Developing Constraint in Bayesian Mixed Models

Julia M. Haaf[1] & Jeffrey N. Rouder[1]

[1] University of Missouri

Author Note

Julia M. Haaf, 210 McAlester Hall, Columbia, MO. This paper was written in R-Markdown with code for data analysis integrated into the text. The Markdown script is open and freely available at https://github.com/PerceptionAndCognitionLab/ctx-indiff. The data used here are not original. We make these freely available with permission of the original authors at https://github.com/PerceptionCognitionLab/data0/tree/master/contexteffects. The analyses presented here were presented in part first at the 57th The Annual Meeting of the Psychonomic Society, 2016. The submission version of this document is archived at https://osf.io/preprints/psyarxiv/ktjnq

Correspondence concerning this article should be addressed to Julia M. Haaf, 205 McAlester Hall, University of Missouri, Columbia, MO 65211. E-mail: jhaaf@mail.missouri.edu

Abstract

"Model comparison in Bayesian mixed models is becoming popular in psychological science. Here we develop a set of nested models that account for order restrictions across individuals in psychological tasks. An order-restricted model addresses the question 'Does Everybody', as in, 'Does everybody show the usual Stroop effect', or 'Does everybody respond more quickly to intense noises than subtle ones.' The crux of the modeling is the instantiation of 10s or 100s of order restrictions simultaneously, one for each participant. To our knowledge, the problem is intractable in frequentist contexts but relatively straightforward in Bayesian ones. We develop a Bayes factor model-comparison strategy using Zellner and colleagues' default g-priors appropriate for assessing whether effects obey equality and order restrictions. We apply the methodology to seven data sets from Stroop, Simon, and Eriksen interference tasks. Not too surprisingly, we find that everybody Stroops—that is, for all people congruent colors are truly named more quickly than incongruent ones. But, perhaps surprisingly, we find these order constraints are violated for some people in the Simon task, that is, for these people spatially incongruent responses occur truly more quickly than congruent ones! Implications of the modeling and conjectures about the task-related differences are discussed."

Keywords: Bayesian mixed models, Bayes factors, Individual differences, Order constraints, Equality constraints, Priming

Developing Constraint in Bayesian Mixed Models

Many experimental tasks in psychology have a certain character where participants perform many trials in a small number of conditions. For example, in most social-cognitive priming tasks, participants perform hundreds of trials in the primed and unprimed conditions (Amodio et al., 2004). In reading tasks, participants may read hundreds of words (e.g. Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004); in perceptual tasks, participants may identify hundreds of items (e.g. Swagman, Province, & Rouder, 2015); in memory tasks, participants may be asked to recognize hundreds of previously-studied memoranda. Examples in decision making and attention spring immediately to mind. We may call such tasks *massively repeated*.

One of the common questions researchers ask in these tasks is whether there is an effect of a manipulation on an outcome variable. For example, in a priming task, a researcher may ask if responses when primed are faster than those when not primed. Likewise, massively repeated tasks may be used to ask whether the strength of a stimulus, the context in which it is presented, or the attention directed toward it affects the response.

If we limit our attention to tasks with two conditions, say as in the priming case, the usual course is to aggregate across the repetitions to produce a participant-by-condition mean score. These mean scores may be subtracted across the conditions to produce a participant-specific observed effect, and these effects may be tested with a t-test. If the t-test null is rejected, then the researcher may conclude that there is evidence for an average effect across the population of participants.

An advantage of this aggregate approach are its simplicity; it may be performed by virtually anyone. A limitation, however, is that researchers are unable to address individual variation in the participants. A significant t-test does not guarantee that all individuals display a priming effect. Some truly may while others truly may not. Even more alarming, while some may truly have the usual priming effect, others may truly have the opposite, negative effect.

Figure 1 shows two scenarios. Plotted are hypothetical distributions of individuals' true effects. In the first scenario, displayed in Figure 1A, all of these true effects are positive, and they come from a truncated normal distribution. In the second scenario, displayed in Figure 1B, individuals' effects are not constrained to be positive, and, instead, follow the usual normal distribution. Importantly, a sizable minority of participants have truly negative effects.

We think the differences between these scenarios are theoretically important. In the first scenario, 1A, the mean effect is interpretable as a proxy for what all participants do. This commonality of direction may be used to specify common mechanisms, say that priming is automatic and beyond strategic control (Greenwald, Klinger, & Schuh, 1995), or that stimulus strength has a simple, common neurological correlate, say that more strength corresponds to greater neural firing rates in certain cell assemblies (Roitman & Shadlen, 2002).

In the second scenario, 1B, effects are mixed in direction. A real-world example is handedness. Imagine a task where people throw balls with their right and left hands, and we ask for which hand did the ball go further. Here, right-handed people are almost always stronger with their right hand; left-handed people tend to be stronger with their left hand. In the population, more individuals are right-handed. So on average, people may be stronger with their right hand. But this average strength does not convey any meaningfull information about all individuals in the sample. Thus, this case corresponds to more complex theoretical implications that assume more than one underlying mechanism. Indeed, handedness is complicated, and it is better described as a syndrome than a single phenomenon (Coren, 1993). Other possible examples of complicated phenomena from the literature are different routes of attitude formation (Sweldens, Corneille, & Yzerbyt, 2014) or different stategies in decision theory (Kahneman & Tversky, 1972).

Unfortunately, it is impossible to tell which scenario holds and which theories are implicated from a t-test. To see the problem, consider Figure 1C and D. The vertical lines in

Figure 1C and D denote the mean of the distributions, and the dashed line shows the distributions of sample means. These distributions are the basis for the t-test, and the fact that these are identical illustrates the problem.

Assessing whether true effects are all in one direction may seem simple, but the presence of sampling noise makes it quite hard. For an illustration of this difficulty, consider the following Stroop interference experiment from Von Bastian, Souza, and Gade (2015). In the classic Stroop interference task (Stroop, 1935), participants are asked to name the color of displayed words. The words themselves are color names, such as RED and GREEN, and the meaning of the word may be congruent with the color, i.e., RED displayed in red, or incongruent with the color, i.e., GREEN displayed in red. Stroop interference refers to the slowdown of color identification when the meaning is incongruent, and it is a robust phenomenon (MacLeod, 1991). The question at hand is not the existence of Stroop interference, but whether all people display the effect in the usual direction.

Figure 2 shows the ordered observed Stroop interference effects, d_i, for 121 individuals from Von Bastian et al. (2015). People vary considerably in the size of their effects: they range from -100 ms to 181 ms. The problem is that it is difficult to distinguish between noise variation and true variation from the observed values. The 95% confidence intervals around invidivduals' sample effects are plotted, and these, unfortunately, do not provide any direct way of assessing whether all effects are positive. CIs are about each individual in isolation and cannot be used to answer questions about a group of individuals. We note that no CI in the figure is located exclusively below zero. Even though no specific individual is identified as definitively negative, there may be enough evidence across the collection of individuals to state that some of these individuals have true negative effects without identifying which one.

The problem at hand is known as order-restricted inference, and there is a voluminous frequentist literature on it (see Robertson, Wright, & Dykstra, 1988). Order-restricted inference is straightforward for one-dimensional cases, say whether the grand mean is greater than zero or not. The usual frequentist solution is to adjust rejection regions as is done in

one-sided t-tests. There are comparable implementations with Bayesian model comparison, where models on group means incorporate order restrictions (Klugkist & Hoijtink, 2007). Klugkist and colleagues introduced a Bayesian approach to order-restricted inference in ANOVA and ANCOVA (Klugkist, Kato, & Hoijtink, 2005), and repeated measure settings (Mulder, Klugkist, Schoot, Meeus, & Hoijtink, 2009). But the question here is whether a separate order-restriction holds simultaneously *for all individuals*. Therefore, we must assess 121 order restrictions simultaneously, and this problem is complicated. Of note, usual AIC and BIC approaches are not applicable because penalties reflect the number of parameters rather than their direction (Klugkist & Hoijtink, 2007).

Our approach is to analyze Bayesian mixed models with order and equality constraints on individuals.[1] There are two advantages of Bayesian methods in this context: 1. It is tractable. Bayesian analysis is conceptually straightforward and computationally feasible. 2. Bayesian methods, particularly Bayes factors, offer a calibration for inference that is appropriate for order constraints (see also Klugkist & Hoijtink, 2007).

Before describing the models we developed to assess constraints like the one in Figure 1A, it is worthwhile to ask if these constraints are useful. We consider two critiques. The first is that the order constraints are so natural and obvious that they assuredly hold *a priori*. For example, it is hard to imagine that anyone identifies dim flashes of light more quickly than bright flashes, that anyone reads long novel nonwords faster than short novel nonwords, or that anyone forgets repeated items at a greater rate than unrepeated items. Yet, we can think of examples where such restrictions do not hold such as the handedness example provided previously. A second critique is the diametric opposite of the above

[1] All analyses were conducted using R (3.3.1, R Core Team, 2016) and the R-packages *abind* (1.4.5, Plate & Heiberger, 2016), *BayesFactor* (0.9.12.2, Morey & Rouder, 2015), *coda* (0.19.1, Plummer, Best, Cowles, & Vines, 2006), *curl* (2.6, Ooms, 2016), *devtools* (1.12.0, Wickham & Chang, 2016), *fields* (8.10, Douglas Nychka, Reinhard Furrer, John Paige, & Stephan Sain, 2015), *gmm* (1.5.2, Chaussé, 2010), *LaplacesDemon* (16.0.1, Statisticat & LLC., 2016), *maps* (3.1.1, Richard A. Becker, Ray Brownrigg. Enhancements by Thomas P Minka, & Deckmyn., 2016), *MASS* (7.3.45, Venables & Ripley, 2002), *Matrix* (1.2.8, Bates & Maechler, 2017), *MCMCpack* (1.3.9, Martin, Quinn, & Park, 2011), *msm* (1.6.4, Jackson, 2011), *mvtnorm* (1.0.6, Genz & Bretz, 2009; Wilhelm & G, 2015), *papaja* (0.1.0.9485, Aust & Barth, 2017), *plotrix* (3.6.4, J, 2006), *sandwich* (2.3.4, Zeileis, 2004, 2006), *spam* (1.4.0, Furrer & Sain, 2010; Gerber & Furrer, 2015), *spatialfil* (0.15, Dinapoli & Gatta, 2015), and *tmvtnorm* (1.4.10, Wilhelm & G, 2015).

critique. It is that these order constraints can never hold exactly. There must be somebody, somewhere who has a negative true Stroop effect. This critique reminds us of the one that the null is never true (Cohen, 1994), for which there are several salient rebuttals (Rouder & Morey, 2012). We consider statements like "Everyone Stroops" to be of high value even if they hold approximately or platonically. If broadly applicable in common settings, they serve as important statements of constraint on theory. In summary, we consider the two extreme positions—that order constraints always hold or that order constraint never hold—to be intellectually unsatisfying. The best course then is to assess the evidence from data for these constraints.

Models of Individual Variation

We consider experiments with two conditions, which are generically termed *control* and *treatment* here. To model individual variability in massively repeated designs, we adopt the following notation: Let Y_{ijk} be a response variable, say response time (RT), for the kth replicate for the ith participant, $i = 1, \ldots, I$ in the jth condition, $j = 1, 2$ with $k = 1, \ldots, K_{ij}$. We place a linear model on Y_{ijk}:

$$Y_{ijk} \overset{iid}{\sim} \text{Normal}\left(\mu + \alpha_i + x_j\theta_i, \sigma^2\right). \tag{1}$$

Here, the term $\mu + \alpha_i$ serve as intercepts with μ being the grand mean of intercepts and α_i being individual deviations. The term x_j codes the condition, with $x_j = 0$ for the control condition and $x_j = 1$ for the treatment condition. The effect is the slope, θ_i, and the collection of these parameters across participants are the target of interest. And finally, σ^2 is the variance of the responses.

We develop a series of four mixed models on these effect parameters. At one end of the spectrum, the most general model with the least constraint simply posits that the individual effects follow a normal distribution with full support. At the other end, the most constrained model specifies that there is no effect for every individual. The following models are ordered

from the least constrained to the most, and the full set provides a useful tool for assessing the constraint in effects. We then use Bayesian model comparison, discussed subsequently, to assess what level of constraint is most appropriate for a set of data.

The Unstructured Model

The unstructured model is denoted $\mathcal{M}_u$ and places no order or equality constraints on the collection of effects:

$$\mathcal{M}_u : \quad \theta_i \overset{iid}{\sim} \text{Normal}(\nu, \eta^2),$$

where ν and η^2 are population-level parameters describing the mean and variance of effects across the population. This model is quite flexible in that all combinations of individual effects are plausible. Figure $3A_1$ shows this flexiblity for two participants. The true-effect value for the first individual is shown on the x-axis, the true-effect value for the second is shown on the y-axis. As can be seen, combinations of effects in the same and opposite directions for the two individuals have plausibility. In application, ν and η^2 are treated as parameters that are free to vary as a function of data. The specifications of of these parameters are described in the section "Additional specification".

The Positive-Effects Model

The positive-effects model, denoted $\mathcal{M}_+$, captures the constraint that each individual has a true effect in the predicted direction.

$$\mathcal{M}_+ : \quad \theta_i \overset{iid}{\sim} \text{Normal}_+(\nu, \eta^2),$$

where Normal_+ denotes a truncated normal distribution with a lower bound at zero. The probability density distribution of θ_i for two participants is shown in Figure $3B_1$. Note that the distribution in the figure is darker, that is more *dense*, than the distribution in Figure

$3A_1$. The reason for that is the reduction of space. The space of valid parameter values in this illustration is one fourth the space of valid parameter values in the unstructured model as the positive-effect model occupies only one of the four quandrants. This reduction of space relative to the unstructued model becomes larger as additional participants are considered.

The Common-Effect Model

The common-effect model, denoted $\mathcal{M}_1$, captures the constraint that each individual has the same effect:

$$\mathcal{M}_1 : \quad \theta_i = \nu.$$

This common-effect model seems *a priori* unlikely as it posits a constant effect, for example a common priming effect, across all individuals. Yet, we include it here for two reasons: First, it is a logical null that can be used to benchmark claims of individual differences. If this common-effects model provides a superior description and yet we view it as unlikely, we may then turn our attention to the adequacy of the design to capture individual differences. In this regard, the model serves as a valuable design check. Second, we think commonalities should be given more *a priori* weight. Important invariances, even if they hold approximately, serve as good structure to build theory. Hence, if individual effect vary to a very small extend, $\mathcal{M}_1$ could still be the preferred model.

For this model, $\mathcal{M}_1$, the distribution of θ_i for two participants follows the diagonal as shown in Figure $3C_1$. The geometry of the constraint reduces the volume of the distributions of the previous models to a single line.

The Null Model

The null model, denoted $\mathcal{M}_0$, specifies that each participant's true effect is identically zero

$$\mathcal{M}_0 : \quad \theta_i = 0.$$

As everyone's effect is fixed to zero in this model, the distribution of θ_i for two participants shown in Figure $3D_1$ is simply a point at zero.

Additional Specifications

We analyze the above four models in a Bayesian framework. Additional specifications are needed for parameters μ, the grand mean of intercepts, σ^2, the variance of responses in each participant-by-condition cell, ν, the mean of effects, η^2, the variance of effects, and the collection of α_i, the individual intercepts. Our choices are motivated by the following two considerations: First, we adopt specifications that lead to computationally convenient algorithms for model comparison. Second, we adopt specifications that adhere to a subjective Bayesian philosophy (DeGroot, 1982; Goldstein, 2006; Rouder, Morey, & Wagenmakers, 2016), where priors are weakly informative and reflect a reasonable ranges of beliefs. Here, the prior settings reflect our *a priori* general substantive knowledge of generic data observed in these types of tasks.

We adopt a Zellner g-prior specification (Zellner, 1986). This specification is well studied and has been influential in linear modeling in statistics (e.g. Bayarri & Garcia-Donato, 2007; F. Liang, Paulo, Molina, Clyde, & Berger, 2008; Overstall & Forster, 2010). It underlies most Bayes factor development in psychology for regression and ANOVA models (Rouder & Morey, 2012; Rouder, Morey, Speckman, & Province, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels & Wagenmakers, 2012). In this context, we follow the ANOVA development by Rouder et al. (2012) as follows:

In the g-prior specification, non-informative reference priors are placed on parameters that are common to all models. In this case, this specification applies to μ and σ^2:

$$f(\mu, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Next, consider the following specification on each individual's baseline, α_i:

$$\alpha_i | g_\alpha, \sigma^2 \overset{iid}{\sim} \text{Normal}(0, \sigma^2 g_\alpha). \tag{2}$$

In the g-prior specification, the prior on effect parameters, in this case α_i, is a function of σ^2, the overall variance of responses, and a new parameter, g_α. The role of σ^2 in the prior may not be transparent. It is helpful to present an equivalent specificiation in *standardized effect sizes* rather than in effects. Let α_i^* be the ith individual's intercept effect size, where $\alpha_i^* = \alpha_i / \sigma$. The model reparameterized in this effect-size parameter may be given as $Y_{ijk} \sim \text{Normal}(\mu + \sigma\alpha_i^* + x_j\theta_i, \sigma^2)$, where the prior on $\alpha_i^* \sim \text{Normal}(0, g_\alpha)$. Here we see there is no mystery placing the σ^2 in the prior variance in (2). It allows us to focus on g_α, which is the prior variance on effect sizes. Effect sizes are convenient because psychologists have developed extensive intuition about how these should vary across manipulations and domains.

In g-prior specification, priors may be placed on the g parameter. The usual specification (Zellner & Siow, 1980), which we follow, is a scaled inverse-χ^2 with one degree-of-freedom.[2] Using the inverse-χ^2 as the prior distribution on g is less assumptive and represents the analyst's uncertainty.

$$g_\alpha \sim \text{Inverse-}\chi^2(r_\alpha^2).$$

The quantity r_α^2 is a scaling setting on variability (in standardized units), and it must be set *a priori*. The distribution on g is fat-tailed, and adding any additional uncertainty, say by adding variability to the scale, may result in posteriors with poor properties (Hobert & Casella, 1996).

How to set r_α^2? We first start by considering the square-root, r_α as this quantity is on the scale of effect sizes. The meaning of this scale is shown in Figure 4A. The figure shows a broad bivariate density centered at $(0, 0)$. This density is the marginal prior for two different

[2]The density of the inverse χ^2 distribution is $f(\sigma^2; b) = \frac{\sqrt{b}}{\Gamma(1/2)} \left(\sigma^2\right)^{-3/2} \exp\left(\frac{-b}{\sigma^2}\right)$, where b is a scale parameter.

individuals' intercepts α_1 and α_2. The centering at zero assures that μ is the grand mean. The value of r_α sets the scale, and the plot shows the case for $r_\alpha = 1$. We show the values in effect-size units (bottom and left axes), and the expectation is that variability in baseline across people is about the same as the variability of repeated trials. Why did we choose this setting? In our experience with these types of tasks, the standard deviation of repetitions for a participant within a condition is around 200 ms to 400 ms for repeated trials (Luce, 1986). We do not use this value to set σ though it is an estimate. But we do use it to think about the setting for r_α. Also in our experience with these types of tasks, people's means also tend to vary from about 200 ms to 400 ms. For example, one person's mean might be 500 ms while another's might be 800 ms. With this knowledge in mind, we set r_α to 1.0, encoding the belief that the variation of individual baselines has a characteristic scale of σ. These prior settings are not so critical but they should be reasonable. We explore their effects on inference subsequently. The same rationale may be used in other domains—researchers should have a defensible notion about what ranges of effects sizes they think are reasonable.

We use a similar rationale and g-prior structure on parameter ν:

$$\nu | g_\nu \sim \text{Normal}(0, \sigma^2 g_\nu)$$
$$g_\nu \sim \text{Inverse-}\chi^2(r_\nu^2).$$

To set values of r_ν, we note that average effects in tasks like these tend to be on the order of tens of milleseconds, and we use 50 ms, or one sixth of the standard deviation for repeated trials, as the scale of the inverse-χ^2 prior ($r_\nu^2 = (1/6)^2$).

To complete the specification, we let $g_\theta = \eta^2/\sigma^2$, the standardized variance of the effects. This allows for a g-prior specification of θ_i, similar to the specification of α_i. As

before, the prior on g_θ is

$$g_\theta \sim \text{Inverse-}\chi^2(r_\theta^2).$$

To set r_θ we consider individual variability around the mean effect. It is difficult to know what the true variation of individual effects is because the observed variability is confounded by trial noise. Our working assumption is that the scale should be also on the order of tens of milleseconds. Moreover we expect the variability to be no larger than the size of the average effect. Using this rationale, we set the scale on the standard deviation of individual differences to be 30ms, or one tenth of the standard deviation ($r_\theta^2 = (1/10)^2$).

Figure 4B shows the marginal priors for two individuals' effects, θ_1 and θ_2. The correlation between θ_1 and θ_2 in the marginal priors arises from the hierarchical nature of the specification in the unstructued and positive models. The variability in ν, is shared between the two individuals and induces the correlation. The degree of correlation is a function of the amount of shared variance, reflecting the setting of r_ν, and the amount of unique variance, reflecting the setting of r_θ.

It is important to know how these substantive choices for setting r_α, r_ν, and r_θ affect model comparison results. We visit this issue after presenting the results of the real-world application.

Model Comparison

A critical question is how to state evidence for the four theoretically-informed models. There are two leading approaches in Bayesian analysis that roughly can be described as estimation-based and model-comparison-based. In the estimation approach, say that in Kruschke (2011), posterior parameters are estimated in general model, and credible intervals for these parameters are inspected to see if restrictions are plausible. We will provide parameter estimates from the unstructured model for first inspection though, as will be shown, these are ultimately unhelpful for the questions at hand. The other approach, the one which we advocate here, is model comparison through Bayes factors (Edwards, Lindman,

& Savage, 1963; Jeffreys, 1961; Kass & Raftery, 1995; Laplace, 1986; Rouder et al., 2009). Bayes factors are the direct and unavoidable consequence of applying Bayes rule to assess the relative plausibility of models. Bayes rule for two model, $\mathcal{M}_a$ vs. $\mathcal{M}_b$, is

$$\frac{P(\mathcal{M}_a \mid \boldsymbol{Y})}{P(\mathcal{M}_b \mid \boldsymbol{Y})} = \frac{P(\boldsymbol{Y} \mid \mathcal{M}_a)}{P(\boldsymbol{Y} \mid \mathcal{M}_b)} \times \frac{P(\mathcal{M}_a)}{P(\mathcal{M}_b)}, \tag{3}$$

where the term on the left-hand side is the posterior odds for the models, and the term on the far right, $\frac{P(\mathcal{M}_a)}{P(\mathcal{M}_b)}$, is the prior odds. The term $\frac{P(\boldsymbol{Y}|\mathcal{M}_a)}{P(\boldsymbol{Y}|\mathcal{M}_b)}$ is the Bayes factor. Rouder and Morey (2012) provide guidance for reporting prior odds, Bayes factors and posterior odds. Prior odds describe *a priori* beliefs about the usefulness or plausibility of the models. Analysts and readers may hold different prior odds, and the analysts' prior odds may be made public, if one wishes to suggest or defend them, private, or not considered at all. The Bayes factor describes how the data have changed the analysts' beliefs. We, along with several others, argue the Bayes factor is the *evidence* from data for competing models, and is the appropriate target of inquiry. The Bayes factor is always made public as a statement to the community about how a reader's beliefs should be updated. Prior odds are useful when the analyst has to make a decision about which model is best, say as in making a policy recommendation. With prior odds specified, the analyst can compute posterior odds, evaluate these posterior odds within the context of a loss function, and make a principled decision. For scientific communication, however, decisions are not necessary, and in this context, the Bayes factor serves as graded continuous evidence much like odds of events. As with odds, we do not need to make decisions to appreciate their value. For instance, it is not necessary to decide if 5-to-1 odds is large, a long-shot, or moderate to comprehend the value of 5-to-1. Bayes factors in our view should be understood analogously. Expanded discussion is provided in Edwards et al. (1963), Jeffreys (1961), Morey, Romeijn, and Rouder (2016), and Rouder and Morey (2012).

The numerator and denominator of the Bayes factor describe the probability of observing data conditional on model $\mathcal{M}_a$ and model $\mathcal{M}_b$, respectively. These probability

statements may be termed the *predictions* of the respective models. They may be expressed for all possible data points before the data are observed. The Bayes factor therefore denotes how well the observed data are predicted under one model relative to another. In practice, the Bayes factor has two roles: It describes the evidence as change of beliefs and describes how well the models predicted the observed data. The equivalence of these roles leads to the deep insight that in Bayesian model comparison, evidence for competing models is the ratio of how well each predicts the observed data (Rouder, Morey, Verhagan, Swagman, & Wagenmakers, 2016).

To display the predictions of the models, we focus on two hypothetical individuals much as we did in specifying the models. The left column in Figure 3 shows the model specification for set values of ν and η. However, when marginalized across ν and η, there is correlation among individual effects as shown in Figure 4B. The right column of Figure 3 shows the corresponding predictions for sample effects, which shows the correlation. Panel A_2, shows the predicted sample effects for the unstructured model $\mathcal{M}_u$ from panel A_1; panel B_2 shows the predicitons for the positive-effects model, $\mathcal{M}_+$; panel C_2 shows the predicted individual effects for the common-effect model, $\mathcal{M}_1$ and, finally, D_2 shows the distribution of predicted effects for the null model, $\mathcal{M}_0$. The more constraint within a model, the greater the predictive density on concordant points and the less density on discordant ones.

Once the predictions are derived, model comparison is as simple as comparing these predictive densities at observed values. Suppose the individuals have observed effects of $d_1 = 150$ ms and $d_2 = 100$ ms. These values are shown by the point in panels A_2 through D_2. The density for $(150, 100)$ under the unstructured model $\mathcal{M}_u$ is 0.00103. The density under the positive-effects model for the same data, $\mathcal{M}_+$, is 0.00204. The remaining densities are 0.00169 and 0.00014 for models $\mathcal{M}_1$ and $\mathcal{M}_0$, respectively. We use the ratios of these densities to compare the models. The ratio of the densities for models $\mathcal{M}_+$ and $\mathcal{M}_u$, for example, is 2-to-1 in favor of $\mathcal{M}_+$. This ratio is the Bayes factor and is denoted B_{+u}. The comparison of all models may be done in the same manner. For example, the Bayes factor

between the positive-effects model and the null-effect model is 14-to-1 in favor of $\mathcal{M}_+$.

Although Bayes factors are conceptually simple, they are often computationally inconvenient in real-world applications with mixed models. The target quantity, the probability of data conditional on a model, is rexpressed using The Law of Total Probability as

$$P(\boldsymbol{Y} \mid \mathcal{M}) = \int_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{Y}|\boldsymbol{\xi}) P(\boldsymbol{\xi}) d\boldsymbol{\xi}, \tag{4}$$

where $\boldsymbol{\xi}$ is a vector of parameters from parameter space Ξ. The terms in the integrand are straightforward to compute. The difficulty comes from evaluating the integral itself. The integration is multidimensional ranging over the dimensionality of the parameters. There are often no closed-form solutions, and brute-force numerical methods fail in large-data settings because the to-be-integrated multivariate function is very peaked and narrow. The numerical method must find the "needle in the haystack." Often, it is difficult to assess whether the outputs approximate the true value of the integral. As a result, obtaining computationally convenient Bayes factor algorithms in mixed settings remains timely and topical in Bayesian research.

One approach to this integration problem is to specify models so that many of the parameters may be integrated symbolically. Zellner and Siow (1980) provide the seminal advance here. They showed that the g-prior specification used here allowed researchers to symbolically integrate all the parameter except g! It is for this reason that the g-prior specification has been popular (see Bayarri & Garcia-Donato, 2007; F. Liang et al., 2008; Rouder et al., 2012; Zellner, 1986). The particular form we adopt was first proposed by Rouder et al. (2012) who developed models with multiple g-parameters for mixed ANOVA designs with fixed and random effects. The key advantage of this form, and the reason we use it here, is that we can ensure accurate evaluation of the needed integrals. In application, this symbolic integration is hard-coded into the BayesFactor R package (Morey & Rouder, 2015). The approach is advantageous, because it avoids the computational uncertainties

inherent in more general-purpose but numerically intensive algorithms such as those implemented in STAN (eg. Stan Development Team, 2016).

Integration proceeds as follows: The full vector of parameters is the concatenation of the g-parameters, denoted $\boldsymbol{g} = (g_\alpha, g_\nu, g_\theta)$, and the remaining parameters, denoted $\boldsymbol{\lambda} = (\boldsymbol{\alpha}, \boldsymbol{\theta}, \mu, \nu, \sigma^2)$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are the collection of individual intercept and slope parameters, respectively. With this notation,

$$P(\boldsymbol{Y} \mid \mathcal{M}) = \int_{\boldsymbol{g}} \int_{\boldsymbol{\lambda}} P(\boldsymbol{Y}|\boldsymbol{g}, \boldsymbol{\lambda}) P(\boldsymbol{\lambda}) d\boldsymbol{\lambda} P(\boldsymbol{g}) d\boldsymbol{g}.$$

The inner integral, denoted $T(\boldsymbol{g}) = \int_{\boldsymbol{\lambda}} P(\boldsymbol{Y}|\boldsymbol{g}, \boldsymbol{\lambda}) P(\boldsymbol{\lambda}) d\boldsymbol{\lambda}$ may be solved analytically. The solution is developed in Rouder et al. (2012, p. 361-362) and implemented in Morey and Rouder's BayesFactor package for R (Morey & Rouder, 2015). Hence,

$$P(\boldsymbol{Y} \mid \mathcal{M}) = \int_{\boldsymbol{g}} T(\boldsymbol{g}) P(\boldsymbol{g}) d\boldsymbol{g}.$$

This integral is over only three dimensions in the current design. More importantly, $T(\boldsymbol{g})$ varies smoothly across the the paramater space of $\boldsymbol{g}$, and the evaluation may be performed to high precision by Monte Carlo sampling (see Rouder et al., 2012 for expanded analysis). This approach in implemented in BayesFactor package using the nWayAOV command separately for the unstructured model, the common-effect model, and the the null model.

Unfortunately, for the positive model, it is not possible to integrate $\boldsymbol{\lambda}$ analytically due to the range restrictions on $\boldsymbol{\theta}$. Instead, we follow the *encompassing* approach discussed by Klugkist and colleagues (Klugkist & Hoijtink, 2007; Klugkist et al., 2005). In our case, where the prior on the positive model is a truncated version of the prior on the unstructured model, a simple counting approach within MCMC-sampling works well. The main idea is as follows:

As seen in (3), the Bayes factor is the ratio of the posterior odds to the prior odds.

These odds may be given by

$$\frac{P(\mathcal{M}_+|\boldsymbol{Y})}{P(\mathcal{M}_u|\boldsymbol{Y})} = P(\boldsymbol{\theta} > \mathbf{0}|\mathbf{Y}, \mathcal{M}_\mathbf{u}),$$

and

$$\frac{P(\mathcal{M}_+)}{P(\mathcal{M}_u)} = P(\boldsymbol{\theta} > \mathbf{0}|\mathcal{M}_\mathbf{u}),$$

where $\boldsymbol{\theta} > \mathbf{0}$ refers to event that each element of $\boldsymbol{\theta}$ is greater than zero. The Bayes factor, the ratio, is given by

$$B_{+u} = \frac{P(\boldsymbol{\theta} > 0|\boldsymbol{Y}, \mathcal{M}_u)}{P(\boldsymbol{\theta} > 0|\mathcal{M}_u)}.$$

Restated, the Bayes factor is the posterior probability that all effects are positive relative to the prior probability of the same event.

Computation of these probabilities is straightforward in MCMC sampling. Let $\boldsymbol{\theta}[m]$ denote a vector of samples on the mth iteration under the unstructured model. The mth iteration is considered evidential of the positive-effects model if all I elements of $\boldsymbol{\theta}[m]$ are positive and not otherwise. Let n_1^+ be the number of evidential iterations conditional on data, and let n_0^+ be the same from the prior. Then, the Bayes factor is,

$$B_{+u} = \frac{n_1^+}{n_0^+}.$$

Additional derivations and development of this encompassing approach may be found in Klugkist et al. (2005). To compute the Bayes factor of the positive-effect model to the remaining models, we use the well-known transitivity of Bayes factors (Rouder & Morey, 2012).

Seven Data Sets

To demonstrate the need for the development of models like the above, we use them to assess the structure of effects in seven existing data sets that cover three common

experimental tasks with massively repeated designs. The tasks are common inference phenomena. The first phenomenon is the aforementioned Stroop interference effect. We analyzed three data sets, one from Von Bastian et al. (2015) and two from Pratte, Rouder, Morey, and Feng (2010).

The second phenomenon we analyzed is Simon interference (Simon, 1969). In a Simon interference task, participants are asked to identify a property of a stimulus, say its color. There are two responce options for each presented item. For the color example, patches may be presented in red or green. Participants press buttons either on the left or right, and an example might be to press a left key for red and a right key for green. The interference comes from the spatial placement of the patch. Patches are placed on the left or right: a congruent patch in this case is a red patch on the left and the congruency is between the spatial placement and the left-key response. Likewise a green patch on the right is congruent. Incongruent patches are red patches on the right or green patches on the left because the side of stimulus placement is opposite the side of the response. Simon effects are relatively small, say about 30 ms, but the phenomenon is nonetheless robust and well studied (C. H. Lu & Proctor, 1995). We analyzed three existing Simon interference data sets including one from Von Bastian et al. (2015) and two from Pratte et al. (2010).

The third phenomenon we analyzed is an Eriksen flanker task (B. A. Eriksen & Eriksen, 1974) from Von Bastian et al. (2015). In the Eriksen flanker task, the interference comes from distractors that are placed around a target stimulus. For example, participants identify a centrally located character, e.g. a consonant like “H”. Those targets are surrounded by distractors that could either match the target character, say another consonant “K”, or they mismatch, say a vowel like “E”.

Table 1 provides an overview of all seven data sets. For the derivations of the used F-statistic and the confidence intervals in Figures 5 to 7, see Appendix A. The cleaning strategy for the RT data for all seven data sets is provided in Appendix B.

Data Set 1

Set 1 is the Stroop task data from Von Bastian et al. (2015). The task used is commonly called a *number Stroop task* (West, Jakubek, Wymbs, Perry, & Moore, 2005 refer to the task as counting task), and it goes as follows: On each trial, participants saw a string of digits. In each string, the digits were always replicates, say "22" or "444", and the lengths varied from one digit to four digits. The participants task was to report the length, for example, the correct report for "444" is 3. In the congruent condition, the length and the digits matched; e.g., "22" and "4444." " In the incongruent condition, the length and digits mismatched, e.g., "44" and "2222." Additionally, participants saw neutral trials that consisted of unrelated symbols, e.g. "###". In total, 121 participants each responded to 48 congruent, 48 incongruent and 48 neutral trials. For the following analysis, we only used data from the congruent and the incongruent conditions.

The mean effect and the individual effects of all 121 participants are shown in Figure 5A$_1$. The dashed line is the mean effect $\bar{d} = 65$ ms ($SD = 47$ ms), which corresponds to an effect size of $d = 1.37$. The points show ordered individual effects, and the gray outer surface denotes the 95% confidence intervals for each individual (see the Appendix for derivations). The graph shows the same information as Figure 2. A one-way random-effects ANOVA (see Appendix for derivations) reveals individual variability, $F(120, 11003) = 1.30$, $p \approx 0.016$.

Data Set 2

Set 2 was conducted by Pratte et al. (2010) (Experiment 1, Stroop task). The task they used is a classical Stroop task: Participants were asked to identify the color of the presented color words, e.g. the word "RED" presented in blue. In the congruent condition, presentation color and word meaning matched, e.g. "BLUE" presented in blue. In the incongruent condition, they did not match, e.g. "RED" presented in blue. In the neutral condition, a neutral word was presented in a color, e.g. "XXXX" presented in blue. A total of 38 participants responded to 168 trial for each of those conditions. For the following

analysis, we only used data from the congruent and the incongruent condition.

The mean effect and the individual effects are shown in Figure $5B_1$. The dashed line is the mean effect $\bar{d} = 91$ ms ($SD = 50$ ms), with a corresponding effect size of $d = 1.81$. The points in the figure show ordered individual effects, and the gray outer surface denotes the 95% confidence intervals for each individual. A one-way random-effects ANOVA shows individual variability, $F(37, 11038) = 2.60$, $p \approx 0$.

Data Set 3

Set 3 was originally conducted by Pratte et al. (2010). The data used for this analysis stems from the five Stroop task blocks of Experiment 2 of the original study. In this task, the stimuli were the words *left* and *right*, presented on the left or the right side of the screen. Participants were asked to identify the position of the word while ignoring the meaning of the word. A congruent trial occured when position of the word and word meaning corresponded; an incongruent trial emerged when position and word meaning did not correspond. In total, 38 participants responded to 180 congruent and 180 incongruent trials.

Mean and individual effects are shown in Figure $5C_1$. The dashed line represents the mean effect $\bar{d} = 12$ ms ($SD = 20$ ms), with a corresponding effect size of $d = 0.60$. The points in the figure show ordered individual effects, and the gray outer surface denotes the 95% confidence intervals for each individual. A one-way random-effects ANOVA reveals no significant individual variability, $F(37, 12489) = 1.25$, $p \approx 0.14$.

Data Set 4

Set 4 is the Simon task data from Von Bastian et al. (2015). In this task, either a red or a green circle was presented to the participants. The circles were either presented on the right or on the left side of the screen. Participants were asked to respond with the left-arrow key to green circles and with the right-arrow key to red circles. In the congruent condition, position of the circle and response position match. In the incongruent condition, position of

the circle and response position do not match. In total, 121 participants responded to 150 congruent trials and 50 incongruent trials.

Figure $6A_1$ shows mean and individual effects. The dashed line shows the large mean effect $\bar{d} = 12$ ms ($SD = 36$ ms), with a corresponding effect size of $d = 2.22$. The points in the figure show ordered individual effects, and the gray outer surface denotes each individual's 95% confidence interval. A one-way random-effects ANOVA reveals individual variability, $F(120, 23211) = 2.65$, $p \approx 0$.

Data Set 5

Data Set 5 is from experiment 2 conducted by Pratte et al. (2010) (Simon task). In this task, participants saw either a green or a red target stimulus on each trial. These targets were either presented on the left or the right side of the screen. The participants were instructed to press a key with their right hand for green and another key with their left hand for red stimuli. In a congruent trial, target position and response position match. In an incongruent trial, target position and response position mismatch. For this set, 38 participants completed 252 incongruent and 252 congruent trials.

Mean and individual effects are shown in Figure $6B_1$. The dashed line is the mean effect $\bar{d} = 17$ ms ($SD = 24$ ms), with a corresponding effect size of $d = 0.72$. The points in the figure show ordered individual effects. A one-way random-effects ANOVA shows individual variability, $F(37, 17267) = 1.82$, $p \approx 0.002$.

Data Set 6

This set is from experiment 1 conducted by Pratte et al. (2010) (Simon task). In this task, the target stimuli were the words "left" and "right". These words were either presented on the left or the right side of the screen. The participants were instructed to press a key with their right hand for the word "right" and another key with their left hand for the word "left". In a congruent trial, target position and word meaning match. In an incongruent trial,

target position and word meaning mismatch. In total, 38 participants completed 180 incongruent and 180 congruent trials.

Figure $6C_1$ shows mean and individual effects. The dashed line represents the mean effect $\bar{d} = 30$ ms ($SD = 30$ ms), with a corresponding effect size of $d = 1.02$. The points in the figure show ordered individual effects. A one-way random-effects ANOVA reveals individual variability, $F(37, 12190) = 2.29$, $p \approx 0$.

Data Set 7

Set 7 is the Eriksen flanker task in Von Bastian et al. (2015). In the task used, participants saw a string of seven characters. All of these characters were identical except for the one in the middle. The task was to decide whether this centrally located character was a vowel, e.g. "S", or a consonant, e.g. "A" or "E". The flanking six characters were as well either vowels, consonants or neither, e.g. "#". In the congruent condition, the flanking characters match the central character, e.g. "AAAEAAA". In the incongruent condition, the flanking characters and the central character mismatch, e.g. "SSSESSS". In the neutral condition, the central character was flanked by "#", e.g. "###E###". In total, 121 participants responded to 48 trials for each of those conditions. For the following analysis, we used data from the congruent and the incongruent condition.

The mean effect and the individual effects are shown in Figure 7A. The dashed line represents the mean effect $\bar{d} = 2$ ms ($SD = 32$ ms), with a corresponding effect size of $d = 0.07$. The points in the figure show ordered individual effects, and the gray outer surface denotes the 95% confidence intervals for each individual. A one-way random-effects ANOVA shows no significant individual variability, $F(120, 10973) = 0.98$, $p \approx 0.538$.

Analyses and Results

To illusrate the advantages of the Bayesian modeling approach, we analyze the seven data sets with the four Bayesian models: the unstructured model, $\mathcal{M}_u$; the positive-effects model, $\mathcal{M}_+$; the common-effect model, $\mathcal{M}_1$; and the null model, $\mathcal{M}_0$.

Parameter Estimation

Posterior distributions for all parameters in the unstructured, the common-effect and the null models were sampled with Markov chain Monte Carlo methods (MCMC). In all cases, conditional posterior distributions of parameters may be derived straightforwardly from Bayes rule (Gelman, Carlin, Stern, & Rubin, 2004; Jackman, 2009; Rouder & Lu, 2005). Priors were chosen to leverage conjugacy, and consequently in all cases posterior distributions may be sampled from known distributions as Gibbs steps (Gelfand & Smith, 1990). This approach is implemented in the BayesFactor package in R (Morey & Rouder, 2015). There are alternative packages for computing posterior distributions such as STAN (eg. Stan Development Team, 2016) and JAGS (Love et al., n.d.). These alternatives do a fine job as well, and estimation in this setting is computationally convenient. The BayesFactor package is not as general as these alternatives, but it is perfectly tuned for the problem at hand.

Mixing. MCMC chains are guaranteed to provide samples from the joint posterior of all parameters in the large-sample limit. In practice, however, outputs from successive iterations are often correlated. If this correlation is severe, then there is no guarantee that the posterior has been well explored. This problem of excessive correlation, when it exists is known as a problem of *mixing*. Chains without excessive correlation from iteration to iteration are said to mix well, chains with excessive correlation are said to mix poorly.

MCMC chains mixed well for the three estimated models. This rapid convergence is expected here as the models are linear and not overly saturated. The slowest converging parameters were the variances of the effects, η^2. Figure 8A and B provide a snippet of a chain and the autocorrelation function, respectively, for this parameter for the unstructured model in the worst case across all data sets (which was from Data Set 1). As can be seen, mixing is acceptable even in this worst case. The correlations here are not consequential for long runs of several thousand iterations (our posterior estimates come from 20,000 iterations).

The mixing in the same case for Data Set 2 was considerably better. Figure 8C and D show mixing for this Data Set, also is for η^2. Even though this is the worst-case for the Data

Set, the mixing here is quite good.

Results. The critical parameters are the individuals' effects, θ_i. The posterior means of these parameters are shown in the right-hand columns in Figures 5, 6, and 7 for the seven data sets. The dark gray lines that have the largest spread are the observed effects, d_i, and these are included for comparison to the model estimates. The points are the estimates θ_i from the unstructured model. The red horizontal line is the common effect estimate from the common-effect model $\mathcal{M}_1$. Note that the right-hand columns in the figures do not have the same scaling on the y-axis when comparing effects between different paradigms.

There are three main findings:

1. There is a sizable degree of shrinkage in all seven data sets. In all cases, much of the variation in the observed effects is due to sample variation at the trial level rather than true variation across individuals. This shrinkage is especially striking for Data Set 1, 3, 5, and 7.

2. According to the unstructured model, most individuals have positively-valued effects. This positivity holds even when the observed effects are negative. This positivity is a direct result of shrinkage to the overall mean, which is positive for Data Sets 1 through 6.

3. The hierarchical models provide a reasoned estimate of individual variation while simultaneously accounting for sample noise at the trial level. The results are that there is dramatically less variation at the individual level. Consequently, the effect size, the magnitidue of the mean effect relative to the variation in individuals, is larger after accounting for trial-by-trial variation. For Data Set 1, for example, the observed effect size is 1.37 reflecting a mean effect of 65 ms and an observed standard deviation around this mean of 47.37 ms. For the unstructured model estimates, in contrast, the effect size is 9.68, and this increase reflects the decreased variation in individuals' effects. It is our belief that in within-subject designs, appropriately measured

hierarchical model-based effect sizes are much larger than is typically reported. The observed and model-based effect sizes for all data sets are shown in Table 1.

In addition to these three main findings, the severe limitations of estimation become apparent. For our main question, "does everyone Stroop in the same direction," there is no single parameter that answers the question. In fact, we cannot even note that for many of the graphs, all individual parameters are positive. These parameter estimates are highly correlated due to the prior specification, and cannot be used en masse to draw any conclusion, certainly not by inspection. This application shows that there are questions that cannot be addressed by estimation alone (Morey, Rouder, Verhagen, & Wagenmakers, 2014).

Model Comparison

To compare the four models, we use the Bayes factor approach discussed previously. The values for all seven data sets are provided in Table 2. The asterisk in each column marks the preferred model for each data set. The values in the table are the Bayes factor between the respective model and the preferred model. For example, for Data Set 1, the positive-effects model is the preferred model. The table shows how much worse the other models perform compared to the preferred one. The Bayes factor for the runner-up, the common-effect model, over the positive-effects model for Data Set 1 is 1-to-11. The Bayes factor for the unstructured model over the positive-effects model is 1-to-12 for the same data set. These Bayes factors indicate that the two models, $\mathcal{M}_1$ and $\mathcal{M}_u$, predict the data worse than the positive-effects model by about an order of magnitude. The Bayes factor for the null model is much worse, $B_{0+} = 10^{-62}$.

There are two major findings:

1. The preferred model varies across the tasks. For the Stroop tasks, the positive-effects and the common effect models are preferred, indicating that all people Stroop in the same, expected direction. For the Simon tasks, in contrast, the unstructured model was slightly preferred for two of the three sets. This result provides modest evidence that

perhaps some people truly have reverse Simon effects, where spatially incompatible responses are speeded relative to spatially compatible ones. For the Flanker task, there was very little effect in the data and, not surprisingly, the null model was preferred.

2. In Data Sets 3 and 7 there is no evidence for individual variablility. In these sets the mean effect is quite small and presumably any true individual differences, should they exist, would even be smaller. Hence, finding them would require larger numbers of trials per individual. The combination of the estimation results and the Bayes factor model comparisons provide the following interpretation: Although there is far less variability in true individual effects than in observed effects, the degree of true variation is nonetheless substantial for five of the seven data sets. Also, the concordance between shrinkage in estimation and the Bayes factors is noteworthy—the greater the shrinkage, the better the relative performance of the common-effect model to the unstructured and positive-effects models.

There is also a subtle paradox to reconcile. Consider the results of Data Set 5, the Simon experiment in Figure $6B_2$. Here we see that all the hierarchical model estimates of the effect are positive for all individuals. Yet, the Bayes factor analysis indicates that the unstructured model is preferred to the positive model, albeit slightly. The combination of results may strike some as paradoxical—how can all estimates be positive and yet the positive model not be preferred? The key to resolving this paradox is to note that these estimates are not independent. They are heavily influenced by the mean effect, ν. This influence is a desirable property in hierarchical modeling, and its benefitial effect is to smooth or regularize each individual's estimate. While these estimates are our best point estimate of individuals' effects, they do not preserve global properties, such as whether all are positive. The only way we know to assess these global properties is through Bayes factors, which are the direct and immediate consequences of Bayes rule for the questions we ask.

Sensitivity to Prior Settings

In analysis, it is necessary to specify the prior scales r_α, r_ν, and r_θ, which are the standardized scales on individual intercepts, the population mean effect, and individual variation around this mean. Of these three, the setting of r_α is relatively inconsequential as individual intercepts are specified in all four models. The other two settings are consequential as they define the dimensionality of the models on θ_i. If, for example, r_θ is large, then there is little *a priori* correlation among the individual effects, and the dimensionality of this part of the model is large. Likewise, if r_θ is small, then all people have nearly the same effect, and the dimensionality of this part of the model is small.

Before exploring the effects of different settings of r_ν and r_θ, we provide some guidance as context. It may seem natural to view dependence of the Bayes factor on these settings as difficult or undesirable. We think, however, the situation is far more nuanced. In hierarchical models, the settings are more akin to model specification inasmuch as they set the dimensionality of the model on individual differences. Setting the scale too small makes the unstructured model resemble too closely the common-effect model; setting the scale too large makes the unstructured model needlessly complex. To us, dimensionality should be a key consideration in specification. And where we would expect inference about a model to depend quite heavily on its specification including its complexity, we expect the Bayes factors to depend markedly on these settings. We first highlight the dependency and then provide some interpretation.

Table 3 shows the Bayes factors for different settings of r_ν and r_θ for Data Set 1. The first line provides the settings we used for our analysis, $r_\nu = 1/6$ and $r_\theta = 1/10$. If we assume 300 ms of overall variability, the settings translate into about 50 ms and 30 ms, respectively. For these settings, the estimates of θ_i (posterior means) from the unstructured model have a standard deviation of 7 ms, which is a data-driven estimate of the true individual variability under the assumption that individuals truly vary. One way to view this state is that we had expected around 30 ms of individual variation *a priori* but observed only 7 ms in the data.

Bayes factors here favor the positive-effects model over the unstructured and common-effect model indicating that while these 7 ms are small, they are detectable as individual variation.

The next two lines of Table 3 show the case that the settings are either halved or doubled. In the second line, we expect a smaller effect overall and smaller individual variation around this effect. We see here attenuation—the estimated individual variation is 4 ms—and still this small estimate is detectable. The next line shows the case of doubling the expectations; here we see that although all the conclusions remain, the evidence for the positive-effects model is not as great as before. It predicts the data 2.7 times better than the common-effect model. If we continue to increase r_θ, the scale of individual variation, the common-effect model gains, and is preferred. Indeed, as shown in the fifth and the seventh lines, this trend can be quite extreme.

To us, these dependencies are reasonable and desirable. First, in the range that we consider reasonable, the Bayes factors have modest variation. For example, if we take lines two and three as defining a reasonable range for effect scales, the Bayes factors between the winning model, the positive-effects model, and the common-effect model varies from 2.7-to-1 to 14.2-to-1. While this variation is moderate, it covers in our opinion the full range of variation across reasonable researchers. This amount of variation may be small when compared to variation from other subjective elements in research. Second, if researchers make unreasonable commitments, then the Bayes factors change dramatically. An r_ν or r_θ of 1.0 is unreasonable because nobody expects to find an effect that is equal in size to the variation in response times across repeated trials. If this were so, no sane researcher would use 50 or 100 repeated trials per person per condition. In this case, the unreasonable specification leads to predicted effects that are far too variable compared to the data. Consequently, the Bayes factors indicate that these unreasonable specifications are too complex.

General Discussion

In this paper we develop a set of Bayesian mixed models for assessing multiple equality and order constraints in simple experimental paradigms. There are four models: the null model, where no individuals have an effect; the common-effect model, where all individuals share a common effect; the positive-effects model, where individual effects are constrained to be positive; and the unstructured model, where no such constraints are placed on individual effects. We compare these four models with Bayes factors, which are principled and convenient in this context. We note that estimation without Bayes factors does not address the questions at hand, because there is no single parameter that captures the relevant constraints.

From a psychological perspective, perhaps the most important consideration for well-established effects is whether the order constraint holds. In the Stroop case, for example, we ask whether all individuals have true Stroop effects in the same direction where congruent colors are named more quickly than incongruent ones. This constraint is compatible with the leading explanation that Stroop interference results from the fact that reading is quick and automatic for competent readers (MacLeod, 1991). And assuming that all of the college participants are competent readers, then the constraint should hold. Indeed, we suspect that many tasks will plausibly obey an order constraint. The interpretation is that they are mediated by nearly universal, automatic processes that do not admit a reverse ordering. We do not expect, however, that all tasks will order as such, and in those that do not, it is reasonable to search for differing strategies and processing among participants.

Stroop, Simon, and Eriksen Flanker Interference

In the course of development, we chose Stroop, Simon, and Eriksen flanker interference as our first application. We are familiar with these types of interference (Pratte et al., 2010; Rouder, Yue, Speckman, Pratte, & Province, 2010; Speckman, Rouder, Morey, & Pratte, 2008) and were fairly certain *a priori* that Stroop intereference would obey the order

constraint. Hence, we are not at all surprised that in all three Stroop data sets, the positive-effects, or positive common-effect models were preferred.

Stroop effects follow a common pattern when response time distributions are considered (Rouder et al., 2010). Figure 9 shows *delta plots* for Data Sets 1-6. A delta plot, first introduced by De Jong, Liang, and Lauber (1994), is a rotated version of a QQ plot (Zhang & Kornblum, 1997). Each point denotes a RT percentile rank, and for the nine points displayed on each line, the ranks are the $10^{th}, 20^{th}, \ldots, 90^{th}$ percentiles. The x-axis value of a point is the average RT at that percentile across the congruent and incongruent condition; the y-axis value is the difference, that is, the effect.[3] There is a common pattern to these plots where effects tend to be smaller for lower percentiles (the faster responses) and larger for higher percentiles (the slower responses) (Luce, 1986; Rouder et al., 2010; Wagenmakers & Brown, 2007). Indeed, Stroop interference effects in the literature show this common pattern (see Pratte et al., 2010), and it is present in the analyzed data sets as well (Figure 9A).

Simon interference has intrigued researchers for decades because the phenomenology seems idiosyncratic. Delta plots commonly have a markedly different pattern. They start positive but have a negative slope. In several experimental reports, they actually cross the centerline implying that the slowest responses to incongruent stimuli are faster than the slowest responses to congruent ones! The negative slope has been observed regularly (e.g. Burle, van den Wildenberg, & Ridderinkhof, 2005; De Jong et al., 1994; Styrkowiec & Szczepanowski, 2013). In Figure 9B, delta plots of the three Simon interference data sets are shown. Two of them show the typical negative-slope pattern. These two data sets are the ones where the unconstrained model was preferred in the Bayes factor model comparison.

[3]Following Zhang and Kornblum (1997), we compute the points on the delta plot lines as follows: Let $\bar{y}_j^p$ denote the average reaction time for the pth percentile and the jth condition across individuals and trials. This is a reasonable values as long as the assumption holds that the individuals' RT distributions do not vary in shape. We can now compute the average RT for the pth percentile, $\frac{\bar{y}_1^p + \bar{y}_2^p}{2}$ and the average pth percentile effect, $\bar{y}_2^p - \bar{y}_1^p$. For a delta plot, we plot those values against each other. If the slope of the line is positive, the fastest responses have the smallest, or even a negative effect. If the slope is negative, the fastest responses have the biggest effect.

Data Set 5 shows the reversal that characterizes data from Burle et al. (2005) as well. There is no other effect that we are aware of that has this negative-slope pattern. Based on this result as well as electrophysiology results, the leading theory of Simon intereference is that it reflects a quick automatic process followed by a separate, slower, compensation process (Ridderinkhof, 1997).

The plausible presence of two opposing processes for Simon interference sets up the possiblity that individual mean Simon effects are not order constrained. Most participants may have a larger early automatic positive component, but some may have a larger, later negative component. Hence, the unstructured model may provide the best account.

We considered predictions before analysis and speculated that Stroop interference would obey the order constraint—indeed, everyone Stroops. We decided not to speculate about Simon interference beforehand. If the order constraint held, it could be that the positive component was always larger than the negative one. Likewise, if it did not, as it may not here, then there is individual variation in the relative sizes of the components.

The combination of results do lead to a new conjecture. There may be a profitable link between the delta plot pattern and the order restriction. Perhaps whenever the slope is negative—an indication for two opposing processes—the order constraint may be violated.

The null Eriksen flanker interference result is surprising because Eriksen flanker effects are well known and often reported. We did find an Eriksen flanker effect in accuracy in Data Set 7 (see Von Bastian et al., 2015) but, as shown, not in RT. Given the prevalence of RT effects in the literature, and that we did not collect these data, we do not further speculate on the null RT result.

The Interpretability of a Common Effect

Two of our four models, the null model and the common effect model, are novel in that they specify no individual differences whatsoever. To our knowledge, individual differences researchers rarely consider a lack of individual differences. In this regard, these models serve

as important controls. Indeed, in our data, these models with no individual differences outperformed the others in two of the seven data sets.

The null model seems plausible or at least theoretically useful as a bound on human behavior. There are simply some effects that do not occur for anyone, perhaps the ability to predict lottery outcomes above chance. Hence, this model is fairly interpretable. But what about the common-effect model? The notion that there is a natural constant for Stroop, that each person has the same exact effect, say 60 ms, is not too believable. If we deem this model unbelievable, then how may we interpret the event that it outperforms the other models?

Perhaps the most fruitful position is to interpret the strong performance of this model in the context of the experimental design. When the data are few in number, then simpler models are preferred to more complicated ones. If researchers are interested in studying the structure of individual variation, then they need to employ designs with larger sample sizes. The analyses here indicate that, from a design perspective, the number of trials per individual per condition is critical. When this number is small, it is difficult to separate true individual variation from sample noise, and in light of this difficulty, the common-effect model provides a more parsimonious description (as it should). As this number becomes larger, the separation of sources of variability is more stable, and evidence for true individual variation may be observed through model comparison. Hence, we recommend the individual-difference researchers to consider the common-effect model as a check that designs have sufficient trials per individual to resolve true individual differences.

Alternatively, if researchers are only interested in model selection, and the common-effect model is deemed implausible, it may be downweighted, or eliminated completely, through setting lower or zero prior odds, respectively.

Computational Considerations

The main computational issue in analyzing these models is computing Bayes factors. The Bayes factor is the relative probability of data, and computing this probability requires

accurate integration across all parameters with respect to the priors. To make the integration convenient, we used a g-prior setup that is common in linear models (Bayarri & Garcia-Donato, 2007; F. Liang et al., 2008). In this setup, models are placed on standardized effect sizes rather than on effects themsleves. With this setup, all parameters save the variabilities on effect sizes ($g_\alpha, g_\nu, g_\theta$) may be integrated in closed form, greatly simplifying the problem.

The g-prior setup was not our first choice. Instead, we placed models on effects themselves, and tried a brute-force approach known as the *Savage-Dickey Density Ratio* (Dickey, 1976) for computing the Bayes factor between the common-effect and unstructured models. This method is precendented and recommended in psychology (Morey, Rouder, Pratte, & Speckman, 2011; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Wetzels, Grasman, & Wagenmakers, 2010). Unfortunately, the posterior density ratio estimates are too unstable to compute accurately with MCMC outputs. Samples varied by some 40 orders of magnitude and convergence of the running mean was not achieved even with ten million MCMC iterations.

Limitations

Although the g-prior approach with symbolic integration of most parameters is suitable here, there are two substantial limitations affecting future generalizations. One of our goals at the outset of this project was to include mixture models, say those where each individual had identically no effect or came from a positive distribution. These types of mixture models are common in Bayesian analysis and go under the moniker of "spike-and-slab" priors (George & McCulloch, 1993). The usual goal with these models is to categorize each individual as being in the spike, that is having no effect, or being in the slab, that is, having an effect. In this setting, a separate Bayes factor for each individual can be computed. To our knowledge, there has been no consideration of the more natural goal in this setting—to assess whether the constraints from the mixture model, taken globally, provide a good

description of the structure in the data. We seek to compare the mixture model as a whole to the positive-effects model, the equivalent model without the spike. Whether the approach here generalizes to the mixture case remains unclear.

Another of our goals was developing realistic three-parameter lognormal models on response time such as those in (Rouder, Province, Morey, Gomez, & Heathcote, 2015). Currently, we use the normal with constant variance. The lognormal is desirable because it captures the fact that response time distributions are skewed and that effects tend to be manifest in the scale rather than in the shift or shape of the distribution. It is not clear, however, how to integrate log-normal parameters in the three-parameter version.

Although these generalizations do not yield convenient closed form solutions for the integration, there are still avenues for future development. Fortunately, the computational toolbox for Bayes factors in mixed models is sizable and growing. Candidates for future work include bridge sampling (Meng & Wong, 1996), importance sampling (Kass & Raftery, 1995) and Laplace approximation (Kass & Raftery, 1995; Raftery, 1996). Hopefully, progress in computational issues will allow for useful generalizations of the models developed here.

Appendix

Appendix A.

In this appendix we derive confidence intervals and the F-test. Let Y_{ijk} denote the kth response for the ith participant in the jth conditions, $i = 1, \ldots, I$, $j = 1, 2$, $k = 1, \ldots, K_{ij}$.

Individuals' Confidence Intervals. The goal here is to derive confidence intervals for each individual in isolation without any model-based pooling. Let M_{i1} and M_{i2} denote the sample means in the congruent and incongruent conditions, respectively, for the ith participant. Let V_{i1} and V_{i2} denote the corresponding sample variances. The sample effect, d_i, is $d_i = M_{i2} - M_{i1}$ and standard error of this effect, denoted s_{di} is

$$s_{di} = \left(\frac{V_{i1}}{K_{i1}} + \frac{V_{i2}}{K_{i2}} \right)^{1/2}.$$

This standard error may be used to compute individuals' CIs in the usual manner (Hays, 1994).

***F*-test.** The goal is to derive a one-way, random-effects F-test to assess whether there is any variation across individuals. This derived value can be interpreted as a proper F if equal sample sizes are assumed (i.e. equal trial number in congruent and incongruent conditions). The effect, $d_i = M_{i2} - M_{i1}$, is distributed as

$$d_i \sim \text{Normal}\left(\mu_{i2} - \mu_{i1}, \frac{\sigma^2}{K_{i1}} + \frac{\sigma^2}{K_{i2}} \right),$$

where μ_{i2} and μ_{i1} are true conditions means for the ith person, and σ^2 is the variability in any observation around its true mean. This expression may be expressed as

$$d_i \sim \text{N}\left(\mu_{i2} - \mu_{i1}, \frac{\sigma^2}{K_i^*} \right), \tag{5}$$

where K_i^* is the *effective sample size* for the ith individual:

$$K_i^* = \frac{K_{i1}K_{i2}}{K_{i1} + K_{i2}}.$$

Consequently,

$$\sqrt{K_i^*}d_i \sim \text{Normal}(\sqrt{K_i^*}(\mu_{i2} - \mu_{i1}), \sigma^2).$$

Let $d_i^* = d_i\sqrt{K_i^*}$. Consider the case that $\sqrt{K_i^*}$ is approximately constant for all individuals, that is the design is approximately balanced. Then, under the null that $\mu_{i2} - \mu_{i1}$ is constant for all individuals, the following expression is a between-participant estimator of σ^2:

$$s_1^2 = \frac{\sum_i (d_i^* - \bar{d^*})^2}{(I-1)},$$

where $\bar{d^*}$ is the mean of all d_i^*. The degrees-of-freedom associated with this estimator is $I - 1$.

The expression for within-subject variation follows from the usual considerations:

$$s_2^2 = \frac{\sum_{ijk}(Y_{ijk} - M_{ij})^2}{\sum_{ij}(K_{ij} - 1)} = \frac{\sum_{ijk}(Y_{ijk} - M_{ij})^2}{N - IJ},$$

where $N = \sum_{ij} K_{ij}$ is the total number of observations. The degrees-of-freedom associated with this estimator is $N - IJ$.

With these expressions, the F statistic is

$$F = \frac{s_1^2}{s_2^2},$$

which under the null is distributed as an F distribution with $\nu_1 = I - 1$ and $\nu_2 = N - IJ$ degrees-of-freedom. Of course, this computation only holds for nearly balanced designs. Fortunately, deviations from balance only occurs when participants make errors, which is not that common here (see Appendix B).

Appendix B.

Neither of the authors discussed their strategies for cleaning the data in the original articles, even though there were clear outliers in the data. The cleaning code of Pratte et al. (2010) was available to us. The authors used three relatively strict criteria to decide which data points should be excluded and we followed all three criteria for Data Sets 2, 3, 5, and 6 and two of the criteria for Data Sets 1, 4, and 7 (data from Von Bastian et al. (2015)). The criteria were

I. All incorrect trials were removed.
II. We removed all trials with RTs less than .2 sec on the grounds that these times are too fast to be related to the processes of interest. We removed all trials with RTs greater than 2 sec on the grounds that these times are too slow to be related to the processes of interest.
III. We removed the first five trials in each experimental block, accounting for the familiarization with the task.

These removals comprised between 3 % and 13 % of the originally collected trials in each Data Set. Table 4 shows those removals for all Data Sets broken down for the three criteria. The data of all seven sets and the cleaning code are available at https://github.com/PerceptionCognitionLab/data0/tree/master/contexteffects.

References

Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science*, *15*, 88–93.

Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283–316.

Bates, D., & Maechler, M. (2017). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Bayarri, M. J., & Garcia-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, *94*, 135–152.

Burle, B., van den Wildenberg, W. P. M., & Ridderinkhof, K. R. (2005). Dynamics of facilitation and interference in cue-priming and Simon tasks. *European Journal of Cognitive Psychology*, *17*, 619–641.

Chaussé, P. (2010). Computing generalized method of moments and generalized empirical likelihood with R. *Journal of Statistical Software*, *34*(11), 1–35. Retrieved from http://www.jstatsoft.org/v34/i11/

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Coren, S. (1993). *The left-hander syndrome: The causes and consequences of left-handedness*. New York: The Free Press.

De Jong, R., Liang, C. C., & Lauber, E. (1994). Conditional and unconditional automaticity: A dual-process model of effects of spatial stimulus-response concordance. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 731–750.

DeGroot, M. H. (1982). Lindley's paradox: Comment. *Journal of the American Statistical*

Association, *77*(378), 336–339. Retrieved from http://www.jstor.org/stable/2287246

Dickey, J. M. (1976). Approximate posterior distributions. *Journal of the American Statistical Association*, *71*(355), 680–689. Retrieved from http://www.jstor.org/stable/2285601

Dinapoli, N., & Gatta, R. (2015). *Spatialfil: Application of 2D convolution kernel filters to matrices or 3D arrays*. Retrieved from https://CRAN.R-project.org/package=spatialfil

Douglas Nychka, Reinhard Furrer, John Paige, & Stephan Sain. (2015). Fields: Tools for spatial data. Boulder, CO, USA: University Corporation for Atmospheric Research. doi:10.5065/D6W957CT

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. doi:10.3758/BF03203267

Furrer, R., & Sain, S. R. (2010). spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. *Journal of Statistical Software*, *36*(10), 1–25. Retrieved from http://www.jstatsoft.org/v36/i10/

Gelfand, A., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis (2nd edition)*. London: Chapman; Hall.

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. Heidelberg: Springer-Verlag.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*, 881–889.

Gerber, F., & Furrer, R. (2015). Pitfalls in the implementation of Bayesian hierarchical

modeling of areal count data: An illustration using BYM and Leroux models. *Journal of Statistical Software, Code Snippets*, *63*(1), 1–32. Retrieved from http://www.jstatsoft.org/v63/c01/

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, *1*, 403–420. Retrieved from http://dx.doi.org/10.1214/06-ba116

Greenwald, A., Klinger, M., & Schuh, E. (1995). Activation by marginally perceptible ("subliminal") stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General*, *124*, 22–42.

Hays, W. L. (1994). *Statistics* (fifth.). Ft. Worth, T.X.: Harcourt Brace.

Hobert, J. P., & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, *91*, 1461–1473.

J, L. (2006). Plotrix: A package in the red light district of r. *R-News*, *6*(4), 8–12.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Chichester, United Kingdom: John Wiley & Sons.

Jackson, C. H. (2011). Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, *38*(8), 1–29. Retrieved from http://www.jstatsoft.org/v38/i08/

Jeffreys, H. (1961). *Theory of probability (3rd edition)*. New York: Oxford University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454. Retrieved from http://www.sciencedirect.com/science/article/pii/0010028572900163

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*(12), 6367–6379.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing

priors. *Statistica Neerlandica*, *59*, 57–69.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299–312.

Laplace, P. S. (1986). Memoir on the probability of the causes of events. *Statistical Science*, *1*(3), 364–378. Retrieved from http://www.jstor.org/stable/2245476

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423. Retrieved from http://pubs.amstat.org/doi/pdf/10.1198/016214507000001337

Love, J., Selker, R., Verhagen, J., Smira, M., Wild, A., Marsman, M., ... Wagenmakers, E.-J. (n.d.). *JASP (version 0.40)*.

Lu, C. H., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychonomic Bulletin and Review*, *2*(2), 174–207.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

MacLeod, C. (1991). Half a century of research on the stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163–203.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, *42*(9), 22. Retrieved from http://www.jstatsoft.org/v42/i09/

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 831–860.

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

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*, –.

Retrieved from http://www.sciencedirect.com/science/article/pii/S0022249615000723

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*, 368–378. Retrieved from http://dx.doi.org/10.1016/j.jmp.2011.06.004

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming. *Psychological Science*, 1289–1290.

Mulder, J., Klugkist, I., Schoot, R. van de, Meeus, W. H. J., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *54*(530-546).

Ooms, J. (2016). *Curl: A modern and flexible web client for r*. Retrieved from https://CRAN.R-project.org/package=curl

Overstall, A. M., & Forster, J. J. (2010). Default bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, *54*(12), 3269–3288.

Plate, T., & Heiberger, R. (2016). *Abind: Combine multidimensional arrays*. Retrieved from https://CRAN.R-project.org/package=abind

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from https://journal.r-project.org/archive/

Pratte, M. S., Rouder, J. N., Morey, R. D., & Feng, C. (2010). Exploring the differences in distributional properties between Stroop and Simon effects using delta plots. *Attention, Perception & Psychophysics*, *72*, 2013–2025.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in

generalised linear models. *Biometrika*, *83*, 251–266.

Richard A. Becker, O. S. code by, Ray Brownrigg. Enhancements by Thomas P Minka, A. R. W. R. version by, & Deckmyn., A. (2016). *Maps: Draw geographical maps*. Retrieved from https://CRAN.R-project.org/package=maps

Ridderinkhof, K. R. (1997). A dual-route processing architecture for stimulus-response correspondence effects. In B. Hommel & W. Prinz (Eds.), *Theoretical issues in S-R compatibility* (pp. 119–131). Amsterdam, the Netherlands: Elsevier.

Robertson, T., Wright, F., & Dykstra, R. (1988). *Order restricted statistical inference.* Wiley, New York.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, *22*, 9475–9489.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review*, *12*, 573–604.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877–903. Retrieved from http://dx.doi.org/10.1080/00273171.2012.734737

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological sciencecollabra. *Collabra*, *2*, 6. Retrieved from http://doi.org/10.1525/collabra.28

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374. Retrieved from http://dx.doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Morey, R. D., Verhagan, A. J., Swagman, A., & Wagenmakers, E.-J. (2016). Bayesian analysis of factorial designs. *Psychological Methods*.

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The

lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225–237. Retrieved from http://dx.doi.org/10.3758/PBR.16.2.225

Rouder, J. N., Yue, Y., Speckman, P. L., Pratte, M. S., & Province, J. M. (2010). Gradual growth vs. shape invariance in perceptual decision making. *Psychological Review*, *117*, 1267–1274.

Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of Experimental Psychology*, *81*, 174–176.

Speckman, P. L., Rouder, J. N., Morey, R. D., & Pratte, M. S. (2008). Delta plots and coherent distribution ordering. *The American Statistician*, *62*, 262–266.

Stan Development Team. (2016). RStan: The R interface to Stan. Retrieved from http://mc-stan.org/

Statisticat, & LLC. (2016). *LaplacesDemon: Complete environment for bayesian inference*. Bayesian-Inference.com. Retrieved from https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Styrkowiec, P., & Szczepanowski, R. (2013). Space positional and motion src effects: A comparison with the use of reaction time distribution analysis. *Advances in Cognitive Psychology/University of Finance and Management in Warsaw*, *9*(4), 202.

Swagman, A., Province, J., & Rouder, J. N. (2015). Evidence for discrete-state processing in perceptual word identification. *Psychonomic Bulletin & Review*, *22*, 265–273.

Sweldens, S., Corneille, O., & Yzerbyt, V. (2014). The role of awareness in attitude formation through evaluative conditioning. *Personality and Social Psychology Review*,

18(2), 187–209.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Von Bastian, C. C., Souza, A. S., & Gade, M. (2015). No evidence for bilingual cognitive advantages: A test of four hypotheses. *Journal of Experimental Psychology: General*, *145*(2), 246–258.

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*, 830–841.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*, 158–189.

West, R., Jakubek, K., Wymbs, N., Perry, M., & Moore, K. (2005). Neural correlates of conflict processing. *Experimental Brain Research*, *167*(1), 38–48.

Wetzels, R., & Wagenmakers, E. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*, 1057–1064.

Wetzels, R., Grasman, R. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage-Dickey density ratio. *Computational Statistics and Data Analysis*, *54*, 2094–2102.

Wickham, H., & Chang, W. (2016). *Devtools: Tools to make developing r packages easier*. Retrieved from https://CRAN.R-project.org/package=devtools

Wilhelm, S., & G, M. B. (2015). *tmvtnorm: Truncated multivariate normal and student t distribution*. Retrieved from http://CRAN.R-project.org/package=tmvtnorm

Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, *11*(10), 1–17. Retrieved from http://www.jstatsoft.org/v11/i10/

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of*

Statistical Software, *16*(9), 1–16. Retrieved from http://www.jstatsoft.org/v16/i09/.

Zellner, A. (1986). On assessing prior distirbutions and Bayesian regression analysis with *g*-prior distribution. In P. K. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honour of Bruno de Finetti* (pp. 233–243). Amsterdam: North Holland.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics: Proceedings of the First International Meeting held in Valencia (Spain)* (pp. 585–603). University of Valencia.

Zhang, J., & Kornblum, S. (1997). Distributional analyses and De Jong, Liang and Lauber's (1994) dual-process model of the Simon Effect. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 1543–1551.

Table 1
Characteristics of the Data Sets

	Data Set						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Participants	121	38	38	121	38	38	121
Trials Per Cell	48	168	180	≈ 100	252	180	48
Mean Effect (ms)	65	91	12	79	17	30	2
Observed effect size	1.37	1.81	0.6	2.22	0.72	1.02	0.07
Model-based effect size	9.68	2.88	1.84	3.5	1.65	1.79	0.57
F-value	1.3	2.6	1.25	2.65	1.82	2.29	0.98
p-value	0.016	0	0.14	0	0.002	0	0.538
Preferred model (BF)	$\mathcal{M}_+$	$\mathcal{M}_+$	$\mathcal{M}_1$	$\mathcal{M}_+$	$\mathcal{M}_u$	$\mathcal{M}_u$	$\mathcal{M}_0$

Note. The model-based effect size is from the unstructured model (see text). The F- and p-value are appropriate frequentist tests for individual differences (see Appendix A). Data set 4 has unequal trial numbers per condition: 50 incongruent and 150 congruent trials.

Table 2
Bayes factor model comparison

	Data Set						
	1	**2**	**3**	**4**	**5**	**6**	**7**
$\mathcal{M}_0$	1 to 10^{62}	1 to 10^{75}	1 to 379	≈ 0	1 to 10^{7}	1 to 10^{21}	$*$
$\mathcal{M}_1$	1 to 11	1 to 10^{7}	$*$	1 to 10^{19}	1 to 14	1 to 2784	1 to 8
$\mathcal{M}_+$	$*$	$*$	1 to 2.97	$*$	1 to 2	1 to 1.3	≈ 0
$\mathcal{M}_u$	1 to 12	1 to 9	1 to 1.57	1 to 11	$*$	$*$	1 to 27

Note. Asterisks mark the preferred model for each data set. The remaining values are the Bayes factors between a model and the preferred model for each data set.

Table 3
Sensitivity of Bayes factors to scale settings

	r_ν	r_θ	SD(θ_i)	B_{1u}	B_{+u}
1	0.167 (50ms)	0.1 (30ms)	7 ms	1.04 to 1	11.63 to 1
2	0.08 (25ms)	0.05 (15ms)	4 ms	0.83 to 1	11.82 to 1
3	0.33 (100ms)	0.2 (60ms)	10 ms	3.41 to 1	9.34 to 1
4	0.33 (100ms)	0.33 (100ms)	13 ms	19.22 to 1	10.04 to 1
5	1 (300ms)	1 (300ms)	22 ms	10^7 to 1	0.39 to 1
6	1 (300ms)	0.1 (30ms)	7 ms	0.88 to 1	3.35 to 1
7	0.167 (50ms)	1 (300ms)	22 ms	10^7 to 1	1.67 to 1

Note. Sensitivity analysis of Bayes factor computation for Data Set 1. Shown are the Bayes factors and the standard deviations for estimates of θ_i from the unstructured model. For ease of comparison, the values in parentheses show translations into variability when an overall standard deviation of 300ms is assumed. The first row shows the settings used for the analysis.

Table 4
Percentage of excluded observations.

	Data Sets						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Criterion I.	0.42	2.43	1.02	0.07	0.48	0.75	0.4
Criterion II.	2.78	3.98	0.85	3.04	2.38	3.16	3.14
Criterion III.	—	8.8	8.33	—	8.33	8.33	—
Total	3.19	12.95	8.15	3.09	9.45	10.34	3.45

Note. The three criteria for exclusion are: I. All incorrect trials. II. All trials with RTs less than .2 s and greater than 2 s. III. The first five trials in each experimental block.

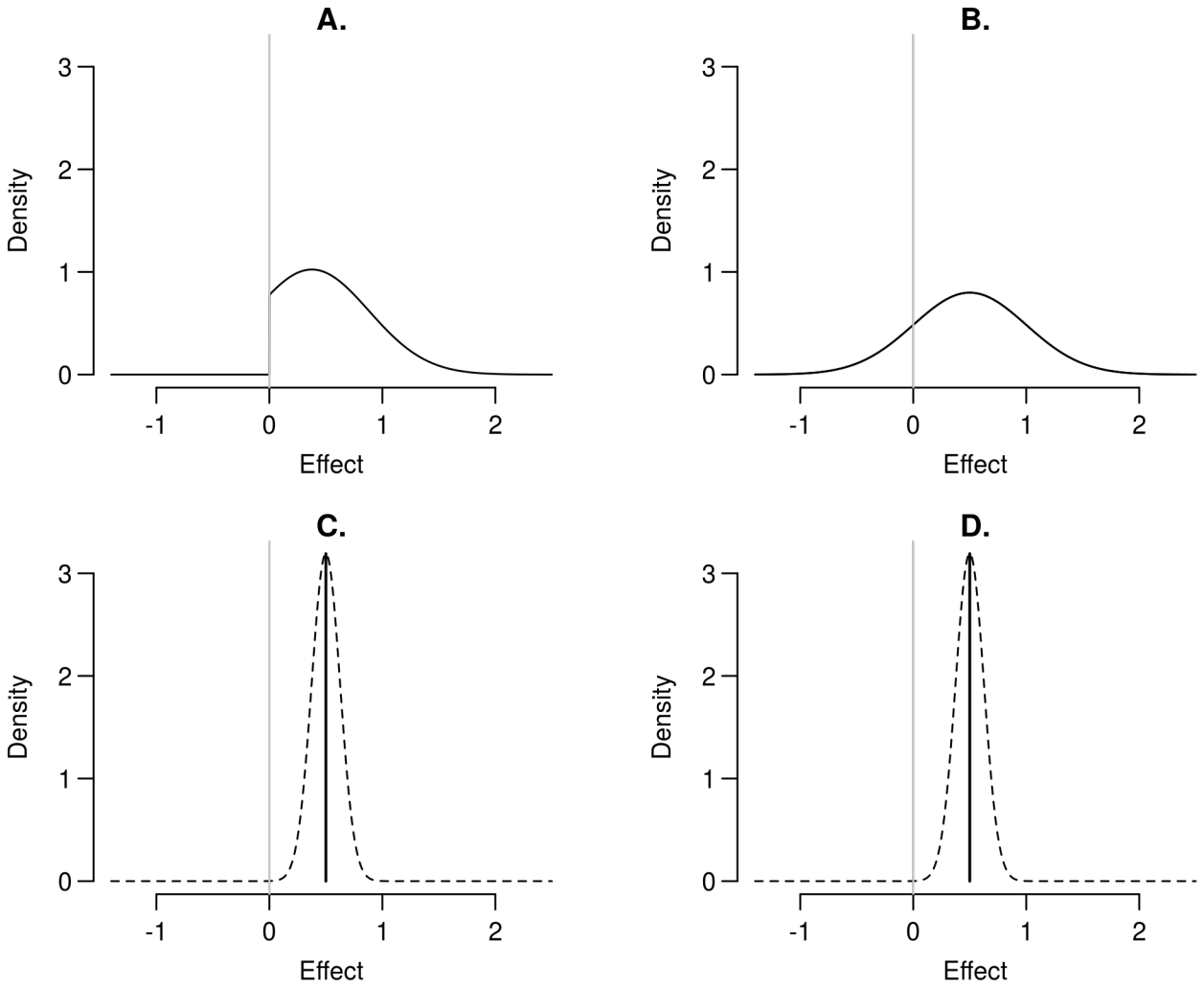

Figure 1. Hypothetical distributions of individuals' true effects (A - B) and the resulting sampling distributions of the mean (dashed lines in C - D). Panel A shows the "Everone Stroops" case where individuals' effects are constrained to be positive. Panel B shows the case where individuals' effects are unconstrained and there is no special role for the direction of the effect. The resulting sampling distribution for the average effect is the same for both cases as shown in panels C and D.

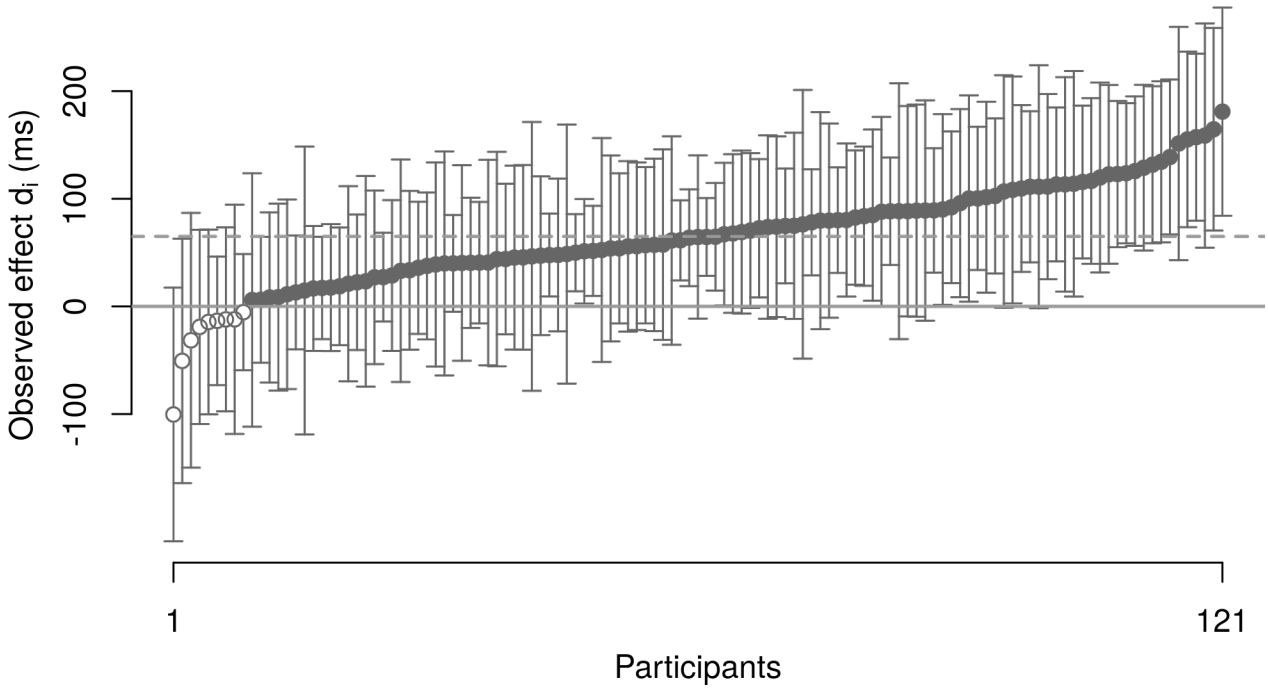

Figure 2. Observed difference scores, d_i, for 121 individuals. The scores are ordered from smallest to largest, and the open circles denote negatively valued differences. The dashed horizontal line is the mean difference, 65 ms, and the error bars are 95% confidence intervals for each individual.

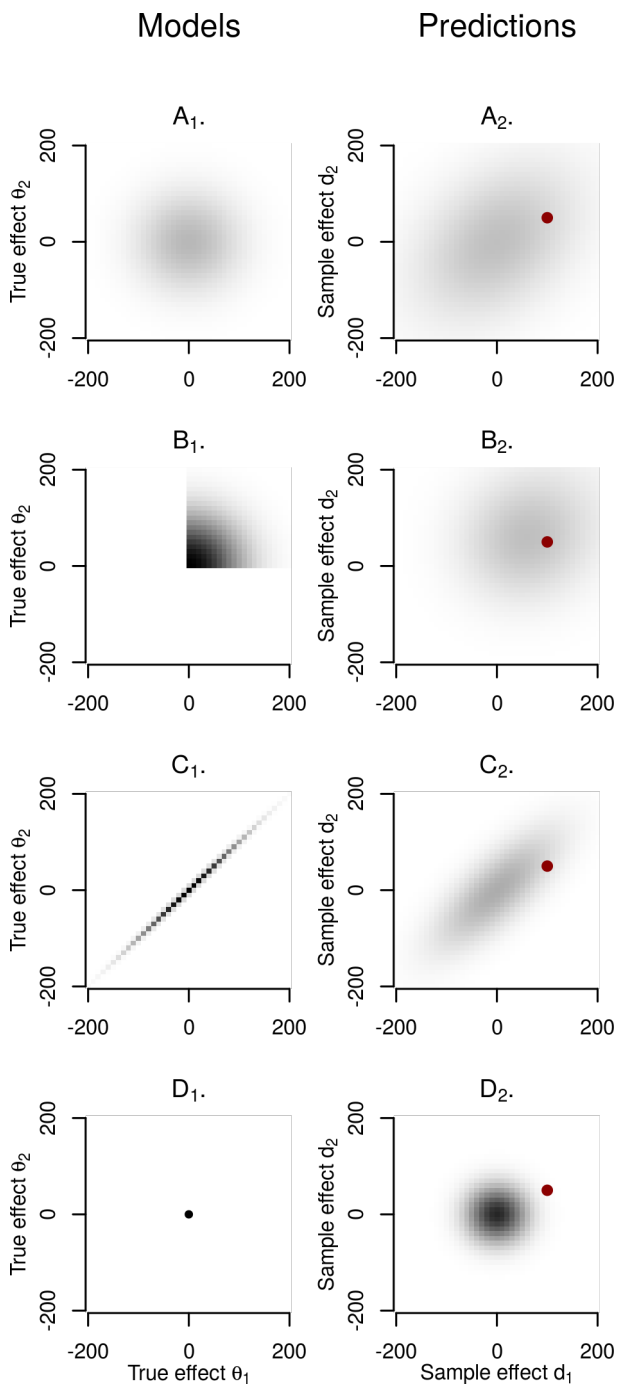

Figure 3. Model specification and resulting predictions for two participants. Panel A_1 shows the specification for the unstructured model $\mathcal{M}_u$ Here, ν and η are fixed and set to 0 ms and 230 ms, respectively. Without structure, the bivariate density is that of a normal. Panel B_1 shows the specification for the positive-effects model, and the bivariate density is restricted to the upper-right quandrant where effects are positive for both participants. Panel C_1 shows the specification for the common-effect model; when the two participants are constrained to have the same effect, the resulting density is a line on the bivariate space. Panel D_1 shows the specification for the null model, the point denotes that both participants have zero effect. The right column, Panels A_2 - D_2, show the bivariate predictions for observed effects for the respective models. Correlations in panels A_2 and B_2 result from prior variation in the overall mean for the hierarchical specification.

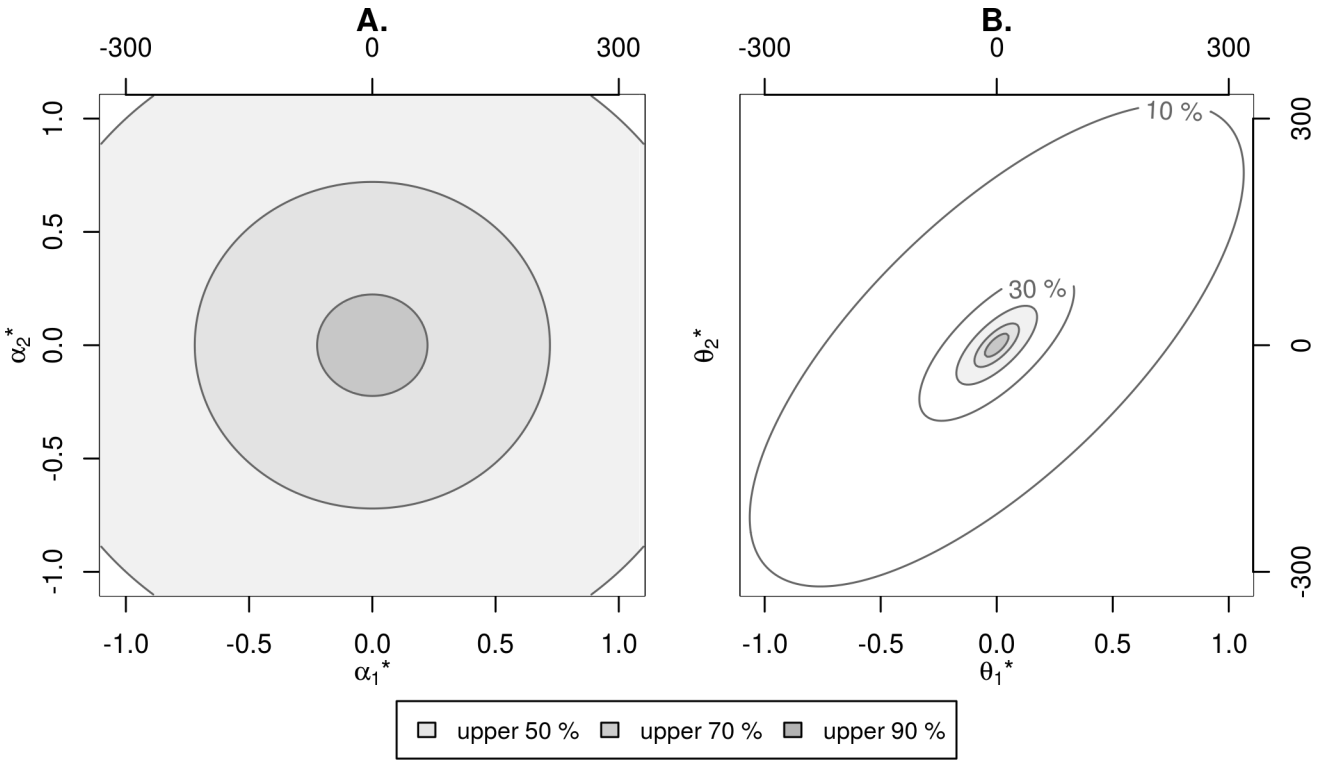

Figure 4. Marginal prior distributions for two different individuals. Values are shown in time units (ms) and in effect-size units (relative to $\sigma = 300$ ms). A. Distribution of α parameters (intercepts), B. Distribution of θ parameters (effects). Correlation in B comes from the hierarchical structure on effects.

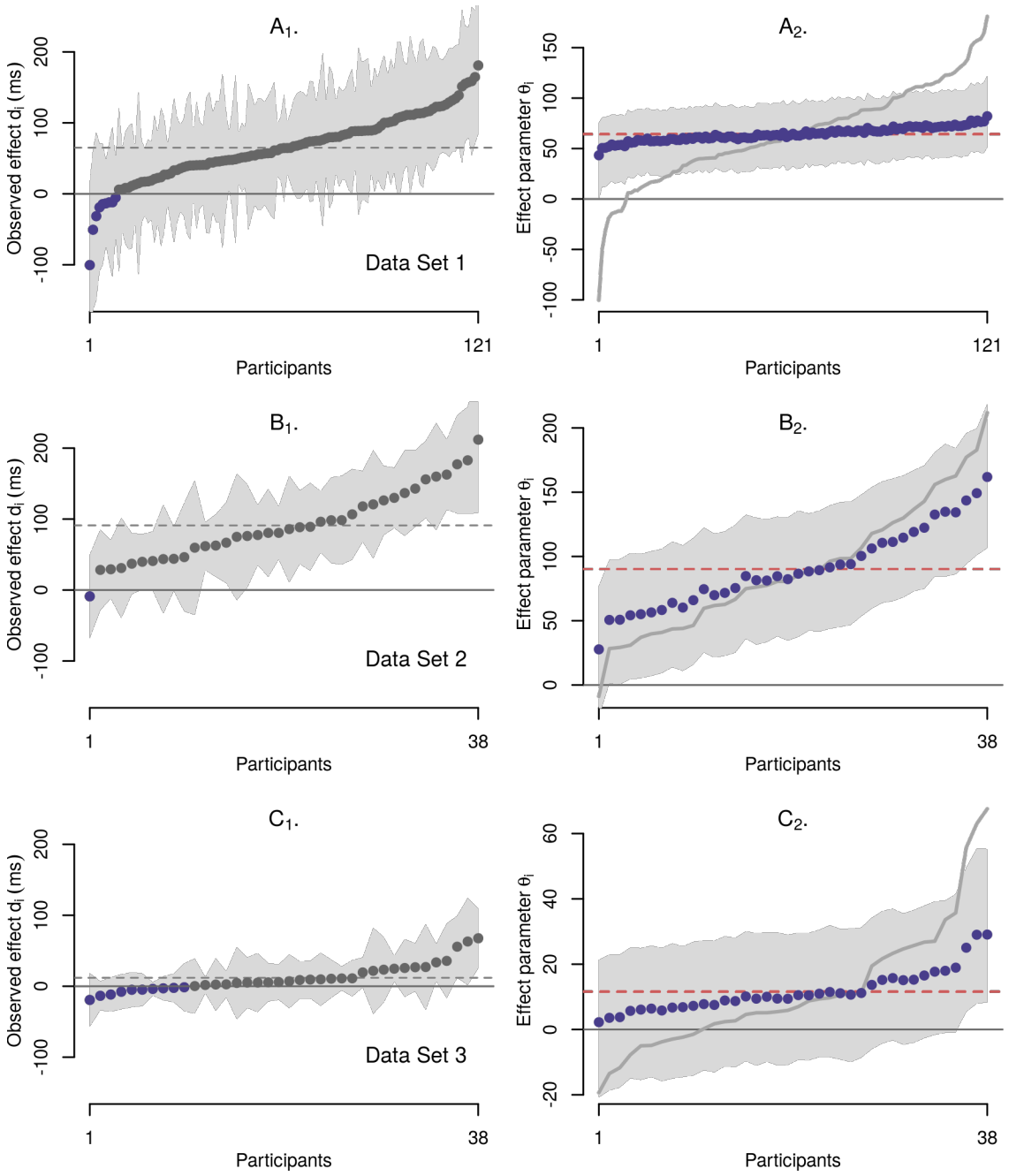

Figure 5. Empirical and Bayesian analyses for the Stroop paradigm (Data Sets 1-3.) Each row shows one data set. The left column shows the ordered observed effects, d_i, as points. The shaded area denotes the associated 95% confidence intervals, and the dashed line is the mean effect. The right column shows the ordered Bayesian estimates of individual effects, θ_i. The points are the estimates from the unstructured model. The red line is the estimate from the constant-effect model. The observered effects, d_i are included as gray line for comparison. In Panel A_2, the Bayesian analysis for Data Set 1, there is a sizable difference between the observed and estimated effects, and this difference is indicative of hierarchical shrinkage. Almost all the variability in the individual effects is seemingly from sample noise.

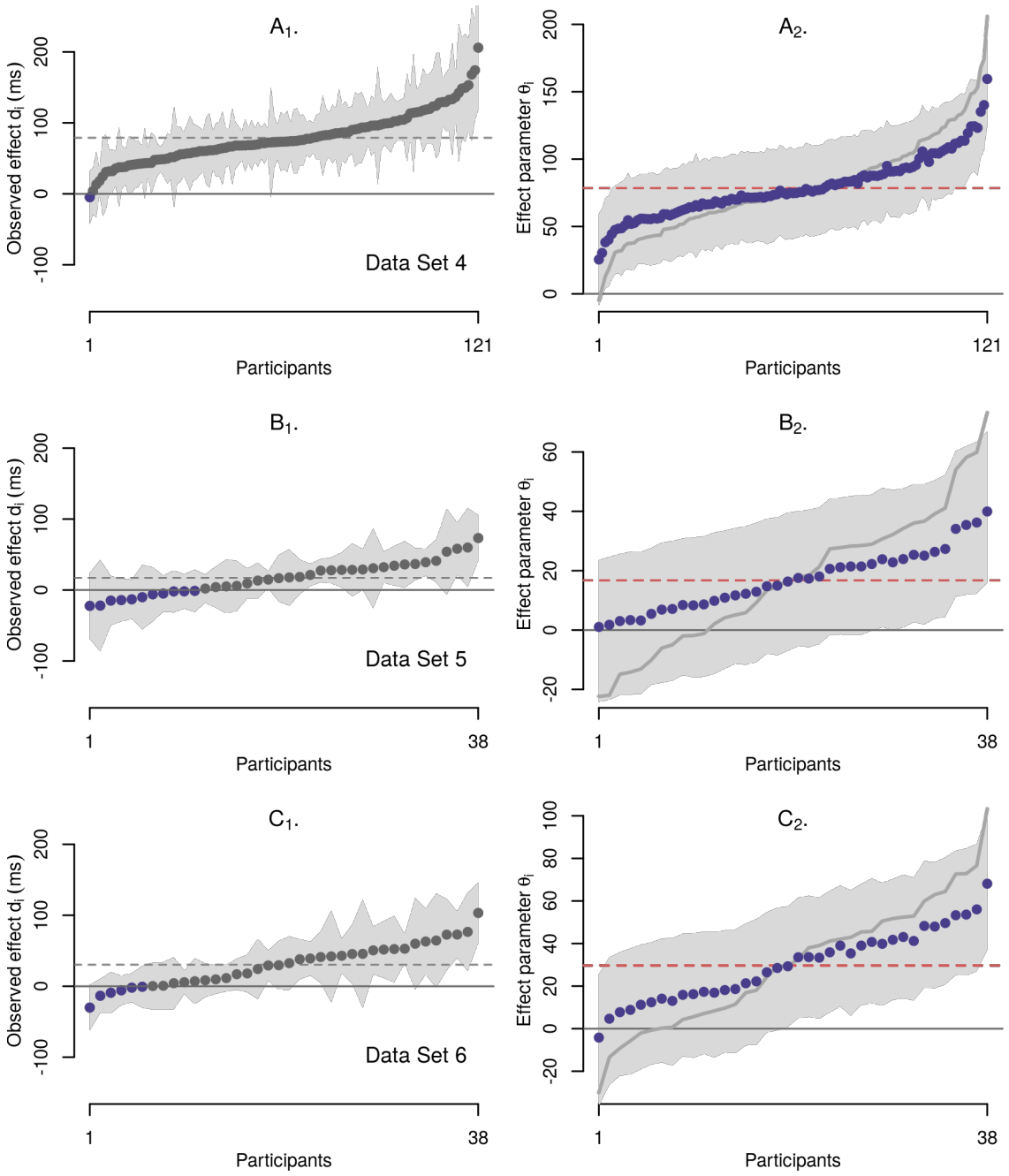

Figure 6. Empirical and Bayesian analyses for the Simon paradigm (Data Sets 4-6). The figure has the same format as Figure 5. The left column shows the ordered observed effects, d_i. The right column shows the ordered Bayesian estimates of individual effects, θ_i.

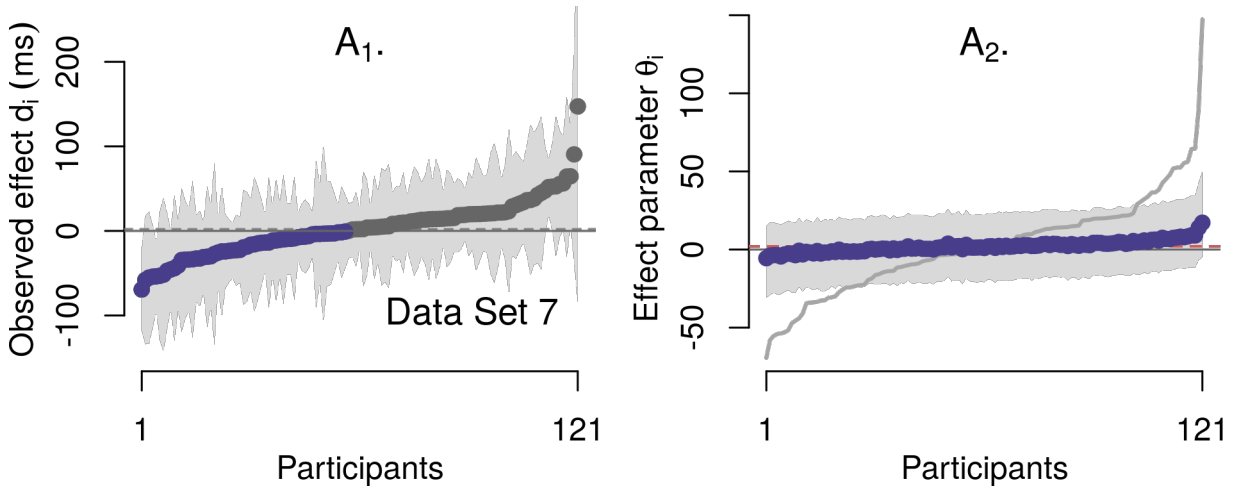

Figure 7. Empirical and Bayesian analyses for the Eriksen flanker paradigm (Data Set 7). The figure has the same format as Figure 5. The left column shows the ordered observed effects, d_i. The right column shows the ordered Bayesian estimates of individual effects, θ_i. The mean effect in panel A_1 is very close to zero. The parameter estimates for the unstructured model in panel A_2 are close to zero as well, indicating that almost all the variability in the individual effects is seemingly from sample noise.

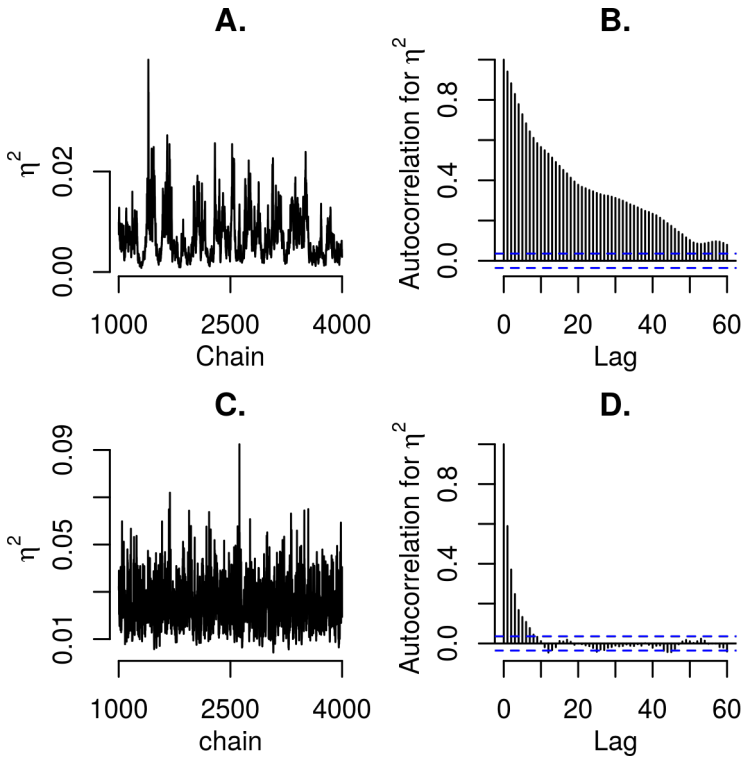

Figure 8. Examination of mixing in MCMC chains for Data Sets 1 and 2. Panels A (Data Set 1) and C (Data Set 2) show snippets of chains for η^2, the slowest converging parameter, for the unstructured model. Panels B and D show the autocorrelation function for the same two cases. Mixing is quite good for Data Set 2 and acceptable for Data Set 1.

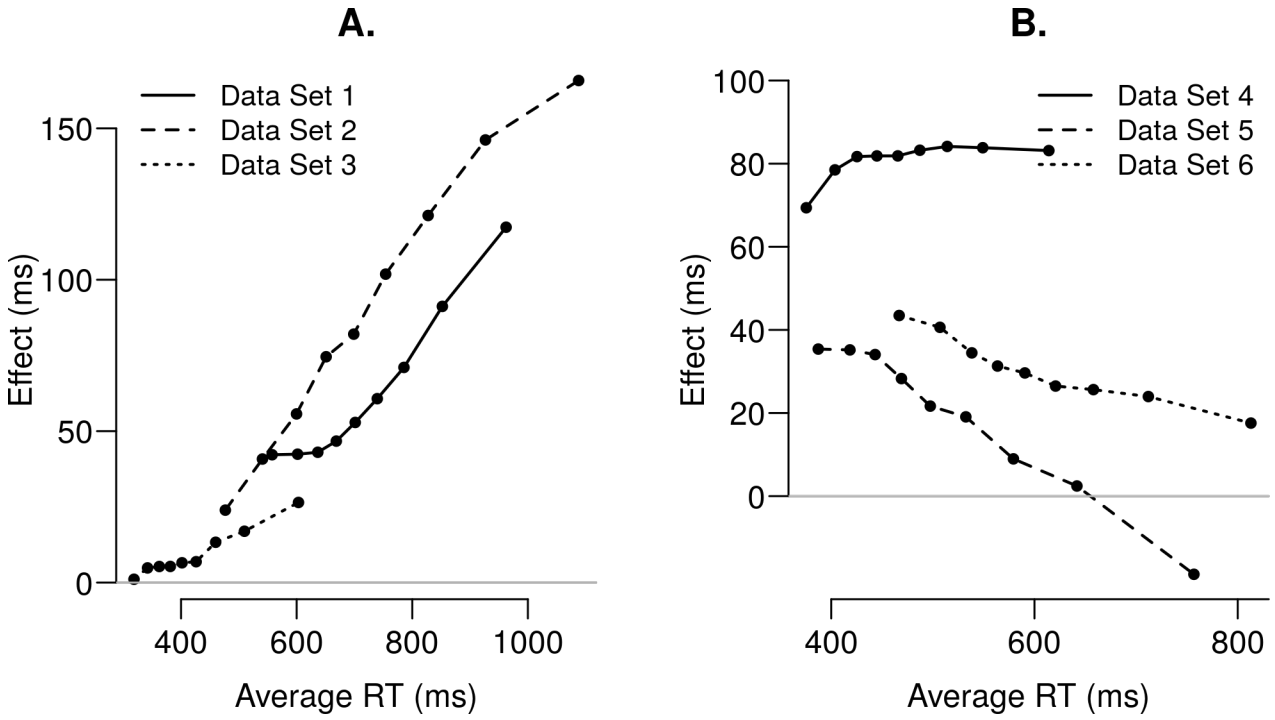

Figure 9. Delta plots for Data Sets 1-6. A. The three Stroop task data (Sets 1-3). All Stroop interference data sets show a positive slope as is often found in these kinds of tasks. B. The Stroop task data (Sets 4-6). Two of the three data sets show a negative slope.