THEORETICAL REVIEW

Some do and some don't? Accounting for variability of individual difference structures

Julia M. Haaf[1] · Jeffrey N. Rouder[1,2]

Abstract

A prevailing notion in experimental psychology is that individuals' performance in a task varies gradually in a continuous fashion. In a Stroop task, for example, the true average effect may be 50 ms with a standard deviation of say 30 ms. In this case, some individuals will have greater effects than 50 ms, some will have smaller, and some are forecasted to have negative effects in sign—they respond faster to incongruent items than to congruent ones! But are there people who have a true negative effect in Stroop or any other task? We highlight three *qualitatively different* effects: negative effects, null effects, and positive effects. The main goal of this paper is to develop models that allow researchers to explore whether all three are present in a task: Do all individuals show a positive effect? Are there individuals with truly no effect? Are there any individuals with negative effects? We develop a family of Bayesian hierarchical models that capture a variety of these constraints. We apply this approach to Stroop interference experiments and a near-liminal priming experiment where the prime may be below and above threshold for different people. We show that most tasks people are quite alike—for example everyone has positive Stroop effects and nobody fails to Stroop or Stroops negatively. We also show a case that under very specific circumstances, we could entice some people to not Stroop at all.

A prevailing folk wisdom is that different people do things differently; and in psychological science, the study of individual differences has a long and storied tradition (Cattell, 1946). A modern target of inquiry has been to understand the covariation of individuals' performance in common cognitive tasks that tap perceptual, attention, and mnemonic abilities (e.g., Miyake et al., 2000). In the usual course of studying these relationships, individuals are almost always considered as coming from a continuous, graded distribution, most often the normal (e.g., Bollen, 1989).

Yet, the tasks that researchers study often have a natural zero point. Take for example, a common number-priming task (Naccache & Dehaene, 2001; Pratte & Rouder, 2009). Here, participants are asked to determine whether a target

digit is greater than or less than 5. Before participants see the target digit, they are exposed to a briefly flashed prime. This prime is another digit, also greater than or less than 5. The prime may be *congruent* with the target, i.e., both numbers are less than (greater than) 5, or it may be *incongruent* with the target, where each has a different status from the number 5. People respond more slowly to targets when the prime is incongruent than congruent, and this slow down defines the priming effect. Note that there is a natural zero point where there is no priming, that is, responses to targets are the same for congruent and incongruent primes. A positive priming effect occurs when responses to targets following congruent primes are faster than those following incongruent primes; a negative priming effect occurs when responses to targets following congruent primes are slower than those following incongruent primes. Here, negative and positive priming effects are qualitatively different because they lead to different theoretical implications. A positive priming effect may indicate the presence of response activation from the prime (Eriksen & Eriksen, 1974). A negative priming effect may indicate temporal segregation and suppression (Dixon & Di Lollo, 1994).

✉ Julia M. Haaf
    jhaaf@mail.missouri.edu

1   University of Missouri, Columbia, MO, USA

2   University of California, Irvine, CA, USA

Cognitive psychologists are well aware of the importance of the qualitative and theoretical distinction between positive and negative effects. When experimentalists can find manipulations that switch an effect from positive to negative, the switch itself becomes a target of study. For example, Eimer and Schlaghecken (2002) show a reversal of the priming effect. In their Experiment 2, they manipulate the prime presentation duration. For longer presentation durations, say between 60 and 100 ms, the regular, positive priming effect is observed. For shorter presentation durations, say between 16 and 50 ms, the authors document a negative priming effect. The combination of results is interpreted as evidence for two different priming processes: a subliminal process leading to inhibition, and a supraliminal process leading to facilitation of the primed response. A second example for the reversal of an effect comes from Rouder and King (2003). The authors reversed the usual Eriksen flanker effect by using morphed targets. When targets were morphed between two letters, there was a contrast effect where responses to targets incongruent with their surrounds were speeded. When clear letters were used, there was the typical assimilation effect where responses to incongruent targets were slowed down. Rouder and King (2003) interpreted the combination of results as implicating two separate and opposing effects in perception and response selection.

Although these examples show that it may be possible to reverse specific effects in tightly controlled contexts, doing so remains quite rare. To our knowledge, nobody has reversed the Stroop effect where responses are faster to incongruent items than to congruent ones. Likewise, it is unlikely that a strength effects can be reversed—that is, we doubt, for example, that there are conditions where responses may be quicker to dim lights than to bright ones, and because these phenomena seemingly cannot be reversed, simple explanations are warranted. In the strength case, it may indicate a rather direct link between stimulus strength and neural activation without much opportunity for modulation from top-down processes. We suspect for a wide class of phenomena, the zero point is never crossed (Haaf & Rouder, 2017).

These examples, where zero is crossed and where it is not, show the theoretical importance of the zero point in understanding cognition. Yet, considerations of individual differences, where individuals' abilities are assumed to come from a graded distribution, do not respect this importance. If an analyst assumes say that individuals' true Stroop abilities come from a normal distribution, then some people by definition are assumed to have a truly negative Stroop effect. What a fantastic finding that would be! The problem with graded distributions that cross the zero point is that they deny the possibility that true-negative effects do not occur. In doing so, they miss the possibility of important

global constraints like the impossibility of negative Stroop effects.[1]

Another problem with graded distributions of individuals' effects is that they explicitly assume that no individual has a true effect of exactly zero.[2] Yet, the idea that some individuals do not display an effect is compelling from a theoretical point of view. Are there effects where some people are immune, and if so, how could we tell?

The main goal of this paper is to develop models that encompass several configurations of individual variability. Do individuals follow a graded distribution? Are there order-constraints such that nobody can have a negative effect? And are there individuals with truly no effect? The work builds on our previous development where individuals could be all positive, all negative, all null, or follow the traditional graded normal distribution (Haaf & Rouder, 2017; Thiele et al., 2017). None of our previous work, however, captures the notion that "some do and some don't," that is that some individuals show a positive effect while others show no effect.

Previously, researchers have attempted to solve this problem with classification: They assume there are do-ers and don't-ers, and classify people as such. We show next why this approach does not work well. Then we develop our models and apply them to a number of extant context and priming effects. The results are quite surprising—we show that there are cases where "everyone does" and other cases where "some do and some don't." We have yet to find a case where some people have a negative effect while others have a positive effect. Perhaps this unicorn is out there, and the tools developed here define the state-of-the-art for finding it.

To understand the nature of the problem and the solution, it is critical to distinguish between observed and true effects. Take, for example, data from a priming task shown in Fig. 1.[3] Plotted are each individual's *sample effect*—that is, the difference between the mean response for incongruent primes and congruent primes. The sample effects are ordered from smallest to largest in the figure, and as can be seen, some are negative, others are near zero, and a few are substantially positive. However, this does not mean that any particular person has a *true effect* that is negative, near zero, or positive. True effects are the underlying latent effects we could observe if we had infinitely many trials per individual. They are the target of interest; we wish to know whether they are negative, zero, or positive. The sample effects reflect the true effects, but they are also perturbed

---

[1]This negative Stroop effect is not to be confused with the *reversed Stroop effect* (e.g., Logan and Zbrodoff, 1979), where participants are asked to respond to the word instead of the ink color of the word.

[2]By definition, the probability of any point in a continuous distribution is identically zero.

[3]Data comes from Pratte and Rouder (2009), Experiment 2. Details on the data set may be found in the Application section.
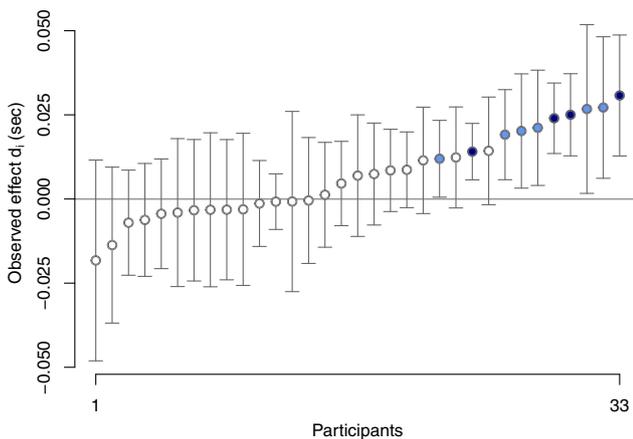
Fig. 1 Individual observed effects from a priming task ordered from lowest to highest. *Shading* of the points indicates the direction of the effect according to two criteria. *Dark blue points* indicate a positive priming effect for the criterion of BF > 2. *Light or dark blue points* indicate a positive priming effect for the criterion of 80%CIs excluding zero. *White points* indicate a null effect according to both criteria

by trial noise. Hence, some analysis is needed to infer the qualitative status of the true effects.

The argument against classification

One approach that has been developed for the problem of capturing individuals' positive, negative and null effects is classification. An example comes from Fific et al. (2008), who classified individuals as using serial, parallel, or coactive processing based on, among other analyses, the direction and magnitude of an interaction contrast. In their Experiment 1, they conclude that five participants engage in serial processing, two participants engage in parallel processing, and none of the participants engages in coactive processing. We can apply a simple classification approach to the data in Fig. 1 considering each individual's confidence interval (CI) around their sample effect. The idea is that the true value is likely somewhere in the CI (cf. Morey et al., 2016a, b). The bars show the 80% confidence intervals, and the key observation is whether these CIs include zero or not. We see no individuals are highly likely to have truly negative effects, some may have null effects, and a few others are squarely positive. Of course, we need not use CIs; we can even calculate a Bayes factor per individual, and once again, we conclude that some individuals show a positive effect and some show no effect.

Yet, the approach in Fig. 1 has a flaw—it is prone to overstating the amount of structural diversity across people. Sample noise will make it seem that different people are in different regions of negative, null, and positive. For example, even if everyone was positive, by sample noise we would expect to observe some people as having a negative

observed value, others near zero, and still others as having a positive observed value. This critique comes from Lee and Webb (2005) who note that it is important to *model* the variability across individuals as well as within an individual. The question, then, is what type of model should be used?

For this purpose, Lee and Webb (2005) proposed a non-parametric Bayesian modeling approach similar in spirit to cluster analysis. If there is a natural group structure, individuals are clustered together. The number of clusters, their locations, and their variances are estimated. Lee and Webb's approach is noteworthy in many respects, but it does not take into account the qualitative distinctions among negative, null, and positive effects. For example, individuals with true effects of −20 ms and 20 ms may be lumped together while those with true effects of 20 ms and 200 ms may be lumped apart. This is problematic because if individuals truly differ in the sign of effects, a far more complicated and nuanced set of processes is implicated.

To address this issue, Haaf and Rouder (2017) and Thiele et al. (2017) developed an approach based on model comparison. Instead of using a single model, say a clustering-like model used by Lee and Webb (2005), Haaf and Rouder (2017) and Thiele et al. (2017) propose that researchers compare several different models where each model instantiates different possible configurations of individual differences. We extend this approach here to assess the folk wisdom that some do and some don't. Figure 2 shows different possible configurations of models. Panel A shows a strong null model. All people have a true null effect, and this null effect is indicated by the spike at zero. Panel B shows a common-effect model; all people have the same true positive effect. Panel C shows a case where individuals' effects vary, but, importantly, they do not cross zero. Moreover, there is no mass at zero, and this model is a flexible version of the "everyone does" position. Panel D shows a model where some do and some don't show an effect. It is a mixture model, and with a certain probability individuals have no effect. Models of this type are called spike-and-slab models (George & McCulloch, 1993; Mitchell & Beauchamp, 1988), with the spike referencing the point mass at zero and the slab referencing the positive distribution. People who are truly in the spike have no true effect and those who are in the slab have a true effect in the positive direction.

Panel E shows a three-component mixture: some people show a true positive effect, some show a true negative effect, and some show a true null effect. This model is again a spike-and-slab model with one spike at zero and two slabs, one on each side of zero. Note that there are only a few cases where this model may be theoretically implied. One such case is Fific et al. (2008), where all three components of the model may be mapped to different processing architectures. Yet, we will not carry this model for the
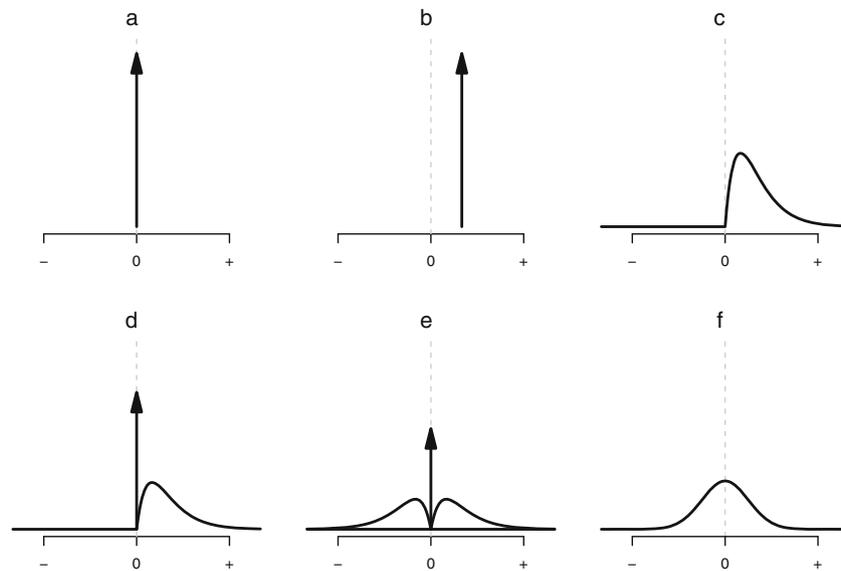
Fig. 2 Models capturing different configurations of individual differences in true effects. **a**, **b** A null model and a common-effect model without individual variation. **c** A model with individual variability where all effects are positive. **d** A model where some individuals have no effect and others have a positive effect. **e** A model where some individuals have no effect and others have a positive effect, and again others have a negative effect. **f** Common random-effects model that captures the case where individuals come from a graded, continuous distribution

current analysis. One reason is that it is not theoretically useful for the applications chosen here. A second reason is that it is computationally inconvenient (We will return to the computational difficulties in the discussion section). A third, and perhaps most important reason, is that if this case holds, the normal model in Panel F will fare quite well. We show this in simulation subsequently.

Panel F shows the usual case where individuals' true effects follow a normal distribution. Even though the model has a convenient mathematical form, it does not take the theoretical importance of the zero point into account. It may be used to highlight the differences between conventional approaches and our development.

The goal here is to assess the evidence from the data for the various models in Fig. 2. If models in the top row are favored, then we may favor accounts where each individual is behaving with the same strategies and processes. Alternatively, if models in the bottom row are favored, then we may favor accounts with qualitative differences in processing and strategies among individuals. The approach is different from categorization because the goal is not to categorize individuals but to compare models of configural relations that embed important theoretical distinctions.

From a classical perspective, the analyses of and comparison among the models in Fig. 2 is difficult. The analyst must account for the possible range restrictions on true effects, and doing so is known as *order-restricted inference*. Order-restricted inference is a difficult topic in statistical analysis (Robertson et al., 1988; Silvapulle & Sen, 2011), and we know of no tests appropriate for the comparison of the null model of Fig. 2a to the "everyone does" model in Fig. 2c. In contrast, Bayesian model comparison through Bayes factors is conceptually straightforward and computationally convenient. Gelfand et al. (1992) provide the conceptual insights; Haaf and Rouder (2017) and Klugkist et al. (2005) provide the computational implementations for all the models except the spike-and-slab versions. The development of the spike-and-slab model that encodes the "some do and some don't" position is novel. We are proud of the development because the model addresses an important element of folk wisdom, and serves as a precursor for more advances mixture models. Moreover, we are proud to have made the analysis computationally convenient.

In the next section, we provide a brief formal overview of the models depicted in Fig. 2. Following this, we informally outline the Bayes factor model comparison strategy. With the Bayes factors developed, we analyze priming and Stroop interference data. We document at least one case where the some-do-and-some-don't wisdom seems to be a good description.

Models of constraints

The tasks we consider here have two conditions that can be termed *compatible* and *incompatible*, or more generally, *control* and *treatment*. It is most convenient to discuss the models in random-variable notation. We start with a basic linear regression model. Let Y_{ijk} denote the response time (RT) for the ith participant, $i = 1, \ldots, I$, in the jth condition, $j = 1, 2$, and the kth trial, $k = 1, \ldots, K_{ij}$.[4] The linear regression model is

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

Here, α_i is each individual's true intercept and θ_i is each individual's true effect. The term x_j is an indicator for the condition, which is zero for compatible trials and one for incompatible trials. The parameter σ^2 is the variance of repeated trials within a cell. The critical parameters in the model are the true individuals' effects, θ_i. Placing constraints on these *true* effect parameters results in the models depicted in Fig. 2.

Null model The null model is denoted as $\mathcal{M}_0$ and specifies a true effect of zero for all individuals:

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

This null model is more constraining than the usual null where the average across individuals is zero. Here, in contrast, each individual truly has no effect. An illustration of the model is shown in the first panel in Fig. 3. The figure illustrates the dimensionality of the models for two participants, and it is a guide useful for the following models. Shown are two hypothetical participants' true effects, θ_1 and θ_2, shown in the figure. For the null model, θ_1 and θ_2 have to be exactly zero. As a result, the density of the distribution of θ_i is a spike at zero, corresponding to the dark point at zero in the figure. The model also corresponds to Fig. 2a.

Common-effect model The common-effect model, denoted $\mathcal{M}_1$, corresponds to the spike in Fig. 2b, and it is less constrained than the null. Individuals share a common effect with no individual variability,

$$\mathcal{M}_1: \quad \theta_i = \nu^+,$$

where ν^+ denotes a constant, positive effect. The first panel in the second row of Fig. 3 shows that both θ_1 and θ_2 are restricted to the diagonal line, depicting that individual participants' effects have to be equal. The diagonal is restricted to be positive to ensure that the model only accounts for effects in the expected direction. Every

individual has the exact same true effect, but this effect is only restricted to be positive, not fixed to a specific value. The diagonal line results from this a priori uncertainty about the size of the effect.

Positive-effects model The positive-effects model is denoted $\mathcal{M}_+$, and it is the first model that introduces true individual variability. True individuals' effects may vary, they are, however, constrained to be positive:

$$\mathcal{M}_+: \quad \theta_i \sim \text{Normal}^+(\nu, g_\theta \sigma^2),$$

where Normal^+ refers to a normal distribution truncated below at zero, ν is the mean parameter for this distribution, and $g_\theta \sigma^2$ is the variance term. The model is illustrated in the first panel of the third row of Fig. 3.[5] Both θ_1 and θ_2 are restricted to be positive, but can be different. Values closer to zero are more plausible. The model roughly corresponds to Fig. 2c. In both cases, the distribution on θ_i is restricted to positive values. Yet, the shape in the figure is different from the one for the positive-effects model specified here.

Spike-and-slab model The spike-and-slab model is denoted $\mathcal{M}_{SS}$. Here, the distribution on θ_i consists of two components, the spike and the slab. Whether an individual's effect is truly in the slab or in the spike is indicated by the parameter z_i. If an effect is truly null, $z_i = 0$; if an effect is truly positive $z_i = 1$. The distribution of θ_i conditional on z_i is

$$\mathcal{M}_{SS}: \quad \begin{array}{l} \theta_i | (z_i = 1) \sim \text{Normal}^+(\nu, g_\theta \sigma^2), \\ \theta_i | (z_i = 0) = 0, \end{array}$$

Here, the spike corresponds to the null model and the slab corresponds to the positive-effects model. In model specification, every individual has some probability of being in the spike and a complementary probability of being in the slab. The first panel in the fourth row of Fig. 3 shows the model specifications for two participants. For these hypothetical individuals, four combinations of true effects are plausible: 1. Both individuals are in the spike. In this case, θ_1 and θ_2 have to be zero, indicated in the figure by the dark point at (0,0). 2. Both participants are in the slab. θ_1 and θ_2 can take on any positive value, restricting the true effects to the upper right quadrant in the figure, just as with the positive-effects model. 3. θ_1 is in the slab and θ_2 is zero. This case is represented by positive θ_1 values on the horizontal line at $y = 0$. 4. θ_2 is in the slab and θ_1 is zero. This case is represented by positive θ_2 values on the vertical line at $x = 0$.

---

[4]Due to data cleaning or design choices, the number of trials per person and condition may vary.

[5]For illustration, mean, and variance of the slab are set to fixed values at $\nu = 0$ and $g_\theta \sigma^2 = .07^2$ (in seconds).
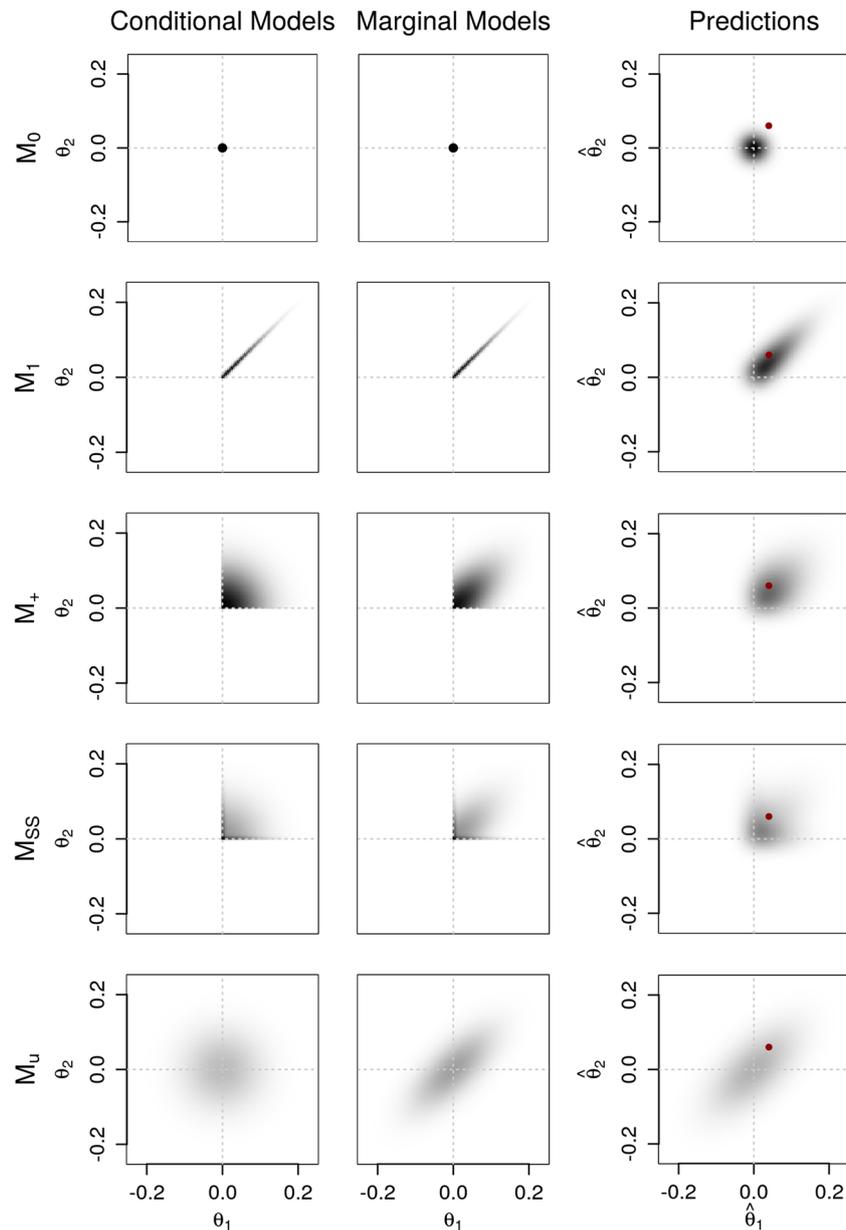
Fig. 3 Model specification and predictions for two exemplary participants. *Left column*: Model specifications conditional on specific prior settings. *Middle column*: Marginal model specifications integrated over prior distributions show correlation between individuals' effects. *Right column*: Resulting predictions from each model for data. The *red dots* show a hypothetical data point that is best predicted by the common-effect model (*second row*)

Unconstrained model The unconstrained model, denoted $\mathcal{M}_u$, is the random-effects model in Fig. 2f. Here, a normal distribution without any constraint is placed on the individual's true effects:

$$\mathcal{M}_u : \quad \theta_i \sim \text{Normal}(\nu, g_\theta \sigma^2).$$

The first panel in the last row of Fig. 3 shows these model specifications. True individuals' effects can take on any values, and values closer to zero are more plausible. With

this model, there is no explicit way of taking differences in the sign of the effect into account. The model serves as a none-of-the-above option capturing when some individuals' effects are truly negative.

Prior specifications and hierarchical constraints

The five models are analyzed in a Bayesian framework. Bayesian analysis requires a careful specification of prior

distributions on parameters. These priors are needed for parameters α_i, the individual intercepts; σ^2, the variance of responses in each participant-by-condition cell; the collection of z_i, each individual's indicator of being in the spike or the slab; ν, the mean of effects; and g_θ, the variance of effects in effect-size units. The priors parameters that are common to all models are not of particular concern. They do not affect model comparison, and we follow Haaf and Rouder (2017) in specification.[6] Several of the models ascribe individual differences across true effects. In this regard, individuals should be treated as random, and a hierarchical treatment is appropriate (Lee, 2011; Rouder & Lu, 2005; Rouder et al., 2008). We model individual differences as coming from either a normal or truncated normal with free mean and variance parameters. Prior settings on these parameters, ν and $g_\theta \sigma^2$, may affect inference. In the following, we describe the reasons for this influence. We show the effects of reasonable ranges of prior settings on these two parameters in the Discussion section.

The shared mean parameter, ν, induces correlation between the individuals' effects. Take, for example, the unconstrained model. We can recast the model on θ_i as $\theta_i = \nu + \epsilon_i$, where ν remains the population mean and $\epsilon_i \sim \text{Normal}(0, g_\theta \sigma^2)$ is the independent residual variation specific to an individual. The parameter ν is not given. Just as the shared effect in the common-effect model, ν must be estimated. It has variability in this regard and this shared variability induces a correlation between individuals' effects. We take this variation into account by computing a marginal model on θ_i. The marginal models are shown in the second column of Fig. 3. For the unconstrained model and the other models that specify variability, the correlation is apparent in the figure. This correlation induces dependency between θ_is, and the resultant of this dependency is a reduction in the dimensionality of the models. This reduction makes the unconstrained model, for example, more similar to the common-effect model, which is important for model comparison.

Estimation model

The above five models describe possible constraints on individuals' effects. Assessing how applicable these models are to data is the core means to determining whether all individuals' true effects are positive, some individuals' true

effects are null, or some individuals' true effects are even negative. In the next section, we discuss a formal inferential approach—Bayes factors—for model comparison. Even though model comparison is the main target, estimating parameters and visualizing them remains a tool for understanding structure in data. When constructing an estimation model here, we have two goals: One is to have relatively few constraints on the parameters; the second is to respect the possibility of true null effects. To meet these goals, we place a generalized spike-and-slab model on θ_i.

The model has a spike at zero and a normal distribution as slab. It may be viewed as a mix between panel a and panel e in Fig. 2. The distribution of each individual's effect, θ_i, is

$$\theta_i | (z_i = 1) \sim \text{Normal}(\nu, g_\theta \sigma^2),$$
$$\theta_i | (z_i = 0) = 0.$$

This spike-and-slab model, just as the unconstrained model in Fig. 2, is agnostic toward the direction of individuals' effects. It is, however, appropriate for estimating posterior spike and slab probabilities and the collection of θ.

Evidence for constraints

In the previous sections, we develop five models: the null model, the common-effect model, the positive-effects model, the spike-and-slab model, and the unconstrained model that embed various meaningful constraints. Here, we provide a discussion on how to state evidence for these five models in the Bayesian framework. Rather than providing a formal discourse, which may be found in Jeffreys (1961), Kass and Raftery (1995), and Morey et al. (2016a, b), we provide an informal discussion that we have previously presented in Rouder et al. (2016) and Rouder et al. (2018a, b). Informally, evidence for models reflects how well they predict data.

The predictions for data from each of the five models here are shown in the right column of Fig. 3. These predictions are for observed effects, $\hat{\theta}$, for each of the two exemplary participants. Note that predictions are defined on data while model specifications are defined on true effects, and this difference is reflected in the plotted quantities in the figure. For the null model, for example, *true* effects, left column, have to be exactly zero, and the *observed* effects, right column, are predicted to be near (0,0). The predictions are affected by sample noise, inasmuch as sample noise smears the form of the model.[7] The remaining rows of

---

[6]An exception are prior settings on z_i, the indicators of whether an individual is truly in the spike or the slab. We set $z_i \sim \text{Bernoulli}(\rho)$, where ρ is the probability of being in the slab. We placed a hierarchical prior on $\rho \sim \text{Beta}(a, b)$, where $a = b = 1$. These prior settings represent an equal prior probability of being in the spike or the slab, and changing them may influence model comparison greatly. For this application, we decided not to explore other settings, because we do not have any theoretical implications of higher slab or spike prior probability.

[7]More technically, the predictions are the integral $\int_\theta f(\mathbf{Y}|\theta)\pi(\theta)d\theta$ where $f(\mathbf{Y}|\theta)$ is the probability density of observations conditional on parameter values and $\pi(\theta)$ is the probability density of the parameters of

Fig. 3 show the predictions for the common-effect, positive-effects, spike-and-slab, and unconstrained models. In all cases, the predictions are smeared versions of the models.

Once the predictions are known, model comparison is simple. All we need to do is note where the data fall. The red dots in the right column of Fig. 3 denote hypothetical observed participants' effects. These observed effects, 40 ms for participant 1 and 60 ms for participant 2, are both positive and about equal, and we might suspect that the common-effect model does well. To measure how well, we note the density of the prediction at the observed data point. The densities for the models have numeric values, and we may take the ratio to describe the relative evidence from the data for one model vs. another. For example, the best fitting model in the figure, the common-effect model, has a density that is three times the value of that of the unconstrained model. Hence, the data are predicted three times as accurately under the common effect model than under the unconstrained model. This ratio, 3 to 1, is the *Bayes factor*, and it serves as the principled measure of evidence for one model compared to another in the Bayesian framework.

Bayes factors are conceptually straightforward—one simply computes the predictive densities at the observed data. Nonetheless, this computation is often inconvenient in practice. It entails the integration of a multidimensional integral, which is often impossible in closed form and may be slow and inaccurate with numeric methods. For the five models here, we follow the development by Haaf and Rouder (2017) using two methods to compute Bayes factors: An analytic approach pioneered by Zellner and Siow (1980) and expanded for ANOVA by Rouder et al. (2012), and the *encompassing approach* introduced by Klugkist et al. (2005). Figure 4 shows which method is used for which of the models.
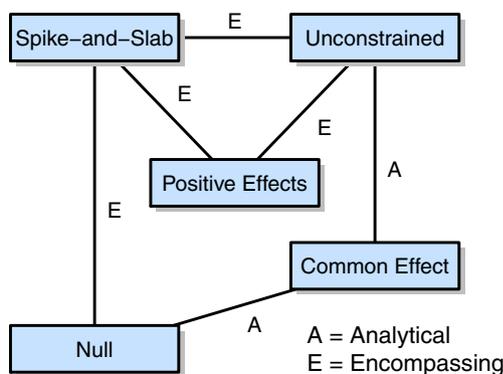

Fig. 4 Bayes factor computations for the five models. Bayes factors between the unconstrained, common-effect, and null model can be computed using analytical solutions. Bayes factors between the spike-and-slab model, the unconstrained model, the positive-effects model, and the null model can be computed using the encompassing approach. All other Bayes factors can be computed utilizing the transitivity property of Bayes factors

The encompassing approach is used for Bayes factor computations for the positive-effects model and the spike-and-slab model. The computations for the positive-effects model are detailed in Haaf and Rouder (2017). New to this paper are model comparisons with the spike-and-slab model. Although the spike-and-slab model is precedented and popular, we are unaware of any prior development for comparing it as a whole to alternatives. Our approach is a straightforward application of the encompassing approach. The encompassing approach is a simple counting method within Markov chain Monte Carlo (MCMC) estimation. Take, for example, the Bayes factor between the null model and the spike-and-slab model. The target parameters for the Bayes factor computation are the collection of z_i, the individuals' indicators of being in the slab. Here we use z to denote the vector of z_i. Using these parameters, the Bayes factor between the null model and the spike-and-slab model can be expressed as

$$B_{0SS} = \frac{P(z = 0|Y, \mathcal{M}_{SS})}{P(z = 0|\mathcal{M}_{SS})},$$

where the event $z = 0$ indicates that every individual is in the spike. This Bayes factor is the posterior probability that all individuals are in the spike relative to the prior probability that all individuals are in the spike. The same approach can be used for comparing the spike-and-slab model to the positive-effects model, using the posterior and prior probabilities that every individual is in the slab.

Using the encompassing approach in MCMC sampling, one can count the number of iterations where $z = 0$ when z_i are sampled from the posterior; likewise, one can count the number of iterations where all $z = 0$ when the z_i are sampled from the prior. Let $z[m]$ denote a vector of i samples of z (one for each individual) on the mth iteration under the spike-and-slab model. The mth iteration is considered evidential of the null model if all I elements of $z[m]$ are zero, that is, on this iteration, every individual's effect θ_i is sampled from the spike. Let n_{01} be the number of evidential iterations from the posterior, and let n_{00} be the number of evidential iterations from the prior. Then, the Bayes factor is

$$B_{0SS} = \frac{n_{01}}{n_{00}}.$$

To compute the Bayes factor of the spike-and-slab model to the remaining models, we use the well-known transitivity of Bayes factors (Rouder & Morey, 2012). Figure 4 provides an illustration for this transitivity: Say the common-effects model predicts the data three times better than the unconstrained model, and the common-effect model predicts the data 20 times better than the null model. We can use these two Bayes factors to calculate the Bayes factor for the unconstrained model over the null model as follows: $B_{u0} = \frac{B_{10}}{B_{1u}} = \frac{20}{3} = 6.67$.

Application

We apply the five models to three different data sets: A priming data set provided by Pratte and Rouder (2009), and two Stroop experiments provided by Pratte et al. (2010). The goal here is to answer the question of whether some participants show an effect while others do not. We provide estimation and model comparison results for the three data sets and discuss them in the light of the experimental paradigms.[8]

Priming data set

The priming data used here, reported by Pratte and Rouder (2009), come from a number-priming task where the primes are flashed briefly before the target stimulus is presented.[9] From our experience with these tasks, we suspect the spike-and-slab model to perform well with some participants being affected by briefly presented primes and others not being affected.

In the task, numbers were presented as primes, followed by target digits that had to be classified as greater or less than five. There is a critical congruent and incongruent condition: The congruent condition is when the prime and the target are both on the same side of five, e.g., the prime is three and the target is four; the incongruent

condition is when the prime and the target are opposite, e.g., the prime is eight and the target is four. The priming effect refers to the speed-up in responding to the target in the congruent versus the incongruent condition. Prime presentation was brief by design, and the goal was to bring it near the threshold of detection. Yet, it is well known that this threshold varies considerably across people. For example, Morey et al. (2008) report high variability in individual threshold estimates for prime perception. Other researchers use adaptive methods to change presentation duration individually for each participant until identification of primes is on chance (e.g., Dagenbach, Carr, and Wilhelmsen, 1989). For any given presentation duration, some individuals may be able to detect the prime and others may not. This difference may lead to variability in processing with some people processing the primes and others not.

Results Figure 5a provides two sets of parameter-estimation results. The first set, denoted by the crosses that span from −0.02 s to 0.03 s, are the observed effects for the individuals, and these are the same points that are plotted in Fig. 1. Observed effects in this context are the differences in individuals' sample means for the incongruent and congruent conditions. Crosses are colored red or gray to indicate whether the observed effects are negative or positive, respectively. Overall, effects are relatively constrained with no participant having more than a 31-ms effect in absolute value. Estimates from the hierarchical estimation model are shown in blue circles. These estimates are posterior means of θ_i where the averaging is across the spike and slab components. The posterior weights of being in the slab are denoted by the shading of the points with lighter shading corresponding to greater weights. The 95% credible intervals, again across the spike and slab components, are shown by the shaded region.

We focus on the contrast between the sample effects and the model-effect estimates. Although the sample effects subtend a small range of about 50 ms, the model-based estimates subtend a much smaller range from almost no effect to an 11-ms effect. These hierarchical estimates reflect the range of true variation after sample noise is accounted for. The compression is known as regularization or shrinkage, and prevents the analyst from overstating evidence for heterogeneity. Hierarchical regularization is an integral part of modern inference (Efron & Morris, 1977; Lehmann & Casella, 1998), and should always be used wherever possible (Davis-Stober et al., submitted). The individuals' posterior probability of being in the slab ranges from 0.30 to 0.64.

From the model estimates, it is evident that individual effects are tightly clustered and slightly positive with a mean of 4 ms. Yet, these results are not sufficient to answer the

---

[8]All analyses were conducted using R (Version 3.4.2; R Core Team, 2016) and the R-packages *abind* (Version 1.4.5; Plate and Heiberger 2016), *BayesFactor* (Version 0.9.12.4.2; Morey and Rouder, 2015), *coda* (Version 0.19.1; Plummer et al., 2006), *colorspace* (Stauffer et al., 2009, Version 1.3.2; Zeileis et al., 2009), *curl* (Version 3.2; Ooms 2017), *devtools* (Version 1.13.6; Wickham and Chang 2016), *diagram* (Version 1.6.4; Soetaert 2014a), *dotCall64* (Gerber et al., 2015, Version 0.9.5.2; 2016), *fields* (Version 9.6; Nychka et al., 2015), *ggplot2* (Version 3.0.0; Wickham 2009), *gmm* (Version 1.6.2; Chaussé, 2010), *gridBase* (Version 0.4.7; Murrell 2014), *maps* (Version 3.3.0; Becker, Ray Brownrigg. Enhancements by Thomas P Minka, and Deckmyn., 2016), *MASS* (Version 7.3.47; Venables and Ripley 2002), *Matrix* (Version 1.2.14; Bates and Maechler, 2017), *MCMCpack* (Version 1.4.3; Martin et al., 2011), *msm* (Version 1.6.6; Jackson 2011), *mvtnorm* (Version 1.0.8; Genz and Bretz, 2009; Wilhelm and Manjunath, 2015), *papaja* (Version 0.1.0.9709; Aust and Barth, 2017), *plotrix* (Version 3.7.2; Lemon, 2006), *RColorBrewer* (Version 1.1.2; Neuwirth 2014), *reshape2* (Version 1.4.3; Wickham, 2007), *sandwich* (Version 2.4.0; Zeileis, 2004, 2006), *shape* (Soetaert, 2014b; Version 1.4.4; Wickham, 2007), *spam* (Version 2.2.0; Furrer and Sain, 2010; Gerber and Furrer 2015), *spatialfil* (Version 0.15; Dinapoli and Gatta, 2015), and *tmvtnorm* (Version 1.4.10; Wilhelm and Manjunath, 2015).

[9]We analyze the data from Pratte and Rouder's Experiment 2. In the original experiment, primes were shown for durations of 16, 18, or 20 ms. We combined data from the 16- and 18-ms conditions and disregarded the difference in duration for this analysis. There were no apparent differences in individuals' effects across the included conditions.
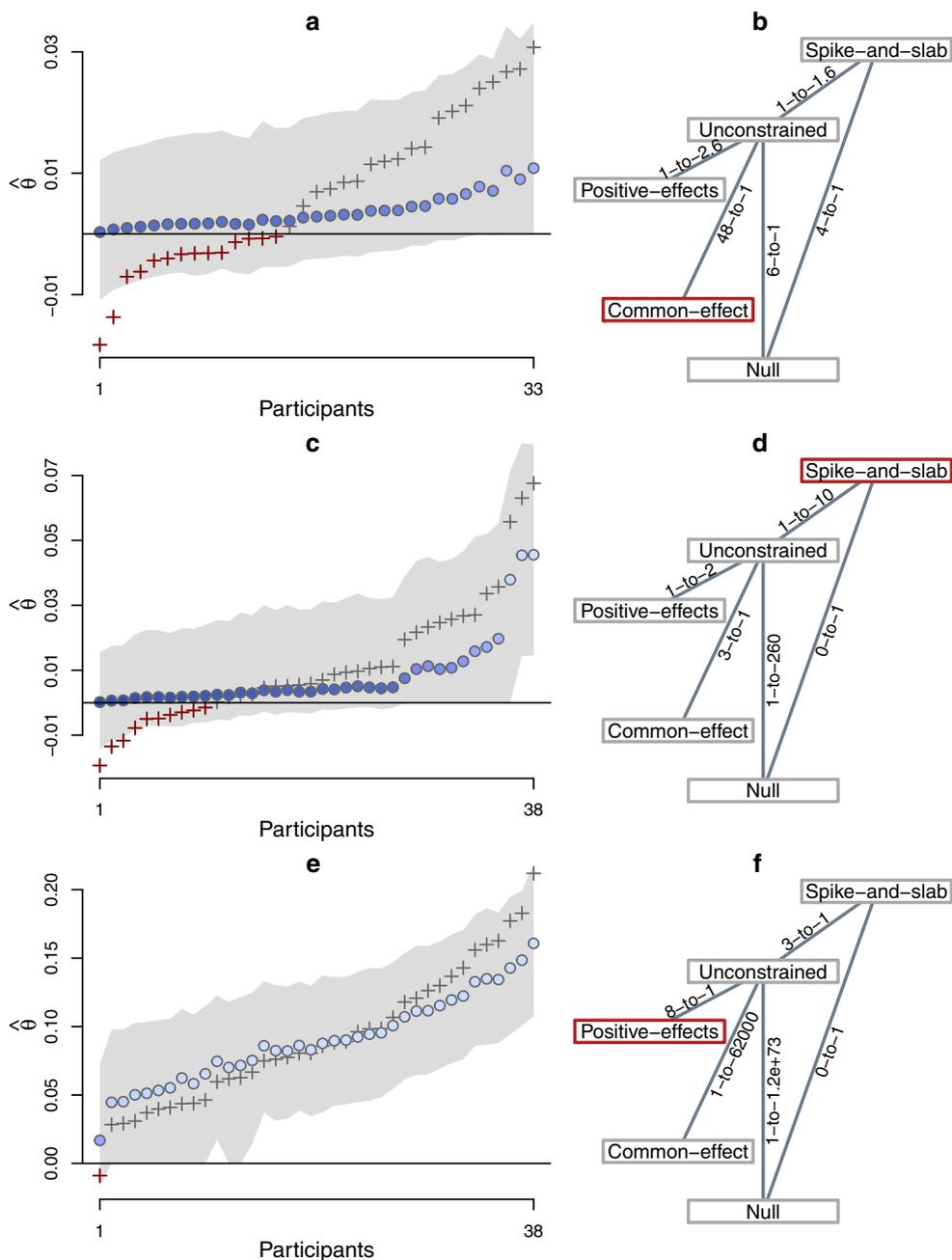
Fig. 5 Model estimates (*left column*) and Bayesian model comparison results for (**a**, **b**) the priming data set; (**c/d**) the location Stroop task; **e/f** the color Stroop task. *Left column*: *Crosses* show observed effects with *red crosses* indicating negative effects. *Points* show model estimates with lighter shading indicating larger posterior weights of being in the slab. *Right column*: Bayes factors for all five models. The *red frames* indicate the winning model

question whether "some do and some don't". It is unclear whether everyone has a small effect or some people have no effect while others have a slightly larger one. To answer this question we analyze the above models and compare them with Bayes factors. The results are shown in Fig. 5b. The common-effect model is preferred, indicating that everyone

has a single, common effect. The next most parsimonious model is the null model, where all individuals have no effect, and it predicts the data about 8.60 times worse than the common-effect model. The Bayes factor between the spike-and-slab model and the common-effect model is 30-to-1 in favor of the common-effect model. We take this

Bayes factor as evidence that all do: everybody has a small priming effect.

A location Stroop experiment

Pratte et al. (2010) ran a series of Stroop and Simon interference experiments to assess distributional correlates of these inference effects. As part of their investigations, they constructed stimuli that could be used in either task, and with this goal, they presented the words "LEFT" and "RIGHT" to either the left or right side of the screen. In the Stroop tasks, participants identified the location; in the Simon task, they identified the meaning.

In their first attempt to use these stimuli, Pratte et al. (2010) found a 12-ms average Stroop effect. This effect is rather small compared to known Stroop effects, and was too small for a distributional analysis. To Pratte et al., the experiment was a failure. At the time, Pratte et al. speculated that participants did not need to read the word to assess the location. They could respond without even moving their eyes from fixation, and even though reading might be automatic at fixation, it may not be in the periphery. To encourage participants to read the word, Pratte et al. subsequently added a few catch trials. On these catch trials, the word "STOP" was displayed as the stimulus to the left or right of fixation, and participants had to withhold their response. This manipulation resulted in much larger Stroop effects.

Here we analyze data from the failed experiment where there was a small Stroop effect of 12 ms (Experiment 2 from Pratte et al., 2010). Our question is whether some participants shift their attention to the word in the periphery while others do not. In this scenario, we would find that the spike-and-slab model would perform well. The alternative is that all participants exhibited a small Stroop effect similar to the priming effect above.

Results Observed effects are shown by the crosses in Fig. 5c. Of the 38 participants, ten show an observed negative Stroop effect, shown by red crosses in the figure. The average effect is 11.90 ms with individuals' effects ranging from −19 ms to 68 ms.

Estimates from the hierarchical estimation model are shown in blue circles, and 95% credible intervals are shown by the shaded region. Again, hierarchical shrinkage is large, reducing the range from 87 ms for observed effects to 45 ms for the model estimates. Of note is also that the individuals' posterior probability of being in the slab varies considerably, ranging from 0.20 to 0.99. This difference in posterior weight suggests that some individuals are better described by the spike while others are almost definitively in the slab.

The model comparison results in Fig. 5d confirm this consideration: the Bayes factor between the spike-and-slab model

and the runner-up common-effect model is 3.50-to-1 in favor of the spike-and-slab model. This Bayes factor provides slight evidence for the "some do and some don't" wisdom in this particular Stroop experiment.

A color Stroop experiment

Pratte et al. (2010) ran another experiment, a more standard Stroop task with color terms (Experiment 1 from Pratte et al., 2010). For this experiment, in contrast to the failed Stroop experiment, we expected that everyone shows a positive Stroop effect.

Results Parameter estimates are shown in Fig. 5e. Individuals' observed effects are fairly large with an average of 91 ms with only one participant showing an observed negative effect. There is less shrinkage than for the other data sets. The range for the observed effects is 221 ms; the range for the hierarchical estimates is 144 ms. Posterior probabilities of being in the slab are high with only one person having a lower probability than .85.

The model comparison results are shown in Fig. 5f. Overall, there is the most evidence for the positive-effects model. The second-best model is the unconstrained model. The Bayes factor between these two models is 8-to-1 in favor of the positive-effects model, and this Bayes factor can be interpreted as evidence that all do. The spike-and-slab model fares even worse with a Bayes factor of 1-to-23 compared to the positive-effects model. The results suggest that everyone shows the expected Stroop effect—if targets are presented at fixation.

Concerns

The Bayesian modeling approach developed here requires judicious choices in model and prior specification. An attentive reader may have some concerns about our choices. It is reasonable to inquire about alternative models that were not included for analysis here; specifications of the normal and truncated normal; sensitivity of the results to prior settings; and computational convenience of the approach. We take these concerns in turn.

Alternative models

The five models developed here are designed to capture the following theoretical positions: 1. No person may have a true effect whatsoever; 2. Everyone has the same positive true effect; 3. Individuals' effects vary, but everyone has a positive effect; 4. Some people show a true positive effect while others truly show no effect at all; and 5. Individuals'

effects follow a normal, graded distribution where some people can have a true negative effect. These five, of course, are not the only choices. Here we discuss alternatives.

One area of potential concern is the unconstrained model. Here we use a graded normal, and this is our only specification accounting for the possibility that some people have truly positive effects while others have truly negative effects. There are other model instantiations, however, that capture this state of affairs. One useful alternative is the mixture model shown in Fig. 2e. Here, there are three groups of individuals: those that are positive, those that are negative, and those that are null. Another possible model is one with two slabs but no spike. If we are willing to speculate, we can come up with a variety of models that instantiate positive and negative effects.

Rather than implementing all of these possible alternatives, we decided to simulate how well our unconstrained normal model fared when the data followed these mixture alternatives. Figure 6a shows the normal model and two alternatives: a two-slab mixture model (labeled "Mixture 1"), and a spike-and-two-slab mixture model (labeled "Mixture 2"). Our aim in choosing particulars for these truths was to equate the overall mean and variance. The true individuals' effects for each study were the ticks at the top of the panel. We simulated data from these true effects 100 times for each of the three models. The Bayes factor between the normal unconstrained model and the positive model is computed for each of the simulated data sets. The Bayes factor distributions from the simulation are shown in the violin plots of Fig. 6b. The unconstrained model is favored as frequently for the mixture truths as for the normal truth.
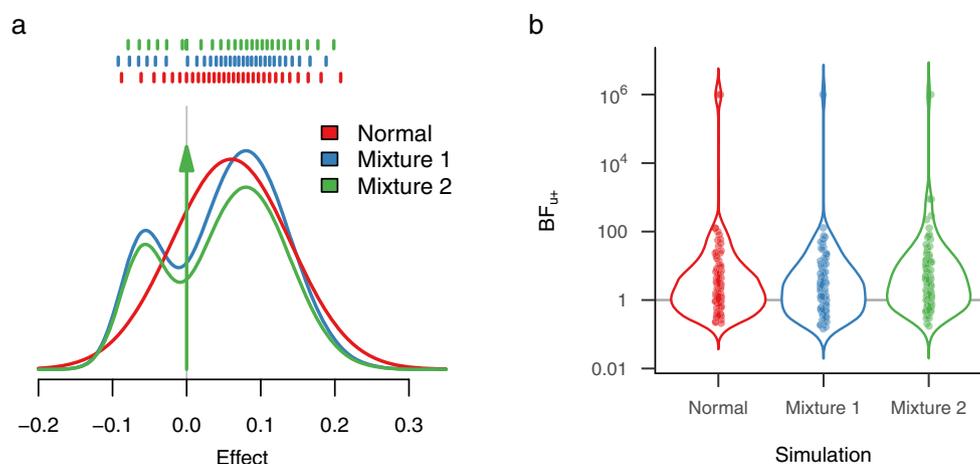
The critical point to emerge from this simulation study is that the unconstrained normal model is a useful instantiation of the unconstrained position, even when misspecified. Here is why: The goal is to detect a few negative true effects against a background of many true positive ones. The normal for this configuration would have a positive mean and sufficient variance so that there is noticeable negative mass (as in Fig. 6a). The distribution of the negative part is not only small in mass, but is skewed such that small negative effects are weighted. The normal therefore is well-suited to detect the most difficult case—the one where negative effects are few and more likely to be clustered near zero. Moreover, with the little negative mass that we expect in these cases, there is little to distinguish a mixture model from the unconstrained.

Another area of potential concern are the point-mass specifications for the null model, the common-effect model and the spike in the spike-and-slab model. Alternative specifications are not just limited to mixture models. A recent trend is to use small equivalence regions instead of point mass (Kruschke & Liddell, 2017; Rogers et al., 1993). Those researchers who are convinced these are helpful models are free to use them, and the Bayes-Factor computations are no challenge (Morey & Rouder, 2011). We do not recommend these models because they provide for less theoretical clarity than either point-mass models or distribution models. The point null, in contrast, is theoretically constrained and useful. This argument is made by Gallistel (2009), Jeffreys (1961), Rouder et al. (2016), and Rouder and Morey (2012), just to name a few.

Fig. 6 Simulation from three different true models. **a** True unconstrained models. The *red line* shows a normal unconstrained model, the *blue line* shows a mixture model of true negative and true positive effects. The *green line* shows a mixture of true negative, null, and positive effects. The *ticks* at the top of the panel show true study effects chosen for simulation. **b** Resulting Bayes factor distributions for the unconstrained model vs. the positive-effects model for the simulation study. The unconstrained model is preferred most of the time for all three true models

Normal specification

Another concern with the proposed approach is the reliance on normal parametric model specifications. The advantage of the normal specification is computational convenience. With this specification, the many dimensions of the high-dimension integrals that define the Bayes factor may be computed symbolically to high precision. Without this specification, we suspect numeric integration would be exceedingly slow and inaccurate. Yet, researchers may be concerned about the misspecification of the normal. We focus here on applications with response times, and RTs are skewed rather than symmetric. Moreover, the standard deviation tends to increase with the mean (Luce, 1986; Rouder et al., 2010; Wagenmakers & Brown, 2007).

We think the concern about the normal specification is misplaced. The main reason is that we focus on the analysis of ordinal relations among true means. If we knew individual's true means, then we could answer questions about the direction of the effects without any consideration the true shapes or true variances. The inference therefore inherently has all the robustness of ANOVA or regression, which is highly robust for skewed distributions, so long as the left tail is thin. Indeed, RTs tend to have thin left tails that fall off no slower than an exponential (Burbeck & Luce, 1982; Van Zandt, 2000; Wenger & Gibson, 2004).

Thiele et al. (2017) addressed this concern through simulation. They considered highly similar models and performed inference with similarly computed Bayes factors. In a simulation, they generated data from a shifted log normal with realistic skewness and with means and variances that varied across individuals and the manipulation. As expected, they found exceptional robustness, and the reason is clear. The main inferential logic is dependent only on true means, and the normal is a perfectly fine model for assessing this quantity even when the data are not normally distributed.

Prior sensitivity

Another concern, perhaps a more pressing concern in our view, is understanding the role and effects of the priors on inference. In general, Bayesian models require a careful choice of priors. These priors have an effect on inference as noted by many Bayesians. A general idea in research is that, if two researchers run the same experiment and obtain the same data, they should reach the same if not similar conclusions. Yet, the priors may be chosen differently by different researchers, and this choice may lead to differing conclusions. To harmonize Bayesian inference with the idea of similar conclusions, many Bayesian analysts actively seek to minimize the effects by picking likelihoods, prior parametric forms, and heuristic methods of inference so

that variation in prior settings have marginal effects (Aitkin, 1991; Gelman et al., 2004; Kruschke, 2012; Spiegelhalter et al., 2002). In contrast, Rouder et al. (2016) argue that the goal of analysis is to add value by searching for theoretically meaningful structure in data. Vanpaemel (2010) and Vanpaemel and Lee (2012) argue that the prior is where theoretically important constraints are encoded in the model. In our case, the prior provides the critical constraint on the relations among individuals. We think it is best to avoid judgments that Bayes factor model comparisons depend too little or too much on priors. They depend on it to the degree they do.

Here we focus on understanding the dependence of Bayes factors on a reasonable range of prior settings and the resulting diversity of opinions. Indeed, Haaf and Rouder (2017) took this tactic in understanding the diversity of results with all the models except for the spike-and-slab-model which was not developed at the time. Here we use a similar range of prior settings to understand the dependency on these settings.

The critical prior settings for understanding the diversity of conclusions come from the priors on ν and ϵ_i (or g_θ). Although they are not the primary target of inference, the prior settings on these parameters do affect Bayes factor results. A full discussion of the prior structures on these parameters is provided in Haaf and Rouder (2017), and here we review the main issues. The critical settings are the *scales* on ν and ϵ_i. These scale settings are relative to σ, the residual noise. Our considerations for these scale settings go as follows: In tasks like this, with sub-second RTs, a standard deviation of repeated response times for a given participant and a given condition may be about 300 ms, and we can use this value to help set the scales. For example, for priming and Stroop tasks, we may expect an overall effect of 50ms, and the scale on ν might be $50ms/300ms$, or 1/6th of the residual noise. Likewise, if the take the variability of individuals' effects depicted by ϵ_i, we may expect this variation to be about 30 ms, or $1/10$ of the residual noise.

With these reasonable ranges of variation, we are ready to explore the effects of prior specification on Bayes factors. We explore the effects of halving and doubling these settings, which represents a reasonable range of variation. The results are shown in Table 1. There is a fair amount of variability in Bayes factors, and in our opinion, there should be. The range of settings define quite different models with quite different predictions. Nonetheless, there is a fair amount of consistency. For the priming data, the common-effect model is preferred for all settings, with the null-model and the spike-and-slab models as the next contenders. For the color Stroop data, the positive-effects model is preferred for all settings, and the ordering for the remaining models stays relatively constant. The only data set where the preferred model varies with prior settings

Table 1 Sensitivity of Bayes factors to prior settings

Scale on ν	Scale on ϵ	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_+$	$\mathcal{M}_{SS}$	$\mathcal{M}_u$
Priming						
1/6	**1/10**	0.12	*	0.01	0.03	0.02
1/12	1/20	0.06	*	0.04	0.06	0.04
1/12	1/5	0.14	*	0.02	0.72	0.04
1/3	1/20	0.06	*	4.79e -8	0	1.03e -5
1/3	1/5	0.13	*	2.25 e -8	7.7e -4	0.98e -5
Location Stroop						
1/6	**1/10**	3.68e -4	0.29	0.05	*	0.1
1/12	1/20	5.4e -4	0.54	0.33	*	0.21
1/12	1/5	2.59e -4	0.17	0.04	*	0.07
1/3	1/20	270.51e -5	29.28	0	*	118.87e -3
1/3	1/5	15.02e -3	9.69	17.25e -5	*	0.04
Color Stroop						
1/6	**1/10**	1e -74	1.9e -6	*	0.04	0.12
1/12	1/20	1.12e -74	7.86e -7	*	0.02	0.05
1/12	1/5	1.45e -74	3.03e -6	*	0.17	0.16
1/3	1/20	8.04e -75	5.62e -7	*	0.00	0.01
1/3	1/5	1.06e -74	2.23e -6	*	0.02	0.06

Note. Sensitivity analysis of Bayes factor computation for all three data sets. Different settings of the scales on ν and ϵ represent a reasonable range of priors around the setting used for the main analysis (*bold*). The *asterisks* mark the winning model for each data set for the original analysis, and Bayes factors are computed for comparison to this model

is the location Stroop data: The spike-and-slab model is preferred for the chosen settings and when the scale on ν is halved. These settings indicate that small average effects are expected for all models. When the scale on ν is doubled, i.e., larger, about 100 ms effects are expected, the Bayes factor between the common-effect model and the spike-and-slab model is about 1, indicating that none of the two models is preferred over the other. This Bayes factor, however, was not large from the beginning, only about 3-to-1 in favor of the spike-and-slab model. This example illustrates how useful this type of sensitivity analysis can be to understand the range of conclusions that may be drawn from the data. In this case, the evidence for the spike-and-slab model is small, and largely dependent on prior settings. For a convincing result, more evidence for a mixture of effects would be needed.

Computational issues

In previous work we developed the null, common-effect, positive-effects and unconstrained models (Haaf & Rouder, 2017; Thiele et al., 2017). Here we add the spike-and-slab model to capture the folk-wisdom that some do and some don't show an effect. We show it is a worthy competitor

in at least one application. The former four models are computationally convenient both in estimation and in Bayes factor computations. As shown in Fig. 4, Bayes factors can be computed quickly using a combination of Rouder et al."s (2012) symbolic integration as implemented in the `BayesFactor` package for R (Morey & Rouder, 2015) and Klugkist and colleagues" encompassing approach (e.g., Klugkist and Hoijtink, 2007).

Computational convenience for the spike-and-slab model is more nuanced. Estimation of this model is quick and stable. The computation of the Bayes factors for the spike-and-slab model, however, is more difficult. The difficulty here is that Bayes factor computation relies on the estimation of each individual's spike indicator, z_i. This collection of parameters is estimated using Markov chain Monte Carlo methods. For each iteration in the chain, we note that $z_i[m]$ is either zero or one, indicating that the ith persons' effect is either in the spike or in the slab, respectively. Bayes factor estimation is difficult whenever the majority of the individuals' effects are in the spike or in the slab. For example, most individuals' effects in the color Stroop task are estimated to come from the slab distribution. In this case, the critical event is when all z_i are zero, and this event is rare. Hence, it is necessary to run long

chains to guarantee enough rare events to estimate its rate of occurrence.

The good news here is that we can assess the accuracy of the Bayes factor estimation for the spike-and-slab model by leveraging the transitivity of Bayes factors. Figure 4 illustrates this check. We can compute the Bayes factor between the unconstrained model and the null model either with the encompassing approach using the spike-and-slab model (Rouder, Haaf, and Vandekerckhove, 2018), or directly with symbolic integration methods (Rouder et al., 2012). We find comparable results from both methods with between 50000 and 100000 iterations in the MCMC chain, indicating good estimations of rates of rare events. Analyzing all five models takes about 45 min on one data set.

Discussion

In this paper, we address the question whether some people show a positive effect, others show a negative effect, and again others show no effect. The example we use here is priming, and we trichotomize the outcome into three basic relations: responses to congruent targets are faster than to incongruent ones (positive effects), responses to congruent targets are slower than to incongruent ones (negative effects), and responses to congruent targets are equally fast as to incongruent ones (no effects). Whether the natural zero point is crossed or not has many theoretical implications. Obvious applications of our approach include context effects (e.g., Stroop, Eriksen, Simon etc.) and strength effects (e.g., stimulus strength, mnemonic strength, etc.).

The approach we take here is Bayesian model comparison across five models: a null model, a common effect model, a positive-effects model, a spike-and-slab model, and an unconstrained normal model. The novel element here is the usage of the spike-and-slab model. Although spike-and-slab models are frequently used in statistics, their most common application is to categorize which covariates (people in our case) are in the spike and which are in the slab. Our usage is novel—we ask how well this spike-and-slab structure predicts the data relative to other models.

Several psychologists have previously asked the related question of whether mixtures account for data. In cognitive psychology, the most common application is whether responses on trials are mixtures of two bases. Falmagne (1968) was perhaps the earliest to formally explore this notion. He asked whether response times for a given individual are the mixture of a stimulus-driven process and a guessing process. Indeed, this type of query has been explored in a number of domains (e.g., Klauer and Kellen, 2010; Province and Rouder, 2012; Yantis, Meyer, and Smith, 1991).

Our approach differs markedly from these previous queries. Our focus is not on characterizing trial-by-trial variability but on characterizing variability across individuals. We do not make as detailed commitments to specific cognitive architectures, but provide a general approach based on ordinal relations of less-than, same-as, and greater-than. In this regard, our approach is more similar to latent class models used in structural equation modeling (Skrondal & Rabe-Hesketh, 2004). In these models, vectors of outcome measures are assumed to come from the mixture of latent classes of people, and the goal is to identify the classes and categorize people into these classes (see also Lee and Webb 2005; Navarro et al., 2006). One critique of this approach is that the models are so weakly identified that it is difficult to reliably recover class structure (Bauer & Curran, 2003). We avoid this problem by restriction. We restrict our classes into three that are well defined as the sign of the outcome measure. In summary, while our approach is similar in some regards to previous latent-class modeling, the statistical development is novel in critical ways.

We apply this approach to three exemplary data sets and find, at least for one case, some support for the claim that some individuals show an effect while others do not. We think, however, these mixtures are relatively rare in cognitive psychology where experimental paradigms tend to isolate the cognitive processes of interest. Only in cases where this isolation is not successful mixtures may occur. This was the case in our location Stroop example, where participants were able to avoid reading the target words by fixating the center of the screen.

If such a mixture occurs, then this result licenses more complicated inquiries. The mixture implies that there are classes of people that have qualitatively different behavior. Why? There are many possibilities including demand characteristics (perhaps some people did not understand the task), strategies, and pathologies. For example, Parkinson patients fail to display foreperiod effects (Jurkowski et al., 2005). We think that once a mixture is documented, the next logical step is to explore person-level covariates—are there any performance, personality, or other covariates that correlate well with whether a person has high probability of being in a particular component of the mixture? Depending on the substantive domain, the presence of mixtures licenses an exploration for rich patterns and structure in data.

In many cases, modeling approaches can be localized on a continuum of applicability: On the one end of the spectrum, models are widely applicable, but they only coarsely test theories. An example for this end would be ANOVA or t tests. On the other end of the spectrum, models are custom tailored to measure specific processes in specific

tasks. Our approach is in the sweet spot between these extremes. It is widely applicable in cognitive psychology where priming and strength tasks are prominent, and it addresses a question more complex than "is there an effect" without making a detailed commitment to specific processes. Knowing whether all do or "some do and some don't" remains timely and topical in perception, action, attention and memory.

Author Note 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-mixture. 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.

References

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(1), 111–142. Retrieved from http://www.jstor.org/stable/2345730

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

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

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*(3), 338.

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

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Burbeck, S. L., & Luce, R. D. (1982). Evidence form auditory simple reaction times for both change and level detectors. *Perception & Psychophysics*, *32*, 117–133.

Cattell, R. B. (1946). Description and measurement of personality.

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/

Dagenbach, D., Carr, T., & Wilhelmsen, A. (1989). Task-induced strategies and near-threshold priming: Conscious influences on unconscious perception. *Journal of Memory and Language*, *28*, 412–443.

Davis-Stober, C., Dana, J., & Rouder, J. (submitted). When are sample means meaningful? The role of modern estimation in psychological science. Retrieved from https://osf.io/mpw8z/.

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.

Dixon, P., & Di Lollo, V. (1994). Beyond visual persistence: An alternative account of temporal integration and segregation in visual processing. *Cognitive Psychology*, *26*, 33–63.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

Eimer, M., & Schlaghecken, F. (2002). Links between conscious awareness and response inhibition: Evidence from masked priming. *Psychonomic Bulletin and Review*, *9*, 514–520.

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*, 143–149.

Falmagne, J.-C. (1968). Note on a simple fixed-point property of binary mixtures. *British Journal of Mathematical and Statistical Psychology*, *21*, 131–132.

Fific, M., Nosofsky, R. M., & Townsend, J. T. (2008). Information-processing architectures in multidimensional classification: A validation test of the systems factorial technology. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 356–375.

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/

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453. Retrieved from http://psycnet.apa.org/doi/10.1037/a0015251

Gelfand, A. E., Smith, A. F. M., & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*(418), 523–532. Retrieved from http://www.jstor.org/stable/2290286

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

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

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/

Gerber, F., Moesinger, K., & Furrer, R. (2015). Extending R packages to support 64-bit compiled code: An illustration with spam64 and GIMMS NDVI3g data. *Computer & Geoscience*.

Gerber, F., Moesinger, K., & Furrer, R. (2016). dotCall64: An efficient interface to compiled C/C++ and Fortran code supporting long vectors. *R Journal*.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*(4), 779–798.

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 ed.). New York: Oxford University Press.

Jurkowski, A. J., Stepp, E., & Hackley, S. A. (2005). Variable foreperiod deficits in Parkinson's disease: Dissociation across reflexive and voluntary behaviors. *Brain and Cognition*, *58*(1), 49–61.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. Retrieved from http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572

Klauer, K., & Kellen, D. (2010). Toward a complete decision model of item and source recognition: A discrete-state approach. *Psychonomic Bulletin & Review*, *17*(4), 465–478.

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., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*(4), 477.

Kruschke, J. K. (2012). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*. Retrieved from http://link.springer.com/article/10.3758/s13423-016-1221-4.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605–621.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation*, (2nd ed.). New York: Springer.

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

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition*, *7*(3), 166–174.

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

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/

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*, 1023–1032.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419. Retrieved from https://doi.org/10.1037/a0024377

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., Rouder, J. N., & Speckman, P. L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology*, *52*, 21–36.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016a). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016b). 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.

Murrell, P. (2014). GridBase: integration of base and grid graphics. Retrieved from https://CRAN.R-project.org/package=gridBase.

Naccache, L., & Dehaene, S. (2001). Unconscious semantic priming extends to novel unseen stimuli. *Cognition*, *80*, 215–229.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*(2), 101–122. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022249605000969

Neuwirth, E. (2014). RColorBrewer: ColorBrewer palettes. Retrieved from https://CRAN.R-project.org/package=RColorBrewer.

Nychka, D., Furrer, R., Paige, J., & Sain, S. (2015). *Fields: Tools for spatial data*. Boulder: University Corporation for Atmospheric Research.

Ooms, J. (2017). Curl: A modern and flexible web client for R. Retrieved from https://CRAN.R-project.org/package=curl.

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. (2009). A task-difficulty artifact in subliminal priming. *Attention, Perception, Psychophysics*, *71*, 276–283.

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.

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences*, *109*, 14357–14362.

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/.

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

Rouder, J. N., & King, J. W. (2003). Flanker and negative flanker effects in letter identification. *Perception & Psychophysics*, *65*(2), 287–297.

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 https://doi.org/10.1080/00273171.2012.734737

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate the equivalence between two experimental groups. *Psychological Bulletin*, *113*, 553–565.

Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General*, *137*, 370–389.

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.

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 https://doi.org/10.1016/j.jmp.2012.08.001

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

Rouder, J. N., Haaf, J. M., & Aust, F. (2018a). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, *85*, 41–56. Retrieved from https://doi.org/10.1080/03637751.2017.1394581

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

Silvapulle, M. J., & Sen, P. K. (2011). *Constrained statistical inference: Order, inequality, and shape constraints* Vol. 912: Wiley.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton: CRC Press.

Soetaert, K. (2014a). Diagram: Functions for visualising simple graphs (networks), plotting flow diagrams. Retrieved from https://CRAN.R-project.org/package=diagram.

Soetaert, K. (2014b). Shape: Functions for plotting graphical shapes, colors. Retrieved from https://CRAN.R-project.org/package=shape.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *64*, 583–639.

Stauffer, R., Mayr, G. J., Dabernig, M., & Zeileis, A. (2009). Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations. *Bulletin of the American Meteorological Society*, *96*(2), 203–216. https://doi.org/10.1175/BAMS-D-13-00155.1.

Thiele, J. E., Haaf, J. M., & Rouder, J. N. (2017). Bayesian analysis for systems factorial technology. *Journal of Mathematical Psychology*, *81*, 40–54.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin and Review*, *7*, 424–465.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.

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

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.

Wenger, M. J., & Gibson, B. S. (2004). Using hazard functions to assess changes in processing capacity in an attentional cuing paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 708–719.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. New York: Springer. Retrieved from http://ggplot2.org

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

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

Yantis, S., Meyer, D. E., & Smith, J. E. K. (1991). Analysis of multinomial mixture distributions: New tests for stochastic models of cognitive action. *Psychological Bulletin*, *110*, 350–374.

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/

Zeileis, A., Hornik, K., & Murrell, P. (2009). Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, *53*(9), 3259–3270. https://doi.org/10.1016/j.csda.2008.11.033.

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