Psychological Methods

ISSN: 1082-989X

2024, Vol. 29, No. 1, 117–136
https://doi.org/10.1037/met0000535

Who Is and Is Not “Average”? Random Effects Selection With Spike-and-Slab Priors

Josue E. Rodriguez, Donald R. Williams, and Philippe Rast
Department of Psychology, University of California, Davis

Abstract

Mixed-effects models are often employed to study individual differences in psychological science. Such analyses commonly entail testing whether between-subjects variability exists and including covariates to explain that variability. We argue that researchers have much to gain by explicitly focusing on the individual in individual differences research. To this end, we propose the spike-and-slab prior distribution for random effect selection in (generalized) mixed-effects models as a means to gain a more nuanced perspective of individual differences. The prior for each random effect is a two-component mixture consisting of a point-mass “spike” centered at zero and a diffuse “slab” capturing nonzero values. Effectively, such an approach allows researchers to answer questions about particular individuals; specifically, “Who is average?”, in the sense of deviating from an average effect, such as the population-averaged slope. We begin with an illustrative example, where the spike-and-slab formulation is used to select random intercepts in logistic regression. This demonstrates the utility of the proposed methodology in a simple setting while also highlighting its flexibility in fitting different kinds of models. We then extend the approach to random slopes that capture experimental effects. In two cognitive tasks, we show that despite there being little variability in the slopes, there were many individual differences in performance. In two simulation studies, we assess the ability of the proposed method to correctly identify (non)average individuals without compromising the mixed-effects estimates. We conclude with future directions for the presented methodology.

Translational Abstract

Mixed-effects models are often used in psychology to study how individuals vary in their behavior, cognition, or personality. Typically, their use involves determining whether individuals differ from population-averaged values, and if they do, using predictors to explain why. We propose a method that provides a more nuanced perspective into individual differences by allowing a two-component representation of each individual—a “spike” and a “slab.” The spike component captures individuals who do not differ (or differ very little) from the population-averaged values whereas the slab component captures individuals who may differ substantially from the average. Thus, if an individual is best represented by the spike, then they can be considered to be “average,” and if they are best represented by the slab, then they can be considered to be “nonaverage.” In addition to allowing novel inferences in individual differences research, our proposed method may be useful for identifying individual who may differ from, say, an average intervention response. We display the utility and flexibility of the proposed method in two applied examples from cognitive psychology. The first lays out the foundation of the spike-and-slab method in a relatively simple setting. The second example provides an extension to common experimental settings in psychology. Results from two simulation studies suggest that the spike-and-slab method is capable of accurately classifying individuals without incurring any unwanted bias in the model estimates. We conclude with future directions for the presented methodology.

Keywords: Bayesian, mixed-effects model, spike-and-slab, individual differences

This article was published Online First November 3, 2022.

Josue E. Rodriguez https://orcid.org/0000-0002-9092-4869
Donald R. Williams https://orcid.org/0000-0001-6735-8785
Philippe Rast https://orcid.org/0000-0003-3630-6629
Donald R. Williams is now at NWEA, Portland, Oregon, United States.

Research reported in this article was supported by a National Science Foundation Graduate Research Fellowship under Grant 1650042 to Donald R. Williams and the National Institute on Aging of the National Institutes of Health under Award R01AG050720 to Philippe Rast. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. This article has been posted as a preprint at https://psyarxiv.com/4d9tv/. We thank Sara van Erp for helpful comments.

Correspondence concerning this article should be addressed to Josue E. Rodriguez, Department of Psychology, University of California, Davis, 1 Shields Avenue, Davis, CA 95616, United States. Email: jerrodriguez@ucdavis.edu

Mixed-effects models are being increasingly used in the social-behavior sciences. Their use spans many areas in psychology from observational inquiries that track individuals over an extended period of time, to controlled settings that can include hundreds of experimental trials for each person. Their rise in popularity is mainly due to their ability to partition and account for different sources of variation, for instance, in experimental effect (Aarts et al., 2014), stimulus type (Wolsiefer et al., 2017), or group membership (Raudenbush & Bryk, 2001). Adequately accounting for these sources of variability leads to the desired inference by ensuring that nominal error rates are maintained (Aarts et al., 2014; Barr et al., 2013; Judd et al., 2012; Williams et al., 2017; Wolsiefer et al., 2017). The idea is that variance components are often considered nuisance parameters that must be controlled or corrected for in order to draw valid inferences. Consequently, the primary inferential targets from mixed-effects models tend to be concerned with population averages, or fixed effects, while variance components play a secondary role. For instance, in a review of articles employing linear mixed-effects models, it was found that less than 10% reported the random effect variances (Meteyard & Davies, 2020), and similarly, only 32% of articles using generalized mixed models reported these variance components (Bono et al., 2021). On the other hand, however, these same sources of variations can provide valuable insights into individual differences in psychological processes (e.g., Haaf & Rouder, 2017; Liu et al., 2012; Williams et al., 2019).

When individual differences are of central interest, it is customary to test the variance of the random effects. For example, in determining whether there is variation between individuals in a random intercepts model, one would fit two (nested) models—one with and one without the random intercepts—and perform a likelihood ratio test. If the test is not rejected, one would settle with the simpler model without the random intercept term (i.e., no individual differences). Conversely, if the test is rejected then the random effect term is retained in the model. In order to explain the individual differences, the latter scenario may be followed up with the inclusion of covariates. In this work, we find a common ground between these two options. Because some individuals are best described by the fixed effect while others may differ drastically from it, we propose a method wherein some individual effects are allowed to deviate from the average and others are not. For example, it may be useful to describe which, if any, individuals depart from a typical learning trajectory (Estrada et al., 2018). As such, we propose a method that offers a more nuanced view of individual differences compared to the classical mixed effect vs fixed effect duality.

The need for more refined views of individual differences is reflected in recent efforts to extend methodological approaches for understanding individual differences. For instance Grice et al. (2020) point out that even though study results, when taken in aggregate, reflect theoretical expectations, it may be that only a few individuals actually behaved in the expected manner. One could imagine that an intervention is shown to alleviate depression on average, but this does not necessarily imply that the intervention is effective for a given individual. As a step toward understanding whether individuals behaved in a hypothesized manner, they propose adopting person-centered effect sizes, wherein effects are computed for each individual. These effects can in turn be used to quantify the proportion of observed effects that were in line with the hypothesized outcome.

In a similar spirit, Rouder and Haaf (2020) advocate for a Bayesian model comparison approach to distinguish situations where: all individuals have true effects in the same direction, individuals have true effects in differing directions, or all individual effects are equal to an average effect (also see Haaf & Rouder, 2017). This method involves fitting mixed-effects models that reflect each of these settings and comparing them. The central aim is to determine if there is support for individual differences in the data, and if so, which model best describes them.

To date, however, no general approach has been provided to formally address the individual in individual differences. For instance, the person-centered effect sizes are general in that they can be applied across a wide variety of settings, but are computed in a somewhat ad-hoc manner with a focus on description. The approach in Rouder and Haaf (2020) allows analysts to quantify evidence for whether individual differences align with a particular pattern, but ultimately relies on global descriptions of individual differences in linear models. Thus, it is desirable to have a framework that fulfills the desiderata of being applicable across the multitude of settings encountered in psychological science while simultaneously allowing researchers to rigorously evaluate individual effects.

Main Contribution

The main contribution of this work is the introduction of a Bayesian mixed-effects framework that may allow novel inferences in individual differences research. In mixed-effects models, there are fixed effects (averages across individuals), and there are random effects (deviations away from those averages). The main advantage of our proposed methodology is that it allows a more nuanced view of individual differences by quantifying evidence for or against *individual* random effects. In addition, because it can be fit using standard statistical software, it is flexible enough to be applied to a broad class of models (i.e., generalized linear mixed models).

With this framework we explicitly address the individual by providing a tool that is capable of answering which individuals are "average" and which ones are not. Intuitively, if β is a fixed effect, θ_j is the corresponding random effect for the jth individual, and $\beta_j = \beta + \theta_j$ is the total effect for the jth individual, then the problem we are interested in can be thought of as evaluating whether $\beta_j = \beta$ or $\beta_j \neq \beta$. As for implementation, the models we describe in this paper can easily be fit in the common programming languages R (R Core Team, 2021) and Python (Van Rossum & Drake, 2009), or by using the R package SSranef.[1]

To answer the question of who is "average," we build upon spike-and-slab priors for Bayesian variable selection (George & McCulloch, 1993; Kuo & Mallick, 1998; Mitchell & Beauchamp, 1988). Traditionally used in the canonical regression setting to select predictors that are likely to have a nonzero effect, our innovation is to apply the spike-and-slab to select which *random effects* are likely nonzero in a mixed-effects model. A similar approach has been applied in psychological settings (e.g., Williams et al., 2021), but was restricted to random intercepts in linear

[1] The SSranef R package can be downloaded from GitHub at https://github.com/josue-rodriguez/SSranef. An example illustrating how to use SSranef can be found in Appendix A.

mixed models whereas, in practice, the primary interest is often the random slopes. Further, it is common to fit models with non-Gaussian likelihoods (e.g., mixed-effects logistic regression). Thus, a novel aspect of this work is the extension of the spike-and-slab to random effects on slopes and generalized linear mixed models.

Overview

In what follows, we first present a motivating example where we introduce the central ideas underlying the spike-and-slab prior in the context of a generalized mixed-effects model. We show the value in using the spike-and-slab on random intercepts and that it is trivial for this approach to be applied to a variety of model types. We then demonstrate how the idea of random effect selection can be extended to random slopes. This allows researchers to, for example, answer how many individuals differed from a common experimental effect. This approach is illustrated in two empirical psychological data sets where we show how individual differences in the random slopes can be comprehensively disentangled. In two simulation studies, we assess the ability of our proposed method to correctly identify (non)average individuals without compromising the mixed-effects estimates. We conclude with a discussion on the implications of the current work and future directions.

Background

We employ the spike-and-slab approach for variable selection. In this approach, the selection problem is formulated in terms of a two-component mixture: (a) a "spike" that is either a distribution centered narrowly around zero (George & McCulloch, 1993, 1997) or a Dirac measure at zero (Kuo & Mallick, 1998; Mitchell & Beauchamp, 1988); and (b) a diffuse "slab" component surrounding zero. The former allows the shrinkage of small effects to zero and the latter prevents heavy shrinkage of larger effects. A central aspect of this approach is the addition of an indicator variable (Kuo & Mallick, 1998), which allows for switching between the spike and the slab throughout the Markov chain Monte Carlo (MCMC) sampling process (i.e., transdimensional sampling; Heck et al., 2019). The proportion of MCMC samples spent in each component can then be used to approximate the respective posterior model probabilities or the marginal Bayes factor for whether an effect should be included. In the context of random effects selection, this Bayes factor expresses the evidence for whether the random effect for a given individual should be included in the model. Interested readers can find an excellent introduction to the spike-and-slab prior for psychology in Rouder et al. (2018) and in-depth overview of its various specifications in O'Hara and Sillanpää (2009).

Importantly, much of the literature on spike-and-slab priors has been concerned with model selection and comparison (George & McCulloch, 1997; O'Hara & Sillanpää, 2009). This is distinct from our application in this article as we do not focus on model selection in a traditional sense. Our goal is not to make judgments with respect to quality of fit among models with different variables, prior distributions, or functional forms, but rather we seek to use spike-and-slab priors as a means of understanding which individuals' effects deviate from a population-average estimate.

Illustrative Example

We begin our exposition by considering the work of Frühwirth-Schnatter and Wagner (2011) who used spike-and-slab priors with the overarching goal of "[making] unit-specific selection of random effects in order to identify units which are "average" in the sense that they do not deviate from the overall mean" (p. 2).

Specifically, they provided examples of random effect selection with a focus on logistic models. However, their approach relied on custom MCMC sampling schemes, rendering the techniques inaccessible to all but those who are comfortable implementing the algorithms on their own. Williams et al. (2021) introduced the idea of selecting unit-specific random effects to psychology with the goal of determining individual reliability, but they did not consider models outside of a classical random intercepts model. Because, to our knowledge, these are the only works to consider random effect selection, we view this as a good place to begin our exposition of the spike-and-slab. Using a random intercepts logistic regression model, we highlight key ideas relevant to our approach for random effect selection.

Model Formulation

For our illustrative example we use data from a linguistics experiment that were first reported in Caplan et al. (2021, Experiment 1). The participants (N = 128) in this study were presented with acoustically ambiguous audio involving minimal pairs of words (e.g., time/dime) along with disambiguating information that biased the audio to be interpreted as /t/ or /d/. The outcome for the ith trial and jth person is coded as a 1 or a 0 and represents whether participants heard a /t/ (1) or a /d/ (0) for a given word during the test phase (see original text for full details). For illustrative purposes, we adopt a simpler version of the full analysis in that we only consider a random intercept without covariates or additional random effects. To facilitate spike-and-slab selection, we employ the noncentered parameterization (Papaspiliopoulos et al., 2007), that is,

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}(\pi_j) \\ \text{logit}(\pi_j) &= \alpha + \tau z_j \\ \alpha &\sim \text{Normal}(0,1) \\ \tau &\sim \text{St}^{+}(\nu = 3, 0, 1) \\ z_j &\sim \text{Normal}(0,1), \end{aligned} \tag{1}$$

where y_{ij} is the outcome, α is the overall intercept, τ is the standard deviation of the random effects, z_j is a standardized effect size, and the product τz_j constitutes the random effect. Here, we are not modeling the random effects directly, but rather inferring them from a latent variable z_j. There are two main reasons for this: (a) it may lead to more efficient sampling of the posterior and (b) it allows us to think about the random effects in terms of standardized effect sizes. Further, we set standard normal priors for α and z_j, and a Half-Student-t distribution with three degrees of freedom for τ. Our choice for the Half-Student-t distribution stems from it having better properties than common alternatives for variance parameters in hierarchical models (e.g., inverse-gamma; Gelman, 2006). The model in Equation 1 estimates the baseline log-odds of hearing a /t/ (intercept), but allows for each individual to deviate away from it (random effect).

In such an analysis, it might be natural to ask whether each individual does indeed deviate from the overall log-odds, α, in hearing a /t/. This question can be addressed by adding an indicator variable $\gamma_j \in \{0, 1\}$ to the above model that governs, for each individual, whether the random effect is in the spike ($\gamma_j = 0$) or the slab ($\gamma_j = 1$) portion of the model in each MCMC iteration. Introducing this variable only requires the following modifications to Equation 1:

$$\begin{aligned} \text{logit}(\pi_j) &= \alpha + \tau(z_j \gamma_j) \\ \gamma_j &\sim \text{Bernoulli}(\delta) \end{aligned} \tag{2}$$

while everything else remains the same. In Equation 2, δ represents the *prior* inclusion probability, or the a priori probability that the *j*th random effect is nonzero. Choosing $\delta = .5$ expresses a lack of a priori preference for whether a random effect should be included or excluded, and it is the choice we make throughout this article. Notice that when $\gamma_j = 0$, the random effect for the *j*th individual random effect drops out of the model, and when $\gamma_j = 1$, it is retained. If there is prior information that indicates whether individuals are more or less likely to deviate away the average, then this information can be included in the analysis by modifying δ to be greater than or less than .5.

The proportion of MCMC samples in which γ_j is equal to one is referred to as the *posterior* inclusion probability (PIP) of the *j*th random effect,

$$\Pr(\gamma_j = 1 \mid \mathbf{Y}) = \frac{1}{S}\sum_{s=1}^{S} \gamma_j^{(s)}, \tag{3}$$

where $s = 1, \ldots, S$ indexes the MCMC samples and $\mathbf{Y}$ denotes the data. When there is strong support for including the *j*th random effect, its PIP will be large, and when there is little support for inclusion, its PIP will be small. PIPs of 0 and 1 indicate complete posterior support for excluding and including the *j*th random effect, respectively. Additionally, Bayes factors (Kass & Raftery, 1995) can be computed based on PIPs. Assuming equal prior odds, the Bayes factor in favor of the random effect being nonzero rather than zero can be calculated as

$$BF_{10} = \frac{\Pr(\gamma_j = 1 \mid \mathbf{Y})}{1 - \Pr(\gamma_j = 1 \mid \mathbf{Y})}. \tag{4}$$

The ability to compute posterior inclusion probabilities and Bayes factors allows for the direct quantification of evidence for whether an individual's baseline log-odds are different than the "average" baseline log-odds of hearing a /t/.

Although it is not the only way to formulate a spike-and-slab prior in a Bayesian model (O'Hara & Sillanpää, 2009), our approach carries some distinct advantages. First, by using a point-mass at zero for the spike instead of a continuous distribution with small variance, we explicitly consider whether a given random effect is equal to zero instead of just nearly zero. Further, the prior probability of drawing a one for γ_j (i.e., the prior inclusion probability) is fixed at .5. This is equivalent to setting equal prior odds for whether a random effect is nonzero or zero, and simplifies the expression for the Bayes factor. Note that allowing the prior probability δ to be a random variable by endowing it with a prior (e.g., Beta) may result in superior selection for point-mass spikes (Ley & Steel, 2009). For these reasons, the above formulation of the spike-and-slab is the one we use throughout the article.

Software and Estimation

We fit the model using the JAGS language in R (Plummer, 2003) because of its ability to easily fit spike-and-slab models (Ntzoufras, 2002; O'Hara & Sillanpää, 2009).[2] The fitted model used four chains of 25,000 iterations after a burn-in period of 5,000 iterations which resulted in a total of 100,000 samples from the posterior distribution. This number of samples provided a good quality of the parameter estimates (all $\hat{R}$s = 1; Brooks & Gelman, 1998).

Results

The results are displayed in Figure 1. Panel A shows the prior distribution for the random effects and Panel B shows the posterior for the random effect of the 56th and 78th participants, respectively. Note that the spike (black arrow) and slab (blue bars) both constitute roughly half of the prior density. Panel C displays the point estimates of the random effects for all 128 participants and their respective 90% credible intervals (CrIs). The individuals from panel B are represented by the green (Participant 78) and orange (Participant 56) dots.

Recall that the goal of fitting this model was to determine the evidence for whether a given individual deviates from the overall log-odds, or intercept. If an individual does not differ from the intercept, then most of the of posterior mass should be in the spike for the random effect. If an individual does differ from the intercept, then there should be a lot of posterior mass in the slab. This can be clearly seen in Figure 1 where most of the posterior mass is in the spike for participant 56 and, conversely, none at all for participant 78. For the former, there was a .23 posterior inclusion probability, or a Bayes factor of roughly 3 in favor of the spike. This can be considered moderate evidence in favor of the participant being "average" (Lee & Wagenmakers, 2013). For the latter, the posterior inclusion probability was 1 and is equivalent to a Bayes factor of infinity that this individual differs from the "average."

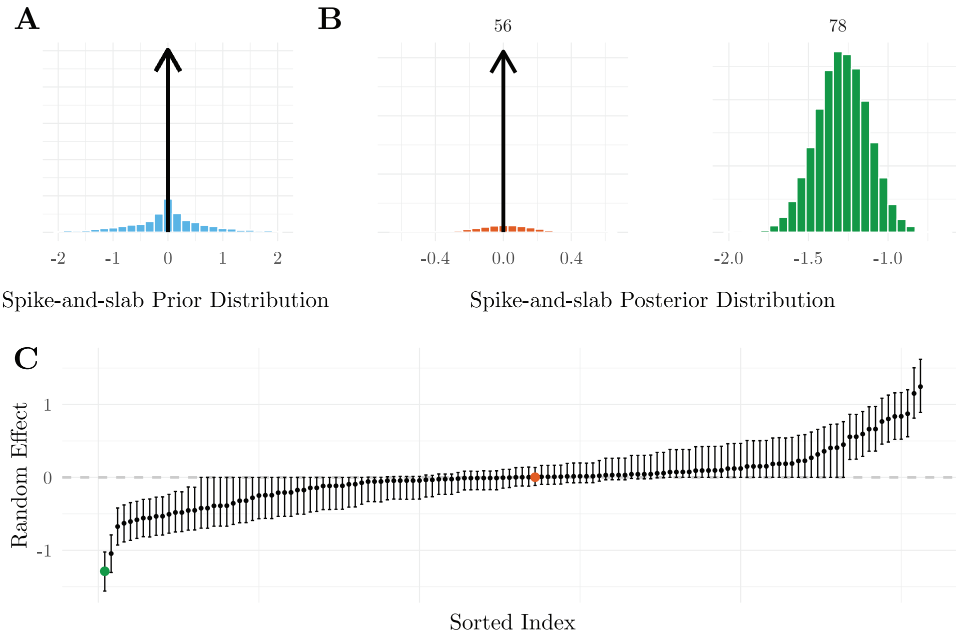
The shapes of these posteriors have a straightforward relationship with the size of the random effect. This correspondence is shown in Figure 1 (Panel C) where the orange dot (Participant 56) is near zero and the green dot (Participant 78) is far away from zero. This makes sense intuitively; if a random effect is near zero, then there will be little to no evidence that a participant differs from the intercept, and conversely, there will be stronger evidence that a participant differs from the intercept with larger random effects.

Summary

The purpose of this illustrative example was to build the foundation for the following methodology. We highlighted the central idea behind the spike-and-slab prior, and in particular, how it can be leveraged to select individual random effects. The results indicated that this methodology can be profitably applied to determine which individuals differ from the overall intercept in a logistic

[2] All code to reproduce the analyses and figures in this article are available on the Open Science Framework at https://osf.io/n2z49/.

Figure 1
Illustration of the Spike-and-Slab Prior Distribution



Note. (A) The point-mass spike-and-slab prior distribution. The spike (arrow) and the slab (blue bars) each take up half the prior density. When a random effect is sampled from the spike, it is set zero and the effect for that individual is equal to the fixed effect. When it is sampled from the slab, it will take a nonzero value and the individual effect will deviate from the fixed effect. The proportion of MCMC iterations that a random effect is sampled from the slab is called its posterior inclusion probability (PIP). (B) The posterior distribution for the random effects of the 56th and 78th participant. For the former, the majority of the posterior mass is in the spike (PIP = 0.23) where there is little mass in the slab (orange bars). For, the latter the entire posterior is in the slab (green bars, PIP = 1). Thus, Participant 56 can be considered "average," and Participant 78 can be considered not "average." (C) Posterior means and respective 90% credible intervals for the random effects. The orange point (Participant 56) is centered near zero and the green point (Participant 87) is far from zero. This matches the corresponding posterior mass in the spike for each of these random effects. See the online article for the color version of this figure.

regression setting. The remainder of this article will extend this idea to include random slopes to determine whether individuals differ from the average experimental effect.

Extension to Random Slopes

In psychology, it is more common for the random slopes to be of focal interest, not the random intercepts. The reason for this is that the slopes often corresponds to the effect of condition or manipulation in experimental settings. Accordingly, random effects in the slope encode individual differences in experimental effects. Thus, we seek to extend the application of the spike-and-slab prior to the random effects in slopes. Placing the spike-and-slab on the slopes allows evidence to be obtained for which individuals differ from the average experimental effect and which do not. As above, our exposition of this extension will be through applied examples.

It has recently been argued that there is low reliability in popular cognitive tasks for studying individual differences (Hedge et al., 2018; Rouder et al., 2019). The main explanation for low reliability among such tasks is that there exists little individual differences. In this context, individual differences are defined in reference to the ratio of between-subjects variance to total variance. In what follows, we are not interested in individual differences in this sense, but whether there are individual differences in these tasks with respect to who deviates from the overall experimental effect, and then determining the kind of insights that may follow.

Empirical Application

We apply the proposed methodology to data from two classical inhibition tasks. These data were first analyzed in Hedge et al. (2018) and again in Rouder et al. (2019).

Dataset 1: Stroop Task

Participants ($N = 47$) responded to the color of a centrally presented word which was red, blue, green, or yellow. The word could be the same as the font color (congruent condition), different from the font color (incongruent condition), or one of four non-color words (neutral condition). Each participant completed 240 trials for each condition with the primary outcome being reaction time (RT). For illustrative purposes, we focus simply on the congruent and incongruent conditions.

Dataset 2: Flanker Task

The same 47 participants responded to the direction of a centrally presented arrow (left or right). On each trial, the central arrow was flanked above and below by two other symbols. Flanking stimuli were arrows pointing in the same direction as the central arrow (congruent condition), arrows in the opposite direction as the central arrow (incongruent condition), or straight lines (neutral condition). Again, each participant completed 240 trials for each condition and the primary outcome was RT. As above, we only give consideration to the congruent and incongruent conditions.

Model Formulation

Because these tasks are similar both in what they are thought to measure and in their design, each dataset contains the same variables on which we focus: outcome (RT) and condition (in/congruent). Accordingly, we define a single model formulation that can be seamlessly applied to each dataset without modifying anything except for the data. For the *i*th trial and the *j*th person, we can define the likelihood for the RT as

$$\begin{aligned} y_{ij} &\sim \text{Normal}(\alpha_j + x_{ij}\beta_j, \sigma^2) \\ \alpha_j &= \alpha + \theta_{1j} \\ \beta_j &= \beta + \theta_{2j}, \end{aligned} \tag{5}$$

where for the *j*th person, α_j is the random intercept and encodes the average response time for the congruent condition, and β_j is the random slope which captures the difference in response time in the incongruent condition, relative to the congruent condition. The term x_{ij} encodes the condition (0 = congruent; 1 = congruent), and σ^2 is the residual variance. The terms θ_{1j} and θ_{2j} indicate the random effects for the intercept and slope, respectively.

For the model parameters defined in Equation 5, we set the priors as follows:

$$\begin{aligned} \alpha, \beta &\sim \text{Normal}(0, 1) \\ \boldsymbol{\theta}_j &\sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}) \\ \boldsymbol{\Sigma} &= \boldsymbol{\tau}\boldsymbol{\Omega}\boldsymbol{\tau} \\ \boldsymbol{\Omega} &\sim \text{LKJ}(\eta = 1) \\ \sigma, \tau_{11}, \tau_{22} &\sim \text{St}^+(\nu = 3, 0, 1). \end{aligned} \tag{6}$$

Here, we place uninformative normal priors over the fixed effects and a multivariate normal prior with covariance matrix $\boldsymbol{\Sigma}$ for the random effects. We model the covariance matrix using the separation strategy discussed in Barnard et al. (2000) where $\boldsymbol{\Omega}$ is a 2×2 correlation matrix of the random effects and $\boldsymbol{\tau}$ is a 2×2 diagonal matrix whose elements are the standard deviations of the random effects. The prior for the correlation matrix is the LKJ distribution (Lewandowski et al., 2009) and is governed by a single parameter η. Setting $\eta = 1$ places a uniform prior over all correlation matrices. We set Half-Student-*t* for all variance parameters for the reasons discussed in the Illustrative Example section.

Of central interest is the parameter β_j, which corresponds to the experimental effect for the *j*th person. Recall that we want to know whether each individual differs from the common effect, β, and that we can use a spike-and-slab prior to answer this question. Thus, we can modify the above to include a spike-and-slab prior on the random slopes

$$\begin{aligned} \beta_j &= \beta + \theta_{2j}\gamma_j \\ \gamma_j &\sim \text{Bernoulli}(\delta), \end{aligned} \tag{7}$$

where $\delta = .5$ and everything else remains the same.

Model Selection

Up to this point, we have not discussed a decision rule for actually determining which individuals differ from the average effect. This is because Bayesian inference is not focused on making discrete choices, but rather considering the weight of evidence (Morey et al., 2016). In any case, there are times when it is desirable to do so. For instance, in addition to reporting random effect variances, one can report for example that 30% of the random effects differed from the average effect. Reporting such a number is in the same spirit as the metrics described in Grice et al. (2020), but supported by formal evidence (i.e., posterior inclusion probability). This might be especially insightful in situations with low between-person variance, a scenario that typically implies a lack of individual differences. This type of information can also be useful in other fields such as clinical or educational psychology, where one can identify a subset of individuals who respond differently to an intervention compared with the average response. Identifying individuals who display unusual behavior via random effects can be extended to models of variability as well (e.g., Rast & Ferrer, 2018). For example, in cognitive aging research, random effects in the residual variance can be used to capture differences in behavioral "consistency" of cognitive ability (Rast & Zimprich, 2011; Watts et al., 2016). Here, identifying individuals with above or below average residual variance could serve as an early warning sign to the onset of Alzheimer's disease (Lövdén et al., 2013; MacDonald et al., 2008).

Because in our above example we place the spike-and-slab prior on $N = 47$ random effects, there are 2^{47} distinct combinations of random effects that can be considered for inclusion in the final model. That is, there are 2^{47} possible models from which to choose. Thus, the issue that presents itself is how to choose which model should be used to determine who is "average." An intuitive choice would be to select the highest probability model (HPM), or the model containing the combination of random effects selected most frequently throughout the MCMC sampling process. In fact, it is the *median probability model* (MPM, Barbieri & Berger, 2004; Barbieri et al., 2021) that is more often considered. The MPM, which is used in the present article, is defined to be the one including only those random effects with posterior inclusion probabilities (Equation 3) of at least 0.5. Several motivations underlie the MPM, including that it is the best single-model approximation to Bayesian model averaging and it is optimally predictive for linear models with respect to squared error loss under orthogonal designs. However, this does not mean the HPM should never be used. Indeed, the HPM can be used when the goal is explicitly to compute a Bayes factor of interest for hypothesis testing. That is, if one has a priori predictions about which individuals differ from the fixed effect. Further, once individuals have been classified as "average" or not, then it is straightforward to compute the proportion of the sample that differed from the common effect.

Software and Estimation

We fit the model above to both the Stroop and Flanker data using the pymc3 (Salvatier et al., 2016) package in the Python programming language (Van Rossum & Drake, 2009). This was primarily because it allows the use of more efficient MCMC sampling schemes (e.g., Hoffman & Gelman, 2011) while retaining the ability to accommodate the point-mass spike-and-slab prior.[3] The fitted models used four chains of 10,000 iterations after a tuning period of 2,000 iterations which resulted in a total of 40,000 samples from the posterior distribution. This number of samples provided a good quality of the parameter estimates (all $\hat{R}$s = 1).

Results

The main results are displayed in Figure 2. Panel A shows the point estimates of the slope random effects for all 47 participants and their respective 90% CrIs. Throughout the rest of this section, we will simply use "random effects" as shorthand for the slope random effect. Panel B displays the PIPs as a function of the magnitude of the random effect. Upon visual inspection, it is easy to see which individuals have more evidence supporting that they differ from the average experimental effect. The PIPs make a V-shape in that they decrease as the magnitude of the effect approaches zero and increase again as they move away from zero. This is again unsurprising. Individuals with larger random effects should have more evidence to support that they differ from the average effect.

For the Stroop task, the mean posterior estimate for the overall experimental effect, β, was .07 and had a corresponding 90% CrI of [.062, .076]. That is, on average, participants' RT was slower by .07 s in the presence of incongruent stimuli. Notably, the mean posterior estimates for the random effects ranged from −.02 to .07, and their corresponding PIPs ranged from .22 to .99 (Bayes factors of .28 to over 13,000), in support of including the random effect. This spread of PIPs indicates considerable fluctuations in the level of support for whether individuals differ from the average experimental effect. They span from "moderate" evidence in favor of belonging to the average experimental effect on one end to "extreme" evidence in favor of different from it on the other (Lee & Wagenmakers, 2013). This spread was even wider in the Flanker task, where the PIPs covered values from .19 to 1.

As previously mentioned, it may sometimes be desirable to categorize individuals as being "average" or not. When using the median probability model, individuals with PIPs over .5 can be thought of as being different from the average effect. In Figure 2 (Panel B), these two groups are separated by the dark dotted gray line. It is intriguing that for both tasks, quite a few points lie above this line. Specifically, 12 and 13 participants are above this line for the Stroop and Flanker tasks, respectively.[4] In other words, there is evidence that despite the belief that few individual differences exist in these kind of data, over a quarter of the sample diverged from the average experimental effect in each task.

Taken together, these results not only attest to the existence of individual differences in these two experiments, but speak to which individuals (and how many) differed from the average effect.

Individual Performance Across Tasks

The Stroop and Flanker tasks have long been considered to be measures of inhibition (Friedman et al., 2004). It is consequently natural to think that individuals who differ from the average experimental effect in one task should also differ from the average effect in the other. In contrast, recent work has suggested that the correlations among inhibition tasks are low (Hedge et al., 2018; Rouder et al., 2019). That is to say, that performance on a given task is not necessarily predictive of performance on another. Because we examine individual differences in the sense of differing from a fixed effect and not in terms of the amount of variance, we look at whether the PIPs were comparable for individuals across tasks. Note that it would be possible to fit a multivariate model with the reaction times for both tasks as the outcome, and directly apply the spike-and-slab formulation to the random slopes for each task. In order to keep the exposition manageable, we opt for simple description.

Figure 3 displays a funnel plot containing the PIPs of the random slope effects for individuals on both tasks, sorted in descending order of PIPs for the Stroop model. The idea here is that if performance on these tasks are related, then we should see a funnel shape that starts wide at the top (i.e., individuals who had large PIPs in both tasks) and becomes narrow at the bottom (i.e., individuals who had small PIPs for both tasks). However, upon visual inspection, there is no apparent relation between the PIPs. For instance, Participant 24 had a PIP of .99 for their random effect in the Stroop model, but a PIP of .31 in the Flanker model. On the other hand, Participant 35 had PIPs of near 1 on both tasks. Hence, whether an individual differs from the average experimental effect in one task may not be predictive of whether they differ from the average experimental effect in another.

Posterior Predictive Check

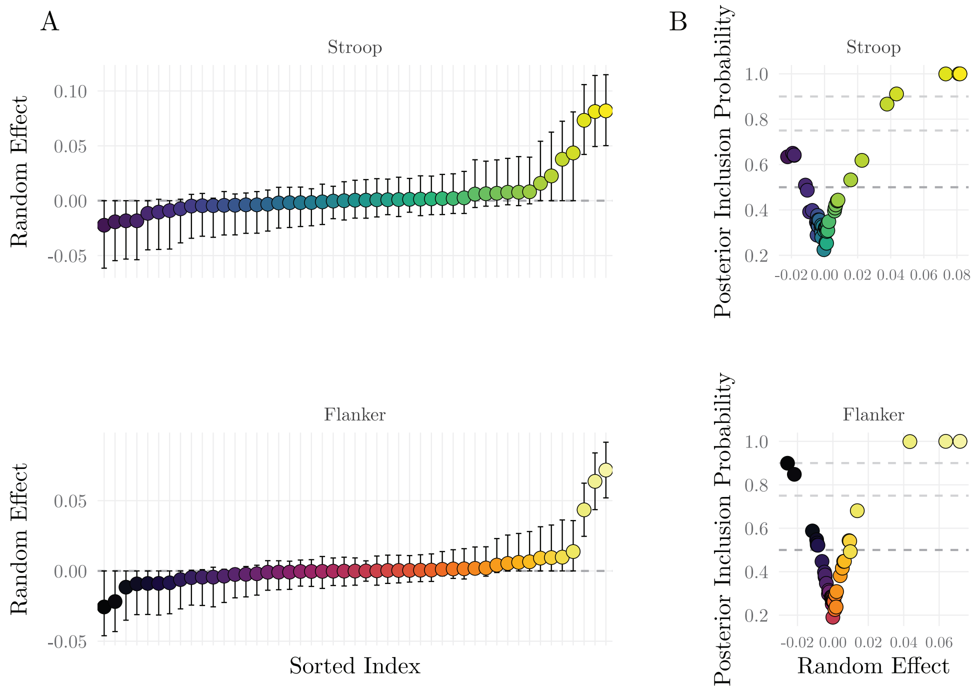
Lastly, an important aspect of Bayesian inference is model checking. This is typically done with the posterior predictive distribution (Gelman et al., 1996; Meng, 1994). The main idea behind a "posterior predictive check" is that data generated from the model should resemble the observed data. The posterior predictive check thus entails generating data sets from the predictive distribution of the fitted model and comparing them to the observed dataset in order to evaluate the model's goodness-of-fit. Importantly, posterior predictive checks should capture aspects of the model which are of particular interest (Gelman & Hill, 2006, p. 514).

A principal quantity here is the Bayesian *p*-value, which can be defined as the proportion of times a quantity of interest calculated from the posterior predictive distribution exceeds the observed quantity. If the model is adequately capturing the data, then the *p*-values should be relatively close to .5 (Gelman, 2013). Values near 0 or 1 would indicate systematic misfits of the model to the data. Because the models we fit are focused on the mean differences in RT between two experimental conditions, as opposed to,

[3] This model converged in JAGS without issues, but we fit it in pymc3 to demonstrate how to employ these models when more efficient samplers are desired.

[4] We also examined PIP cut-offs of 0.75 and 0.9 (light gray dotted lines). For the former, this corresponds to 11% of the sample differing from the average experimental affect and roughly 7% for the latter.

Figure 2
Main Results From the Empirical Application



Note. (A) Posterior means and corresponding 90% CrIs for the random effects for the slopes (or experimental effects) in the Stroop and Flanker data, sorted in ascending order. (B) The corresponding posterior inclusion probabilities for each random effect. The dark gray dotted line indicates a PIP of 0.5. The two light gray lines denoted PIPs of 0.75 and 0.90. Random effects that are closer to zero have lower PIPs. If one were to use the median probability model as a decision then everyone above the dotted gray line would be considered as different than "average." For both the Stroop and Flanker tasks, over 25% of the points lie above the dotted line. This clearly demonstrates individual differences in these tasks. Across both panels, distinctly colored points refer to the same random effect. See the online article for the color version of this figure.

say, the shape of the RT distributions, we perform a posterior predictive check on the subject-specific mean differences. If the model adequately captures these mean differences, the p-values should be dispersed around .5.

For each of 2,000 draws from the predictive distribution, we calculated the mean difference in RT between conditions for each of the 47 subjects. The resulting values were then compared with the empirical mean differences. The results of the posterior predictive checks are shown in Figure 4. The empirical mean differences are represented by red points and posterior predictive mean differences are indicated by the black points. The numbers on the right-hand side are the corresponding Bayesian p-values. Across both tasks, the p-values span from .16 to .84, with most of them between .25 and .75. These results can be viewed as evidence that the fitted model adequately captures mean differences between conditions in the data and "passes" this posterior predictive check.

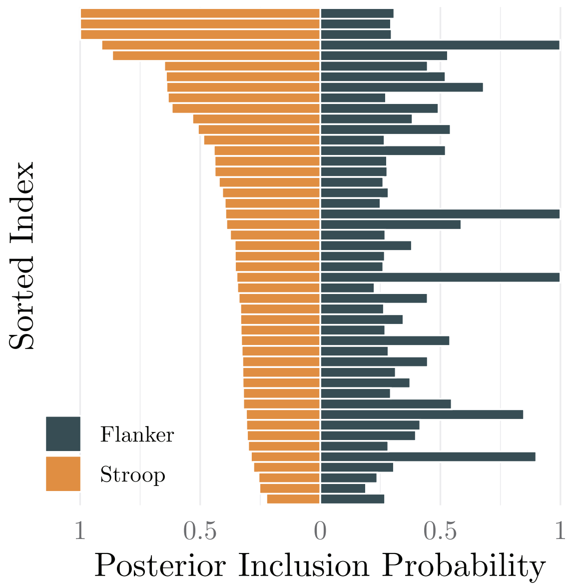
Simulation Studies

Up to this point, we have demonstrated how the spike-and-slab prior can be applied to gain new insights into individual differences in psychology. We now focus on better understanding the properties of the spike-and-slab prior when placed on random effects by way of two simulation studies. The first aims to support our claim that the spike-and-slab prior on the random effects is indeed capable of correctly identifying those who differ from the average and those who do not. In the second simulation study, we address a potential issue noted by reviewers. As shrinkage is already an inherent part of mixed-effects models (Gelman & Hill, 2006; Raudenbush & Bryk, 2001), the inclusion of a spike-and-slab prior could incur "double shrinkage." That is, the random effects may be biased due to shrinkage in both the slab (as in a typical mixed-effects model) and spike components of the prior. The second simulation study investigates this possibility. To situate the findings within a familiar context, we include a standard mixed-effects model (i.e., a normal prior on the random effects) for comparison in both simulation studies.

Study 1

The goal of this study was to assess the classification performance of the spike-and-slab prior with respect to average and nonaverage random effects. Accordingly, we simulated data for a random intercepts model with n = 100 units of interest (e.g., people) and varied the number of observations per unit n_j = 5, 10, 25.

Figure 3
PIPs Across the Stroop and Flanker Tasks



Note. Funnel plot of the PIPs in the Stroop and Flanker tasks. For each individual, the orange bar indicates the PIP for their random effect in the Stroop task. The opposite-side blue bar indicates that individual's PIP for their random effect in the Flanker task. Because the plot does not produce a funnel shape, this suggests that whether an individual deviated from the average experimental effect in one task may not be predictive of whether they deviated in the other. Hence, although there were individual differences insofar as who was "average" in each task, it seems that reliability was low. See the online article for the color version of this figure.

For each $j = 1, \ldots, n$ unit, each $i = 1, \ldots, n_j$ observation, y_{ij}, was generated as

$$\begin{aligned} y_{ij} &= \alpha_j + \epsilon_{ij} \\ \alpha_j &= \alpha + \theta_j \\ \epsilon_{ij} &\sim \text{Normal}(0, 1), \end{aligned} \tag{8}$$

where $\alpha = 1$ and θ_j captures the random effect for the jth person. The θ_j were systematically varied to be either 0, +1, or −1. The proportion of random effects that were exactly zero was set to be either .94, .74, or .5. The remaining random effects were set to either +1 or −1 in equal proportions.[5] These proportions translate to between-unit variances τ^2 of approximately .05, .2, and .25. Further, by setting $\sigma^2 = 1$, the resulting intraclass correlation coefficients (ICCs) are approximately .05, .15, and .2, respectively, where the ICC is defined as (Raudenbush & Bryk, 2001)

$$\text{ICC} = \frac{\tau^2}{\tau^2 + \sigma^2}. \tag{9}$$

The ICC plays a key role in mixed-effects models because it captures test-retest and interrater reliability (Shrout & Fleiss, 1979; Weir, 2005), and is the proportion of total variance accounted for by the between-group variance, τ^2. As we will discuss below, the ICC is also of particular interest because it determines, in part, the amount of shrinkage that occurs.

For each of 200 iterations, data were generated as previously described and two mixed-effects model were fit: one employing the spike-and-slab prior on the random effects and another using the customary normal prior on the random effects. For the spike, a point-mass at zero was used whereas a diffuse normal distribution was used for the slab (as in Equations 1 and 2). The latter distribution was also used as the prior for the random effects in the standard mixed-effects model. For both models, the likelihood and remaining priors were specified as

$$\begin{aligned} y_{ij} &\sim \text{Normal}(\alpha + \theta_j, \sigma^2) \\ \alpha &\sim \text{Normal}(0, 1) \\ \sigma, \tau &\sim \text{St}^+(\nu = 3, 0, 1). \end{aligned} \tag{10}$$

All models were fit in R using the JAGS language. The fitted models used four chains of 5,000 iterations after a burn-in period of the same length.

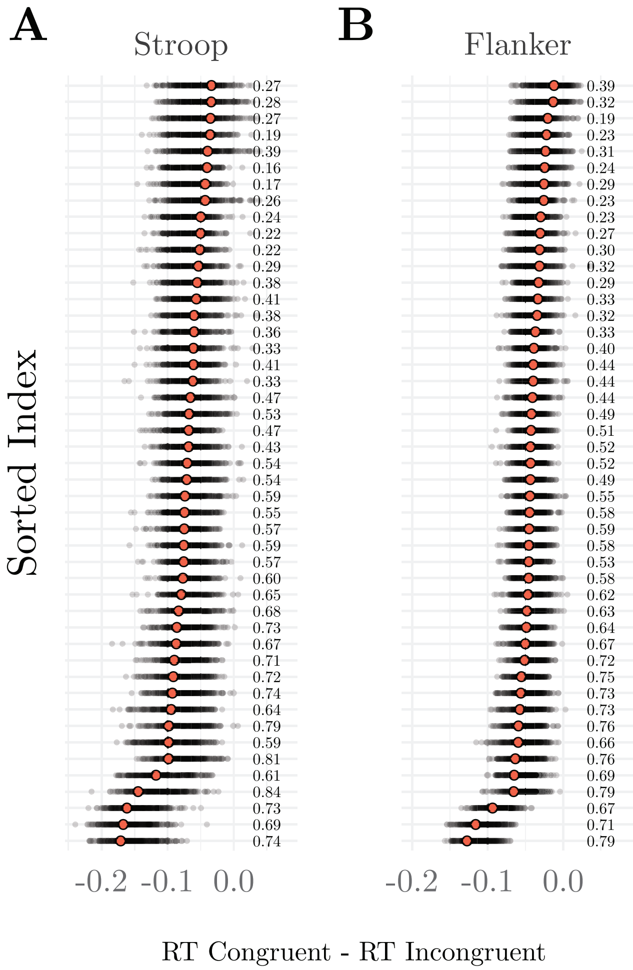
Once the models were fit, each random effect θ_j was classified as average or differing from the average. A correct classification occurred when a nonzero random effect was included in the final model or when a zero random effect was excluded. For the model with the spike-and-slab prior, we considered two thresholds for inclusion: (a) a posterior inclusion probability (PIP) of .5 (i.e., the median probability model); and (b) a PIP of .75 (i.e., a Bayes factor of 3). For the model with the normal prior on the random effects, a 90% credible interval was used to classify the random effects. If the interval for the jth random effect included 0, then it was excluded from the final model, and included otherwise. Model performance was considered in terms of specificity,[6] the proportion of truly zero random effects that were correctly classified, and sensitivity, the proportion of truly nonzero random effects that were correctly classified.

The results are displayed in Figure 5. Panel A displays the average sensitivity for the random effects across ICCs, observations per unit, and priors. Across ICCs, all priors tended toward a sensitivity rate of 1; however, there were some discrepancies in conditions with few observations per unit. When n_j was either five or ten, the spike-and-slab prior using a PIP of .5 as the inclusion threshold ($SS_{0.5}$) was superior to both the spike-and-slab model using a PIP of .75 ($SS_{0.75}$) and the normal model using the 90% CrI. Interestingly, with relatively little between-unit variance (ICC = .05) and few units per observation ($n_j = 5$), the $SS_{0.5}$ model was 3.5 and 11 times more accurate in detecting nonaverage units than the $SS_{0.75}$ and normal models, respectively. This suggests that the spike-and-slab may be fruitfully applied to detect nonaverage individuals even when between-person variance is low. In sum, with sufficient observations, all models performed comparably well in detecting nonaverage units, but the $SS_{0.5}$ model (i.e., median probability

[5] These values were chosen so that approximately, 50%, 75%, and 95% of the random effects were exactly zero, but an even number of nonzero random effects remained.

[6] Note that (1 – specificity) is the false positive rate.

Figure 4
Posterior Predictive Check for the Stroop and Flanker Tasks



Note. A posterior predictive check on the mean difference in reaction time between the congruent and incongruent conditions for each individual in the Stroop and Flanker tasks. The red points indicate the observed mean difference in reaction time and the black dots are draws from the posterior predictive distribution. The numbers on the right-hand side of each panel correspond to the Bayesian *p*-value for these predictive checks. Bayesian *p*-values that are closer to 0.5 than 0 or 1 suggest the model is successfully capturing the mean differences. As can be seen, the spike-and-slab formulation for these models adequately captures the mean differences. See the online article for the color version of this figure.

model) was superior when either the ICC or number of observations was small.[7]

The average specificity is similarly displayed in Panel B. Across all conditions, the worst specificity was observed for the $SS_{0.50}$ model with the ICC set to .05 and n_j set to 5. Here, the specificity for the $SS_{0.50}$ model was .75, while it was .99 for both the $SS_{0.75}$ and the normal model. As n_j increased, specificity for the normal model decreased and stabilized near a specificity of .9, or a false positive rate of .1. This is unsurprising as the specificity for credible interval approach should be roughly equal to the width of the credible interval (Rubin, 1984). In contrast, the specificity for both spike-and-slab models were stable near one or tended to one. This finding hints at the model selection consistency property the spike-and-slab prior. Recall that, assuming prior equal odds, the PIP for each random effect corresponds to the Bayes factor (see Equation 4). Bayes factors tend to infinity and posterior model probabilities tend to one in favor of the "true" model as the sample size increases (O'Hagan, 1995). Therefore, with a sufficiently large sample size, the spike-and-slab approach will completely avoid false positives and false negatives, whereas the same cannot be said for random effect selection under the credible interval strategy.

Further, the classification results help clarify the trade-off in choosing different values for the PIP. Using a lower threshold, such as PIP $\geq$ .5, results in better sensitivity (i.e., detecting who *is not* average) at the cost of lower specificity (detecting who *is* average). As the PIP threshold increases (e.g., PIP $\geq$ .75), this relationship reverses. Although not included in our results above, a similar relationship would be observed for the credible interval approach. Using a more narrow credible interval would result in higher sensitivity, at the cost of lower specificity, and vice versa for a wider interval. In studying variable selection, Li and Lin (2010) found that for a credible interval approach, a 50% CrI provided the best balance between sensitivity and specificity. Though such narrow intervals are not commonly used in psychological science, Appendix B contains the results from Study 1 using 50% CrIs instead of a 90% CrIs, but they do not shift the main conclusions from our results here. Taken together, our results here suggest that a strategy utilizing a spike-and-slab prior on the random effects is preferable to one using a customary normal prior on the random effects for detecting who is and is not "average."

Study 2

We now tackle the issue of double shrinkage in the random effects. Recall that the potential issue here is that the random effects may be biased toward zero due to shrinkage occurring both within the slab, as is typical in an ordinary mixed-effects model, and in the spike. In a customary random intercepts model with a normal prior on the random effects, the amount of shrinkage that occurs can be precisely determined through the so-called shrinkage factor, ω_j, which is given by

$$\lambda_j = \frac{\tau^2}{\tau^2 + \sigma^2/n_j} \tag{11}$$

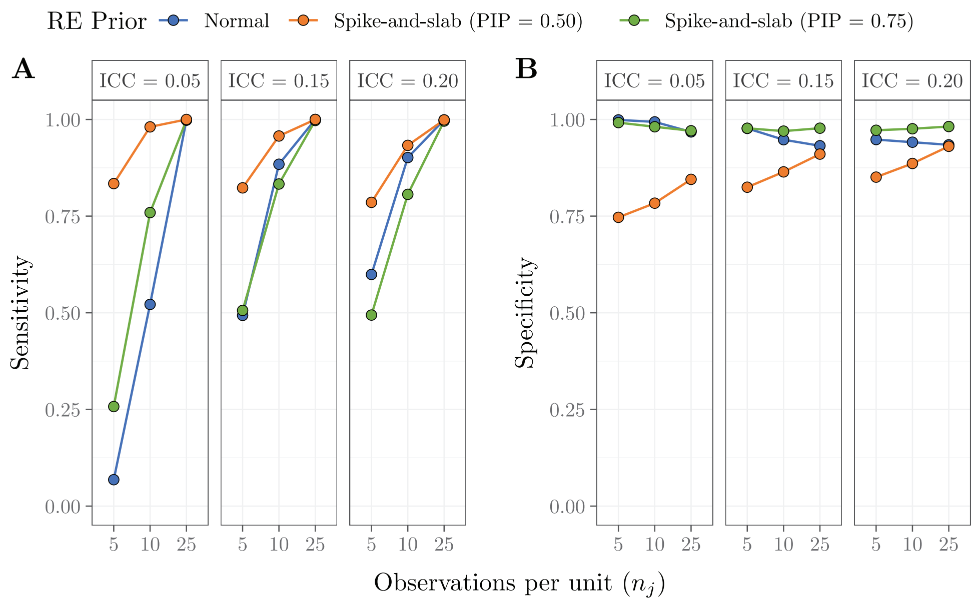
$$\omega_j = 1 - \lambda_j. \tag{12}$$

Notice that λ_j is calculated just as the ICC with the exception that the within-unit variance σ^2 is divided by n_j. Thus, holding n_j constant, larger ICCs imply smaller shrinkage factors and vice versa. Further, units with more observations will have smaller shrinkage factors. When all j units have equal observations ($n_1 = \ldots = n_j$), then there is a constant amount of shrinkage applied to all random effects ($\omega_1 = \ldots = \omega_j$).

When a spike-and-slab prior is instead placed on the random effects, determining the shrinkage involves an additional consideration. For every MCMC iteration, each random effect is either included (slab) or excluded (spike). All else being equal, the slab

[7] Because these results may have been due to the discrete nature of the random effects, an alternative simulation study was conducted using continuous random effects, and its results can be found in Appendix C.

Figure 5
Results From Study 1



Note. Classification rates of random effects under normal and spike-and-slab priors. For the normal prior, a 90% CrI was used to determine whether a random effect was "average" or not. For the spike-and-slab, two thresholds were used: a PIP of 0.50 and PIP of 0.75. (A) Sensitivity between the three methods. Sensitivity tended to one as n_j increased for all methods, but the spike-and-slab combined with a PIP of 0.5 generally had the best sensitivity. (B) Specificity between the three methods. Under the normal prior and 90% CrI, specificity was high with fewer n_j, but decreased to 0.9 as n_j increased. The spike-and-slab with a PIP of 0.75 had specificity of 1 or near 1 across all conditions, whereas using a PIP of 0.50 resulted in worse specificity. However, specificity still tended to 1 for the latter. This is a benefit of using the spike-and-slab prior—it will converge on the "true" model as the sample size grows. See the online article for the color version of this figure.

portion of the prior has the effect of applying stronger shrinkage to larger random effects relative to smaller random effects. Conversely, the spike has the effect of subjecting small random effects to more extreme shrinkage, relative to larger random effects. Dropping the notational dependence on the iteration index s, λ_j is calculated in each MCMC iteration as a piecewise function of the form

$$\lambda_j = \begin{cases} 0 & \text{if } \gamma_j = 0 \\ \dfrac{\tau^2}{\tau^2 + \sigma^2/n_j} & \text{if } \gamma_j = 1 \end{cases}, \tag{13}$$

where γ_j denotes whether the jth random effect is included in the model. The final estimate for each λ_j can be calculated as the average of Equation 13 across all MCMC iterations. Finally, the shrinkage factor can then be computed as $\omega_j = 1 - \lambda_j$. Because the posterior inclusion probability for the jth random effect is defined as the proportion of MCMC iterations where $\gamma_j = 1$, then keeping all else constant, using a spike-and-slab prior results in stronger shrinkage for estimates that have lower posterior inclusion probabilities.

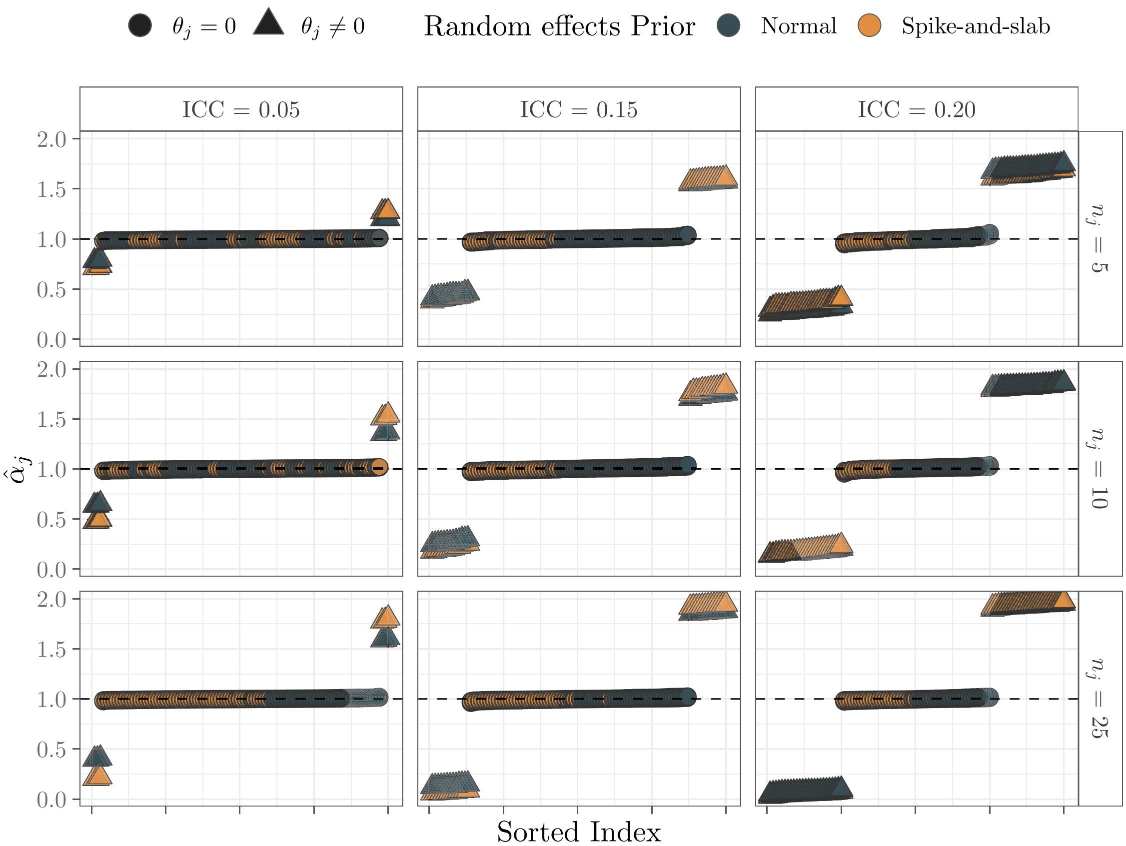
With the shrinkage factors in hand, the estimate of each unit-specific intercept, α_j, can be computed by

$$\hat{\alpha}_j = \omega_j \cdot \bar{y}_{.} + (1 - \omega_j) \cdot \bar{y}_j, \tag{14}$$

where $\bar{y}_{.}$ indicates the grand mean of the outcome and $\bar{y}_j$ denotes the unit-specific mean of y. A shrinkage factor ω_j of 1 indicates total shrinkage toward the grand mean ($\hat{\alpha}_j = \bar{y}_{.}$), and conversely, a shrinkage factor of zero indicates no shrinkage toward the grand mean ($\hat{\alpha}_j = \bar{y}_j$). By comparing the estimated α_j between mixed-effects models with normal and spike-and-slab priors, in addition to the shrinkage factors they produce, we can thoroughly investigate the impact of double shrinkage on the resulting random effects. To accomplish this, we followed the same set up as in Study 1. However, rather than focusing on the classification rates, we recorded the posterior estimates for the random intercept α_j and the shrinkage factors ω_j.

The average estimates for the α_j are displayed in Figure 6. Columns differentiate between ICCs, rows differentiate between n_j, color differentiates between priors and shape differentiates between (non)zero random effects. The dashed line denotes $\alpha = 1$. As expected, the estimated α_j are subject to less shrinkage toward α as the ICC increases, and similarly, as n_j increases, regardless of the prior. Further, for units where $\theta_j = 0$, the estimated α_j were estimated to be near the fixed effect α regardless of ICC, n_j, or prior. On the other hand, there were discrepancies in shrinkage between the spike-and-slab and normal priors when considering random effects that were set to either -1 or $+1$. For these random effects, the spike-and-slab prior often resulted in *less* shrinkage for the $\hat{\alpha}_j$ than the standard normal prior. For example, when the ICC was set to .05 and n_j = 25, the estimates $\hat{\alpha}_j$ for nonzero random effects were approximately .75 and 1.75 under the spike-and-slab prior. Meanwhile, the same estimates were roughly .5 and 1.5

Figure 6
Results From Study 2



Note. Estimates of the random intercepts α_j for mixed-effects models under normal and spike-and-slab priors, sorted in ascending order. The dashed lined denotes $\alpha = 1$. As expected, less shrinkage occurred as ICC and n_j increased regardless of prior, but there were differences in the amount of shrinkage. When the random effects were zero, the $\hat{\alpha}_j$ were highly similar between the priors across all conditions, but there were pronounced differences in the estimates for nonzero random effects. When $\theta_j \neq 0$, the spike-and-slab prior typically applied less shrinkage than the normal prior, such that estimates were closer to their true values. This is especially noticeable with smaller ICCs. See the online article for the color version of this figure.

under the normal prior. That is, the spike-and-slab prior allowed nonzero estimates to be closer to their actual values (0 and 2, respectively) than the normal prior. This result displays a nice property of the spike-and-slab in that the shrinkage is adaptive; larger effects receive little shrinkage whereas there is strong shrinkage for small effects (Rouder et al., 2018).

In order to better understand the differences in shrinkage between the priors, the average shrinkage factors ω_j for each condition are displayed in Table 1. As implied by the $\hat{\alpha}_j$ in Figure 6, the shrinkage factors decreased with increasing ICCs and n_j regardless of the prior. Note, however, that the shrinkage factor under the normal prior is constant in each condition regardless of whether the random effect was actually equal to zero or not. Because shrinkage under the spike-and-slab prior is adaptive, the shrinkage factors were larger when $\theta_j = 0$ than when $\theta_j \neq 0$. Relatedly, under the spike-and-slab prior, there was relatively strong shrinkage for the random effects equal to zero, regardless of ICC or n_j, but for nonzero random effects, the shrinkage dissipated with increasing ICC and n_j. Generally speaking, the spike-and-slab prior applied more shrinkage to random effects that were truly zero and less shrinkage for nonzero random effects, relative to the normal prior.

Part of our results here are due to setting the prior inclusion probability for each random effect to .5 (see Equation 2). In practice, this is the most common choice because it expresses equal prior odds for whether a given random effect should be included or excluded from the model. Choosing alternative values would alter the amount of shrinkage observed in Figure 6 and Table 1. In practice, researchers applying the spike-and-slab prior to random effects should bear this in mind when setting the prior inclusion probabilities. To provide some intuition on the impact of choosing alternative values for the prior inclusion probabilities, we conducted additional simulation studies. The results are included in Appendix D.

In summary, we observed that the double shrinkage induced from the spike-and-slab did not bias the random effects relative to a standard mixed-effects model by applying too much shrinkage. Rather, in many cases, the shrinkage applied by the spike-and-slab prior was preferable in that it applied weak shrinkage to nonzero effects and stronger shrinkage to truly zero random effects.

Table 1
Average Shrinkage Factors (ω_j)

ICC	n_j	Normal $\theta_j = 0$	Normal $\theta_j \neq 0$	Spike-and-slab $\theta_j = 0$	Spike-and-slab $\theta_j \neq 0$
0.05	5	0.80	0.80	0.84	0.76
	10	0.65	0.65	0.74	0.51
	25	0.40	0.40	0.68	0.21
0.15	5	0.45	0.45	0.70	0.48
	10	0.28	0.28	0.68	0.25
	25	0.13	0.13	0.72	0.07
0.20	5	0.29	0.29	0.68	0.43
	10	0.17	0.17	0.69	0.23
	25	0.07	0.07	0.74	0.05

Note. Larger values indicate more shrinkage of the random effects towards zero. The shrinkage applied by the normal prior is constant regardless of whether $\theta_j = 0$ or $\theta_j \neq 0$, but the shrinkage applied by the spike-and-slab prior is adaptive.

Discussion

In this work, we provided a general spike-and-slab formulation for random effect selection in mixed-effects models. The empirical application evidenced the utility of the proposed methodology for addressing individual differences in psychological science. Two simulation studies were conducted that illustrated key properties of the approach. Although spike-and-slab priors are not new in psychology research, their advantages were thought to be limited to exploratory variable selection and big-data contexts, such as fMRI analyses (Rouder et al., 2018). As we illustrated in this article, however, the spike-and-slab is also valuable in the context of "small-data" which is common in the social-behavioral sciences.

In the empirical application, we performed posterior predictive checks on the models for the cognitive tasks in order to inspect their adequacy in capturing important patterns in the data. While model checking is indeed an important part of statistical modeling, an additional motivation for performing the posterior predictive check was to address the concern of whether the spike-and-slab, "taken globally, [can] provide a good description of the structure in the data" (Haaf & Rouder, 2017, p. 794). As was shown in Figure 4, our formulation did a good job of describing the experimental effects, or mean differences between conditions. This ability of the spike-and-slab prior to provide trustworthy estimates was also observed in Study 2. Placing a spike-and-slab prior on the random effects does not necessarily compromise the model estimates.

The data we used in this article came from experiments in psycholinguistics and cognitive psychology. We chose these data because: (a) they are typical representations of research that is done in the realm of individual differences with an emphasis on mixed-effects methodology, and (b) data from cognitive tasks have been recently used in the context of reliability research. Given the history of individual differences in cognitive research, finding little individual differences in these tasks is somewhat unexpected. This perhaps points to the rather restrictive nature of the standard approach for probing individual differences in mixed-effects models. That is, if there is little between-subjects variability, then a researcher might conclude that there are no individual differences. The spike-and-slab approach, in turn, offers a more nuanced view as it allows the differentiation between those who are and are not "average," even in low ICC settings. This was clearly seen in Study 1, where the spike-and-slab prior had good performance in detecting nonaverage units even when the ICC was as low as .05 and in the empirical application, where over a quarter of the experimental effects for individuals did not conform with the average experimental effect.

Future Directions

An oft-overlooked aspect of mixed-effects models is that the residual variance (σ^2) and between-subjects variance (τ^2) are considered to be constant across subjects. This can result in an improper amount of shrinkage (Hoff, 2009, Chapter 8), in essence, distorting the model estimates and their variability. This assumption can be relaxed so that the within- and between-subjects variances can be allowed to vary as a function of predictors. Such models have been introduced to psychology under the name of mixed-effects location-scale models (Hedeker et al., 2012; Rast & Ferrer, 2018; Williams et al., 2019). By allowing nonconstant variances, individual differences may be more pronounced (Williams et al., 2020). Applying the spike-and-slab prior to the random effects in these models remains an interesting direction for future work because of the potential to tease apart individual differences in even finer detail.

The methodology we discuss in this article also has promising potential in clinical fields. In this domain, there has been increasing interest in idiographic methods, or methods focused on individuals (see e.g., the models described in Piccirillo & Rodebaugh, 2019). The motivation for their use is often to identify individuals for whom a treatment may have different levels of efficacy. The use of mixed-effects models (and also mixed-effects location-scale models) in combination with spike-and-slab prior may provide an interesting avenue of research in idiographic studies because information is not lost by fitting separate models, but individuals who deviate from an average treatment effect may still be identified.

Summary

In this work, we discussed a general strategy to apply the spike-and-slab prior to the random effects in mixed-effects models for individual differences research. Importantly, this method allows researchers to gain a more nuanced view of individual differences than traditional approaches. By going beyond the testing of variance components to using the spike-and-slab for random effect selection, researchers can determine which individuals differ from an average effect. The methods discussed in article have been implemented in the R package SSranef.

References

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & van der Sluis, S. (2014). A solution to dependency: Using multilevel analysis to accommodate nested data. *Nature Neuroscience*, *17*(4), 491–496. https://doi.org/10.1038/nn.3648

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, *32*(3), 870–897. https://doi.org/10.1214/009053604000000238

Barbieri, M. M., Berger, J. O., George, E. I., & Ročková, V. (2021). The median probability model and correlated variables. *Bayesian Analysis*, *16*(4), 1085–1112. https://doi.org/10.1214/20-BA1249

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*(4), 1281–1311.

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), 255–278.

Bono, R., Alarcón, R., & Blanca, M. J. (2021). Report quality of generalized linear mixed models in psychology: A systematic review. *Frontiers in Psychology*, *12*, 666182. https://doi.org/10.3389/fpsyg.2021.666182

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Caplan, S., Hafri, A., & Trueswell, J. C. (2021). Now you hear me, later you don't: The immediacy of linguistic computation and the representation of speech. *Psychological Science*, *32*(3), 410–423. https://doi.org/10.1177/0956797620968787

Estrada, E., Ferrer, E., Shaywitz, B. A., Holahan, J. M., & Shaywitz, S. E. (2018). Identifying atypical change at the individual level from childhood to adolescence. *Developmental Psychology*, *54*(11), 2193–2206. https://doi.org/10.1037/dev0000583

Friedman, N. P., Profile, S., Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology; General*, *133*(1), 101–135. https://doi.org/10.1037/0096-3445.133.1.101

Frühwirth-Schnatter, S., & Wagner, H. (2011). Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In J. M. Bernardo (Eds.), *Bayesian statistics 9* (pp. 165–200). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199694587.003.0006

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, *7*, 2595–2602. https://doi.org/10.1214/13-EJS854

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models* (1st ed.). Cambridge University Press.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–760.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.1080/01621459.1993.10476353

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*(2), 339–373.

Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science*, *3*(4), 443–455. https://doi.org/10.1177/2515245920922982

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*(4), 779–798. https://doi.org/10.1037/met0000156

Heck, D. W., Overstall, A. M., Gronau, Q. F., & Wagenmakers, E.-J. (2019). Quantifying uncertainty in transdimensional Markov chain Monte Carlo using discrete Markov models. *Statistics and Computing*, *29*(4), 631–643. https://doi.org/10.1007/s11222-018-9828-0

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, *31*(27), 3328–3336. https://doi.org/10.1002/sim.5338

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer.

Hoffman, M. D., & Gelman, A. (2011). *The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo*. arXiv. https://arxiv.org/abs/1111.4246

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, *60*(1), 65–81.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Ley, E., & Steel, M. F. J. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, *24*(4), 651–674. https://doi.org/10.1002/jae.1057

Li, Q., & Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, *5*(1), 151–170. https://doi.org/10.1214/10-BA506

Liu, S., Rovine, M. J., & Molenaar, P. C. M. (2012). Selecting a linear mixed model for longitudinal data: Repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychological Methods*, *17*(1), 15–30. https://doi.org/10.1037/a0026971

Lövdén, M., Schmiedek, F., Kennedy, K. M., Rodrigue, K. M., Lindenberger, U., & Raz, N. (2013). Does variability in cognitive performance correlate with frontal brain volume? *NeuroImage*, *64*, 209–215. https://doi.org/10.1016/j.neuroimage.2012.09.039

MacDonald, S. W., Hultsch, D. F., & Dixon, R. A. (2008). Predicting impending death: Inconsistency in speed is a selective and early marker. *Psychology and Aging*, *23*(3), 595–607. https://doi.org/10.1037/0882-7974.23.3.595

Meng, X.-L. (1994). Posterior predictive p-values. *Annals of Statistics*, *22*(3), 1142–1160. https://doi.org/10.1214/aos/1176325622

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032. https://doi.org/10.1080/01621459.1988.10478694

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. https://doi.org/10.1016/j.jmp.2015.11.001

Ntzoufras, I. (2002). Gibbs variable selection using BUGS. *Journal of Statistical Software*, *7*(7), 1–19. https://doi.org/10.18637/jss.v007.i07

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 99–118. https://doi.org/10.1111/j.2517-6161.1995.tb02017.x

O'Hara, R. B., & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, *4*(1), 85–117. https://doi.org/10.1214/09-BA403

Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, *22*(1), 59–73. https://doi.org/10.1214/088342307000000014

Piccirillo, M. L., & Rodebaugh, T. L. (2019). Foundations of idiographic methods in psychology and applications for psychotherapy. *Clinical Psychology Review*, *71*, 90–100. https://doi.org/10.1016/j.cpr.2019.01.002

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Plummer.pdf

R Core Team. (2021). *R: A language and environment for statistical computing manual*. R Foundation for Statistical Computing.

Rast, P., & Ferrer, E. (2018). A mixed-effects location scale model for dyadic interactions. *Multivariate Behavioral Research*, *53*(5), 756–775. https://doi.org/10.1080/00273171.2018.1477577

Rast, P., & Zimprich, D. (2011). Modeling within-person variance in reaction time data of older adults. *GeroPsych*, *24*(4), 169–176. https://doi.org/10.1024/1662-9647/a000045

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). SAGE Publications, Inc.

Rouder, J. N., & Haaf, J. M. (2020). Are there reliable qualitative individual difference in cognition? *Journal of Cognition*, *4*(1), 46. https://doi.org/10.31234/osf.io/3ezmw

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102–113. https://doi.org/10.3758/s13423-017-1420-7

Rouder, J., Kumar, A., & Haaf, J. M. (2019). *Why most studies of individual differences with inhibition tasks are bound to fail*. PsyArXiv. https://doi.org/10.31234/osf.io/3cjr5

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151–1172. https://doi.org/10.1214/aos/1176346785

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, *2*, e55. https://doi.org/10.7717/peerj-cs.55

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Watts, A., Walters, R. W., Hoffman, L., & Templin, J. (2016). Intraindividual variability of physical activity in older adults with and without mild Alzheimer's disease. *PLoS ONE*, *11*(4), e0153898. https://doi.org/10.1371/journal.pone.0153898

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research*, *19*(1), 231–240.

Williams, D. R., Carlsson, R., & Bürkner, P.-C. (2017). Between-litter variation in developmental studies of hormones and behavior: Inflated false positives and diminished power. *Frontiers in Neuroendocrinology*, *47*, 154–166. https://doi.org/10.1016/j.yfrne.2017.08.003

Williams, D. R., Martin, S. R., & Rast, P. (2021). Putting the individual into reliability: Bayesian testing of homogeneous within-person variance in hierarchical models. *Behavior Research Methods*, *54*(3), 1272–1290. https://doi.org/10.3758/s13428-021-01646-x

Williams, D. R., Mulder, J., Rouder, J. N., & Rast, P. (2020). Beneath the surface: Unearthing within-person variability and mean relations with Bayesian mixed models. *Psychological Methods*, *26*(1), 74–89. https://doi.org/10.1037/met0000270

Williams, D. R., Zimprich, D. R., & Rast, P. (2019). A Bayesian nonlinear mixed-effects location scale model for learning. *Behavior Research Methods*, *51*(5), 1968–1986. https://doi.org/10.3758/s13428-019-01255-9

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*(4), 1193–1209. https://doi.org/10.3758/s13428-016-0779-0

Appendix A

Example Code

Listing 1: Example R Code

```
# install package
remotes::install_github("josue-rodriguez/SSranef")

# fit mixed-effects model with spike-and-slab prior
# on random slopes
fit1 <- ss_ranef_beta(y = stroop$rt,
                      X = stroop$congruent,
                      unit = stroop$id)


# extract PIPs and calculate what proportion
# of sample differed from average
pips <- ranef_summary(fit1)$PIP
n_non_avg <- sum(pips > .5, na.rm = T)
n_total <- length(unique(stroop$id))
n_non_avg/n_total


# refit model with different prior inclusion probability
# for random effects
priors <- list(gamma = "gamma[j] ~ dbern(0.8)")
fit2 <- ss_ranef_beta(y = stroop$rt,
                      X = stroop$congruent,
                      unit = stroop$id,
                      prior = priors)
```

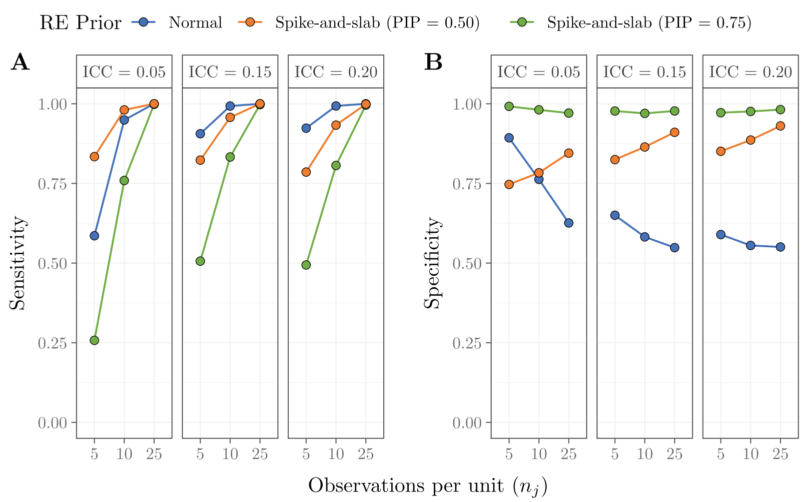
(*Appendices continue*)

Appendix B

Credible Interval Width

Figure B1 displays the results from Study 1, but using 50% CrIs instead of 90% CrIs for the standard mixed-effects model. As described in the main text, random effect selection with narrower intervals leads to higher sensitivity in exchange for lower specificity, and vice versa for wider intervals. Because it has previously been argued that a 50% CrI provides the best balance between sensitivity and specificity (Li & Lin, 2010), we compared the performance of a 50% CrI strategy for random effects selection to the spike-and-slab prior with PIP cut-offs of .5 ($SS_{0.5}$) and .75 ($SS_{0.75}$). However, the core conclusions from Study 1 did not change. In terms of sensitivity, the $SS_{0.5}$ model had the best sensitivity for the lowest ICC condition (ICC = .05), but now the CrI strategy had superior sensitivity as the ICC increased. In terms of specificity, however, the 50% CrI strategy performed worse than both spike-and-slab models in all but one condition (when the ICC was set to .05 and the observations per unit was set to 5). Importantly, the key difference remains that the spike-and-slab models are model selection consistent and will converge on the "true" model with increasing sample size while this statement does not hold for the credible interval strategy, regardless of the width that is chosen.

Figure B1
Results From Study 1 With 50% CrIs Instead of 90% CrIs for the Model With the Normal Prior



Note. See the online article for the color version of this figure.

(*Appendices continue*)

Appendix C

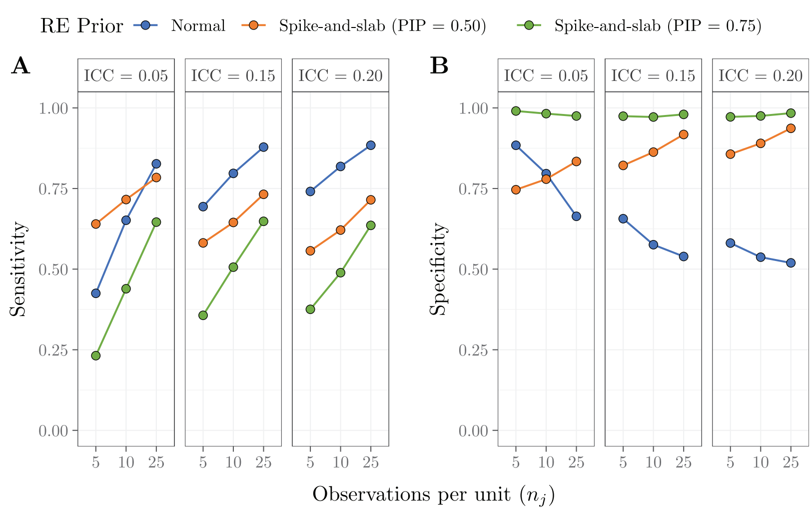
Normally Distributed Random Effects

In Study 1, the random effects were generated in a discrete manner to better assess their classification under a spike-and-slab prior. We repeated this simulation study such that all procedures remained the same, but instead of assigning the nonzero random effects values of ± 1, they were drawn from a standard normal distribution so that very small nonzero random effects would be introduced. Generating the random effects in this way reflects the commonly made assumption about the normality of random effects in mixed-effects models. Because random effects near zero are likely to be absorbed by the spike component of the spike-and-slab prior, its classification performance depends on how well the θ_j's can be distinguished from zero (George & McCulloch, 1997). Therefore, a drop in sensitivity for the spike-and-slab may be expected with normally distributed random effects because small nonzero values may be considered to be zero.

The results are shown in Figure C1. We once again found the $SS_{0.5}$ model to have the best sensitivity with few n_j and a small ICC, but it was the credible interval strategy under the normal prior that had superior sensitivity in all other conditions. As n_j increased, however, so did the sensitivity for all three classification strategies. Although the normal prior generally had the best sensitivity with normally distributed random effects, it still fared poorly with respect to specificity. It was the $SS_{0.75}$ strategy that was the best in this regard. Of particular importance is that the model selection consistency property of the spike-and-slab prior was retained whereas it was still not applicable random effect selection with the credible interval approach. That is, both the sensitivity and the specificity tended to 1 under the spike-and-slab prior as the observations per unit increased, but specificity decreased with increasing n_j under the credible interval approach. The results from this simulation study suggest that if it the random effects are truly normally distributed and the goal is explicitly to maximize sensitivity, then using a credible interval to select nonzero random effects may be used (at the cost of an increased false positive rate). If a balance of good sensitivity and specificity is instead desired, then random effect selection with the spike-and-slab prior is preferable.

Figure C1

Results From Repeating Study 1, but With the Random Effects Generated From a Normal Distribution



Note. See the online article for the color version of this figure.

(*Appendices continue*)

Appendix D

Varying Prior Inclusion Probabilities

To provide further intuition on the role of the prior inclusion probability of the random effects in classification and shrinkage of the random effects, we repeated Study 1 and Study 2 twice each. Once with a prior inclusion probability of .2 for all random effects and once with a prior inclusion probability of .8 for all random effects. The results for classification performance can be viewed in Figures D1 and D2, and the shrinkage results can be viewed in in Figures D3 and D4.

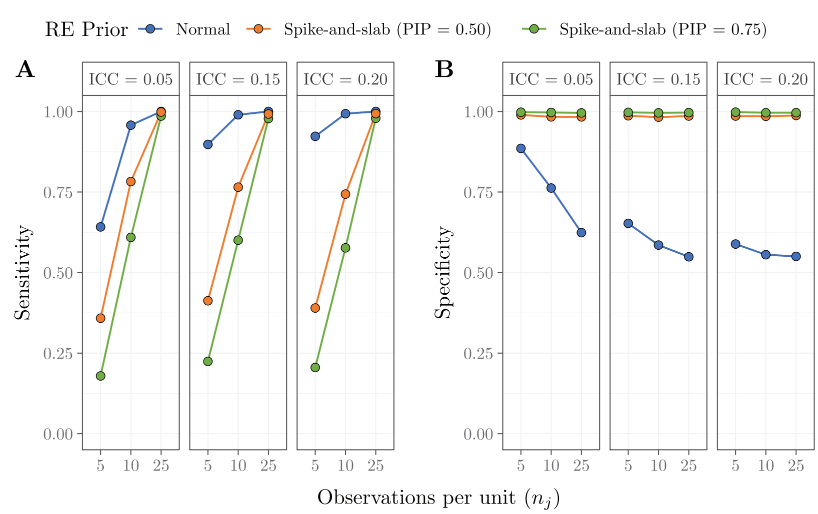
As shown in Figure D1, reducing the prior inclusion probability had the effect of reducing the sensitivity for both spike-and-slab strategies. This resulted in the credible interval selection strategy having the highest sensitivity across all conditions. However, both models under the spike-and-slab prior ($SS_{0.5}$ and $SS_{0.75}$) had a specificity near one. The reverse pattern was true when the prior inclusion probability was .8 (Figure D2). Here, the $SS_{0.5}$ model had a sensitivity of near 1 in all conditions, but a specificity of 0 in nearly all conditions. Again, the model selection consistency property of the spike-and-slab prior was observed. Together, these results show that the prior inclusion probability may be used to adjust the trade-off between sensitivity and specificity in classifying the average and nonaverage random effects.

With respect to the amount of shrinkage incurred by spike-and-slab and normal priors, many of the same general trends emerged occurred as in Study 2. Specifically, regardless of the prior, shrinkage decreased as n_j increased and as the ICC increased. Although there were again differences in shrinkage between the normal and spike-and-slab priors, they showed a different pattern than in Study 2. For instance, the top right panel in Figure D3 shows that when the prior inclusion probability was .2, there was much stronger shrinkage induced under the spike-and-slab prior than the normal prior whereas in Study 2, there was not much difference in shrinkage between the priors. The reason for this additional shrinkage concerns the incongruence between the prior inclusion probability and the proportion of truly nonzeros. Setting the prior inclusion probability to .2 reflects that the expected a priori proportion of nonzero random effects is .2, but the data were generated such that the actual proportion of nonzero random effects was .5. The result of this mismatch is that more zeros were drawn during MCMC sampling for the γ_j's—reflecting the low inclusion probability—and in turn, this induced more shrinkage in the random effects (see Equation 13). This can be contrasted with the top left panel of Figure D3, where the prior inclusion probability (.2) was higher than the proportion of nonzero random effects (.06). Here, the amount shrinkage is *less* than normal prior (mirroring the result in Study 2).

Figure D4 shows the case for prior inclusion probabilities of .8 and so the prior expected proportion of nonzero random effects was higher than the actual proportion of nonzeros in all conditions. Now, the top right panel shows that the shrinkage is almost identical between the normal and spike-and-slab priors. Because the prior probability was relatively high, values of one were sampled more often for the γ_j's, and thus the amount of shrinkage induced under the spike-and-slab prior was nearly identical to that of the normal prior. These findings suggest that the spike-and-slab produces more shrinkage on the random effects than the normal prior when the prior inclusion probability is smaller than the true proportion of nonzero random effects, and less shrinkage when the prior inclusion probability is greater than the true proportion of nonzero random effects. Crucially, though, the influence of the prior inclusion probability vanished with increasing sample size n_j. This was especially pronounced in the lower right panels of Figures D3 and D4, where there was almost no shrinkage applied to the nonzero random effects, regardless of the prior inclusion probability.

Figure D1

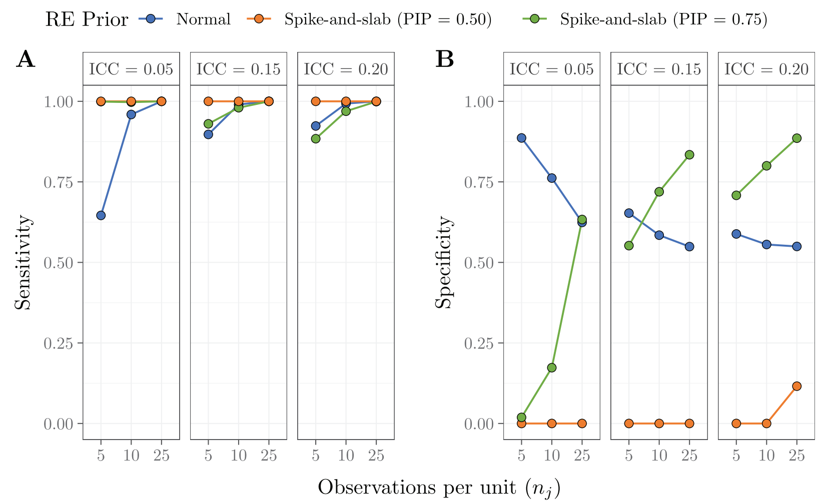
Sensitivity and Specificity for Prior Inclusion Probabilities of 0.2



Note. See the online article for the color version of this figure.

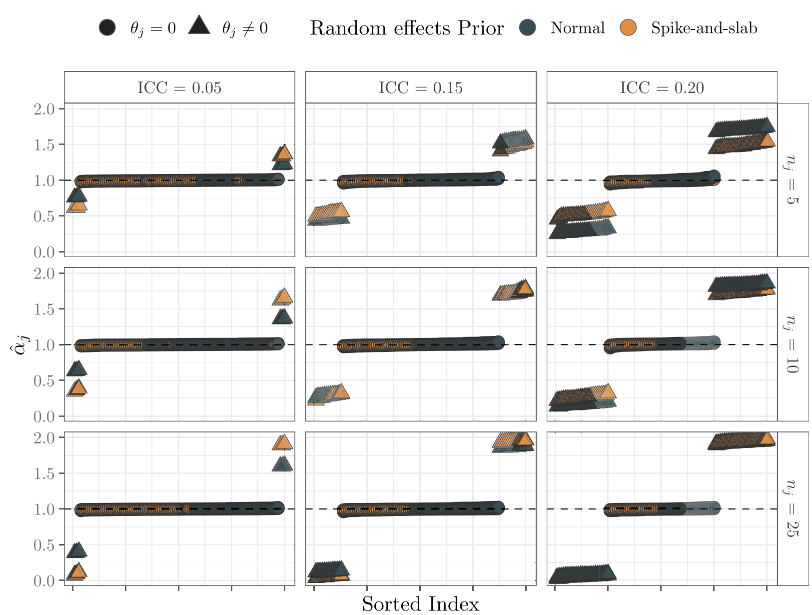
(*Appendices continue*)

Figure D2
Sensitivity and Specificity for Prior Inclusion Probabilities of 0.8



Note. See the online article for the color version of this figure.

Figure D3
Estimates of the Random Intercepts α_j for Mixed-Effects Models Under Normal and Spike-and-Slab Priors

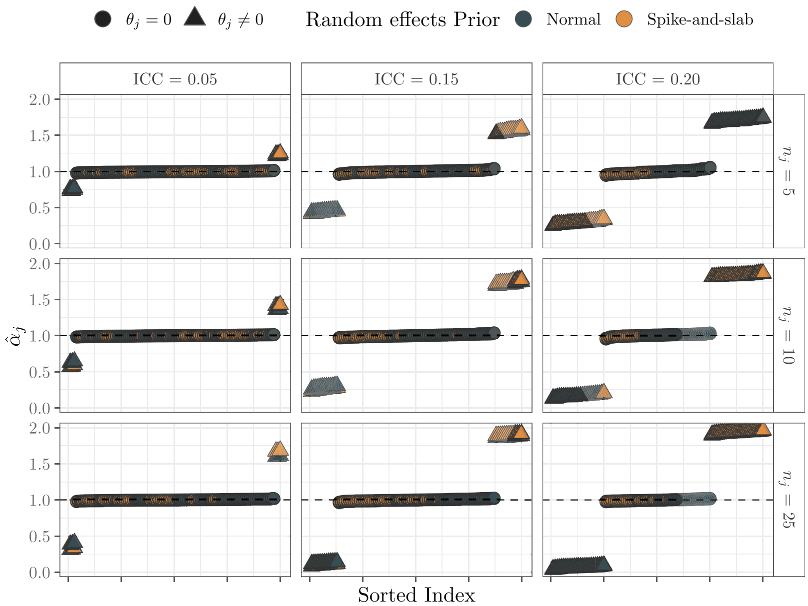


Note. The dashed lined denotes α = 1. The prior inclusion probability for each random effect was set to 0.2. See the online article for the color version of this figure.

(*Appendices continue*)

Figure D4

Estimates of the Random Intercepts α_j for Mixed-Effects Models Under Normal and Spike-and-Slab Priors



Note. The dashed lined denotes $\alpha = 1$. The prior inclusion probability for each random effect was set to 0.8. See the online article for the color version of this figure.

Received May 14, 2021
Revision received June 19, 2022
Accepted August 10, 2022 ■