Bayesian Inference for Psychology, Part IV: Parameter Estimation and Bayes factors.

Jeffrey N. Rouder[a,b,1], **Julia Haaf**[b], **and Joachim Vandekerckhove**[a]

This is a preprint of a manuscript submitted for publication with *Psychonomic Bulletin and Review*.

In the psychological literature, there are two seemingly different approaches to inference: that from estimation of posterior intervals and that from Bayes factors. We provide an overview of each method and show that a salient difference is the choice of models. The two approaches as commonly practiced can be unified with a certain model specification, now popular in the statistics literature, called *spike-and-slab* priors. A spike-and-slab prior is a mixture of a null model, the spike, with an effect model, the slab. The estimate of the effect size here is a function of the Bayes factor, showing that estimation and model comparison can be unified. The salient difference is that common Bayes factor approaches provide for privileged consideration of theoretically useful parameter values, such as the value corresponding to the null hypothesis, while estimation approaches do not. Both approaches, either privileging the null or not, are useful depending on the goals of the analyst.

Bayes factors | Bayesian estimation | Bayesian inference | ROPE | Hypothesis testing

Bayesian analysis has become increasing popular in many fields including psychological science. There are many advantages to the Bayesian approach. Some champion its clear philosophical underpinnings where probability is treated as a statement of belief or information and the focus is on updating beliefs rationally in face of new data (de Finetti, 1974; Edwards, Lindman, & Savage, 1963). Others note the practical advantages—Bayesian analysis often provides a tractable means of solving difficult problems that remain intractable in more conventional frameworks (Gelman, Carlin, Stern, & Rubin, 2004). This practical advantage is especially pronounced in cognitive science where substantive models are designed to account for mental representation and processing. As a consequence, the models tend to be complex and nonlinear, and may include multiple sources of variation (Kruschke, 2011b; Lee & Wagenmakers, 2013; Rouder & Lu, 2005). Bayesian analysis, especially Bayesian nonlinear hierarchical modeling, has been particularly successful at providing straightforward analyses in these otherwise difficult settings (e.g., Rouder, Sun, Speckman, Lu, & Zhou, 2003; Vandekerckhove, Tuerlinckx, & Lee, 2011; Vandekerckhove, 2014).

Bayesian analysis is not a unified field, and Bayesian statisticians disagree with one another in important ways (Senn, 2011).[1] We highlight here two popular Bayesian approaches that may seem incompatible inasmuch as they provide different answers to what appears to be the same question. We discuss these approaches in the context of the simple problem where there is an experimental and control condition and *we wish to characterize the evidence from the data for the presence or absence of an effect.*

In one approach, termed here the **estimation approach**, the difference between the conditions is represented by a parameter, and the posterior density of this parameter is updated using Bayes' rule. Two examples of posteriors on effect size are provided by the curves in Figure 1; in these examples there are 50 observations per curve. The next step is going from these posteriors to inferential statements. We highlight two approaches: 1. Lindley (1965), in his early career, notes that one could examine the *highest-density credible intervals* (HDCIs). These highest-density credible intervals contain a fixed proportion of the mass, say 95%, and posterior values inside the interval are more plausible than those outside the interval. Examples of these HDCIs are shown in Figure 1 as dashed vertical lines. Values outside the intervals may be considered sufficiently implausible to be untenable. By this reasoning, there is evidence for an effect in Figure 1A as zero is outside the credible interval; there is a lack of evidence for an effect in Figure 1B as the zero is inside the credible interval. 2. Kruschke (2012) and Kruschke and Liddell (2017) take a modified version of this approach, in which the credible interval is compared to a pre-established region called a *region of practical equivalence* (ROPE). ROPEs are small intervals around zero containing only values that are considered to be practically the same as zero. An example of a ROPE might be the interval on effect sizes from $-.1$ to $.1$, and this interval is shaded in Figure 1. In Kruschke's approach, one concludes that the null hypothesis is false if the HDCI falls completely outside of the ROPE. If the HDCI falls completely inside of the ROPE, one concludes that the null hypothesis is (for all practical purposes) true. If the HDCI partly overlaps with the ROPE, Kruschke recommends one reserve judgment. By this reasoning, neither posterior in Figure 1 yields a firm decision though the HDCI in Figure 1A comes close to being fully outside the ROPE. The fact that that HDCI both contains the ROPE and is so much wider than it in Figure 1B might indicate that more data are needed. The key commonality of these two estimation approaches is that inference is based on the posterior distributions of key parameters.

Although the posterior-estimation approach seems straightforward, it is not recommended for drawing conclusions about the presence or absence of effect by a number of Bayesian psychologists (Dienes, 2014; Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007). These authors

[1] Perhaps such disagreements should be expected given the contentious history of academic statistics. Even null hypothesis significance testing is a contentious hybrid of Fisherian and Neyman-Pearson schools of thought (Gigerenzer et al., 1989; Lehmann, 1993).

[a]University of California, Irvine; [b]University of Missouri
All authors contributed to the final draft.
[1]To whom correspondence should be addressed. E-mail: jrouder@uci.edu.

instead advocate a **Bayes factor** approach for drawing such conclusions. In Bayesian analysis, it is possible to place probability on models themselves without recourse to parameter estimation. In this case, a researcher could construct two models: one that embeds no difference between the conditions and one that embeds some possible difference. The researcher starts with prior beliefs about the models and then updates these rationally with Bayes' rule to yield posterior beliefs. Evidence from data is how beliefs about the models themselves change in light of data; there may be a favorable revision for either the effects or null-effects model. This Bayes factor approach remains controversial too, and it has been critiqued as well for being too sensitive to the prior density on parameter values (see Liu & Aitkin, 2008; Gelman & Carlin, 2017; Kruschke, 2011a).

From a pragmatic view, Bayes factor and posterior estimation often lead to the same conclusion when observed effects are large. This is not too surprising as large effects should be detected by all approaches. Conclusions may differ, however, when observed effects are small. Consider for example the posterior in Figure 1A where the posterior 95% credible interval does not include zero. If zero is considered sufficiently implausible, this posterior seemingly provides some evidence for an effect. Yet, the Bayes factor, which is discussed at length subsequently, is only 2.8-to-1 in favor of the effect. If we had started with 50-50 beliefs about an effect (vs. a lack of an effect), we end up with just less than 75-25 beliefs in light of data. While this is some revision of belief, this small degree is considered modest rather than substantial (Jeffreys, 1961; Raftery, 1995). Likewise, consider the posterior for Figure 1B where the posterior 95% credible interval is centered around zero. With the ROPE inference, there might be a hint of evidence for no effect, but the width of the credible interval is problematic for firm assessment. With Bayes factor, however, the evidence is 6.8-to-1 in favor of the null model. We may state positive evidence for an invariance of performance across the two conditions.

This divergence in conclusions leaves the nonspecialist in a quandary about whether to use posterior estimation or Bayes factors. Here we address this quandary head-on: We will first draw a sharp contrast between the two approaches and show that they provide for quite different views of evidence. Then, to help understand these differences, we highlight a unification that has run in the Bayesian literature since Jeffreys (1938). We show that the Bayes factor may be represented as part of estimation under a certain model specification known currently in the statistics literature as a *spike-and-slab* model (George & McCulloch, 1993). With this demonstration, a key difference between estimation and a Bayes factor approach comes into full view: it is a difference in model specification. These spike-and-slab models entail different assumptions than more conventional models. Our view is that the assumptions underlying spike-and-slab are the most judicious ones for most scientific questions. Once researchers understand these assumptions, they can make informed and thoughtful choices about which are most appropriate for specific research applications. Guidance is provided subsequently.

Posterior Estimation

Bayesian posterior estimation is performed straightforwardly through updating by Bayes' rule. Let us take a simple example where a set of participants provide performance scores in each of two conditions. For example, consider a priming task where the critical variable is the response time, and participants provide a mean response time in a primed and unprimed condition. Each participant's data may be expressed as a difference score, namely the difference between mean response times. Let Y_i, $i = 1, \ldots, n$ be these difference scores for n participants. In the usual analysis, a researcher might perform a t-test to assess whether these difference scores are significantly different from zero.

In contrast, Bayesian analysis begins with consideration of a model, and in this case, we assume that each difference score is a draw from a normal with mean μ and variance σ^2:

$$Y_i \sim \text{Normal}(\mu, \sigma^2), \quad i = 1, \ldots, n. \qquad [1]$$

In the following development, we will assume that σ^2 is known to simplify the exposition, but it is straightforward to dispense with this assumption. It is helpful to consider the model in terms of effect sizes, δ, where $\delta = \mu/\sigma$ is the true effect size and is the parameter of interest.

Bayesian analysis proceeds by specifying what is known or believed about the effect size parameter δ. This information is expressed as a prior distribution on parameters. In this article, we use the term *prior* and *model* interchangeably as a prior is nothing more than a model of parameters. Model $\mathcal{M}_1$ provides prior beliefs on δ.

$$\mathcal{M}_1 : \qquad \delta \sim \text{Normal}(0, \sigma_0^2). \qquad [2]$$

The centering of the distribution at zero is interpreted as a statement of prior equivalence about the direction of any possible effect—negative and positive effects are *a priori* equally likely. The prior variance, σ_0^2 must be set before analysis.

It is helpful to explore how the value of this setting affects estimation. Figure 2A shows this effect. Ten hypothetical values of Y_i, the difference scores, are shown as small line segments across the bottom (x-axis) of the plot. The sample mean of these ten is shown as the vertical line. The posterior distributions of δ are shown for three different prior settings. The first prior setting, $\sigma_0 = .5$, codes an *a priori* belief that δ is not much different than zero. The second prior setting, $\sigma_0 = 2$, is a fairly wide setting that allows for a large range of reasonable effect sizes without mass on exceedingly large values. The third prior setting, $\sigma_0 = 5000$ indicates that researcher is unsure of the effect size, and holds the possibility that it can be exceedingly large. Even though the priors are very different, the posterior distributions are quite similar. We may say that the posterior is robust to wide variation in prior settings. In fact, it is possible to set $\sigma_0 = \infty$ to equally weight all effect sizes *a priori*, and in this case, the posterior would be indistinguishable from that for $\sigma_0 = 5000$.

This robustness to prior specification in estimation translates to a robustness in making inferential statements. Figure 2B shows the case for Lindley's credible interval approach. Shown is the minimal observed effect size needed such that zero is excluded from the lower end of the credible interval. As can be seen, this value stabilizes quickly and varies little. The same behavior holds for Kruschke's ROPE approach as well (dashed lines).

It may be tempting to think that this robustness is a general property of posterior estimation. We will show it is not. There are useful models where posterior estimates are not robust to prior settings. In these cases, the prior settings become theoretically important parts of the model specification.

Bayes Factors

In Bayesian analysis, it is possible to place beliefs directly onto models themselves and update these beliefs with Bayes' rule. Let

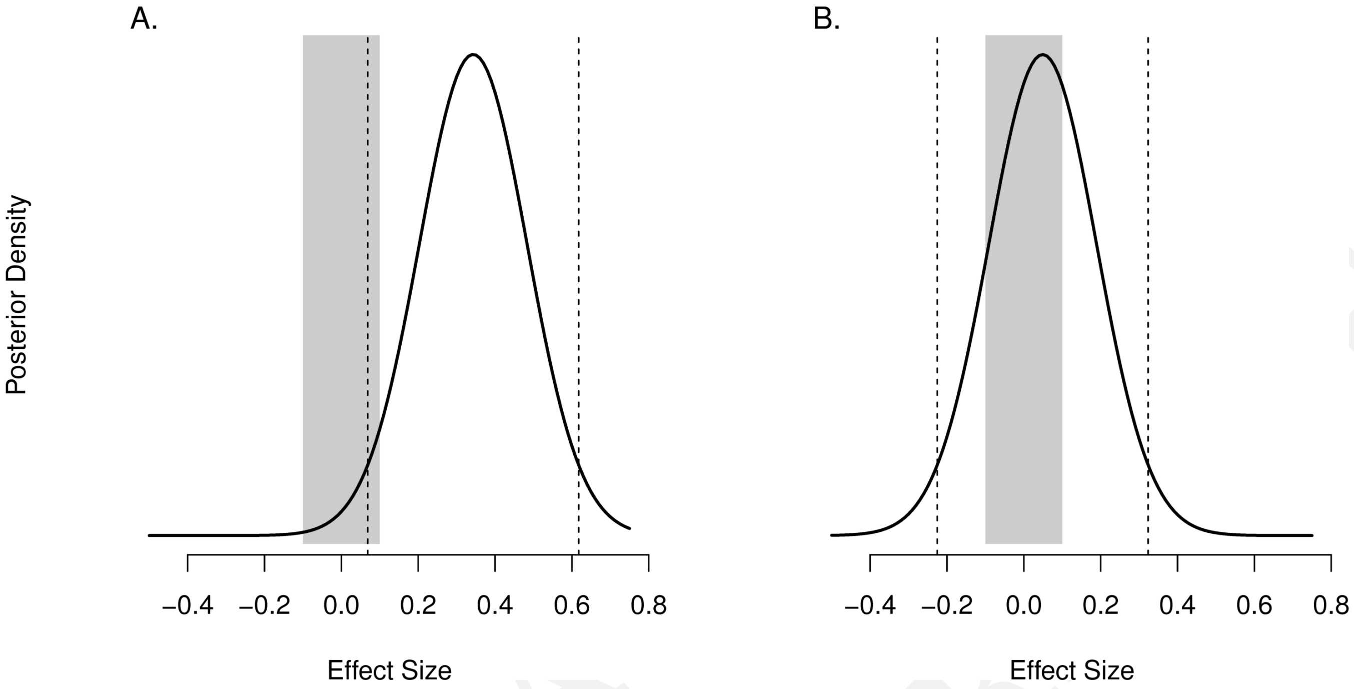

Fig. 1. Examples of posterior distributions of effect size. The intervals between the vertical dashed lines are the 95% highest density credible intervals. The shaded regions are a small region of posterior equivalence (ROPE). **A.** The posterior is localized away from zero though not completely away from the ROPE. **B.** The posterior is localized around zero though it extends past the ROPE. The conclusions drawn from these posteriors depend on heuristical rules used for interpretation.

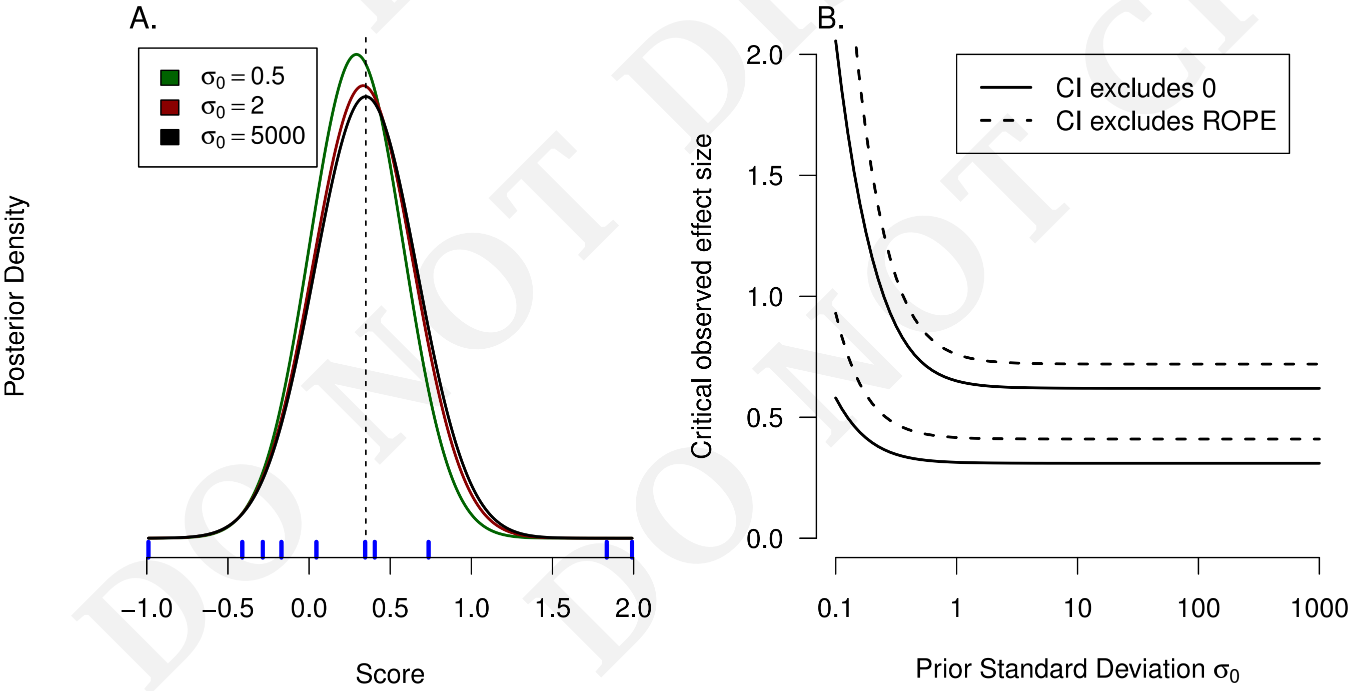

Fig. 2. The dependence of the posterior estimation on prior setting σ_0. **A.** Posterior distributions on effect size δ for $N = 10$ and for a sample effect size of .35. for three settings of σ_0. The small blue bars denote the observations. The value of σ^1 is known and set to 1.0. **B.** Minimum observed effect sizes needed such that the posterior 95% credible interval excludes zero. The two lines are for sample sizes of 10 (top) and 40 (bottom). The results show a robustness to the prior setting of σ_0.

$\mathcal{M}_A$ and $\mathcal{M}_B$ denote any two models. Let $Pr(\mathcal{M}_A)$ and $Pr(\mathcal{M}_B)$ be *a priori* beliefs about the plausibility of these two models. It is more desirable to state relative beliefs about the two models as odds. The ratio $Pr(\mathcal{M}_A)/Pr(\mathcal{M}_B)$ is the relative plausibility of the models, and for example, the statement $Pr(\mathcal{M}_A)/Pr(\mathcal{M}_B) = 3$ indicates that Model $\mathcal{M}_A$ is three times as plausible as Model $\mathcal{M}_B$. Odds such as $Pr(\mathcal{M}_A)/Pr(\mathcal{M}_B)$ are called *prior odds* because they are stipulated before seeing data. They may be contrasted to *posterior odds*, which are the same odds in light of the data and denoted $Pr(\mathcal{M}_A \mid \mathbf{Y})/\mathbf{Pr}(\mathcal{M}_\mathbf{B} \mid \mathbf{Y})$. be the prior and posterior odds, respectively. Bayes rule for updating to posterior odds from prior odds is

$$\frac{Pr(\mathcal{M}_A|\mathbf{Y})}{Pr(\mathcal{M}_B|\mathbf{Y})} = \frac{Pr(\mathcal{M}_A)}{Pr(\mathcal{M}_B)} \times B. \qquad [3]$$

The updating factor,

$$B = \frac{Pr(\mathbf{Y} \mid \mathcal{M}_\mathbf{A})}{Pr(\mathbf{Y} \mid \mathcal{M}_\mathbf{B})},$$

is called the *Bayes factor*, and it describes how the data have led to a revision of beliefs about the models. Several authors including Jeffreys (1961) and Morey, Romeijn, and Rouder (2016) refer to the Bayes factors as the *strength of evidence from data about the models* precisely because the strength of evidence should refer to how data lead to revision of beliefs. This evidence flows strictly though the probability of observing data under the models, a property of inference which is also known as the *likelihood principle* (Berger & Wolpert, 1988).

The Bayes factor has a second interpretation stemming from it being the relative probability of data under models. The probability of data under a model may be thought of as the *predictive accuracy* of that model – the degree to which the model predicted the data. The data in the equation are the observed data we obtain in an experiment. If the probability of observed data is high, then the model predicted the observed data to be where they were observed. If the probability of data is low, then the model did not predict the observations well. The Bayes factor is the relative predictive accuracy of one model over another. The deep meaning of Bayes' rule is that the strength of evidence is the *relative predictive accuracy*, and this is captured by the Bayes factor in Equation 3.

When we write the Bayes factor as B_{AB}, the subscripts indicate which two models are being compared. A Bayes factor of $B_{AB} = 10$ means that prior odds should be updated by a factor of 10 in favor of model $\mathcal{M}_A$; likewise, a Bayes factor of $B_{AB} = .1$ means that prior odds should be updated by a factor of 10 in favor of model $\mathcal{M}_B$. Bayes factors of $B_{AB} = \infty$ and $B_{AB} = 0$ correspond to infinite—total—support of one model over the other with the former indicating infinite support for model $\mathcal{M}_A$ and the latter indicating infinite support for model $\mathcal{M}_B$. We might say in such a case that one of the models is ruled out (i.e., "falsified") by the data.

For the simple example of comparing performance in two experimental conditions, we need one model for an effect and a second model for a lack of an effect (which is also called an invariance). A suitable model for an effect is the previous model, $\mathcal{M}_1$ given in (2). A model for an invariance is given by

$$\mathcal{M}_0 : \qquad \delta = 0.$$

With this setup, the Bayes factor is straightforward to compute.[2]

Inference by Bayes factor is more dependent on the prior setting σ_0^2 than is inference by the preceding posterior-estimation approach. Figure 3A shows the effects of increasing σ_0. As can be seen, the Bayes factor B_{10} favors the alternative when σ_0 is small (say, near 1) but decreases toward zero as σ_0 becomes increasingly large. Of note is the limit as σ_0 gets increasingly large. These diffuse priors on effect size in the alternative leads to total support for the null model over the alternative (Lindley, 1957), and this result contrasts to that for inference with credible intervals where inference is stable when σ_0 becomes increasingly large. This result occurs because the Bayes factor is sensitive to the complexity of the model, and when the $\sigma_0^2 = \infty$, the alternative can account for all data equally well, without constraint. Consequently, it is penalized completely. Figure 3B provides an different view of the effect of prior setting σ_0. It shows the minimum positive effect size need to support a Bayes factor of 3-to-1 in favor of Model $\mathcal{M}_1$ over $\mathcal{M}_0$ and is comparable to Figure 2B. As can be seen, inference by Bayes factor is more sensitive to prior settings than inference by estimation. The key region of difference is for increasing prior variance, σ_0^2. As σ_0^2 becomes large, greater and greater observed effect sizes are needed to evidence an effect. This behavior is in contrast to that for inference by credible intervals (Figure 2B) where there is stability with increasing prior variance.

At first glance, this dependence of the Bayes factors on the prior settings may seem undesirable. One fear is that researchers can seemingly obtain different results by adjusting the prior settings perhaps undermining the integrity of their conclusions (Gelman & Carlin, 2017). This dependence seems all the more undesirable when contrasted to the the robustness of posterior intervals to prior settings as shown in Figure 2 (Kruschke, 2011a). However, the situation is far more nuanced. Both Bayesian parameter estimation and Bayes factor model selection are supported by the same rules of probability (see Etz & Vandekerckhove, this issue), and the differences are more subtle and perhaps even more interesting than they first appear. In the next section we cover a well-known unification, and with this unification can pinpoint the differences and make recommendations for researchers.

Before we do so, we should note that from a mathematical viewpoint, the Bayes factor approach cannot be assailed. The Bayes factors are the natural, direct, and unavoidable consequence of Bayes' rule. They are often critiqued because of the above robustness issue, but the logical consequences of these critiques are that either that one should not place beliefs on models or not use Bayes' rule for updating these beliefs. The estimation of posteriors is also mathematically unassailable as it too is the natural, direct, and unavoidable consequence of Bayes' rule. The critical issue is the step between estimation and using estimates to draw conclusions about the presence or absence of effects. These rules do not come from Bayes rule, and in this sense they may be considered heuristics.

Unification

We follow a classic line of unification that has been well recognized in the statistical literature since Jeffreys (1939). The differences between the estimation and Bayes factor approach can be understood by combining models $\mathcal{M}_0$ and $\mathcal{M}_1$. Figure 4A shows the combination, which is expressed as a mixture. One component of the mixture is the usual normal model on effect size (Model $\mathcal{M}_1$), and this component is denoted by the curve in Figure 4A.

[2] The Bayes factor between Model $\mathcal{M}_1$ and $\mathcal{M}_0$ is

$$B_{10} = \frac{1}{\sqrt{n\sigma_0^2 + 1}} \exp\left(\frac{n^2 d^2}{2(n + 1/\sigma_0^2)}\right) \qquad [4]$$

where d is the observed effect size given by $\bar{Y}/\sigma$.

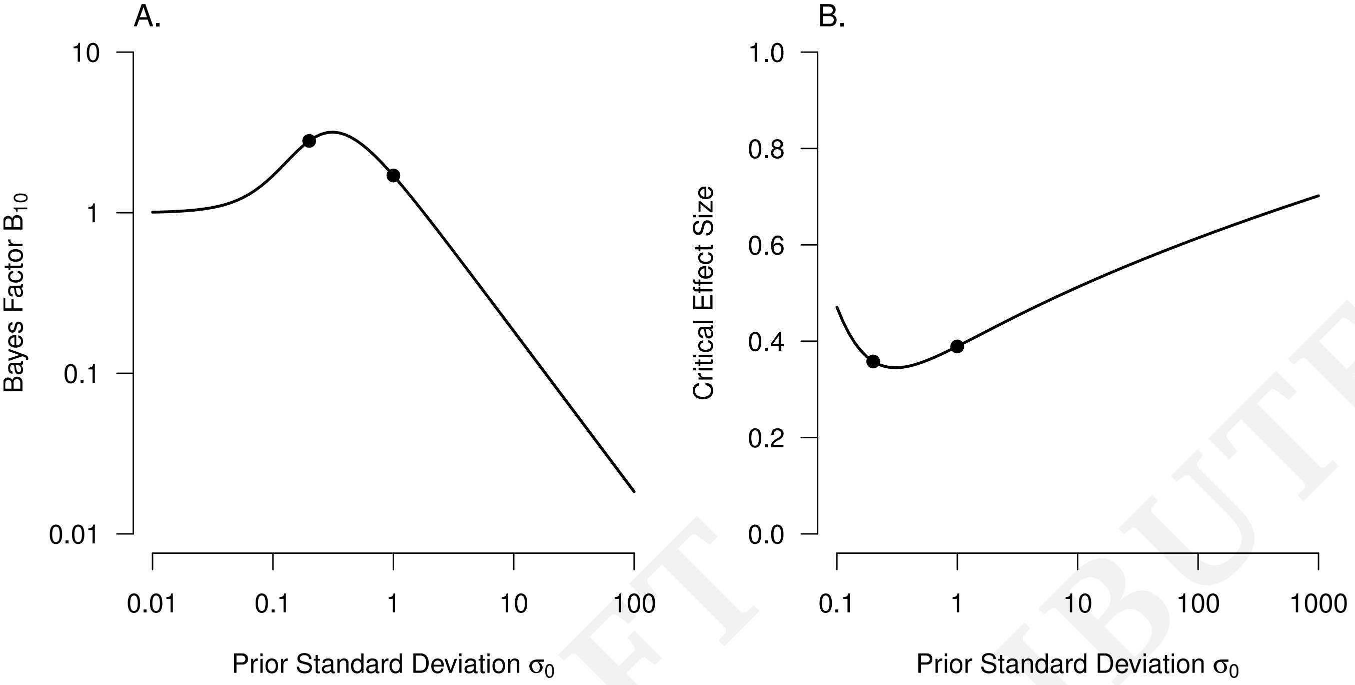

Fig. 3. The dependence of Bayes factor on prior setting σ_0. **A.** Bayes factor as a function of σ_0 for $N = 40$ and for an observed effect size of .35. **B.** Minimum observed effect sizes needed such that Bayes factor favors the alternative by 3-to-1. The filled circles show the lower and upper bounds of reasonable variation in prior standard deviation.

The other component is a placing mass on the point of zero, and this component is denoted by the arrow. In this case, the arrow is half-way up its scale, shown in dashed line, indicating that half of the total mass is placed at zero, and the other half is distributed around zero. This model is well known in the statistics literature as a *spike-and-slab model* (Mitchell & Beauchamp, 1988). We denote it by Model $\mathcal{M}_s$.[3] The spike-and-slab model in Figure 4 has two parameters: the amount of probability in the spike, denoted ρ_0, and the variance of the slab, denoted σ_0^2. Figure 4A shows the case where $\rho_0 = 1/2$ and $\sigma_0^2 = 1$.

It is straightforward to update beliefs about δ in the spike-and-slab model using Bayes' rule.[4] Figure 4B-C show a few examples for different observed effect sizes. In all cases, the resulting posterior is in the spike-and-slab form, but the spike has changed mass and the slab has shifted and rescaled. Figure 4B shows the posterior for a small observed effect size of 0.1. The spike is enhanced as the effect is compatible with a null effect. The slab is attenuated in mass, narrowed, and shifted form 0 to about .1. Figure 4B shows the posterior for a large observed effect size of 0.5. The spike is attenuated as the effect is no longer compatible with the null, and the slab is enhanced, narrowed, and shifted from 0 to about .5.

How shall we interpret the spike-and-slab specification? The spike-and-slab specification instantiates the case where the zero point is theoretically and qualitatively different from the other points. For instance, in the usual testing scenarios, researchers consider the "no-effect" baseline to be theoretically and qualitatively different than effects. The spike-and-slab model instantiates this qualitative difference, and, consequently, licenses the theoretically useful categories of "effect" and "no effect."

There are alternative interpretations that we find somewhat cumbersome. One is that the spike-and-slab can be viewed not as a model but as a model-averaging device. Here, the goal is not so much to define categories of effect and no-effect, but to average across both of them. Another alternative interpretation comes from Kruschke and Liddell (2017). Here, the spike and slab are seen as separate components in a hierarchical model. Accordingly, a focus on Bayes factors denotes a focus on the choice between components; a focus on posterior estimation entails parameter estimation *after* choosing the slab. We find this view difficult inasmuch as there is no *a priori* reason to choose the slab to focus on estimation. If one admits the possibility of the spike, then assuredly it should affect posterior estimation as well.

The spike-and-slab model is useful for examining whether posterior estimation is *always* robust to prior settings. In the previous slab-only model, the prior setting σ_0^2 played only a minimal role so long as σ_0^2 was somewhat large. Figure 4D-E show how two different prior settings, $\sigma_0^2 = 1$ and $\sigma_0^2 = 10$ affect parameter estimates. Here, there is an effect of the prior settings. As as prior variance is increased, more posterior mass is concentrated in the

[3] The density of a spike-and-slab model is given by

$$f(\delta) = \rho_0 s(\delta) + (1 - \rho_0)\phi(\delta/\sigma_0),$$

where s is the density of the spike, defined next, ϕ is the density of a standard normal, ρ_0 is the prior mass on the spike, and σ_0^2 is the variance of the slab. The density of the spike, s, is known as a Dirac delta function and defined as follows: Consider a normal density centered at zero with standard deviation η, denoted $g(\delta) = \phi(\delta/\eta)$. The Dirac delta function, s, is defined as the density in the limit that $\eta \to 0$:

$$s(\delta) = \lim_{\eta \to 0} \phi\left(\frac{\delta}{\eta}\right) = \begin{cases} \infty, & \delta = 0, \\ 0, & \text{otherwise.} \end{cases}$$

[4] The resulting posterior density, $f(\delta|\mathbf{Y})$ is

$$f(\delta|\mathbf{Y}) = \boldsymbol{\rho_1} \mathbf{s}(\delta) + (\mathbf{1} - \boldsymbol{\rho_1})\phi\left(\frac{\delta - \boldsymbol{\mu_1}}{\boldsymbol{\sigma_1}}\right),$$

where

$$\begin{aligned} \sigma_1^2 &= (n + \sigma_0^{-2})^{-1} \\ \mu_1 &= n d \sigma_1^2 \\ \rho_1 &= \frac{\rho_0}{\rho_0 + (1 - \rho_0)B_{01}}, \end{aligned}$$

where d is the observed effect size and B_{01} is the Bayes factor between Model $\mathcal{M}_0$ and $\mathcal{M}_1$.

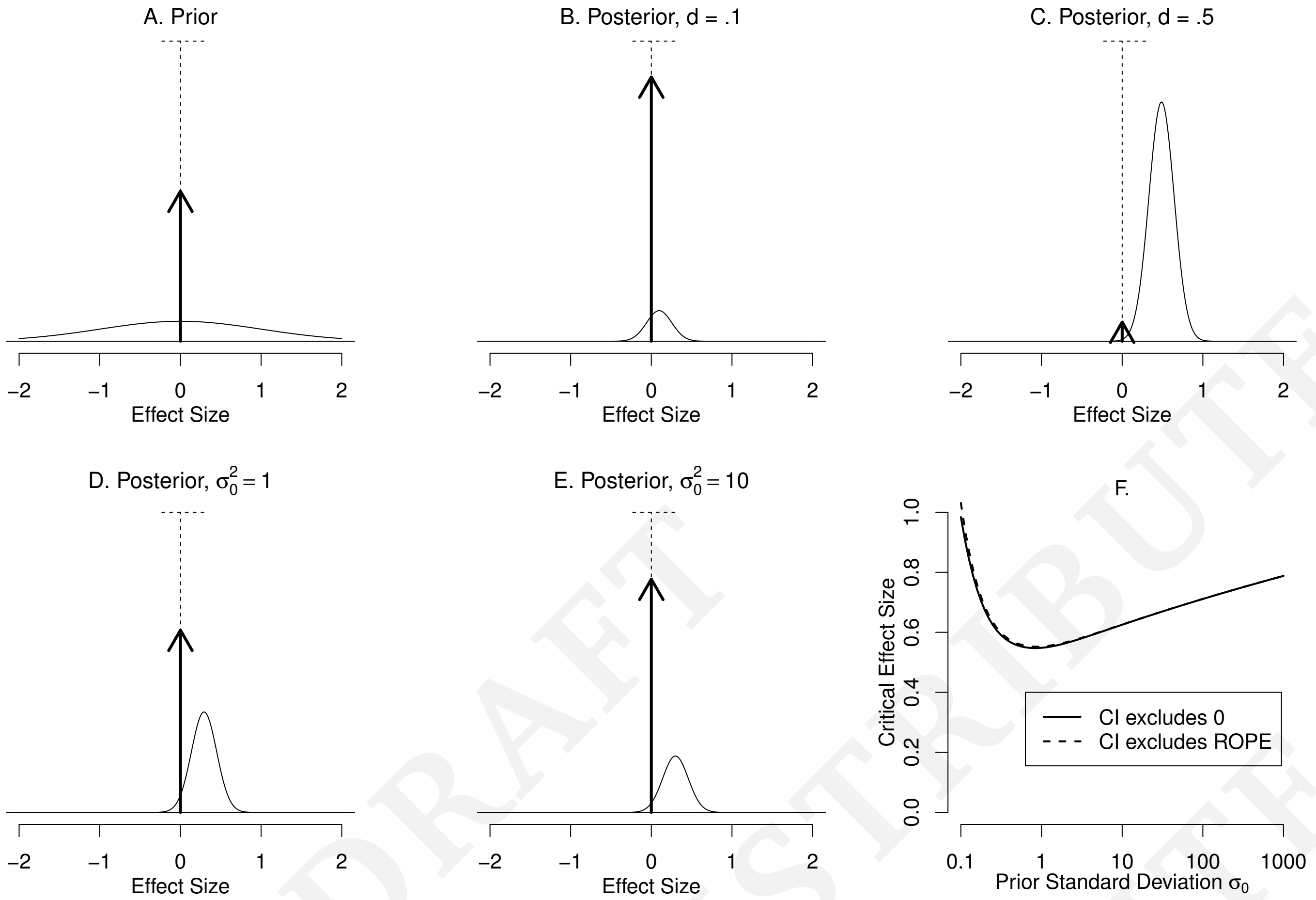

Fig. 4. The spike-and-slab model is a mixture of a spike, shown as an arrow, and slab, shown as the normal curve. **A.** Prior distribution on effect size with half the mass in the spike, and the slab centered around zero. **B-C.** The posterior on effect size δ for observed effect sizes of $d = .1$ and $d = .5$, respectively, for a sample size of 40. **D-E**. The posterior on effect size for prior slab variances of $\sigma_0^2 = 1$ and $\sigma_0^2 = 10$, respectively for $d = .3$. **F.** Critical values needed for stating an effect from posterior CIs as a function of prior slab variance σ_0^2.

spike. The estimate of ρ in particular is sensitive to prior settings. Figure 4F shows the critical effect size needed for inferential statements from posterior estimates. It is the analog of Figure 2B, but for the spike-and-slab model rather than for the slab model. As can be seen, inference with credible intervals, even with ROPES, depends on the prior variance setting. The dependency is similar to that for Bayes factor in Figure 3B.

This similarity in dependency is not too surprising because there is an intimate relationship between the spike-and-slab posterior distribution and the Bayes factor B_{01} for the comparison between models $\mathcal{M}_0$ and $\mathcal{M}_1$: The Bayes factor describes the change in the spike. The prior probability of the spike, ρ_0, can be expressed as odds, $\omega_0 = \rho_0/(1-\rho_0)$. The posterior probability of the spike, ρ_1, can likewise be expressed as odds, $\omega_1 = \rho_1/(1-\rho_1)$. The Bayes factor is the change in odds: ω_1/ω_0. In Figure 4B, for example, the initial odds on the spike were 1-to-1, indicating that equal mass was in the spike as was in the slab. In light of data, the posterior odds were 7.4-to-1, or that 88% of the posterior mass was in the spike and 12% of posterior mass was in the slab. Indeed, the Bayes factor for this case is $B_{01} = 7.4$, and this factor describes the change in odds in the spike in light of data (because originally they were 1-to-1).

One of the more interesting consequences of spike-and-slab models is that they display regularization without recourse to hierarchical structures. The solid curves are posterior means of δ as a function of observed effect size d. For the slab-only specification (Panel A), the estimated mean follows the observed value, and does so for all prior values of σ_0^2. But, for the spike-and-slab specification ($\mathcal{M}_s$, Panel B), there is a pull toward zero. In hierarchical models, this pull is known as *shrinkage*, and we borrow the term here. The degree of shrinkage from the spike-and-slab model is *adaptive* in that shrinkage toward zero is sizable for small observed values while there is hardly any shrinkage for large values. The dynamics are that small observed effect sizes are more compatible with the hypothesis that there is no effect, and therefore, estimates are more influenced by the zero value. Large effects in contrast are more compatible with the hypothesis that there is an effect, and the estimates are more influenced by the sample effect size. The amount of adaptive shrinkage depends on the prior setting σ_0^2. As σ_0^2 increases, there is more shrinkage to zero as the spike is relatively more salient. In this regard, the prior setting σ_0^2 serves as a tuning parameter. This adaptive shrinkage is very much like other regularization approaches such as lasso regression (Lehmann & Casella, 1998), and it protects researchers from overstating the importance of modest effects. In this case, if one believes *a priori* that zero is special, it should result in much conservatism for small effects.

The unification through spike-and-slab priors highlights similari-

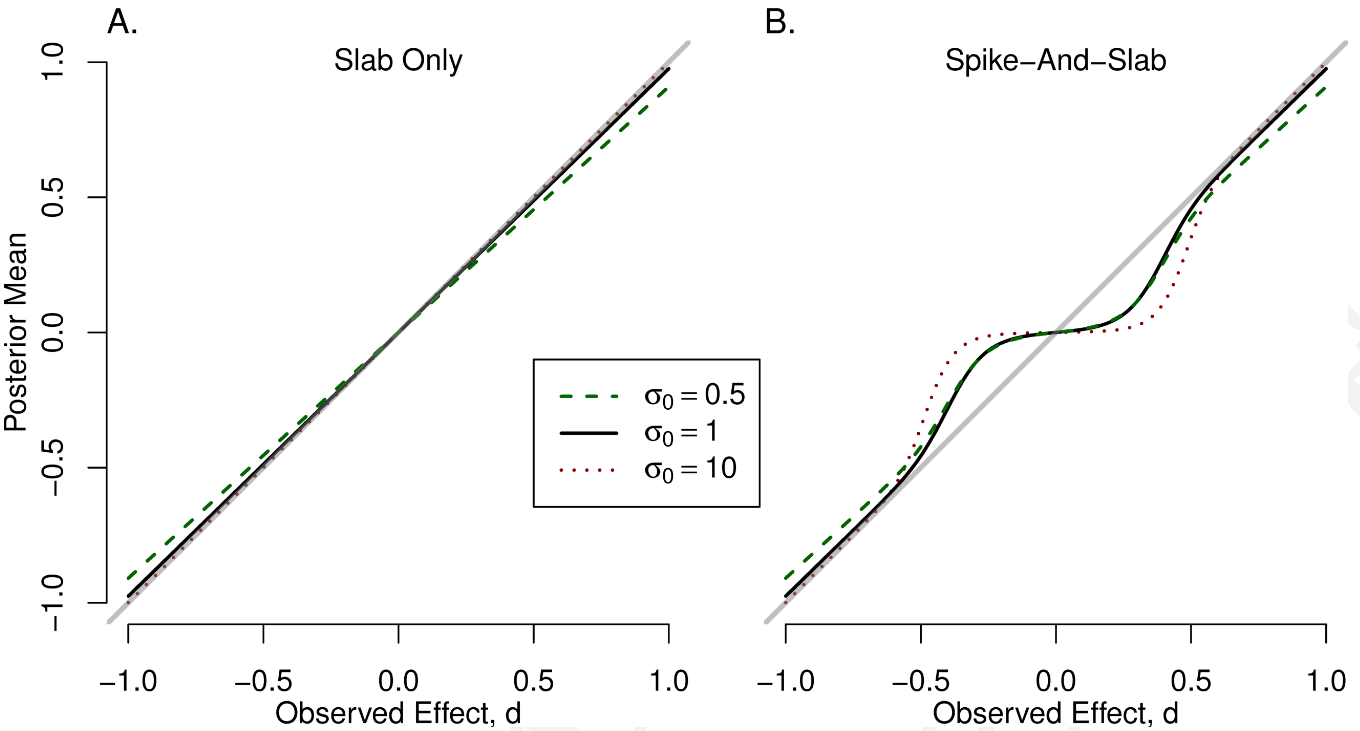

Fig. 5. A comparison of slab-only $(\mathcal{M}_1)$ and spike-and-slab $(\mathcal{M}_s)$ specifications for a moderate sample size of $N = 40$. **A-B**: Posterior mean of δ as a function of d for a few prior settings of σ_0^2. The light grey line is the diagonal, and the posterior mean of the slab-only model approaches this diagonal as the prior becomes more diffuse. The posterior mean in the spike-and-slab model shows *adaptive shrinkage* where small values observed values result in greatly attenuated estimates.

ties and differences between inference from posterior estimation and inference from Bayes factors as they are commonly used in psychology. The similarities are obvious, both methods are sibling approaches in the Bayes' rule family lineage. They rely similarly on specification of detailed models including models on parameters (priors), and updating follows naturally through Bayes' rule. There are differences as well, and the difference we highlight here is that in model specification. The recommended methods of inference by estimation, say those proposed by Kruschke, rely on priors that preclude spikes at set points such as points of invariance. The Bayes factor approaches we have developed in Guan and Vandekerckhove (2016), Rouder et al. (2009), Rouder and Morey (2012) and Rouder, Morey, Speckman, and Province (2012), place point-mass on prespecified, theoretically important values. It is this difference in model specification—rather than the difference in inferential statistic—that leads to some of the most salient practical differences between the Bayes factor and estimation approaches.

Which Model Specification To Use?

A critical question for researchers is then which model specification to use. The answer is that the choice depends on the context of the analysis and the goals of the researcher. As a rule of thumb, if zero is a theoretically meaningful or important quantity of interest it makes sense to consider a point mass on zero. The spike-and-slab model instantiates this qualitative difference, and in the process license the theoretically useful categories of "effect" and "no effect." In the context of this goal of *stating evidence for or against effects*, it is reasonable and judicious to use a spike-and-slab estimation approach.

In the past, we have justified the usage of models with point-mass at zero as corresponding to theoretically useful invariances (Rouder et al., 2009). In most sciences, for example, these invariances correspond to scientific laws or useful conservations. In psychology, however, such justifications may seem abstract. The current state-of-the-art is that we do not have many laws, invariances, or conservations to test (but see, e.g., Donkin, Newell, Kalish, Dunn, & Nosofsky, 2015, for exceptions).

Does the lack of scientific laws preclude the usage of Bayes factors in psychology? We think the theoretical coarseness of the field actually enhances the need for Bayes factors. In our field, there is usually little theoretical consideration of the metric size of phenomena. For example, we know of no theory of Stroop interference that anticipates whether effects will be 20 ms or 200 ms even though these values vary by a factor of 10! Psychologists gain theoretical specificity by exploring what factors affect phenomena and what do not. For example, the size of Stroop interference is affected by the proportion of congruent and incongruent items (Logan & Zbrodoff, 1979); it is not affected by tonic levels of arousal (MacLeod, 1991). To account for the congruency proportion effect, we may posit that people maintain low-level expectations about the stimulus to guide processing, but such a theory does not predict the amount of the effect. Indeed, our theories, which tend to be verbal, rarely anticipate or are challenged by metric-level data. In these common cases, having a statistical model that instantiates the categorical difference between null effects and effects makes sense. The slab models where points remain undistinguished cannot provide for this categorical difference.

Specifications that capture this categorical difference are not limited to spikes. For example, we could declare a region of equivalence, say an interval $-.1 < \delta < .1$, which is theoretically and qualitatively different than values outside this interval. The model specification then becomes a mixture of this region with wider slabs, and its density may be given by $f(\delta) = \rho u(\delta) + (1-\rho)\phi\left(\frac{\delta-\mu}{\sigma}\right)$, where u is a density over the equivalence region. Here, the key is that we assign a probability to each component, and it is that this probability is updated in light of data that provides for the categorical difference. Morey and Rouder (2011) develop Bayes factors for this equivalence region approach and compare it to conventional equivalence testing. Rouder (2016) provides R code in his archived blog post, "Roll Your Own: How to Compute Bayes Factors for Your Priors."

In some cases—perhaps ones where measurement is a main goal and where the zero value has no special meaning—a slab-only approach may be best. Researchers in these measurement contexts, however, should avoid drawing inferences about whether or not there are effects in the data as the model specification does not capture such categorical difference and Bayes' rule does not provide a reallocated probability for either proposition. Any inference about the existence of an effect in the slab-only model should be treated as an informal heuristic not tied directly to Bayes' rule.

There will be some differences among researchers as to which specification is best in any given context. These differences should be welcomed as they are part of the richness of adding value in psychological science (Rouder, Morey, & Wagenmakers, 2016). In all cases, however, researchers should justify their choices in the context of the goals.

Which Approach To Use

While model specification is a critical difference between posterior estimation approaches and Bayes factor approaches, there are other differences as well. Bayes factors are updating factors. They describe how data lead to a revision of belief, and in this sense, they are a measure of the relative strength of evidence from the data for competing propositions. Following Jeffreys (1961) and many others, we find them ideal for scientific communication. Researchers reporting Bayes factors are providing a description of evidence. Bayes factors may be understood naturally as odds without recourse to further qualification. In particular, it is not necessary nor helpful to decide if a Bayes factor is sufficiently big just as it is not necessary to have a criterion for "big" odds. Refraining from making decisions strikes us as advantageous in most contexts. If researchers feel compelled to make a decision, however, then they may be guided by Bayes rule. Accordingly, the posterior quantities become important, and these are combined with loss or utility considerations in making decisions (Savage, 1972). For example, if one wishes to decide whether there is or is not an effect, then ρ_1 the posterior probability of being in the slab, becomes the critical quantity. Criteria that reflect the expected utility of the resulting decision may be placed on this quantity. Hence, Bayes factors serve as a direct evidence measure that may be combined with prior odds and utility considerations in statistical decision making.

Posterior estimation always remains useful in reporting the results of analyses. Posterior means give an overall indication of where we think values are best localized, and posterior intervals describe the precision of this localization. Posterior means and CIs may provide valuable graphical displays, especially for researchers who are used to confidence intervals. Moreover, these analyses are not exclusive; one may graph a posterior and report a Bayes factor as evidence, combine it with prior odds and a loss function to reach a decision. From our vantage point, however, the Bayes factor remains primary as it is most intimately tied to the data—it is the evidence from the data for competing, theoretically relevant positions. We do not endorse using posterior intervals for stating evidence or reaching decisions about competing theoretical positions such as whether or not there is an effect. Instead, these posterior intervals are best used as providing overall summaries of posterior distributions.

Prior Dependency

Researchers who consider Bayes factors may worry about their dependence on prior settings especially when compared to estimation with slab-only models. This worry is assuredly overstated, and a bit of common sense provides for a lot of constraint. It seems to us unreasonable to consider prior settings that are too small or too large as researchers generally know that true effect sizes in psychological experiments are neither arbitrarily small or large. A lower limit of σ_0 is perhaps 0.2 as researchers rarely search for effect sizes smaller than this value and the practical value of such small effects will often be low.[5] Likewise, an upper limit is perhaps 1.0 as the vast majority of effect sizes are certainly smaller than this value and effects much larger than that would often be clear in day-to-day life. Within these reasonable—if context-dependent—limits, Bayes factors vary but not arbitrarily so. We have highlighted the Bayes factor values associated with these limits in Figure 3A as filled circles. Here the Bayes factors differ from 1.7 to 2.8 or by about 40%. This variation is not too substantial—the Bayes factor is reasonably robust—and in both cases the evidence for an effect is marginal. Such variation strikes us as entirely reasonable and part-and-parcel of the normal variation in research (Rouder et al., 2016). It is certainly less than other accepted sources such as variation in stimuli, operationalizations, paradigms, subjects, interpretations and the like.

The Potential of Spike-And-Slab Models In Psychology

Spike-and-slab models are currently timely and topical in the statistics literature. We think they gain popularity in the psychological sciences as psychologists adopt new analytic techniques, especially in big-data applications. Consider applications in imaging where there are a great many voxels or in behavioral genetics where there are a great many nucleotide markers in a SNP array. It is desirable to consider the activity in any one voxel or the contribution to behavior of any one marker, and the resulting models necessarily have a large numbers of parameters, say with one parameter for each voxel or each marker. It is in this context, when there are large numbers of parameters relative to the sample size, that spike-and-slab priors have become popular. The seminal article for assessing covariates in this context is George and McCulloch (1993), and recent conceptual and computational advances, say from Scott and Berger (2010) and Ročková and George (2014), make the approach feasible in increasing large big-data contexts.

In these big-data contexts, the spike-and-slab specification captures the position that many elements will not contribute to the outcome. For example, in the fMRI case, many voxels will have no task-related activity; in the genetics case, many markers will be unrelated to behavior. A spike-and-slab specification is placed on each voxel or marker, and this specification acts as a means of categorization. If the posterior value of spike mass falls below a set criterion, then these voxels are markers may be retained in the final model. Otherwise, the voxel or marker is not retained.

[5] Which is not to say that small effects cannot be *theoretically* meaningful in certain contexts, but we believe interest in very small effects to be generally low.

As an example of big-data applications in psychology, we highlight the recent work of Sanyal and Ferreira (2012) who used spike-and-slab priors for fMRI analysis. These researchers sought to enhance the spatial precision of imaging by improving the spatial smoothing. Typically, researchers smooth the image by passing a Gaussian filter over it. Sanyal and Ferreira instead performed a wavelet decomposition where activation is represented as having a location and a resolution. In this approach there is a separated wavelet coefficient for each resolution and location pairing, and the upshot is a proliferation of coefficients. Sanyal and Ferreira placed a spike-and-slab prior on these coefficients, and used large values of ρ_0, the prior probability that a coefficient is zero. In analysis, the posterior for many of these coefficients remained dominated by the spike, and could be removed. When the activation was reconstructed from only the coefficients for which there was substantial mass from the slab, the image quality improved from the elimination of mostly high-frequency components. The resulting smoothing was spatially adaptive—it was more smooth where activation was spatially homogenous (say within structures) and less smooth where activation was spatially heterogeneous (say at boundaries).

Conclusions

In this paper we discuss a well-known, classic unification between two competing Bayesian approaches—that based on the estimation of posterior intervals and that based on Bayes factors. A salient difference between these two approaches is model specification. It is common in estimation approaches to place broad priors over parameters that give no special credence to a zero point. Common Bayes factor approaches, such as that from Rouder and Morey and colleagues (Rouder et al., 2009; Rouder & Morey, 2012; Rouder et al., 2012; Guan & Vandekerckhove, 2016) are closely related to estimation with a prior that has some point mass at zero. Which model specification a researcher should choose, whether a broad slab or a spike-and-slab, should depend on the context and goals of the analyst. For the usual case where researchers are interested in whether there is or effect or not, the categorical differences provided by point masses that reallocate in light of data is appropriate and useful, while common slab-only specifications do not provide this facility.

References

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)*. Hayward, CA: Institute of Mathematical Statistics.

de Finetti, B. (1974). *Theory of probability* (Vol. 1). New York: John Wiley and Sons.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Quantitative Psychology and Assessment*. Retrieved from 10.3389/fpsyg.2014.00781

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., & Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: [a] case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 933.

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

Etz, A., & Vandekerckhove, J. (in press). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*. Retrieved from https://osf.io/preprints/psyarxiv/q46q3

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

Gelman, A., & Carlin, J. (2017). *Some natural solutions to the p-value communication problem—and why they won't work.*

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

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

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance*. London: Cambridge.

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin and Review*, *23*(1), 74–86. Retrieved from http://www.cidlab.com/prints/guan2015bayesian.pdf

Jeffreys, H. (1938). *Theory of probability*. Oxford: Claredon.

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

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

Kruschke, J. K. (2011b). *Doing Bayesian analysis: A tutorial with R and BUGS*. Academic Press.

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., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, *88*, 1242-1249.

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

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187-192.

Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian point of view, part 2: Inference*. Cambridge, England: Cambridge University Press.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *56*, 362-375. Retrieved from http://dx.doi.org/10.1016/j.jmp.2008.03.002

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.

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

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., 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. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406-419. Retrieved from http://dx.doi.org/10.1037/a0024377

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111-163.

Rouder, J. N. (2016). *Roll your own: How to compute bayes factors for your priors.* Retrieved from https://osf.io/preprints/psyarxiv/nvsm5 (Archived blog posts)

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

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

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

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

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

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*, 587-604.

Ročková, V., & George, E. L. (2014). EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association*, *109*.

Sanyal, N., & Ferreira, M. A. R. (2012). Bayesian hierarchical multi-subject multiscale analysis of functional MRI data. *Neuroimage*, *63*, 1519-1531.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed. ed.). New York: Dover.

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*, 2587-2619.

Senn, S. (2011). You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets and Morals*, *2*, 48-66.

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response time. *Psychological Methods*, *16*, 44-62.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, *14*, 779-804.