU and GPU math expression compiler**. In *Proceedings of the Python for scientific computing conference (SciPy).* Austin, TX, 2010; **4**: 3.
Reference Source

Breiter HC, Etcoff NL, Whalen PJ, *et al.*: **Response and habituation of the human amygdala during visual processing of facial expression.** *Neuron.* 1996; **17**(5): 875–887.
PubMed Abstract | Publisher Full Text

Carp J: **On the plurality of (methodological) worlds: Estimating the analytic flexibility of fmri experiments.** *Front Neurosci.* 2012; **6**: 149.
PubMed Abstract | Publisher Full Text | Free Full Text

Chang LJ, Gianaros PJ, Manuck SB, *et al.*: **A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect.** *PLoS Biol.* 2015; **13**(6): e1002180.
PubMed Abstract | Publisher Full Text | Free Full Text

Clark HH: **The language-as-fixed-effect fallacy: A critique of language statistics**

in psychological research. *J Verbal Learning Verbal Behav.* 1973; **12**(4): 335–359.
Reference Source

Coleman EB: **Generalizing to a language population.** *Psychol Rep.* 1964; **14**(1): 219–226.
Publisher Full Text

Donnet S, Lavielle M, Poline JB: **Are fMRI event-related response constant in time? A model selection answer.** *Neuroimage.* 2006; **31**(3): 1169–1176.
PubMed Abstract | Publisher Full Text

Gelman A, Hill J: **Data Analysis Using Regression and Multilevel/Hierarchical Models**. 2007.
Publisher Full Text

Glasser MF, Sotiropoulos SN, Wilson JA, *et al.*: **The minimal preprocessing pipelines for the Human Connectome Project.** *Neuroimage.* 2013; **80**: 105–124.
PubMed Abstract | Publisher Full Text | Free Full Text

Gorgolewski KJ, Varoquaux G, Rivera G, *et al.*: **NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain.** *Front Neuroinform.* 2015; **9**: 8.
PubMed Abstract | Publisher Full Text | Free Full Text

Hariri AR, Mattay VS, Tessitore A, *et al.*: **Serotonin transporter genetic variation and the response of the human amygdala.** *Science.* 2002; **297**(5580): 400–403.
PubMed Abstract | Publisher Full Text

Hoffman MD, Gelman A: **The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.** *J Mach Learn Res: JMLR.* 2014; **15**: 1351–1381.
Reference Source

Judd CM, Westfall J, Kenny DA: **Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem.** *J Pers Soc Psychol.* 2012; **103**(1): 54–69.
PubMed Abstract | Publisher Full Text

Kanwisher N, McDermott J, Chun MM: **The fusiform face area: a module in human extrastriate cortex specialized for face perception.** *J Neurosci.* 1997; **17**(11): 4302–4311.
PubMed Abstract | Publisher Full Text

Kruschke JK: **Bayesian estimation supersedes the *t* test.** *J Exp Psychol Gen.* 2013; **142**(2): 573–603.
PubMed Abstract | Publisher Full Text

Lieberman MD, Hariri A, Jarcho JM, *et al.*: **An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals.** *Nat Neurosci.* 2005; **8**(6): 720–722.
PubMed Abstract | Publisher Full Text

Morris JS, Frith CD, Perrett DI, *et al.*: **A differential neural response in the human amygdala to fearful and happy facial expressions.** *Nature.* 1996; **383**(6603): 812–815.
PubMed Abstract | Publisher Full Text

Patil A, Huard D, Fonnesbeck CJ: **PyMC: Bayesian Stochastic Modelling in Python.** *J Stat Softw.* 2010; **35**(4): 1–81.
PubMed Abstract | Publisher Full Text | Free Full Text

Penny WD, Holmes AP, Friston KJ: **Random effects analysis.** *Human Brain Function.* 2003; **2**: 843–850.
Reference Source

Poldrack RA, Gorgolewski KJ: **OpenfMRI: Open sharing of task fMRI data.** *Neuroimage.* 2015. pii: S1053-8119(15)00463-2.
PubMed Abstract | Publisher Full Text | Free Full Text

Poldrack RA, Mumford JA, Nichols TE: **Handbook of Functional MRI Data Analysis**. Cambridge University Press, 2011.
Publisher Full Text

Raaijmakers JG: **A further look at the "language-as-fixed-effect fallacy".** *Can J Exp Psychol.* 2003; **57**(3): 141–151.
PubMed Abstract | Publisher Full Text

Raaijmakers JG, Schrijnemakers JM, Gremmen F: **How to Deal with "The Language-as-Fixed-Effect Fallacy": Common Misconceptions and Alternative Solutions.** *J Mem Lang.* 1999; **41**(3): 416–426.
Publisher Full Text

Salvatier J, Wiecki T, Fonnesbeck C: **Probabilistic Programming in Python using PyMC**. *arXiv 1507.08050 [stat.CO],* 2015.
Reference Source

Van Essen DC, Smith SM, Barch DM, *et al.*: **The WU-Minn Human Connectome Project: an overview.** *NeuroImage.* 2013; **80**: 62–79.
PubMed Abstract | Publisher Full Text | Free Full Text

Westfall J, Kenny DA, Judd CM: **Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli.** *J Exp Psychol Gen.* 2014; **143**(5): 2020–2045.
PubMed Abstract | Publisher Full Text

Wickens TD, Keppel G: **On the choice of design and of test statistic in the analysis of experiments with sampled materials.** *J Verbal Learning Verbal Behav.* 1983; **22**(3): 296–309.
Publisher Full Text

Wolsiefer K, Westfall J, Judd CM: **Modeling stimulus variation in three common implicit attitude tasks.** *Behav Res Methods.* 2016; 1–17.
PubMed Abstract | Publisher Full Text

Yarkoni T, Westfall J: **PsychoinformaticsLab/nipymc: v0.0.1-alpha [Data set].** *Zenodo.* 2016.
Data Source

Open Peer Review

Current Peer Review Status: ✗

Version 1

Reviewer Report 20 January 2017

https://doi.org/10.21956/wellcomeopenres.11091.r19186

✗ **Sophie Donnet**
University of Paris-Saclay, Paris, France

In this paper, the authors claim that the conclusions of several studies would be modified if a statistical models taking into account the variability of the presented stimuli had been considered.

Although the problem is interesting, I am not completely convinced by the conclusions and the statistical tools used to assess the results.

First of all, the authors argue that, due to the complexity of the new model, the authors can not use the standard numerical tools to perform the statistical inference (R or SAS) and so will prefer a Bayesian inference, making this choice quite opportunist. However, besides the fact that they use a Bayesian inference (including prior distribution), they base their conclusions on frequentist arguments (comparing test statistics). To my point of view, this is quite confusing. If a Bayesian framework is considered, then the hypothesis testings should be perform using Bayes Factor or any other tools taking into account the prior distribution.

Moreover, when proposing a new model, the first thing to do is to compare the old model and the new model for any given dataset (using statistical tools such as hypothesis testing, in a frequentist or Bayesian framework). I could not see such a test in the paper. The mixed effects model involves much more parameters for the same amount of observations. Before deriving conclusions on the new model, it should be interesting to put in competition the fixed effects model and the mixed effects models, to be sure that the use of the more complex model is supported by the data.

Besides, in a Bayesian context, the greater posterior variance observed on the regression parameters was completely expected (Figure 2) (more parameters for the same quantity of information leads to more incertitude a posteriori). However, without objective criteria (such as Bayes factor or hypothesis testing), I am not able to decide whether the difference between fear faces and anger faces is significant or not. (caption for Figure 2). It is not clearly stated in the paper how the authors were able to do so (even though I could find clues in the simulation section?).

I am aware of the fact that Bayes factor are quantities difficult to estimate and that frequentist

hypothesis testing in mixed effects models can be a tough issue. However, if the authors want to prove that the conclusions of statistical testing are modified when using mixed effects models, then they should perform the adequate and rigorous statistical testings.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 06 Mar 2017

Thomas Nichols, University of Oxford, UK

We'd like to thank Dr. Donnet for taking the time to consider our paper carefully. For ease of reading we've pasted in her comments as boldface, bulleted text.

- **In this paper, the authors claim that the conclusions of several studies would be modified if a statistical models taking into account the variability of the presented stimuli had been considered.**
- **Although the problem is interesting, I am not completely convinced by the conclusions and the statistical tools used to assess the results.**
- **First of all, the authors argue that, due to the complexity of the new model, the authors can not use the standard numerical tools to perform the statistical inference (R or SAS) and so will prefer a Bayesian inference, making this choice quite opportunist.**

We regret that the choice to use a Bayesian model created confusion. We noted in the Methods section that the fMRI random stimulus model "cannot be fit using standard mixed modeling statistical packages, such as lme4 in R or SAS PROC MIXED, because these packages assume that each row of the dataset is associated with one and only one level of each random factor, i.e., a single participant and a single stimulus." Hence, we were forced to develop a custom model, and a Bayesian one developed with PyMC3 was the most convenient approach.

We note that most of the datasets we analyzed could in principle also have been fitted using a modified (and approximate) version of the standard "summary statistics" model; however, this would have required extensive additional analysis, as we would have needed to fit a separate first-level model for each individual subject (adding individual regressors for each individual stimulus and temporarily omitting existing condition regressors, which in the classical formulation are non-identifiable in the presence of the stimulus effects); and then fit a mixed model at the second level that models the individual stimulus effects as random effects. The former steps would have been extremely time-consuming relative to our strategy of fitting a single large model, and the latter step (i.e., modeling stimulus as a random effect) is currently not supported in any existing fMRI package, so we would have had to develop a custom model anyway. Moreover, as we noted in the supplement, "This [summary statistics] model is only an approximate version of the full RSM, as it ignores the varying degrees of uncertainty in the stimulus-level coefficients estimated at the first level (due in particular to factors such as differences in how often the stimuli were presented and how collinear the stimulus-level regressors were for each subject)".

We also note that we did in fact use the summary statistics approximation to obtain estimates for the Chang et al. (2015) dataset (see supplement for details). This was not by choice, and we would have preferred to fit the full RSM; however, as we explain in the supplement, this dataset was available in a form amenable only to the summary statistics approach (i.e., because pre-computed estimates for each individual trial were provided).

- **However, besides the fact that they use a Bayesian inference (including prior distribution), they base their conclusions on frequentist arguments (comparing test statistics). To my point of view, this is quite confusing.**

We do not believe that our conclusions are based on frequentist arguments. Notably, we have not computed any p-values nor refer to any z-scores. In the section "Does the amygdala preferentially respond to emotional faces?" we write: "we define a test statistic $z = \mu/\sigma$, where μ and σ are the mean and standard deviation, respectively, of the posterior samples for the associated parameter estimate". The use of "test statistic" was a bit careless, as we are not "testing hypotheses"; we simply needed a convenient summary of the posterior. We admit that this choice of terminology could be confusing, so in our revision we have renamed this statistic the posterior standardized effect.

Another important point is that our argument is fundamentally qualitative and not quantitative; we focus on how uncertainty surrounding the estimate of interest (denominator of our standardized effect) increases dramatically in most cases. We present this increase in terms of standardized effect because that's what most people are familiar with, but nothing would change if we instead talked about the relative increases in the variances of the estimates.

- **If a Bayesian framework is considered, then the hypothesis testings should be perform using Bayes Factor or any other tools taking into account the prior distribution.**

Bayes Factors in this setting are not trivial, and any rate would not serve the broad audience, who would like to know: How severe is the random stimulus effect problem? Again, we could have measured this directly in terms of variance inflation, but since users are most familiar with signal-to-noise ratio, we felt that this was the best metric.

- **Moreover, when proposing a new model, the first thing to do is to compare the old model and the new model for any given dataset (using statistical tools such as hypothesis testing, in a frequentist or Bayesian framework). I could not see such a test in the paper. The mixed effects model involves much more parameters for the same amount of observations. Before deriving conclusions on the new model, it should be interesting to put in competition the fixed effects model and the mixed effects models, to be sure that the use of the more complex model is supported by the data.**

The reviewer is correct that no formal model assessment was used. This was intentional, as (crucially) we are not testing a hypothesis about the relative fit of different models. Rather, we're pointing out a fundamental mismatch between the conclusions researchers typically draw (which almost invariably involve assuming that it is safe to generalize conclusions over a population of stimuli) and the inferences that are actually licensed by the statistical model (which, in the absence of a variance component reflecting the random sampling of levels from some factor, are only valid for the specific levels used in the experiment). Our position is that it is the design of the experiment and the goals of the analyst--not the observed data-

-that determine which random effects should be included in the model. This position is consistent with the arguments of Barr, Levy, Scheepers, and Tily (2013).

It may be useful to draw an analogy to a random subject effect in a multi-subject model. No reasonable investigator would test for the significance (or BF-based evidence) for a subject random effect; while such tests may come up negative sometimes, we know our subjects are a random sample from a larger population, a population to which we wish to generalise. Likewise, it is now generally accepted that meta-analyses should treat study as a random effect irrespective of formal tests of heterogeneity (Borenstein, et al., 2010).

Hence our paper is not focused on the question "Is there statistical evidence for random stimulus effects?" but rather the assertion "Some stimuli are random, and such stimuli should be so modeled." This is a widely noted point in many other disciplines, as we have referenced in the introduction; our contribution is to raise awareness of this issue in neuroimaging, and provide easy-to-interpret measures of the impact of neglecting random stimulus effects.

- **Besides, in a Bayesian context, the greater posterior variance observed on the regression parameters was completely expected (Figure 2) (more parameters for the same quantity of information leads to more incertitude a posteriori).**

Our reanalysis of the Emotion Regulation dataset and our simulation results both speak to the contrary. The Emotion regulation dataset featured a relatively large number of stimuli with small estimated stimulus variance; consequently the standardized effects (what we called the test statistics in the manuscript) from the RSM and standard models are essentially identical, despite the fact that the RSM technically contains many more parameters. Consistent with this general finding, in the Method subsection describing our simulation results we wrote: "When the true data generating process contained zero stimulus variability, the standard model and RSM yielded very similar [standardized effects] for the stimulus category difference, for all participant and stimulus sample sizes. The exception was that when the stimulus sample size was small (m = 16), the [standardized effects] from the RSM were slightly attenuated (by about 18%) compared to the standard model [standardized effects]. However, this attenuation disappeared with increasing stimulus sample size, so that at m = 32 and m = 64, the [standardized effects] from the two models were essentially identical, as should be the case given the lack of true stimulus variability." Note that as the stimulus sample size increases, the number of parameters in the RSM technically increases as well, but this actually causes the posterior for the category difference to become more narrow, not more wide. We believe that these analyses sufficiently highlight the conditions under which the RSM does and does not give overly pessimistic posterior estimates of the category difference.

- **However, without objective criteria (such as Bayes factor or hypothesis testing), I am not able to decide whether the difference between fear faces and anger faces is significant or not. (caption for Figure 2). It is not clearly stated in the paper how the authors were able to do so (even though I could find clues in the simulation section?).**

We were careful not to claim that the difference was or was not statistically significant, but rather, that there appears to be at most a small effect. In Bayesian estimation terms, this amounts to saying that the posterior distributions for the effects of fearful and angry faces mostly overlap. We have updated the text to clarify these points (see updated section

beginning with “Depending on one’s statistical approach and assumptions…”).

- **I am aware of the fact that Bayes factor are quantities difficult to estimate and that frequentist hypothesis testing in mixed effects models can be a tough issue. However, if the authors want to prove that the conclusions of statistical testing are modified when using mixed effects models, then they should perform the adequate and rigorous statistical testings.**

As expanded on above, we do not think model comparison is necessary or helpful in this context. Our argument is not predicated on the idea that a model with random stimulus effects fits the data better (although we strongly suspect that it often will), but rather on how the experimental design and the desired inferences demand such a model.

REFERENCES

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3). https://doi.org/10.1016/j.jml.2012.11.001

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. Research Synthesis Methods, 1(2), 97–111. http://doi.org/10.1002/jrsm.12

Competing Interests: No competing interests to report.