Some do and some don't? Accounting for possible variation in strategies.

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

[1] University of Missouri

[2] University of California, Irvine

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.

Correspondence concerning this article should be addressed to Julia M. Haaf, 205 McAlester Hall. E-mail: JHaaf@mail.missouri.edu

Abstract

A primary question in experimental psychology is whether different people use different strategies when performing a task. In this paper, we focus on priming and context tasks where different strategies lead to different outcomes or effects. Consider, for example, a Stroop task where the target word is presented in the visual periphery. A strategy-pure case is when all people read the word quickly and automatically such that word identity interferes with color naming. A strategy-mixed case is as follows: Some people may choose to make an eye movement, read the word, and exhibit a Stroop effect; others may not make an eye movement, identify the color without reading, and show no Stroop effect. We define strategies as tied to ordinal relations and constraints among outcomes, and develop a family of Bayesian hierarchical models that capture a variety of these constraints. Doing so leads to new definitions of heterogeneity where some variation across people is due to variation within a common strategy, while other variation across people is due to variation of strategies. 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.

Keywords: Cognitive psychometrics, Individual differences, Bayes factors, Mixture models

Some do and some don't? Accounting for possible variation in strategies.

Some of them want to use you. Some of them want to get used by you. (Eurythmics, 1983)

A prevailing folk wisdom is that different people do things differently. In cognitive sciences this folk wisdom manifests in the concept of *strategy* as being distinct from the concept of *process*. Process refers to a series of cognitive operations that a participant may go through in completing a task. Strategy, on the other hand, implies that the participant may have several different processing paths to complete a task and selects among these. Take, for example, the learning of novel alphabet arithmetic statements such as "A + 3 = D". Useful processes for this task may be counting and recollection from memory (Logan, 1988). We say there are multiple strategies in play if participants use these two processes in different configurations. For example, there may be four different strategies: 1. A participant may solely use counting; 2. A participant may solely use recollection; 3. A participant may use one or the other on any given trial (Rickard, 2004); 4. A participant may use both on every trial, say in a race configuration (Logan, 1992). If every participant uses the same strategy, we call the task a strategy-pure task. Alternatively, if different participants engage in different strategies, we call the task a strategy-mixed task. We note here that the choice of strategy need not refer to a conscious act. People may automatically rely on one strategy or another.

As researchers, we are sometimes critiqued for not accounting for variations in strategies. For example, if we propose a race model for alphabet arithmatic, a reviewer might ask if all people race. This is a difficult critique, especially if reviewers assume *a priori* that multiple strategies exist. The critique raises the important question of how to tell whether a task is strategy-pure or strategy-mixed. And having a principled means of ruling in or ruling out the possiblity of multiple strategies across individuals would be a generally applicable advance for theory development in cognitive psychology. Our goal here is just this—to present a means of addressing the multiple-strategy question across individuals.

Providing empirical evidence for mixtures of strategies is a complicated task. Certainly, there will be cases where it is nearly impossible to do so. For instance, consider two very simiar processing models, say the diffusion model (Ratcliff, 1978) and the linear ballistic accumulator model (Brown & Heathcote, 2008). It strikes us as impossible to detect mixtures of these models across people.

Our approach here is to map simple behavioral outcomes to processing. This mapping is easiest in the context of specific tasks rather than in general. Consider a priming task, for example. In the task, participants are flashed a prime before identifying a subsequent target stimulus. If prime and target share features, that is if they are compatible, then we typically observe a reduction in the time to identify the target compared to incompatible primes. Because this is the typical effect, we refer to it as the positive priming effect. In contrast, a negative priming effect occurs when incompatible primes speed the identification of a target.

To define the strategies that could be in play here, we identify three outcomes: The first outcome is the positive effect, and it is compatible with the prime activating the target. The second outcome is the negative effect. If it occured here, one would think of some contrast effect. The third outcome is no priming effect, and the strategy is that of successful segregation and supression of the prime. Now we are ready to define strategy-pure and strategy-mixed tasks. The priming task is strategy-pure if all participants show the same effect—that is all display true negative priming, all display a lack of priming, or all display true positive priming. A priming task is strategy-mixed if there are differences in the sign of the true effects. Perhaps some participants have a true null effect while others have a true positive effect.

The question then is how can we assess evidence for strategy-pure and strategy-mixed cases when outcomes map to processing? There are two important considerations in pursuing this question. The first is the difference between observed and true effects. The second is the difference between variation within a process and variation across processes. We take these in turn:

Understanding the distinction between observed and true effects is essential. We can certaintly graph observed effects for all individuals in any task such as the aforementioned priming task and determine whether everyones' observed effects are positive, negative or null. This procedure, however, does not solve the issue as the concept of measurement noise is disregarded. Statements about pure and mixed strategies therefore have to refer to true effects, and a model that incorporates noise is needed for inference.

The priming example also illustrates the second consideration of individual variability. The distinction between variability within a process and across processes is necessary in this setup. It is plausible that individuals' effects may vary in size, say, one individual may have a true 30 ms priming effect while another may have a true 300 ms priming effect. This variability is an indicator for heterogeneity within a process. It does not, however, imply mixed strategies. On the other hand, if one individual has a true effect of -30 ms (indicating a negative priming effect) and another individual has a true effect of 30 ms, then the task is strategy-mixed. Variation alone is not critical; The direction of effects is.

Stating evidence for strategies

The strategy question belies a difficult conceptual issue—how to account for parsimony. While it is easy to wonder about multiple strategies, it is critical to realize a multiple strategy account may add unneeded complexity. The simplest explanation of a phenomenon is that all people use the same strategy. In this case, the lack of variation in process leads to simple theories that apply to all people. The more complicated explanation is that there is variation in strategies across individuals. One must not only specify multiple processing possibilities but explane why certain individuals follow one strategy while others follow different strategies. Even though the strategy-pure explanantion is simper, it is our experience that researchers are often quick to make a verbal argument for mixed strategies. Indeed, we are reminded of a well-known quote commonly attributed to Einstein:

"Everything should be made as simple as possible, but not simpler."[1]

We argue that whether mixed strategies should be incorporated in a theoretical model is an empirical question. A precedented way of evidencing such mixed strategies is to classify individuals into different modes of processing. A good example comes from Little, Nosofsky, & Denton (2011) who classified individuals as using serial, parallel or coactive processing based on the direction and magnitude of an interaction contrast. Figure 1 illustrates this classification method. The figure shows the ordered effects of each individual in a priming task.[2] As can be seen, a few people show a negative sample effect, many people have near-zero sample effects, and a few have substantial positive sample effects. One way of classifying people is to draw a confidence interval around each sample effect. In the figure we use the 80%CIs to balance error rates. According to this classification scheme, 23 of the 33 individuals show an effect that is not distiguishable from zero (open dots), and 10 individuals show a positive effect. Another classification scheme is to compute Bayes factors for each indivdual. Figure 1 highlights the four individuals' effects that are more compatible with an effects-model than with the null. According to both approaches, this priming task is strategy-mixed with some people showing an effect and others not. There have been several mixture model approaches in psychology with the goal of classification (e.g. Houpt & Fific, 2017; Little et al., 2011; Rouder, Morey, Speckman, & Pratte, 2007).

Thiele, Haaf, & Rouder (2017) level a difficult critique at using classification schemes to assess the strategy question. In classification, sample noise is misinterpreted as evidence for the mixed-strategy conclusion. Heterogeneity is assumed, and nuisance variation may result in both null and effect interpretations. In fact, with an appropriate analysis to be presented subsequently, we show that the data in Figure 1 are best understood as a single, small, common effect with no variation across people.

To mitigate Thiele et al.'s critique, Haaf & Rouder (2017) and Thiele et al. (2017)

[1]See http://quoteinvestigator.com/2011/05/13/einstein-simple/ for a discussion on the attribution of the quote.

[2]Data comes from Pratte & Rouder (2009), Experiment 2. Details on the data set can be found in the Application section.

developed a Bayesian model-comparison framework where some models are explicitly strategy-pure. In this framework, a single model is placed on the collection of individuals' effects rather than placing separate models on each individual's effect. Take, for example, the strategy-pure model where all individuals have a positive true priming effect. This model is implemented by simulaneously imposing order constraints—effects must be positive—on all individuals. How to assess all these order-constraints simultaneously is not known in conventional statistical frameworks. Fortunately, this assessment is conceptually straightforward in the Bayesian framework (Gelfand, Smith, & Lee, 1992), and computational development is provided in Haaf & Rouder (2017) and Klugkist, Kato, & Hoijtink (2005). This Bayesian framework provides for the assessment of strategy-purity relative to an unconstrained model that is neither strategy-pure nor strategy-mixed. Here, we include a mixture model that is explicitly strategy-mixed.

Figure 2 graphically depicts the core of the models. The top row shows three strategy-pure models. Panel A is the most simple strategy-pure model. All people have a true null effect, and this null effect is indicated by the spike at zero. Panel B shows another simple strategy-pure model. Here, all people have the same true positive effect. Panel C shows a strategy-pure case with individual variability. Here, individuals' effects follow a distribution over positive values rather than a spike at one value. The distribution is restricted to positive values, and it is this restriction that defines the strategy purity. The bottom row shows two strategy-mixed models. Panel D is a mixture model: People either have no effect with a certain probability or they have a positive effect that follows a distribution with a complementary probability. 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 distribution. People who are truly in the spike exhibit a different strategy than those who are truly in the slab. Finally, Panel E shows the usual random-effects model where individuals' true effects follow a normal distribution. Even though the model has a convenient mathematical form, it does not make

a theoretical distinction between strategies on an individual level. We use this model as a none-of-the-above strategy-mixed option in cases where individual variation truly spans positive and negative effects.

The goal here is to assess the evidence from the data for the various models in Figure 2. If models in the top row are favored, then we may favor a strategy-pure account. Alternatively, if models in the bottom row are favored, then we may favor a strategy-mixed account. Fortunately, Bayesian model comparison through Bayes factors is ideal for this application.

In the next section, we provide a brief formal overview over the models depicted in Figure 2. Following this, we outline informally the Bayes factor model comparison strategy and how it penalizes complexity by focussing on predictions. With the Bayes factors developed, we analyze priming and Stroop interference data. While strategy-mixed processing is rare, we can document at least one case where it occurs.

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, $k = 1, \ldots, K_{ij}$.[3] 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

[3]Due to data cleaning, variation in the number of trials per person and condition is possible.

these effect parameters results in the models depicted in Figure 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 Figure 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 Figure 2A.

Common-effect Model. The common-effect model, denoted $\mathcal{M}_1$, corresponds to the spike in Figure 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 Figure 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 still has the exact same true effect, but this effect is only restricted to be positive, not fixed to a specific value.

Positive-Effects Model. The positive-effects model is denoted $\mathcal{M}_+$, and it is the first model that introduces true individual variability. Even so, the model still specifies

strategy-purity. True individuals' effects may vary, but are constrained to be positive:

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

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 Figure 3.[4] 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 Figure 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

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

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 Figure 3 shows the model specifications for two participants. For these hypothetical individuals, four compinations 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,

[4] 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).

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 $x = 0$.

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

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

The first panel in the last row of Figure 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 strategies into account.

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 & Rouder (2017) in specification.[5] 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, Lu, Morey, Sun, & Speckman, 2008). We model individual differences

[5] 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.

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.

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. It must be estimated. It has variability in this regard and this 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 Figure 3. For the 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 a task is strategy-pure or strategy-mixed. 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 Figure 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 Figure 2, cannot distinguish between strategies. It is, however, appropriate for estimating posterior spike and slab probabilities and the collection of $\boldsymbol{\theta}$.

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. Of these five, the spike-and-slab and unconstrained models are strategy-mixed cases; the null, common-effect and positive-effects models are stratgy-pure cases. 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 & Raftery (1995), and Morey, Romeijn, & Rouder (2016), we provide an informal discussion that we have previously presented in Rouder, Morey, & Wagenmakers (2016) and Rouder (2017). Informally, evidence for models reflects how well they predict data.

The right column of Figure 3 shows predictions for data from each of the five models. 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.[6] The remaining rows of Figure 3 show the predictions for the common-effect, positive-effects, spike-and-slab, and unconstrained models. In all cases, the precictions 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 Figure 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 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 innacurate with numeric methods. We follow here the development by Haaf & Rouder (2017) who provide the details of model comparison between the null, common-effect, positive-effects and unconstrained models. Their development builts on analytical solutions pioneered by Zellner & Siow (1980) and expanded for ANOVA by Rouder, Morey, Speckman, & Province (2012). These analytical solutions are used for comparisons among the null, common effect, and unconstrained models. The development

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

also employs the *encompassing approach* introduced by Klugkist et al. (2005). This approach may be used for comparisons with the positive-effects model.

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 Bayes factor between the null model and the spike-and-slab model is given by

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

where the event $\boldsymbol{z} = \mathbf{0}$ 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 of the same event. 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.

Computation of these probabilities is straightforward in MCMC sampling. Let $\boldsymbol{z}[m]$ 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 $\boldsymbol{z}[m]$ 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).

Application

We apply the five models to three different data sets: A priming data set provided by Pratte & Rouder (2009), and two Stroop experiments provided by Pratte, Rouder, Morey, & Feng (2010). The goal here is to answer the question wether the tasks deployed in the three experiments are strategy-pure or strategy-mixed. We provide estimation and model comparison results for the three data sets and discuss them in the light of the experimental paradigms.[7]

Priming Data Set

The priming data used here, reported by Pratte & Rouder (2009), comes from a number priming task.[8] We suspect this task is strategy-mixed 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.

[7]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), *beeswarm* (0.2.3, Eklund, 2016), *coda* (0.19.1, Plummer, Best, Cowles, & Vines, 2006), *colorspace* (Stauffer, Mayr, Dabernig, & Zeileis, 2009; 1.3.2, Zeileis, Hornik, & Murrell, 2009), *curl* (2.8.1, Ooms, 2017), *devtools* (1.13.2, Wickham & Chang, 2016), *diagram* (1.6.3, Soetaert, 2014a), *dotCall64* (Gerber, Moesinger, & Furrer, 2015, 0.9.4, 2016), *fields* (9.0, Douglas Nychka, Reinhard Furrer, John Paige, & Stephan Sain, 2015), *gmm* (1.6.1, Chaussé, 2010), *maps* (3.2.0, Richard A. Becker, Ray Brownrigg. Enhancements by Thomas P Minka, & Deckmyn., 2016), *MASS* (7.3.45, Venables & Ripley, 2002), *Matrix* (1.2.10, Bates & Maechler, 2017), *MCMCpack* (1.4.0, Martin, Quinn, & Park, 2011), *msm* (1.6.4, Jackson, 2011), *mvtnorm* (1.0.6, Genz & Bretz, 2009; Wilhelm & G, 2015), *papaja* (0.1.0.9492, Aust & Barth, 2017), *plotrix* (3.6.5, J, 2006), *sandwich* (2.3.4, Zeileis, 2004, 2006), *shape* (1.4.2, Soetaert, 2014b), *spam* (2.1.1, Furrer & Sain, 2010; Gerber & Furrer, 2015), *spatialfil* (0.15, Dinapoli & Gatta, 2015), and *tmvtnorm* (1.4.10, Wilhelm & G, 2015).

[8]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.

Yet, it is well known that this threshold varies considerably across people. For example, Morey, Rouder, & Speckman (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, & Wilhelmsen, 1989). For any given presenation duration, some individuals may be able to detect the prime and others may not. This difference may lead to variablity in processing with some people processing the primes and others not. Such variability corresponds to our strategy question.

Results. Figure 4A provides two sets of parameter-estimation results. The first set, denoted by the crosses that span from -0.02 seconds to 0.03 seconds, are the observed effects for the individuals, and these are the same points that are plotted in Figure 1. Observed effects in this context are the differences in individuals' sample means for the incongruent and congruent conditions. Crosses are coloured red or gray to indicate whether the observed effects are negative or positive, respectively. Overall, effects are relatively constrained with no participant having a more than 31 millisecond 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 whereever possible (Davis-Stober, Dana, & Rouder, 2017). The individuals' posterior probability of being in the slab ranges from 0.31 to 0.65.

From the model estimates, it is evident that individual effects are tightly clustered and slighty positive with a mean of 4 ms. Yet, these results are not sufficient to answer the strategy question. 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 analyse the above models and compare them with Bayes factors. The results are shown in Figure 4B. The common-effect model is preferred, indicating that everyone has a single, common effect. The next most parsemonious model is the null model, where all individuals have no effect. The best-performing strategy-mixed model is the spike-and-slab model, and the Bayes factor between it and the common-effect model is 31-to-1 in favor of the common-effect model. We take this Bayes factor as evidence for strategy-purity in this task: 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 used the strategy of not shifting their attention to the word in the periphery while others did. In this scenario, the task is strategy-mixed with some participants showing a true Stroop effect and others showing none at all. The alternative is a strategy-pure account where all participants exhibited a small Stroop effect similar to the priming effect above.

Results. Observed effects are shown by the crosses in Figure 4C. Of the 38 participants, 10 show an observed negative priming 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 0.19 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 Figure 4D confirm this consideration: the Bayes factor between the spike-and-slab model and the runner-up common-effect model is 3-to-1 in favor of the spike-and-slab model. This Bayes factor provides slight evidence for mixed strategies 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 task-purity with everyone showing a Stroop effect.

Results. Parameter estimates are shown in Figure 4E. 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 Figure 4F. Overall, there is 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 for the strategy-purity of the task. 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 the Stroop task is strategy-pure — if targets are presented at fixation.

Discussion

In this paper, we address whether people use differing strategies for tasks where the sign of the behavioral outcome measure maps well into different processing 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 priming), responses to congruent targets are slower than to incongruent ones (negative priming), and responses to congruent targets are equally fast as to incongruent ones (no priming). Whenever a behavioral outcome can be trichotomized this way, we can assess whether processing is strategy-pure or strategy-mixed. Obvious applications include context effects (e.g., Stroop, flanker, 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 strategy-pure null model; a strategy-pure common effect model; a strategy pure slab model; a strategy-mixed spike-and-slab model; and a strategy-mixed slab model. The novel element

here is the usage of the spike-and-slab model. Although spike-and-slab models are used frequently in statistics, they are used 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 the 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 (Klauer & Kellen, 2010; Province & Rouder, 2012; e.g. Yantis, Meyer, & Smith, 1991).

Our approach differs markedly from these previous queries. Our focus is not on characterizing trial-by-trial variability but on 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. 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, 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 mixed-strategy claim. We think, however, that mixtures of strategies are relatively rare in cognitive psychology where experimental paradigms are relatively well defined. Only in cases where tasks are ill-defined, i.e. participants have the degrees of

freedom to decide on a strategy that was not anticipated by the researchers, 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 and still successfully completing the task.

Concerns

Normal Specification. One concern with the proposed approach is the reliance on normal parametric model specifications. The advantage of the normal specification is computational convenience. With it, the many dimensions of the high-dimension integrals that define the Bayes factor may be computed symbolically to high precision. Without it, we suspect numeric integration would be exceedingly slow and inaccurate. Yet, researchers may be concerned about the misspecification of the normal. Here, for example, we focus on applications with response times. RT is skewed rather than symmetric, and the standard deviation tends to increase with the mean (Luce, 1986; Rouder, Yue, Speckman, Pratte, & Province, 2010; Wagenmakers & Brown, 2007).

We think this concern 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 the processing questions without any need to know or consider the true shapes or true variances. The inference here 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 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 prior on inference. In general, Bayesian models require a careful choice of priors. These priors have an effects 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 above starting point, many Bayesian analysts actively seek to minimize these effects by picking likelihoods, prior parametric forms, and heuristic methods of inference so that variation in prior settings have marginal effects (Aitkin, 1991; Gelman, Carlin, Stern, & Rubin, 2004; Kruschke, 2012; Spiegelhalter, Best, Carlin, & Linde, 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 & Lee (2012) argue that the prior is where theoretically important constraint is 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 & 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 & Rouder (2017), and here we review the main issues. The critical settings are the *scales* on ν and ϵ_i, and these settings are relative to σ, the residual noise. In tasks like this, with subsecond RTs, a standard deviation of repeated response times for a given participant and a given condition is 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 1/6th or 50/300. 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. We explore the effects of halving and doubling these settings, which represents a reasonable range of variation.

With these reasonable ranges of variation, we are ready to explore the effects of prior specification on Bayes factors. These 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 is the location Stroop data: The spike-and-slab model is preferred for the chosen settings and when the scale non 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 100ms 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 extensively 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 convincind result, more evidence for a mixed-strategy

account would is needed.

Computational Issues. In previous work (Haaf & Rouder, 2017; Thiele et al., 2017) we developed the null, common-effect, positive-effects and unconstrained models. Here we add the spike-and-slab model to the set and show it is a worthy competitor in at least one application. The former four models are computationally convenient, the Bayes factors can be computed quicky using a combination of Rouder et al.'s (2012) symobic integration as implemented in the BayesFactor package for R (Morey & Rouder, 2015) and Klugkist and colleagues encompassing approach (e.g. Klugkist & Hoijtink, 2007). Bayes factors for the spike-and-slab model, however, while conceptually similar, is computationally far more difficult in practice. The difficulty here is that Bayes factor computation is reliant on Marcov-chain-Montecarlo methods. We find that convergence in these methods is slow for the spike-and-slab model and hundreds of thousands of iterations are needed to approximate the Bayes factor. The good news here is that we can assess the accuracy of this approximation fairly readily through transitivity of Bayes factors. Figure 4B/D/F illustrate this check. We can compute the Bayes factor between the unconstrained model and the null model either through the spike-and-slab model directly with symbolic integration methods. We find comparable results from both methods.

Our computational difficulties illustrate that there is much work to be done. Although Bayes factors are convenient in standard cases with normal distributions, moving to the assessment of more custom-tailored models of psychological process remains timely and topical.

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.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

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

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. (2017). When are sample means meaningful? The role of modern estimation in psychological science. *Open Science Framework. April*, *12*.

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

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

Eklund, A. (2016). *Beeswarm: The bee swarm plot, an alternative to stripchart*. Retrieved from https://CRAN.R-project.org/package=beeswarm

Eurythmics. (1983). Sweet dreams (are made of this). UK.

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

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. 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 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/

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*.

Houpt, W., J., & Fific, M. (2017). *A hierarchical bayesian approach to distinguishing serial and parallel processing*.

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

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.

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

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., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69.

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

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

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

Springer.

Little, D. R., Nosofsky, R. M., & Denton, S. (2011). Response time tests of logical-rule-based models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1–27.

Logan, G. D. (1988). Towards an instance theory of automization. *Psychological Review*, *95*, 492–527.

Logan, G. D. (1992). Shapes of reaction time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 883–914.

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 Assocation*, *83*, 1023–1032.

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

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/

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

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

Rickard, T. C. (2004). Strategy execution in cognitive skill learning: An item-level test of candidate models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 65–82.

Rouder, H., J. N. (2017). *From theories to models to predictions: A bayesian model comparison approach*.

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., 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., 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., & Pratte, M. S. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin and Review*, *14*, 597–605.

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., 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.

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., & Linde, A. van der. (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. doi:10.1175/BAMS-D-13-00155.1

Thiele, J. E., Haaf, J. M., & Rouder, J. N. (2017). *Bayesian analysis for systems factorial technology*.

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). Assing 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., & 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

Yantis, S., Meyer, D. E., & Smith, J. E. K. (1991). Anaylsis 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. doi:10.1016/j.csda.2008.11.033

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.

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	*	0.05	0.06	0.04
1/12	1/5	0.14	*	0.02	0.04	0.04
1/3	1/20	0.06	*	7.31e −8	0.002	1.06e −5
1/3	1/5	0.14	*	3.57 e −8	9.58e −4	1.01e −5
Location Stroop						
1/6	**1/10**	4.13e −4	0.31	0.05	*	0.1
1/12	1/20	1.18e −4	0.12	0.12	*	0.08
1/12	1/5	2.69e −4	0.19	0.04	*	0.07
1/3	1/20	9.87e −5	0.98	0	*	4.39e −3
1/3	1/5	2.1e −3	1.36	6.29e −5	*	0.01
Color Stroop						
1/6	**1/10**	1.05e −74	1.92e −6	*	0.04	0.12
1/12	1/20	1.29e −74	8.91e −7	*	0.03	0.05
1/12	1/5	1.42e −74	3.01e −6	*	0.04	0.16
1/3	1/20	7.34e −75	5.06e −7	*	0.03	0.01
1/3	1/5	1.06e −74	2.25e −6	*	0.04	0.05

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 askerisks mark the winning model for each data set for the original analysis, and Bayes factors are computed for comparison to this model.

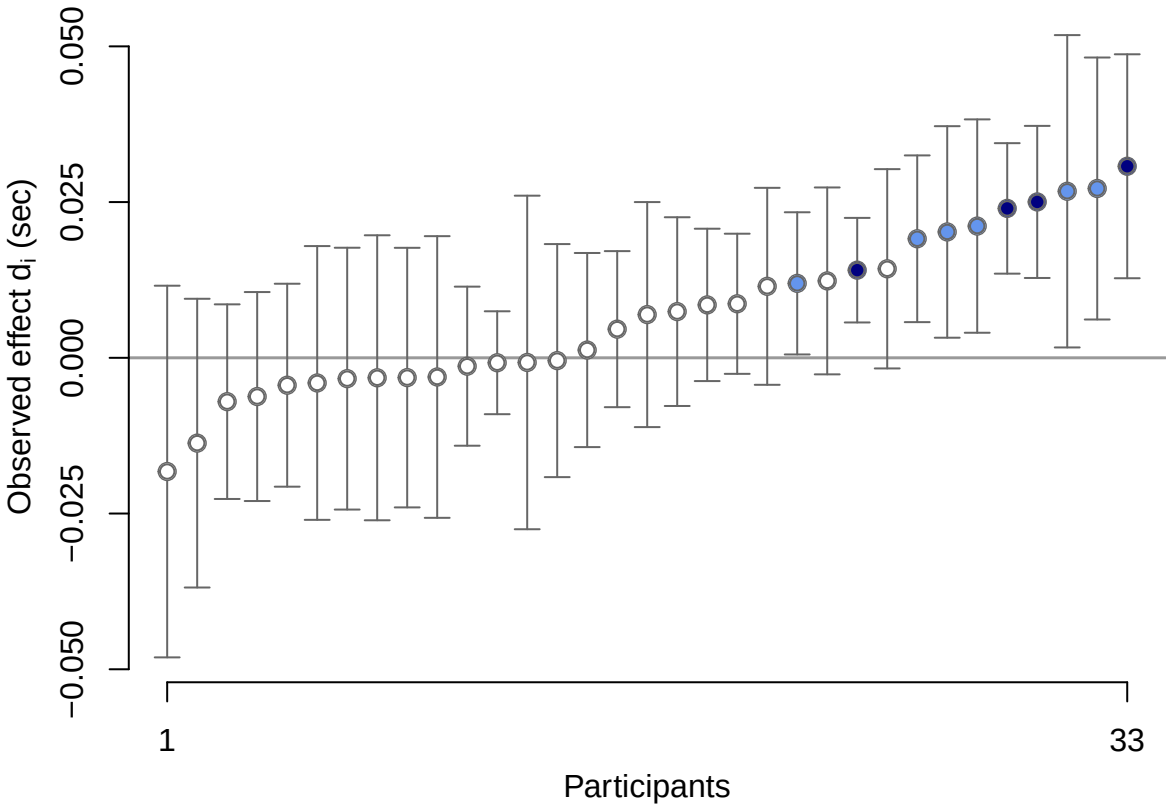

Figure 1. Individual observed effects from a priming task ordered from lowest to highest. Shading of the points indicates strategy according to two criteria. Dark blue points indicate a positive priming effect for the criterion of BF>2. Light or dark blue points indicate a positive priming effect for the criterion of 80%CIs excluding zero. White points indicate a null effect according to both criteria.

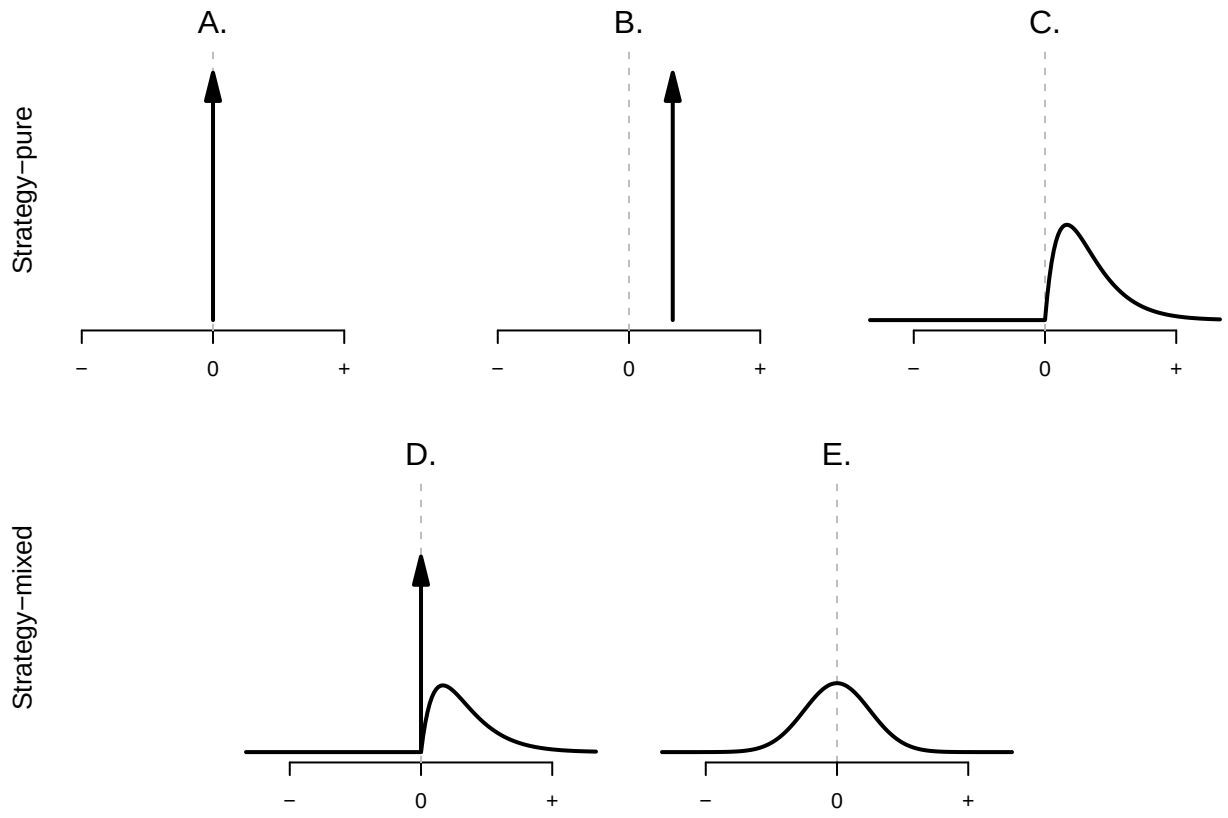

Figure 2. Models for pure and mixed strategies. A./B. Pure-strategy accounts without individual variation. C. Pure-strategy account with individual variability. D. A mixed-strategy account where some individuals have no effect and others have a positive effect. E. Common random-effects model that can neither exclude the possibility of strategy-mixtures nor distinguish between strategies.

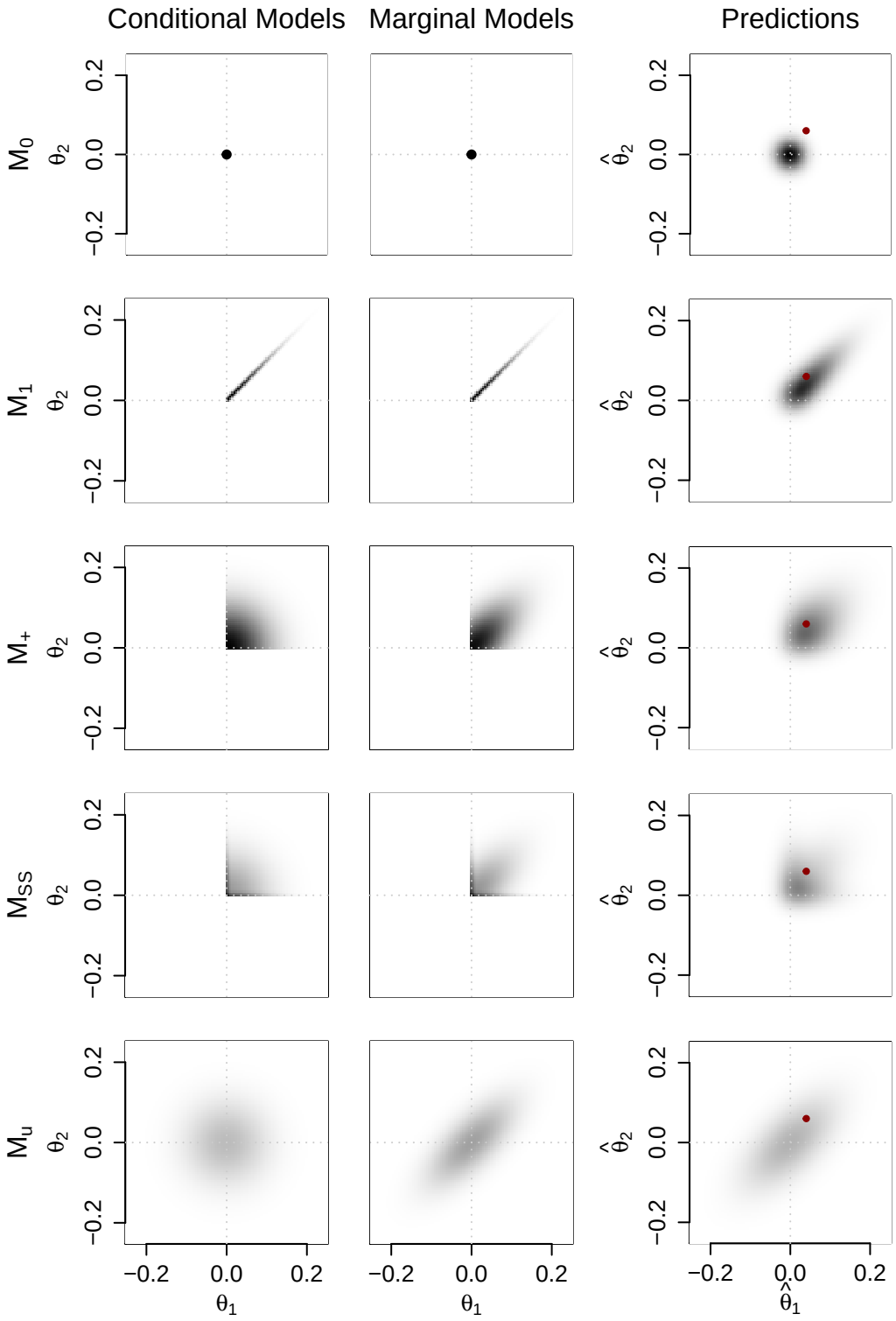

Figure 3. Model specification and predictions for two exemplary participants. Left column: Model specifications conditional on specific prior settings. Middle column: Marginal model specifications integrated over prior distributions show correlation between individuals' effects. Right column: Resulting 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).

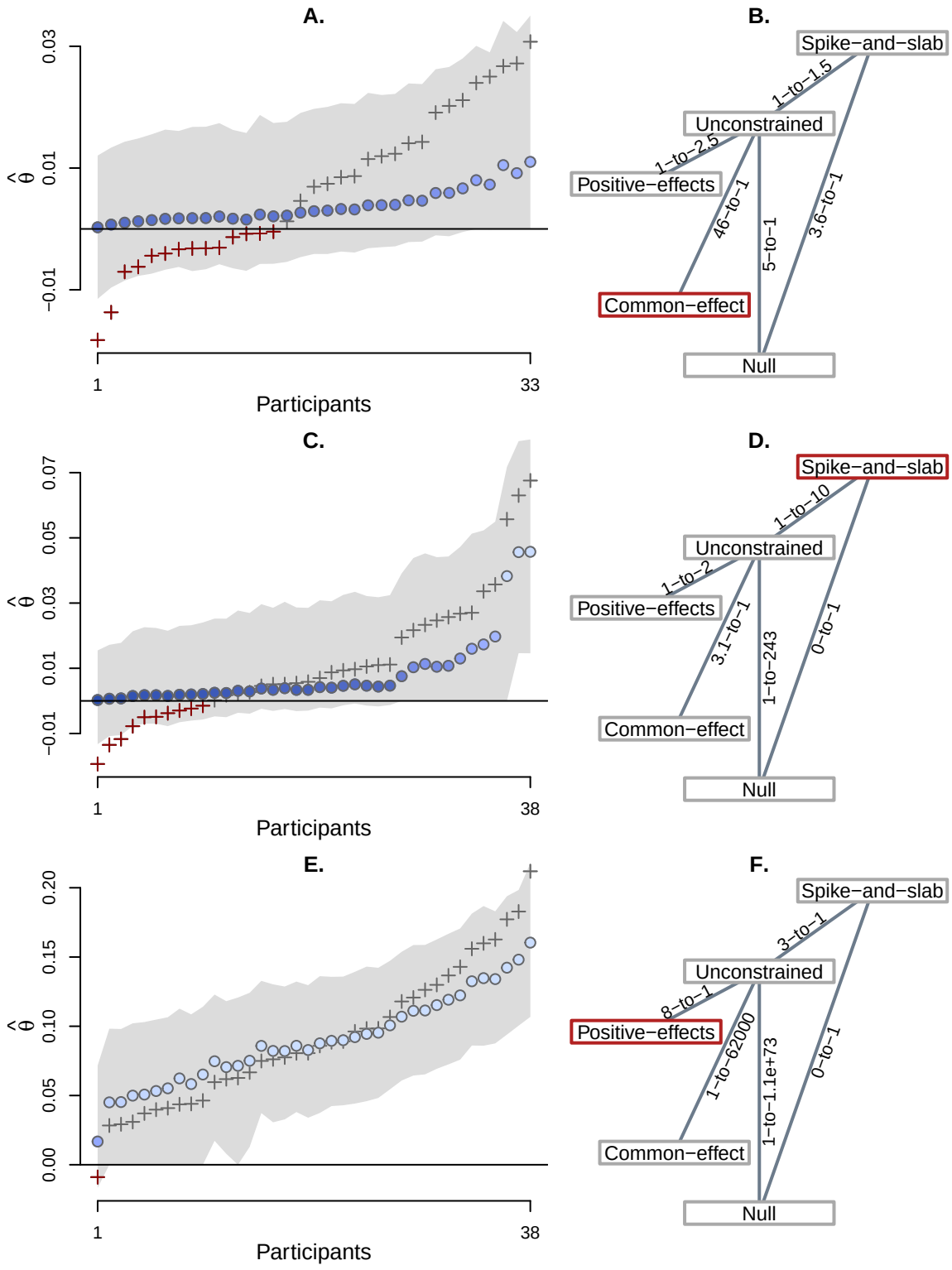

Figure 4. Model estimates (left column) and Bayesian model comparison results for A./B. the priming data set; C./D. the location Stroop task; E./F. the color Stroop task. Left column: Crosses show observed effects with red crosses indicating negative effects. Points show model estimates with lighter 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.