Beyond Overall Effects: A Bayesian Approach to Finding Constraints In Meta-Analysis

Jeffrey N. Rouder[1], Julia M. Haaf[2], Clintin Davis-Stober[2], & Joseph Hilgard[3]

[1] University of California, Irvine

[2] University of Missouri

[3] Illinois State University

Author Note

This paper is developed in RMarkdown with integrated text and code for analysis and figures. An executable source file that downloads the data, performs all analyses, and typesets the manuscript may be found at github.com/PerceptionAndCognitionLab/meta-planned. Version 2.

Correspondence concerning this article should be addressed to Jeffrey N. Rouder, Social and Behavioral Science Gateway, Irvine, CA. E-mail: jrouder@uci.edu

Abstract

Most meta-analyses focus on the behavior of meta-analytic means. In many cases, however, this mean is difficult to defend as a construct because the underlying distribution of studies reflects many factors including how we as researchers choose to design studies. We present an alternative goal for meta-analysis. The analyst may ask about relations that are stable across all the studies. In a typical meta-analysis, there is a hypothesized direction (e.g., that violent video games increase, rather than decrease, agressive behavior). We ask whether all studies in a meta-analysis have true effects in the hypothesized direction. If so, this is an example of a stable relation across all the studies. We propose four models: (i) all studies are truly null; (ii) all studies share a single true nonzero effect; (iii) studies differ, but all true effects are in the same direction; and (iv) some study effects are truly positive while others are truly negative. We develop Bayes factor model comparison for these models and apply them to four extant meta-analyses to show their usefulness.

Keywords: meta-analysis, random-effects meta-analysis, Bayesian models, mixed models, order-constrained inference

Beyond Overall Effects: A Bayesian Approach to Finding Constraints In Meta-Analysis

Most readers at some point have considered the validity of averages. Sometimes, we may have been asked, "What does this average mean?" or, "Who does this average describe?" We may be quick to to gloss over such questions. After all, the average, or sample mean, is a natural measure of the central tendency, and central tendency holds a privileged place in understanding variability.

Yet answers based on naturalness and privilege can be unsatisfying. A better account of the average comes from modeling. A model is an abstract, platonic account that has an irreducible element of uncertainty. According to the model, observations come from a distribution that captures this uncertainty. Part of our goal as analysts is to characterize this distribution. Unlike the data, which are real, the distribution is itself an abstraction (de Finetti, 1974). We often say observations are samples from a distribution. The reader, however, should keep in mind that this saying is somewhat misleading: Although observations are real, the concept of a sample from a distribution is an abstraction.

We may profitably view the sample mean in this context. The sample mean is useful in characterizing this abstract distribution. If we go further and assume that the observations come from a common distribution, say the normal distribution, the sample mean serves as an estimator of another abstraction, a parameter. For the normal, this parameter is called the true mean. Although we use the term "true," we should be careful to remember that it is not a real quantity—rather, it is a mathematical abstraction.[1] Even though parameters are abstractions rather than real, they are nonetheless useful in understanding constraint in data. In this regard, the sample mean is validated as an estimator of a theoretically meaningful parameter in a model.

One area where this validation may be questioned, however, is meta-analysis. The

[1]Some people use the term population mean rather than true mean. The population, as commonly used, is an abstraction. For example, the population of all people is an abstract concept not dependent on who is currently alive. Of course, there can be concrete populations, say the population of the first 44 U.S. presidents, but the typical usage in psychology is for abstract rather than concrete populations.

usual goal of meta-analysis is to combine several similar studies to draw a common conclusion. This conclusion almost always centers on a grand mean or overall effect. Take, for example, the meta-analysis of Anderson et al. (2010). After an extensive review, these authors concluded the meta-analytic average of the link between violent video game exposure and subsequent aggressive behavior was $r = .21$. Yet, to interpret this meta-analytic mean, we need to posit a distribution over experiments and treat this average as an estimate of a true parameter. What does this distribution signify? The distribution surely has something to do violent-video game exposure and aggression, but it also has something to do with how we as a community design, run, and select experiments. Thus, the concept of a meta-analytic mean must be treated with care.

One way of providing this care is to consider the differences between *metric* and *ordinal* properties of the distribution. The metric properties are the usual real-valued parameters that describe the exact location of the probability mass, including the mean, variance, quantiles, and moments. For the Anderson et al. meta-analysis, the value $r = .21$ is a metric property describing the central tendency of the distribution of effect sizes across studies.

Ordinal properties, in contrast, are about orderings. We ask whether basic ordering relations hold across all studies. To start, we note that there is almost always an anticipated direction of relations or effects in meta-analyses. For example, if there is an effect of violent video games on aggressive behavior, theory predicts that such violent video games are positively associated, rather than negatively associated, with aggression. We call this anticipated direction the positive direction. With this emphasis on direction, effects may be classified as positive, null, or negative.

We ask whether all studies in a population of studies have the same ordinal properties. For example, we may expect that if there is a positive effect between aggression and violent video games, all studies with competent methods and measurements will have a true positive effect. This is not to say that every study will yield a positive sample effect, as some negative sample effects are expected from sample noise. Once this noise is modeled, however,

the resulting parameters may be called *true effects*. They denote the noise-free or population value of the study, and we will use the terms *true* and *truly* throughout to refer to these values as not to confuse them with sample or observed values. The constraint is whether all true effects across a class of studies are positive. This all-studies-positive constraint, if it holds, is a strong statement. It means that every experiment in the class has a true positive effect. It is stronger than the usual meta-analytic statement about the averages because it applies to all studies. Likewise, we can also define a strong null constraint—all studies in the class show a true null effect. This null is stronger than the usual null that the average across studies is zero as the average may be zero while constituent studies may be truly positive and negative. Importantly, these ordinal properties are easier to interpret than metric properties because they are less dependent on design choices.

Of course, it may not be that all studies in a corpus have true positive effects or that all have a true null effects or that all have a true negative effects. Perhaps some studies in the meta-analysis show a true positive effect while others show a true negative effect. This case, should it exist, motivates different considerations. If there is a mix of true positive and true negative effects across studies, it may indicate that the individual studies are measuring disparate phenomena or are confounded by some moderator powerful enough to change the sign of the true effect. In this case, researchers may want to study why some effects are truly positive and others are truly negative.

We believe this focus on ordinal properties that are common across all studies matches well with the type of questions researchers are interested in. Do all studies show a true effect in the same direction? Do all studies show a true null effect? Is the effect so heterogeneous that some studies have a true positive and others have a true negative effect? The meta-analytic mean, while convenient, is not helpful in answering these questions. Our goal here is to develop models that account for both meta-analytic metric properties like the mean and variance and ordinal constraints.

It is important to note that the focus on ordinal constraints in meta-analysis is new. It

represents a new set of questions that are different from the usual ones where the meta-analytic mean is the focus. It is also to note that it necessitates new statistical analysis. The usual approach of estimating or testing the mean and variability of study effects is quite different from asking say if all are positive. Certainly, if there is small heterogeneity and a large mean, it is highly likely that all study effects are truly positive, and conversely, if the mean is near zero and the heterogeneity is large, then it is likely that some effects are truly negative and others are truly positive. But for the majority of cases, where there is a moderate mean and some heterogeneity, it is impossible to answer the question of whether all studies show a true positive effect consideration of just the meta-analytic mean and meta-analytic variance.

Although this shift from means to ordinal constraints in meta-analysis is novel, the concept of an ordinal constraint itself is not new. Indeed, significance tests are tests of ordinal constraints on true means. Whereas classical tests are focused on a single order constraint—that of the grand mean, a slope, or a variance—the needed tests here are about whether many order constraints hold simultaneously. There is a classical literature on order constraints and the topic is conceptually complicated (Robertson, Wright, & Dykstra, 1988; Silvapulle & Sen, 2011). To our knowledge, there is no classical solution to the "does every study show a true positive effect" hypothesis in a hierarchical context appropriate for meta-analysis.

Although the problem appears difficult for classical testing, it is straightforward in the Bayesian framework. Bayesian analysis has become popular in part because it makes difficult statistical problems straightforward. Assessing multiple order constraints simultaneously follows fairly readily from Bayes rule (Gelfand, Smith, & Lee, 1992; Klugkist, Laudy, & Hoijtink, 2005). There are many reasons to adopt Bayesian analysis; in this case none is more important than it is the only analysis we know that provides a solution to the "does every study" problem.

In this paper we develop Bayesian meta-analysis with a focus on ordinal constraints.

We apply the analysis to four extant meta-analyses, and in the process illustrate a variety of patterns in the literature. One constraint we document is a strong null where all studies have a zero-valued true effect. We find this constraint holds in a reanalysis of Wagenmakers et al. (2016). These authors performed a registered replication of Strack, Martin, & Stepper (1988), who demonstrated a well-cited instance of embodied cognition. A second constraint we document is one where all studies in the meta-analysis are best described as having one true effect. We document this common effect with a reanalysis of a set of studies from Ebersole et al. (2016). These authors replicated a social-psychological phenomenon called moral credentialism (Monin & Miller, 2001) where prejudice is expressed to a greater degree after participants reject overtly sexist statements. A third constraint we document is one where true effects may differ but all are positive. This case comes from a reanalysis of Haaf & Rouder (2017), who reported the results of three extant Stroop experiments. Finally, we document variability across sites. This is demonstrated by a reanalysis of Corker, Donnellan, Kim, Schwartz, & Zamboanga (2017), who studied how Big Five personality characteristics varied across different universities.

Constraints Among True Effects

Our main goal is to focus on constraints among the constituent experiments themselves. We first illustrate this focus with a reanalysis of Wagenmakers et al. (2016). Participants were asked to rate how humorous cartoons were while either smiling or pouting. Strack et al. (1988) reported a sizable effect where participants rated the cartoons as more humorous when smiling than when pouting. Wagenmakers et al.'s replication set is comprised of data from 17 independent lab sites who each performed the exact same experiment.

We explicitly model the variability within and between the studies with an ordinary mixed linear model. Let Y_{ijk} denote the rating from the ith site, the jth condition, and the kth participant. For example in the Wagenmakers' set, there are $I = 17$ studies, two conditions (pout and smile, $j = 1, 2$, respectively), and about 60 replicates per study per

condition. The base model is

$$Y_{ijk} = \mu + \alpha_i + x_j\theta_i + \epsilon_{ijk}.$$

Here, μ is a grand mean and α_i is an overall study-specific effect. Studies with higher ratings on average will have greater values of α_i. In this regard, α_i is a study-specific *intercept* parameter. The design element x_j is a condition indicator with $x_j = -1/2$ and $x_j = 1/2$ for pout and smile conditions, respectively. The parameter θ_i is the study-specific effect of the pout/smile manipulation, and it is the main target of inquiry. These parameters may be thought of as study-specific *slopes* as they describe the study-specific change in performance as a function of the manipulation.

The term ϵ_{ijk} is a homogeneous noise term,$\epsilon_{ijk} \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$. Note that this homogeneity-of-variance assumption is quite strong. A more relaxed and traditional treatment would be to allow the true variability to depend on the study, $\epsilon_{ijk} \overset{ind}{\sim} \text{Normal}(0, \sigma_i^2)$, where σ_i^2 is study-dependent variation. In this paper, we retain the homogeneity-in-variance specification because it simplifies the development and analysis. Homogeneity is appropriate for the current data sets where the studies are quite similar (the same base study was replicated at several sites). The limitations from this assumption are discussed more broadly in the General Discussion.

Our critical questions are about θ_i, the effect of the smile/pout manipulation. To address these questions, we place a series of models on θ_i that capture various constraints:

The most constrained model is the *null model*. Here, all of the constituent studies have a true effect of zero, and this model is implemented with the constraint $\theta_i = 0$. Note that this model is a much stronger null than the usual meta-analytic null where the true grand average is zero; here, both the average and variance of θ_i are zero. Figure 1, left column, provides a graphical representation of the models. Panel A is for the null model. Depicted is the specification of the effect θ_i for two studies. Since θ_i is zero for both studies, the only

point with mass is at $(0, 0)$.

The next generalization is what we term a *common effect model*. All constituent studies have the same true value, denoted ν. The constraint is simply $\theta_i = \nu$. This common-effect model captures the assumption of homogeneity, and it is sometimes called a fixed-effect meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2010). Figure 1B depicts this model. Because of the equality constraint, there is mass only on the main diagonal. If we further constrain ν to be positive, then there is only mass on positive values of this diagonal.

A generalization of the common effect model is to allow the true effects to vary from study-to-study, but to stipulate they have the same direction. For example, it is reasonable to assume that if facial expression affects humor ratings, all studies would show, to some degree, more humorous ratings while smiling than while pouting. Although that degree may vary, and certain factors may lead to smaller effects (less-sensitive measurements, less-susceptible populations, noisier methodology), no study would have a truly negative effect where pouting led to truly higher humor ratings. We call this model the *positive effects* model, and with it we constraint $\theta_i > 0$. For example,

$$\theta_i \sim \text{Normal}_+(\mu_\theta, \sigma^2_\theta),$$

where Normal_+ denotes a normal distribution truncated below at zero. Here μ_θ and σ^2_θ are population parameters that describe the distribution of effect sizes across studies. Figure 1C shows this model. There is only mass distributed across the quadrant of joint positive effects. Prior settings are needed for μ_θ and σ^2_θ, and we discuss how we chose these and the effects of these choices on inference subsequently.

Finally, we can relax this positivity constraint:

$$\theta_i \sim \text{Normal}(\mu_\theta, \sigma^2_\theta).$$

This model is termed the *unconstrained model*, and it specifies that true effects may be

positive or negative. Figure 1D shows this model, and there is mass across all values of joint effects across participants.

Some readers might be a tad confused that we previously critiqued meta-analytic averages and yet still posit parameter μ_θ, which is the meta-analytic average across studies. The difference, however, is that our focus remains on the collection of θ_i's and not on μ_θ or σ^2_θ. In this sense, the experiment-population parameters serve as auxiliary parameters that improve our estimates of the θ_i's.

These four models provide a means of characterizing the ordinal constraint in data. For example, if the null model best describes the data, then the conclusion is that there is no effect for any study. Likewise if the common-effect model best describes the data, we can talk about a unified phenomenon, and, here, the mean becomes meaningful as it characterizes all effect sizes. If the positive model best describes the data, we may note that while there is variation across studies, all index the same basic ordinal relation. It is this relation that is the main constraint in the data. Finally, if the unconstrained model best describes the data, resulting conclusions are nuanced. Perhaps the most prudent course is to wonder about the coherence of the collection of studies—they may index disparate phenomena.

How do these four model compare to more traditional meta-analytic models? Let's take the case of a researchers asking if there is an effect across a corpus of studies. The typical random-effects meta-analysis corresponds to a generalizatoin of our unconstrained model. At the first level, it is typically assumed the residual variability on obervations varies across studies. In our development, in contrast, a homogeneity-of-variances assumption is used. If we focus on the second level, the typical random effects model is our unconstrained model, and there are no constraints on θ_i other than they are draws from a normal parent distribution The more subtle difference is the test of whether there are effects. In traditional meta-analysis, this test is performed by compariong the effects model to a null model where the grand effect is zero, e.g.,

$$\theta_i \sim \text{Normal}(0, \sigma^2_\theta).$$

Note that this model is not the null model we advocate above. Our null, θ_0 for all studies is much stronger, and we chose it purposefully. The traditional null specifies that each each study has a different true effect, but the mean is zero. Consequently, exactly half are positive and half are negative. Had we collected an unlimited amount of data from all studies, we could observe this perfect centering, say with half studies showing video-game violence leads to increased aggreesion and half showing it leads to decreased aggression. This type of model strikes us as implausible, and we have no desire to interpret it. The stronger null is that no study shows an effect, that is, video-game violence does not affect subsequent aggression. This is an interpretable proposition. So, in summary, there are two main points of departure from the traditional framework: 1. the imposition of a homogeneity-of-variance assumption (which is not problematic here) and 2. the specification of a stronger, far more realistic null model.

Although we carry the above four models into the subsequent analyses, they are not the only possible choices. There are some recent trends in modeling that we have chosen not to follow. One of these is the use of equivalence regions instead of sharp point nulls (Rogers, Howard, & Vessey, 1993; Tryon, 2001). An equivalence region is a small region around a zero point that serves as a practical null. Another modern trend is the use of mixtures of latent classes (Bishop, Fineberg, & Holland, 1975). After developing analyses with these four models, we will revisit our choices in light of other modeling trends.

One critical question is which model of the four best describes the obtained data. We address this question in subsequent sections. Before doing so, we take a brief detour to discuss estimation of effects. Although estimation does not provide a calibrated, formal means of assessing the aforementioned constraints, it certainly provides an appropriate informal visualization of these constraints, and therefore, obtaining principled estimates of study effects is consequential.

Estimating True Effects

A standard course to visualizing effects is a *forest plot*, which summarizes the sample effect and corresponding confidence interval for each study or site. Figure 3A is an example from Wagenmakers et al., and it is quite similar to Wagenmakers et al.'s Figure 4. We find that forest plots place too much emphasis on the sample means, which, in the case of meta-analysis, are poor estimates of the true mean.

A sample mean estimate for a certain study relies on the data from that certain study and not on the data from the other studies in the meta-analysis. At first glance, this property may seen reasonable, but since the 1960s, statisticians have known that the data in the other studies may be used to improve the estimate of a particular study's true effect size (Efron & Morris, 1977; Stein, 1956). The estimator for one study's effect size should depend on the data from that study and from the other studies as well. This approach is now standard in hierarchical modeling.

We use the unconstrained model, defined above, for estimating true study effects.[2] The unconstrained model is a typical mixed linear model, and models of this type are exceedingly popular Bayesian (Gelman, Carlin, Stern, & Rubin, 2004; Jackman, 2009) and frequentist contexts (**???**). This choice is well suited for estimating a single, best value for each study. Figure 3B shows the hierarchical estimates from the unconstrained model (filled circles) for Wagenmakers et al.'s data along with associated 95% credible intervals. Notice that the hierarchical estimates are more compact, closer to the meta-analytic average, and have credible intervals that are smaller and more uniform than for the sample mean and CIs. This effect is called regularization, and the notion here is that once the within-study variability is accounted for, the resulting model estimates better show the true variation across studies. Regularization is a feature of frequentist and Bayesian mixed models, and similar results are

[2]We use conjugate priors so that estimation may proceed through Gibbs sampling, and our setup is documented in Rouder, Morey, Speckman, & Province (2012) and Haaf & Rouder (2017). Estimation is robust to prior settings in the sets we examine. Where prior settings matter most is for Bayes factor computations, and these effects are discussed subsequently.

obtained in frequentist packages such as LMER (cite). For Wagenmaker's et al. data, there is a large degree of regularization. Sample mean estimates are about 3 times as variable as the hierarchical estimates. This degree of regularization is meaningful and substantial, and it needs to be conveyed to readers.

Plots based on sample effects always overstate the variability across the sites. To avoid this overstatement, regularization, whether from frequentist or Bayesian methods, should be used. Sample effects may be plotted, but they should serve as data rather than as a target of inference. Consequently, confidence and credible intervals should be placed on regularized estimates rather than sample means, and Figure 3B provides an example of such a plot. Fortunately, most researchers are familiar with modern estimation and mixed linear models, and their usage is built into meta-analytic software packages such as `metafor` (Veichtbauer, 2010) and Comprehensive Meta-Analysis.

Evidence for Constraints

Previously, we discussed four theoretically-motivated models of constraint: the null model, the common effect model, the positive effects model, and the unconstrained model. Estimating true values for effects is useful in visualizing data, but it provides no direct and calibrated measure of the evidence for the four models. To provide principled measures of evidence, we use Bayes factors. 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, Haaf, & Aust (2017). Informally, evidence for models reflects how well they predict data.

The right column of Figure 1 shows the predictions for each of the four models. We consider the relationship between two hypothetical studies that yield sample effects, $\hat{\theta}_1$ and $\hat{\theta}_2$. Possible sample effects for the first experiment are plotted on the x-axis, and possible values for the second experiment are plotted on the y-axis. Each point in the figure

represents a possible combination of observed effects. For the null model, the observed effects are predicted to be near (0,0), and this case is shown in Figure 1E. The effect of sampling error is to smear the form of the model.[3] Figures 1F-H show the predictions for the common effect, positive effects, and unconstrained models, respectively. One aspect of the predictions that is not obvious is the correlation for the positive and unconstrained models. This correlation comes from the hierarchical structure in these models, and is a direct result of variability of the population mean μ_θ. This variability comes from the prior and is discussed further after presenting the applications.

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 denote a hypothetical observed sample effect for both studies. This value is about equal for both studies, and we might suspect that the common effect model does well. To measure how well, we note the density of the prediction. Here, the density is darkest for the common effect model. These densities have numeric values, and we may take the ratio to describe the relative evidence for one model vs. another. For example, the best fitting model in the figure, the common effect model, has a density that is twice the value of that of the positive effects model. Hence, the data are predicted twice as accurately under the common effect model than under the positive effects model. This ratio is the *Bayes factor*, and it serves as the principled measure of evidence for comparing one model to another in the Bayesian framework.

Bayes factors are conceptually straightforward—one simply computes the predictive densities at the observed data. While this computation is conceptually straightforward, it is often inconvenient in practice. The computation entails the integration of a multidimensional integral which is often impossible in closed form and may be slow or inaccurate with numeric methods. To that end, there has been a voluminous literature on how to compute these integrals in mixed settings such as the one here. We follow a fairly general set of specification and computations known to work well. These have been pioneered by Zellner & Siow (1980)

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

and expanded for ANOVA by Rouder et al. (2012). This development covers comparisons among the null, common effect, and unconstrained models. It does not, however, cover comparisons to the positive effects model. To make these comparisons, we use a different computational approach from Hoijtink and colleagues (Klugkist & Hoijtink, 2007; Klugkist et al., 2005). The specific computational implementation used here comes from Haaf & Rouder (2017), who developed Bayes factor computations for many simultaneous order constraints.

There are many ways to compare the four models besides Bayes factors, but we have not developed these alternatives. We provide coverage of why we prefer Bayes factors to other ways in the General Discussion.

Sensitivity to Prior Settings

Bayesian analysis is predicated on specifying prior distributions on parameters. Analysts should be familiar with how these specifications affect model comparison. A few points of context are helpful. It seems reasonable as a starting point to require that if two researchers run the same experiment and obtain the same data, they should reach the same if not similar conclusions. Yet, almost all Bayesians note that priors have effects on inference. 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 minimal influence (Aitkin, 1991; Gelman et al., 2004; Kruschke, 2012; Spiegelhalter, Best, Carlin, & Linde, 2002). In the context of these views, the effect of prior settings on inference is viewed negatively; not only is it something to be avoided, it is a threat to the validity of Bayesian analysis.

We reject the starting point above including the view that minimization of prior effects is necessary or even laudable. 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) provide a particularly appealing view of the prior in this light.

Accordingly, 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 studies. The choice of prior settings are important because they unavoidably affect the predictions about data for the models (Figure 1). Therefore, these settings necessarily affect model comparison. 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. Whatever this degree, it is the degree resulting from the usage of Bayes rule, which in turn mandates that evidence for competing positions are the degree to which they improve predictive accuracy.

When different researchers use different priors, they will reach different opinions about the data. Rouder et al. (2016) argue that this variation is not problematic. They recommend that so long as various prior settings are justifiable, the variation in results should be embraced as the legitimate diversity of opinion. When reasonable prior settings result in conflicting conclusions, we realize the data do not afford the precision to adjudicate among the positions.

The critical prior specifications are those that define the differences between the models. In our case, the specifications are on μ_θ and σ^2_θ, the population parameters. Although these parameters are not the primary target of inference, the prior settings on them affect the resulting Bayes factors. 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 *scale* on μ_θ and σ^2_θ. The scale on μ_θ calibrates the expected size of the effect. This scale is not a point setting; μ_θ may be free to take on any value that reflects the data. Figure 2A shows a plot of the prior on μ_θ for three different scale settings. In application, we set the scale on μ_θ to be 0.40 in standardized effect size, and this value corresponds to the middle curve in the figure, which is dashed. The other setting is the scale of σ^2_θ, and this setting calibrates the expected amount of variability in effect size across studies. We chose a value of 0.24 (this is a standard deviation on site-specific standardized effect sizes). The expected variation across sites or studies is 60% of the expected effect size, which seems like

a reasonable ratio of scales. Figure 2B shows a plot of the prior on σ_θ for three different scale settings, and the middle one, which is dashed, corresponds to the value we used.

Wagenmakers et al.'s Embodied Cognition

Figure 3B provides the results of the reanalysis of Wagenmakers et al.'s registered replication report. The results from estimation are highly suggestive that there is no effect among any of the studies. The Bayes factor analysis favors the null model. The null is preferred 11-to-1 to the common effect model, the next most preferred model. The strong null is preferred 250-to-1 and 36,000-to-1 to the unconstrained and positive effects models, respectively. Here we see formal support for the strong null model—not only is the *average* effect nearly zero, but the most parsimonious description among the four models is that *all* studies have a true zero effect.

Above we discussed that these Bayes factors are dependent on prior settings. Our choices of scale for mean and standard deviation are shown as the middle densities in Figure 2. These choices are informed by general knowledge about the field. We have also provided a reasonable range of variation in these choices, and these are indicated by the bracketing densities. We explore the effects of using these bracketing priors, and the resulting Bayes factors 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 all settings, the ordering of the models remain: the null model is preferred to the common effect model which is preferred to the unconstrained model which is preferred to the positive effects model. This type of sensitivity analysis can always be performed to understand the range of conclusions that may be drawn from the data. In our case, the range is limited to a single ordering.

Ebersole et al.'s Moral Credentialism and Sexism

To show how the meta-analytic Bayes factor model-comparison system works in a more complex example, we re-analyzed a meta-analytic data set from Ebersole et al. (2016). This

paper was the result of the *Many Labs 3 Project*, which was designed to assess the replicability of ten effects across several sites and across different periods of the semester. We focus here on one particular effect, *the moral credential effect,* which was originally demonstrated by Monin & Miller (2001).

The Monin and Miller study was designed to assess whether participants were more likely to express prejudiced attitudes when their prior behavior suggested that they were not prejudiced. To manipulate prior behavior, Monin and Miller asked participants to consider sexist statements and endorse those they agreed with and reject those they did not. The key manipulation is whether the statement was worded to describe *most* women or *some* women. The main notion is that participants would be more likely to reject sexist statements that described most women rather than some women. Participants were randomly assigned to the *most* and *some* condition, with those in the former rejecting more sexist statements than those in the latter. Next, participants read a vignette that described a hiring opportunity at a manufacturing company. Participants rated how much more or less suitable a man would be for the position relative to a woman. The rating scale was a seven-point scale from strong preference for a woman through neutral to strong preference for a man.

The main hypothesis is that participants who previously rejected sexist statements—those who judged sexist statements referring to *most* rather than *some* women—would be more likely to express that men are more suitable than women for the job. Indeed, Monin and Miller report such an effect, and they also report an interaction such that the effect is prevalent for male participants but not for female participants.

We specify four critical parameters for each site. There is a site-specific intercept parameter, denoted α_i for the ith site. This parameter denotes overall rated suitability of men vs. women for the hypothetical job opportunity. Variation in this parameter across sites accounts for variation of overall expression of gender prejudice. There are three slope parameters to describe the effects. One is a site-specific gender-of-rater effect parameter, denoted θ_{gi}. The gender-of-rater effect is whether male participants rate men candidates

higher than female participants rate men candidates. We refer to this effect as the gender effect for brevity. The remaining two parameters represent a moral credential effect—do participants express more prejudice if they were in the *most*-women condition previously? Because Monin and Miller reported moderation of this moral credential effect by gender, we used separate site-specific parameters for male and female participants, denoted θ_{mi} and θ_{wi}, respectively. This parameterization is well-suited for assessing the question whether any credential effect is stronger for men than for women.

We start with an unconstrained model where all four site-specific parameters are free to vary subject to a hierarchical structure as used above.[4] These hierarchical structures lead to regularization, and the resulting estimates for the slope effects are shown in Figure 4. From these estimates several trends are evident. From Figure 4A, there is an overall tendency to judge men, as compared to women, as more suitable for the job. This tendency seems not to vary among the sites. This tendency is a function of the gender of the rater: as compared to female participants, male participants are more likely to rate men higher than women. The gender effect seems to be stable across sites. Finally, there is a small credential effect for both men and women.

The figure alone does not lead to calibrated inference. There are many possible models corresponding to the placement of null, common effect, positive effects, and unconstrained structures jointly on the four site-specific variables $\boldsymbol{\alpha}$, $\boldsymbol{\theta}_g$, $\boldsymbol{\theta}_m$, and $\boldsymbol{\theta}_w$.

Table 2 shows a comparison of a preferred model, labeled *Common Site + Common Gender + Common Credential*, versus similar alternatives. Perhaps the most theoretically similar alternative is the model where there is only a common credential effect for male participants and none for female participants. This model, labeled *Common Site + Common Gender + Common Men Credential*, fares worse than the above model by a Bayes factor of

[4] A formal statement of the unconstrained model is as follows. Let $Y_{ijk\ell}$ denote the ℓth replicate for the ith site, jth gender-of-rater ($j = 1, 2$), and kth credential condition ($k = 1, 2$). The model is given by $Y_{ijk\ell} \sim \text{Normal}(\mu_{ijk}, \sigma^2)$, where $\mu_{ijk} = \alpha_i + u_j\theta_{gi} + m_{jk}\theta_{mi} + w_{jk}\theta_{wi}$. Here $u_j = -.5, .5$ is an indicator that encodes the gender of the rater; $m_{jk} = 0, 1$ is an indicator that is 1 if the rater is a man and the condition is credentialed (most statements) and 0 otherwise; $w_{jk} = 0, 1$ is an indicator that is 1 if the rater is a woman and the condition is credentialed (most statements) and 0 otherwise.

50, indicating that there is a credential effect for participants of both genders. Likewise, a model with separate credential effects for men and women, labeled *Common Site + Common Gender + Common 2 Credentials* also fares worse, though not as extremely. Table 2 also shows that common gender and credential effects are strictly necessary to predict the data. Removing either results in a drastically lower Bayes Factor value.

We also consider more complex models by adding in positive and unconstrained site-specific effects in intercept, gender and credentials. Adding positive variation to the intercept produced a slightly better Bayes factor value than the preferred model (1-to-0.7). This modest Bayes factor indicates that there is only equivocal evidence as to whether the prejudice against women is constant or variable across sites. Without firm evidence for variation, we prefer to use the common intercept form as our preferred comparison model for its simplicity. We also ran the Bayes factor model comparison statistics for the range of reasonable prior settings. The findings above held constant across this range with one notable exception. The conclusion about variability in the intercept depended markedly on the prior settings. When smaller effects are expected, the positive intercepts model is favored; when larger effects are expected, the common intercept model is favored. Hence, the data are not evidential enough to make statements about the variability in the intercept across sites.

In summary, we find a gender-of-rater prejudice effect and a moral credential effect. Further, we find there were no differences across the sites in these effects. Finally, the moral credential effect was the same for both men and women participants. We are unable to learn from the data whether overall prejudice varied across sites.

Haaf and Rouder's Stroop-Effect Analysis

The above two analyses favored models where there was a single common effect across the sites for the critical slope effects. In some sense, this result is not too surprising as these meta-analyses come from carefully planned replication studies. Each of the constituent studies, which are from different sites, followed the same procedures. Hence, the homogeneity

of the effects across the sites is plausible. In the next two analyses, we highlight cases where this homogeneity is not favored. Models with heterogeneity best describe the data.

Haaf & Rouder (2017) developed the models we use here for repeated-measure tasks. In these tasks, several participants each performed several trials in one of two conditions. Haaf and Rouder analyzed three different Stroop experiments, each independently. We analyze the same data here meta-analytically. For each participant in each experiment, we calculate two scores: a mean response time across all trials in the congruent condition, and a mean response time across all trials in the incongruent condition. We analyze the data with the four basic meta-analytic models: the strong null model that there is no Stroop effect in any study; the common effect model that there is a single, common true Stroop effect for all studies; the positive effects model that true Stroop effects for all studies are in the usual direction; and the unconstrained model where true Stroop effects across studies may have different directions.

One difference in the Haaf and Rouder set is that the critical variable is manipulated in a within-subjects fashion. All participants provide a score in each condition. The four models may be adapted in a straightforward manner for within-subject designs.[5] The only complication is reconsideration of the prior settings on scale, and this reconsideration reflects the increased resolution of within-subject designs to detect variation. To set a scale on overall effects we need to consider the size of the effect in individuals. In our experience, response times on repeated trials for the same individual vary about 300 ms in standard deviation. In these experiments, there are about 100 trials per individual per condition. These two facts combined imply that the per-individual-per-condition sample mean has about 30 ms in variation. Whereas these sample means serve as data in analysis, we can

[5] A formal statement of the unconstrained model is as follows. Let N denote the total number of participants across all the studies, and let $j = 1, \ldots, N$ index these participants. Let i_j be the study that the jth participant is in. Let Y_{jk} denote the response time for the jth participant in the kth Stroop condition ($k = 1, 2$). The model is given by $Y_{jk} \sim \text{Normal}([\alpha_j^* + \alpha_{i_j} + x_k[\theta_j^* + \theta_{i_j}], \sigma^2)$. Here $x_k = 0, 1$ is an indicator that encodes the Stroop condition. Parameters α_i and θ_i are study-specific intercept and effect parameters; parameters α_j^* and θ_j^* are participant-specific deviations from study-specific parameters. The null, common-effect, and positive models are placed on θ_i, the site-specific effect parameters.

ballpark the value of σ at 30 ms. Now, we expect Stroop effects on the order of 50 ms. The implication is that the scale on μ_θ in these designs is about 50/30 or 1.6. We also expected that if there was true variation in the effect across sites, it might be 20 ms or so in standard deviation. This yields a scale for σ_θ at 20/30 or .67. We used these values in analysis.

The first task is plotting the estimates of the Stroop effect from the unconstrained model. These estimates are shown with corresponding 95% credible intervals in Figure 5. There is very little shrinkage here. Because of the massively-repeated character of the within-subjects experimental design, there is far less sample noise to regularize.

Bayes factor analysis reveals that the positive model is most preferred. It is preferred by a factor of 3.6-to-1 over the unconstrained model, by a factor of 10^{11}-to-1 over the common-effect model, and by a factor of 10^{46}-to-1 over the null model. Hence, we may conclude that all studies show a Stroop effect and that there is relatively modest evidence for variability across these studies.

Corker et al's Stability of the Big Five Personality Traits

In the preceding three analyses, the models favored were the null, the common effect, and the positive effects models. We have yet to find a meta-analysis where the unconstrained model is preferred to the positive effects model. We suspect, in fact, that such circumstances are rare in the literature for well-defined phenomena.

In this section, we reanalyze a recent meta-analysis from Corker et al. (2017) who examined the stability of personality data from 30 sites. Their main question is whether the Big Five traits are stable across different university populations. Big Five traits are measured as Likert scale ratings of endorsement of certain statements, and each individual is given a score that ranges from 1 to 5 for each characteristic. Stability across sites means that the average across people for a particular trait does not vary across sites. One might hope *a priori* that site averages do not truly vary as such variation may complicate personality research.

One feature of the Corker et al. application is that there is no concept of a true zero, nor are there positive and negative effects. For this application we focus on models with and without variability across sites. Corker et al. use a mixed linear model analysis to assess the variability across sites and to test whether certain covariates, when included, account for this variability. Their work is exemplary and they highlight the two themes promoted here: (i). that estimates should be regularized by models, and (ii). that model comparison and selection is the primary approach to formally address questions about constraints in data. Their approach, frequentist mixed modeling, at least in this application, is similar in spirit to our Bayesian mixed modeling. Therefore we refit their data as a demonstration that our approach yields similar conclusions to standard mixed models in cases where order constraints are not relevant.

To estimate personality traits among labs, we develop a random slope and intercept estimation model where there are site-specific intercept parameters and site-specific slope parameters for each personality trait. For each lab there is a site-specific intercept denoting on average how people in that lab score across all five factors. If participants in one lab tend to endorse higher ratings than in another, the intercept is higher for the first than for the second. There are also five site-specific personality-characteristic parameters; these are denoted θ_{ij}, where i indexes the site and j indexes the personality characteristic $j = 1, \ldots, 5$.

The personality parameters are the target of interest. There are two theoretical positions: one where the distribution of personality traits is common across sites and another where the distribution indeed varies across sites. We take the common-effect position first. We constrain $\theta_{ij} = \nu_j$, where ν_j is a constant that describes how much of the jth characteristic there is in the population. To add heterogeneity across sites[6], we simply distribute these parameters: $\theta_{ij} \sim \text{Normal}(\nu_j, \delta_j)$. Here δ_j is the variability across the jth

[6]A formal statement of the models are as follows. Let Y_{ijk} denote the kth participants score on the jth characteristic in the ith site. The model is given by $Y_{ijk} \sim \text{Normal}(\alpha_i + \theta_{ij}, \sigma^2)$ where α_i are site-specific intercepts and θ_{ij} are defined above. In the common-effect model, the constraint $\theta_{ij} = \nu_j$ guarantees identifiability. In the unconstrained model, the constraint $\theta_{ij} \sim \text{Normal}(\nu_j, \delta_j)$ is sufficient to guarantee identifiability in this context (Rouder et al., 2012).

trait.

Figure 6 shows the estimates of $\boldsymbol{\theta}$ from the unconstrained model. As can be seen, there is a fair amount of variability for each of the personality characteristics.

There are several approaches to specifying families of models for comparison. We highlight what we consider to be an appropriate minimalist approach based on the comparison among three models. The simplest of these models has slopes and intercepts fixed across labs, the second model has intercepts that may vary but slopes that are fixed, and the third model has intercepts and slopes that may vary across labs.

The results are as follows: The common intercept and slope model is least compatible with the data. It is dominated by the unconstrained intercept and common slope model (Bayes factor of about 10^{39}-to-1), which is in turn dominated by the unconstrained intercept and unconstrained slope model (Bayes factor of about 10^{48}-to-1). These staggering values remain staggering across reasonable variation in prior settings.

The alternative approach, perhaps a maximal approach, is to specify all possible submodels of the unconstrained intercept and slopes model. Accordingly, we include models where some traits vary across sites while others do not. An example of such a model is where *agreeableness* and *openess* vary across sites, but *conscienciousness*, *extraversion*, and *neuroticism* are constant. We have avoided such models because we have no theoretical basis for testing why some but not other traits vary across sites. Hence, to us, assessing all these models is not a well-motivated inferential question. We prefer to reserve testing (through Bayes factor model comparison) for cases where models have immediate theoretical interpretations, as they do for the above three models.

Even when testing is inappropriate, we may still report estimates of the variability of the characteristics across the sites. Figure 7 shows the credible intervals on the standard deviation of personality ratings across sites. As can be seen, the degree of variability is fairly stable across the characteristics. This usage of Bayes factor assessment for theoretically important positions along with estimation for exploration of new phenomena is broadly

useful.

Alternative Models

The four models used here are designed to capture the following theoretical positions: 1. A strong null effect where no study has any effect whatsoever; 2. A homogeneous, common effect across studies; 3. Heterogeneity in study effects subject to the constraint that all studies have a positive true effect; and 4. The negation of this constraint where some studies have true positive effects and others have true negative effects. These four, of course, are not the only choices, and here we discuss alternatives.

Equivalence Testing

One trend in the literature is equivalence testing (Rogers et al., 1993). Equivalence testing is motivated by the concern that the null may be too restrictive in many contexts. Instead, a null region is defined, and true nonzero effects in this region are considered too small to be of practical interest. One of the main advantages of equivalence testing in a classical testing framework is that it provides a vehicle for specifying small effects that may not be of practical interest and classifying observed effects as of or of not practical interest based on their magnitude.

Two fairly similar equivalence-region models are possible in the current context. One is that there is a single common effect that is in the equivalence region:

$$\theta_i = \nu$$
$$\nu \sim \text{Uniform}(-\epsilon, \epsilon),$$

The other is that study effects, though constrained to be in the null region, vary from each other:

$$\theta_i \sim \text{Uniform}(-\epsilon, \epsilon).$$

Bayesian analysis of models with equivalence regions follows the usual form (Morey &

Rouder, 2011). Figure 8, analogous to Figure 1, shows the model specifications for two studies as well as the predictions. As can be seen, these models tend to look a lot like the null if the equivalence region is small and much like the unconstrained model if it is large. We computed the Bayes factor for the common effect version for the embodied-cognition task with an equivalence region that is 1/5 of a point on the Likert scale. The Bayes factor for the equivalence-region model was NA-to-1 when compared to the null model, which indicates modest preference for the null.

The appeal of equivalence regions in classical settings—that one can decide if an effect is large enough to be of practical interest, holds in Bayesian settings. In Bayesian analysis, evidence can be stated for or against any model depending on how well they predict data. Models with equivalence regions offer no special advantage, though, they are just models like the other models in the set under consideration. Equivalence-region models tend to be interstitial between the null and unconstrained model, and in this sense, they may be less interpretable in our view than either of the two extremes which have clear theoretical interpretation.

Robustness to Alternative Specifications

We chose the normal and truncated normal specifications for their computational convenience. The Bayes factor model comparison statistic is the marginal probability of the observed data under a model, and computing this marginal is done through integrating out the parameters. Accurate evaluation of this integration can be problematic. The current distributional assumptions follow from Zellner & Siow (1980) who showed their computational convenience.

It may seem reasonable to wonder what may happen if the data drastically violate these distributional assumptions. One area of concern is the unconstrained model. Here we use a graded normal, and this is our only specification to account for the possibility that some studies have a truly positive effect while others have a truly negative effect. There are

other model instatiations, however, that capture this state of affairs. One is a latent mixture model. One can imagine that there are two (or more) classes of studies, and there is some probability that each study belongs to a class.

Figure 9A shows two unconstrained models: the normal and a mixture model. Our aim in choosing particulars for these truths was to equate the overall mean and variance. The true effects for each study were the ticks at the top of the panel. We simulated data from these true values 100 times for each of the two models. The Bayes factor between the normal unconstrained model and the positive model is computed for each of the simulated data sets. At first glance, one might think the Bayes factor favors the normal unconstrained model over the positive model when the truth is from the normal as there is a match between method and assumption. The Bayes factor distributions from the simulation are shown in the first two violin plots of Figure 9C. Surprisingly, the above intuition is wrong. The Bayes factor between the normal unconstrained model and the positive model favored the unconstrained model when the latent mixture served as truth even more so than when the normal model served as truth. This behavior, though counter intuitive, is worth consideration. The mixture model has a larger fraction of negative true values than the normal model (see the ticks in Figure 9A); hence the resulting data tend to be better predicted by the unconstrained model relative to the positive model.

The critical point to emerge from this simulation study is that the unconstrained normal model is a useful instantiation of the unconstrained position. Here is why: The goal is to detect a few negative true effects against a background of many true positive ones. The normal for this configuration would have a positive mean and sufficient variance so that there is noticeable negative mass (as in Figure 9A). The distribution of the negative part is not only small in mass, but is skewed such that small negative effects are weighted. The normal therefore is well-suited to detect the most difficult case—the one where negative effects are few and more likely to be clustered near zero. The mixture models are much easier cases as negative true effects are more numerous and more negative. And this is why the Bayes factor

favors the unconstrained model with mixture truths more so than with normal truths.

Figure 9B shows a set of simulations where all true study effects are positive. One model is the half normal. A second is a positive truncated normal with positive mean. The third represents a misspecification. Here the true values follow an exponential rather than a truncated normal. These three truths were matched in that they all have the same mean. Here, we might expect the Bayes factor to favor the positive model over the unconstrained model. And indeed, this occurs for the truncated model with positive mean. It does not occur as readily for the half normal truth and hardly ever for the exponential truth. Hence, the way these models are specified, data sets generated from true effect sizes where several of these effect sizes are small are likely to be as compatible with the unconstrained normal model than with the positive truncated model. Therefore, the setup is tuned to detect small violations of positivity even at the risk of a false alarm. In reality, this is a useful tuning as we have yet to find any violations of positivity in any meta-analytic data set.

General Discussion

In this paper, we seek to redefine the fundamental question of meta-analysis. Traditionally, inferences in meta-analysis are performed using means, variances, and CIs. These are properties that help the analyst understand the distribution of outcomes across a collection of experiments. Yet, we argue this distribution itself is difficult to interpret because it reflects in part the variation of design parameters across studies. Consequently, we suggest focusing on basic ordinal properties that may be shared among studies. The strong null model stipulates that no study shows a true effect. The common-effect model predicts that all studies have the same true effect. The positive effects model stipulates that true effects may vary in size but not in sign across studies. The unconstrained model stipulates studies have a true effect in one direction and others truly in the opposite direction. This unconstrained model, if favored, suggests substantial differences between studies, such that certain methods, measures, or populations appear to change the sign of the effect. For

laboratory phenomena, success of this unconstrained model should be cause for careful scrutiny, as even the sign of the effect cannot be predicted in advance.

Alternative Model Comparison Methods

The classical approach to meta-analysis stresses two questions: first, is the meta-analytic mean different than zero, and second, is there heterogeneity among study effects? Test statistics for the first question include δ, its z-value and associated p-value; those for the second include the Q-statistic (Cochran, 1954) and its associated p-value. For each of these test statistics, one can perform a classical hypothesis test to reject the appropriate null. Our main concern, however, is with a new question: Does every study in a meta-analysis plausibly show a true effect in a common direction? We do not know of a classial test for this case. As an alternative to classical tests, we stress model comparison by Bayes factor. We find Bayes factor model comparison advantageous for a number of reasons. The most pertinent one here is feasibility—we can compare models with many order-constraints.

For model selection, it is common to use evaluation of goodness-of-fit statistics without preserving long-term error rates. Two examples are the Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwartz, 1978). These fit statistics have built-in penalties for complexity instantiated by counting parameters. Such an approach, however, is inappropriate for ordinal constraints because the constraint does not limit the number of parameters but the range of valid values (Klugkist et al., 2005). Hence, classical model-selection statistics are inappropriate here.

There are several approaches to model comparison in Bayesian analysis as well. Alternatives include inference by posterior credible intervals (Kruschke & Liddell, 2017), deviance information criteria (DIC; Spiegelhalter et al., 2002), and, most recently, inference based on cross validation (Vehtari, Gelman, & Gabry, 2017). We think, however, that these other approaches are inappropriate for assessing multiple order constraints. Here is why:

Credible Intervals: One can certainly compute and plot credible intervals on each study effect as we do in Figures ~**??** and ~**??**. But it is unclear what to do next. We know of no principled way of comparing the null vs. common vs. positive vs. unconstrained model from these intervals. Any rule, we suspect, is ad hoc. Take, for example, the credible intervals in Figure~**??**D. None of the credible are exclusively on one or the other side of zero, and perhaps if one used the credible intervals then one might conclude the null is best. Yet all posterior mean study effect are greater than zero, and there is clearly an effect.

Posterior-Based Methods Some methods, such as DIC and wAIC, are based on computing model comparison statistics from the posterior distribution of the parameters. We show here that these methods are miscalibrated for assessing ordinal constraints. Figure~**??** shows a simple setup with one ordinal constraint. Is the mean of observations positive? In the upper panel, the observations are shown in the upper part as thick vertical segments. As can be seen, all of these observations are negative, and the constraint that the mean is positive is an inferior description. The posterior distributions of the parameter μ under the unconstrained and positive models are shown as colored histograms, and as can be seen, they are quite different. The posterior of μ is more discordant with the observations under the positive model, and this discordance is captured well in DIC, where the deviance is indeed greater for the positive model. Bayes factors too captures this; here the data are better predicted by the unconstrained model than the positive model. The problem occurs, however, for data that are well-within the constraint. These are shown in the bottom panel, and as can be seen, the posteriors are highly similar under either model. Any model comparison based on these posteriors will be equivalent, as is DIC. The Bayes factor, however, shows the appropriate preference for the constraint. Of course one could adjust the statistic, but that would be done outside of Bayesian theory. Alternatively, one could avoid unconstrained models, say comparing negative and positive models. Yet, the unconstrained model is a perfectly valid and interpretable model, and it seems desirable that a model-comparison method be useful for compariong it to its submodels.

With Bayes factors, the ordinal constraints enter not through the posterior but through the prior. The priors give rise to a prediction in the form of prior distribution on possible outcomes. The constrained model has more mass on positive outcomes; the unconstrained model has less because some of the mass is on negative outcomes as well. So when positive outcomes are observed, the positive model is favored. When negative outcomes are observed, the unconstrained model is favored.

The relative utility of Bayes factor is often a topic of vigorous debate, and interested readers can consult Berger & Berry (1988), Edwards, Lindman, & Savage (1963), Gelman & Shalizi (2013), Kruschke & Liddell (2017), Liu & Aitkin (2008), Rouder & Morey (2012), Sellke, Bayarri, & Berger (2001), and Wagenmakers (2007) among many, many others. We will refrain from hashing the debate here. Our choice is a matter of principle. Bayes factor is a direct consequence of Bayes' rule (Efron, 2005). If one wishes to apply the law of conditional probability, then Bayes factor is the unique multiplier for updating the plausibility of models. Other methods are based on different sets of desiderata, the most common of which is the desire to have minimal sensitivity to changes in the prior. We can respect others' goals in this regard, but as a matter of principle, we are content with rational updating of beliefs.

Limitations

One of the main substantive limitations of this paper is that the development is appropriate for similar studies, say those that use the same dependent measure. The reason is that we use a homogeneity of variance assumption which would be grossly violated if the studies used different dependent variables. This assumption is quite reasonable here where the studies in each corpus had the same dependent measure (or, in the case of the personality data, were assessed on the same 5-point scale). It may not hold more broadly, and to the degree it does not, it is a limitation.

We think those who use Bayes factors should be mindful of their practical limitations.

We view Bayes-factor model comparison as a powerful tool that must be wielded with expertise, wisdom, transparency, and restraint. Researchers should have well-conceived questions that are instantiated in well-specified models. Not all model comparisons are helpful and some are even misleading. For example, we decided to not test variability across each of the Big Five personality characteristics separately because we were unsure of the theoretical ramifications of the results.

Another limitation with Bayes factor model comparison comes from considerations in specifying the prior. We expect analysts to differ in their choices, though these differences should fall in a range of reasonable values rather than be arbitrarily broad. Restraint is practiced when we respect this range. In our case, for example, we were unable to assess whether or not there was variation in prejudice toward women across sites in Ebersole et al's data set. The conclusion depended too heavily on how much variation one expected, and different reasonable values resulted in qualitatively different conclusions. The most prudent course is to note that the question could not be answered with the data in hand.

Future Directions

We view the development here as only a first step in implementing this ordinal approach where we ask what is common among a population of studies. There are many possible extensions that would increase the usefulness of this approach. First, the current development is based on a homogeneity-of-variance assumption that limits applications to the analysis of similar studies. Dispensing with this assumption would be quite useful. Second, the analysis in its current form requires all of the data from the constituent studies. In many cases, however, the meta-analyst has access only to the summary statistics. It will be useful to adapt the analysis so that the summary statistics from each study may be used as input. Third, an extension is needed to account for the role of moderators and other covariates. We may still assess whether all studies are positive or null or varied even when demographics and other covariates are accounted for. Fourth, at some point it may prove

useful to develop more realistic models of incoherent phenomena. Incoherent phenomena are those where some studies yield truly negative effects while other studies yield truly positive effects. In these cases, perhaps studies should be modeled as belonging to latent classes as in the above simulation study. Then, the analyst can assess whether this more flexible model better accounts for the data than the four presented. Fifth, it may prove useful to model publication bias. Guan & Vandekerckhove (2016) provide a new model-based approach, and their treatment of publication bias may conceivably be incorporated into our meta-analytic models.

We hope this new approach to meta-analysis proves 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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., ... Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: A meta-analytic review. *Psychological Bulletin*, *136*(2), 151–173. Retrieved from http://psycnet.apa.org/doi/10.1037/a0018251

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*, 159–165.

Bishop, Y. M. M., Fineberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129. Retrieved from 10.2307/3001666

Corker, K. S., Donnellan, M. B., Kim, S. Y., Schwartz, S. J., & Zamboanga, B. L. (2017). College student samples are not always equivalent: The magnitude of personality differences across colleges and universities. *Journal of Personality*, *85*(2), 123–135.

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

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... Nosek, B. A. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*,

68–82. Retrieved from http://ezid.cdlib.org/id/doi:10.17605/OSF.IO/QGJM5

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242. Retrieved from http://dx.doi.org/10.1037/h0044139

Efron, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, *100*(469), 1–5.

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

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., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 57–64.

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

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

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

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

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. Retrieved from

http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572

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

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

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

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

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

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*(1), 33.

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

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

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

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

Rouder, J. N., & 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., Haaf, J. M., & Aust, F. (2017). *From theories to models to predictions: A bayesian model comparison approach.*

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., 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

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician*, *55*, 62–71. Retrieved from http://dx.doi.org/10.1198/000313001300339950

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

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.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distributions. In *Proceedings of the third berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 197–206).

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*(5), 768–777.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy

using inferential confidence intervals: An integrated alternative method of conducting null hypothesis significance tests. *Psychological Methods*, *6*, 371–386.

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.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and wAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Veichtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3). Retrieved from http://www.jstatsoft.org/v36/i03/

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, *14*, 779–804. Retrieved from https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R., ... others. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928.

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
Effect of Prior Variation on Bayes Factor values with the Wagenmakers et al. data set

Mean	SD	Null-to-Common	Null-to-Unconstrained	Null-to-Positive
0.40	0.24	11.1-to-1	264-to-1	47686.1-to-1
0.40	0.40	10.7-to-1	1974.5-to-1	2497771.1-to-1
0.40	0.12	11.3-to-1	54.1-to-1	1399.2-to-1
0.80	0.48	22.3-to-1	10314.4-to-1	Inf-to-1
0.80	0.80	22.1-to-1	370311.3-to-1	Inf-to-1
0.80	0.24	22.5-to-1	506.7-to-1	67260.6-to-1
0.20	0.12	5.9-to-1	28.4-to-1	671.2-to-1
0.20	0.20	5.6-to-1	83.5-to-1	12773.3-to-1
0.20	0.06	5.8-to-1	12.9-to-1	86.7-to-1

Note. Infinite values exceed our precision

Table 2
Bayes factors for select models with the Ebersole et al., data set

Model	Bayes Factor
Common Site + Common Gender + Common Credential	1-to-1
Common Site + Common Gender + Common Men Credential	1-to-53.2
Common Site + Common Gender + Common 2 Credentials	1-to-6.5
Common Site + Common Gender	1-to-105.7
Common Site + Commen Credential	1-to-28041.8
Positive Site + Common Gender + Common Credential	1-to-0.8
Unconstrained Site + Common Gender + Common Credential	1-to-4.5
Common Site + Positive Gender + Common Credential	1-to-4.8
Common Site + Common Gender + Positive Credential	1-to-12.6
Common Site + Common Gender + Unconstrained Credential	1-to-Inf
Comon Site + Unconstrained Credential + Common Credential	1-to-3.4

Note. Infinite values exceed our machine precision of $10^3 04$

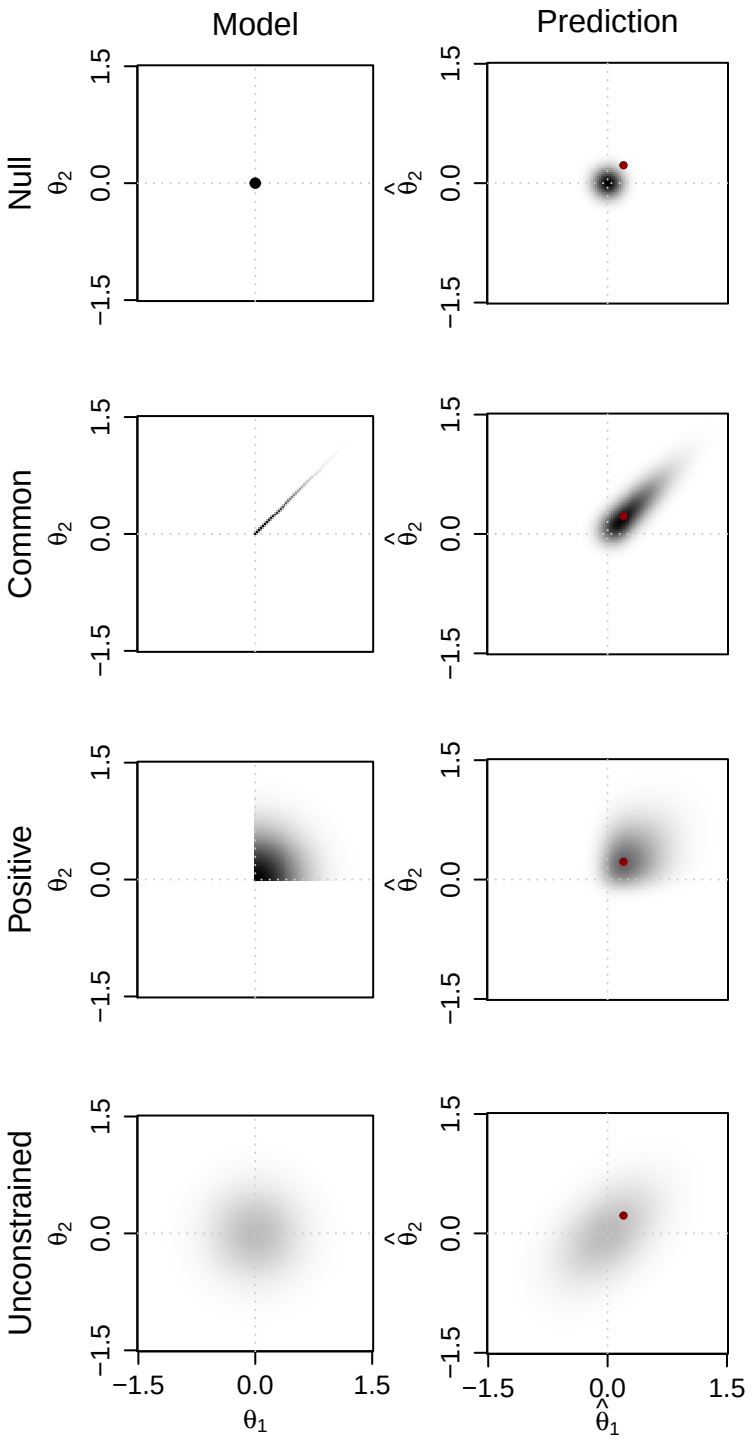

Figure 1. The four meta-analytic models as shown as bivariate distributions across two hypothetical studies. The left column shows model specifications. In each panel, the x-axis is the true value of the effect for Study 1; the y-axis is the true value of the effect for Study 2. The plots show the bivariate distributions of true study effects and darker points correspond to greater density. The right column shows the resulting predictions on observed effects. The format of the plots are the same as in the left column.

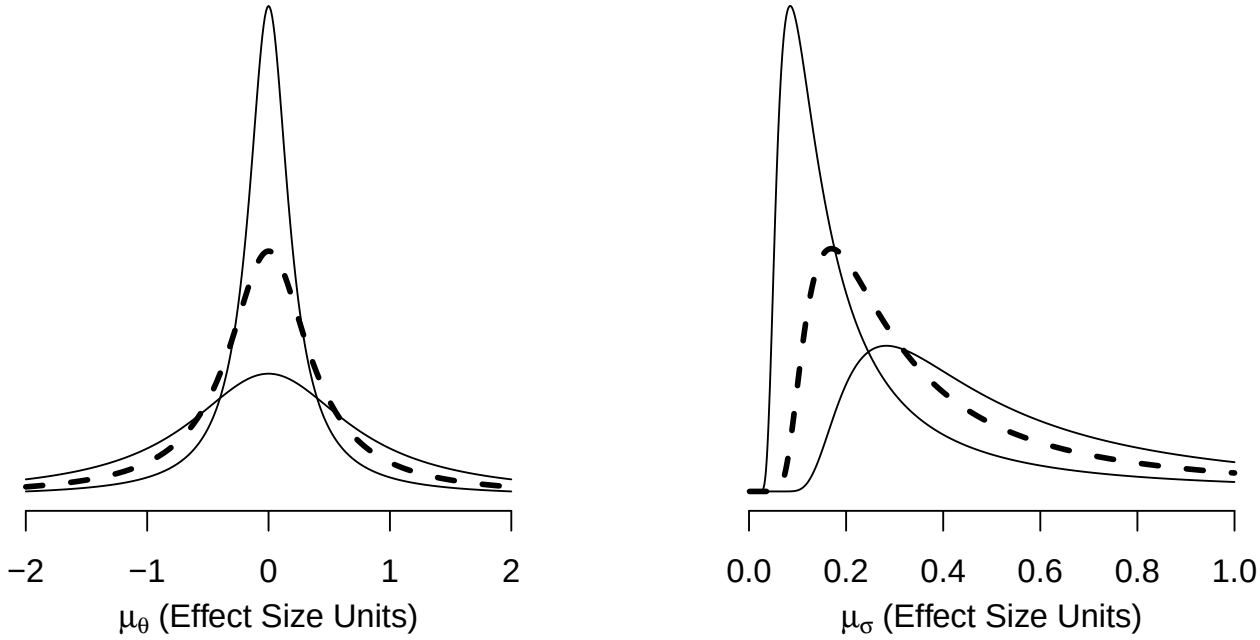

Figure 2. Prior distributions on critical parameters for different scale settings. **A** Priors for μ_θ with scale factors that range from 0.20 to 0.80. Our choice is the dashed curve for scale value of 0.40. **B**. Priors for σ_θ with scale factors that range from 0.12 to 0.48. Our choice is the dashed curve for scale value of 0.24.

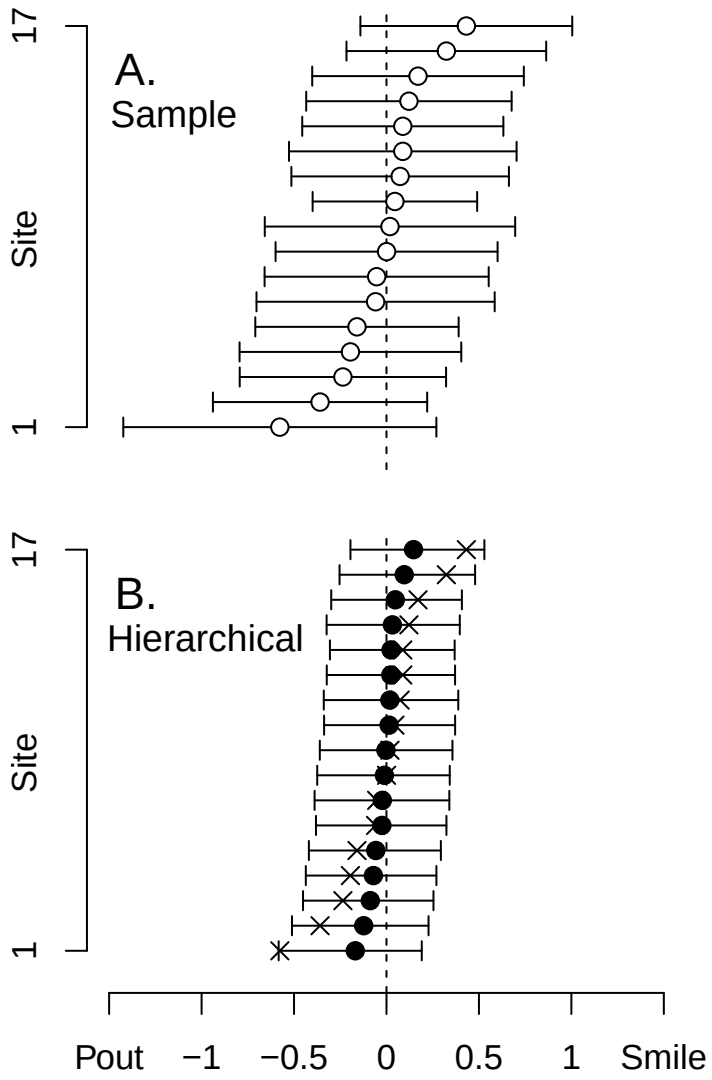

Figure 3. Reanalysis of Wagenmakers et al. (2016) registered replication report. **A**. Sample means with 95% confidence intervals for the 17 sites. **B**. The filled points are hierarchical model estimates (unconstrained model) with 95% credible intervals. The X's are the sample means from Panel A, and these are shown for comparison purposes. These estimates show that after sample noise is accounted, there is not much heterogeneity.

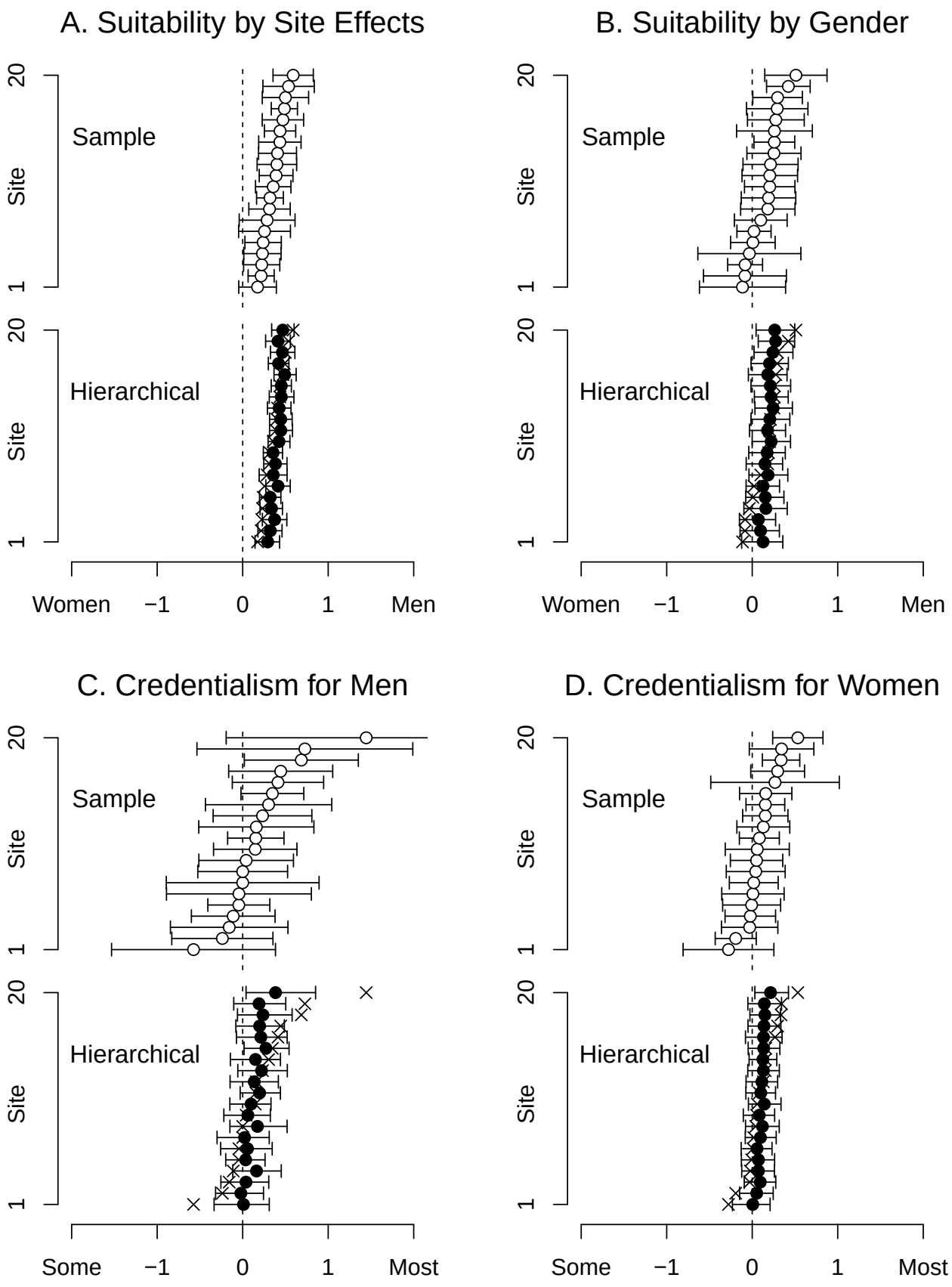

Figure 4. Reanalysis of Ebersole et al.'s (2016) replication of the moral-credentialism effect. **A**. Overall suitability effects by site. The top panel shows the sample effects, the bottom panel shows the hierarchical model estimates (unconstrained model). The filled points are posterior means; the error bars are 95% credible intervals; the Xs are the sample effects. **B**. Sutiability effects by the gender of the respondent. **C, D** Moral Credentialism effect for men and women respondents.

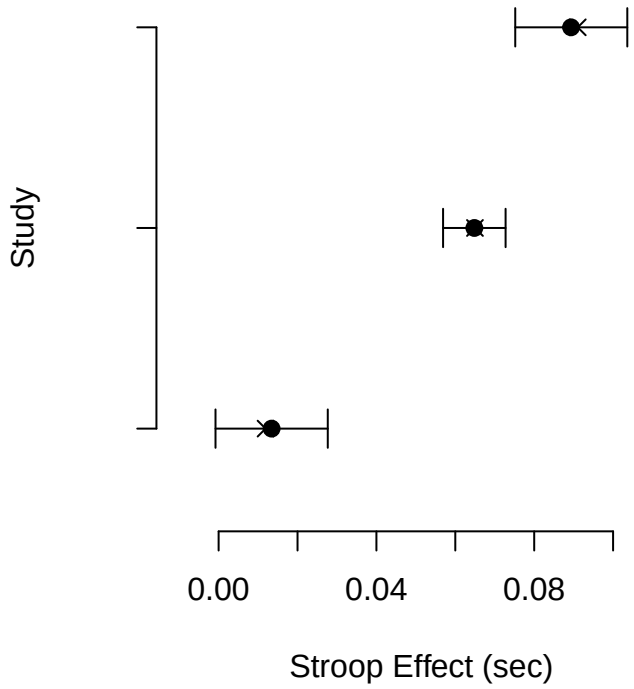

Figure 5. Meta-analytic hierarchical-model (unconstrained) estimates for Haaf and Rouder's (2017) collection of Stroop experiments. The X's denote sample effects which, in this case, are quite similar to posterior means from the model. The effects varies from study to study, but the sign is consistently positive.

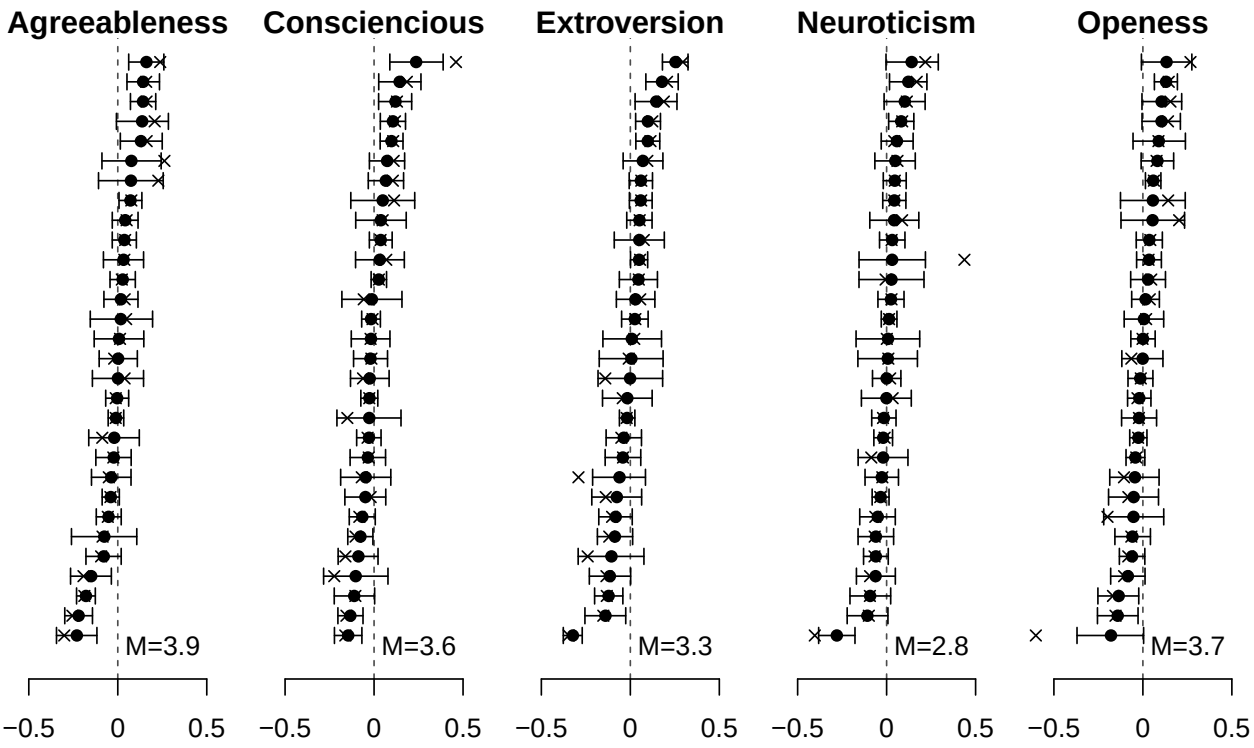

Figure 6. Meta-analytic hierarchical-model (unconstrained) estimates for Big Five personality characteristics across the 30 sites in Corker et al. (2017). The Xs denote sample effects. It is apparent that there is substantial variation across the labs in all five characteristics.

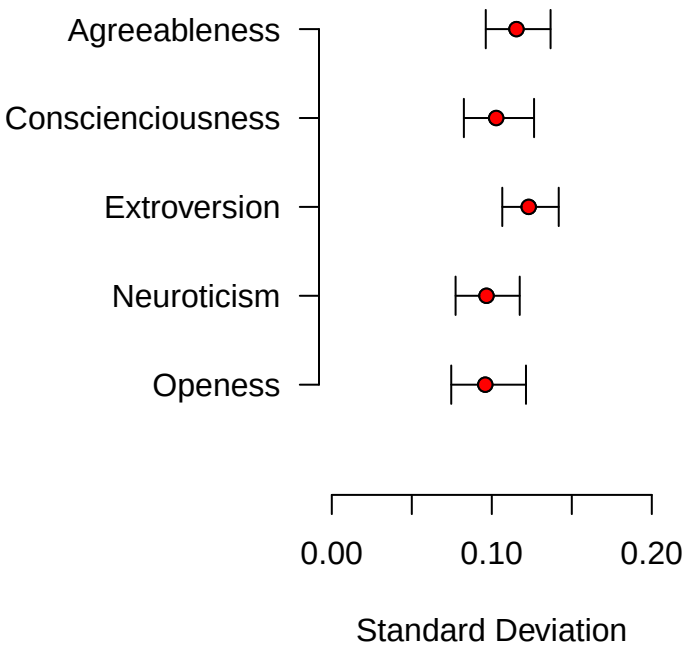

Figure 7. Estimates of variability of personality characteristis. Posterior distribution of standard deviations were computed and plotted are the means and 95% credible intervals of these standard-deviation distributions.

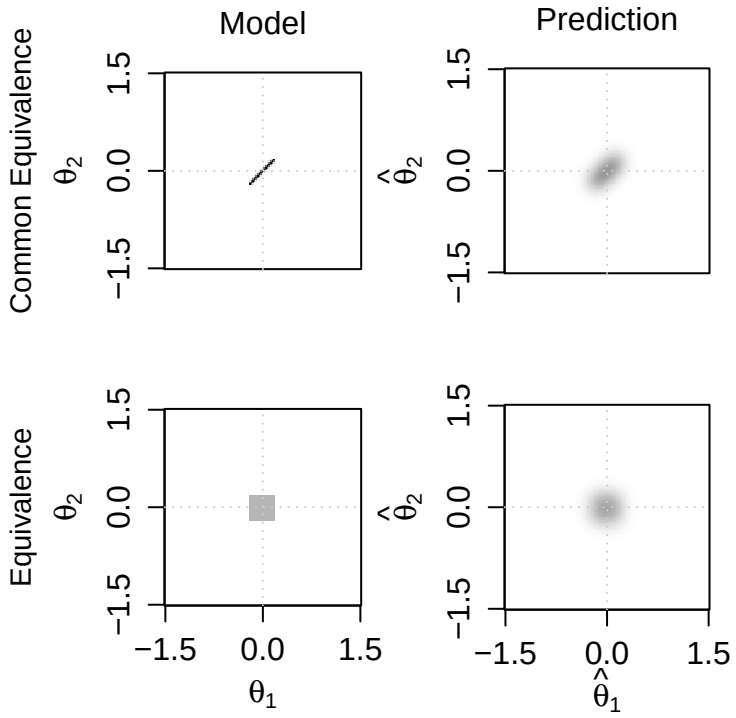

Figure 8. Model specifications for two studies for the common-effect equivalence model (top row) and the regular equivalence model (bottom row). The left column shows model specifications, the right column shows the predictions for data from the models. The format is the same as Figure 1.

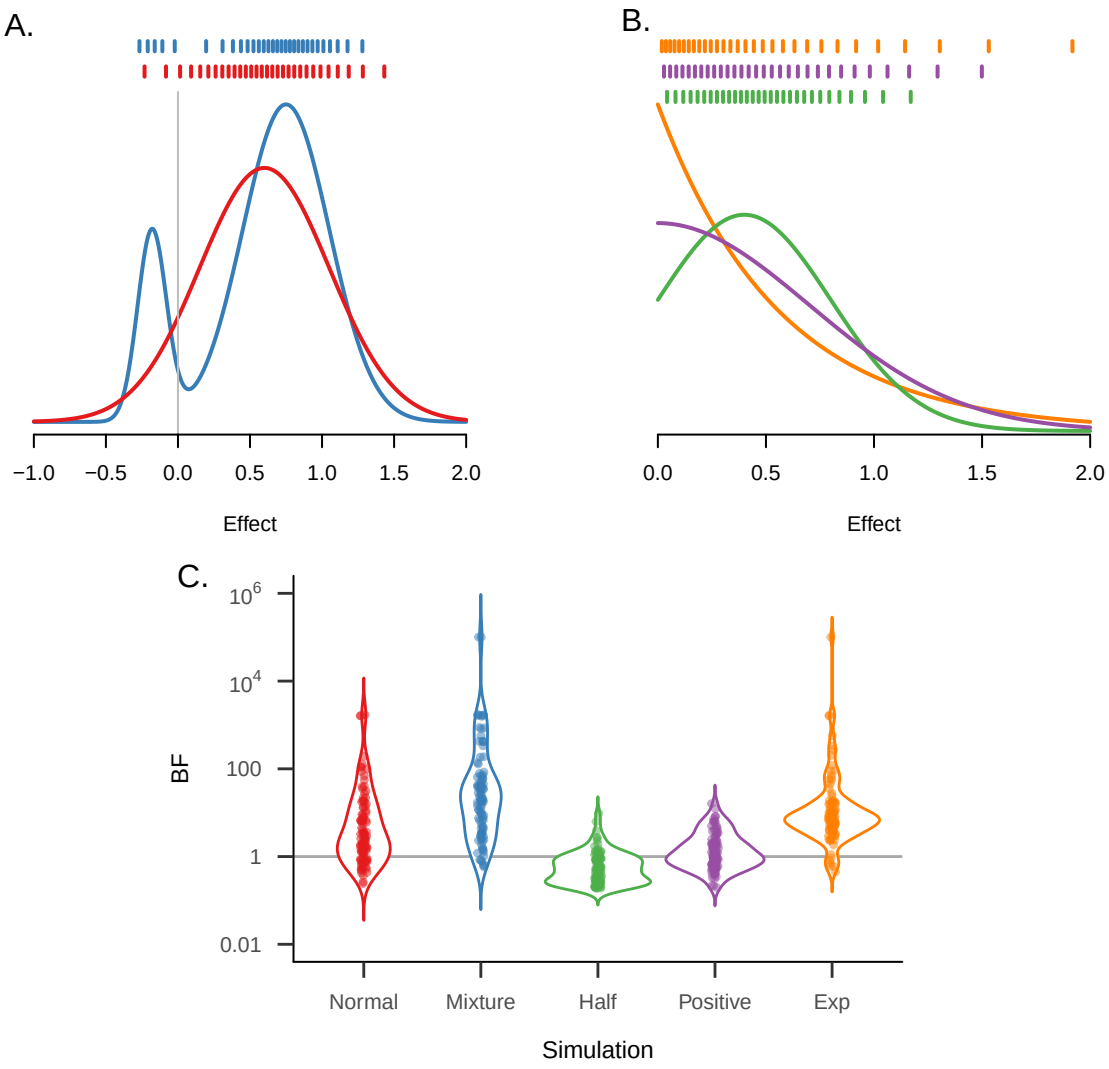

Figure 9. Simulation from five different true models. A. Two true unconstrained models. The red line shows a normal unconstrained model, the blue line shows a mixture model. The ticks at the top of the panel show true study effects chosen for simulation. B. Three true positive models. The purple line shows a half-normal; the green line shows a normal truncated at zero with a positive mean; the orange line shows an exponential. C. Resulting Bayes factor distributions for the unconstrained model vs. the positive model for the simulation study.

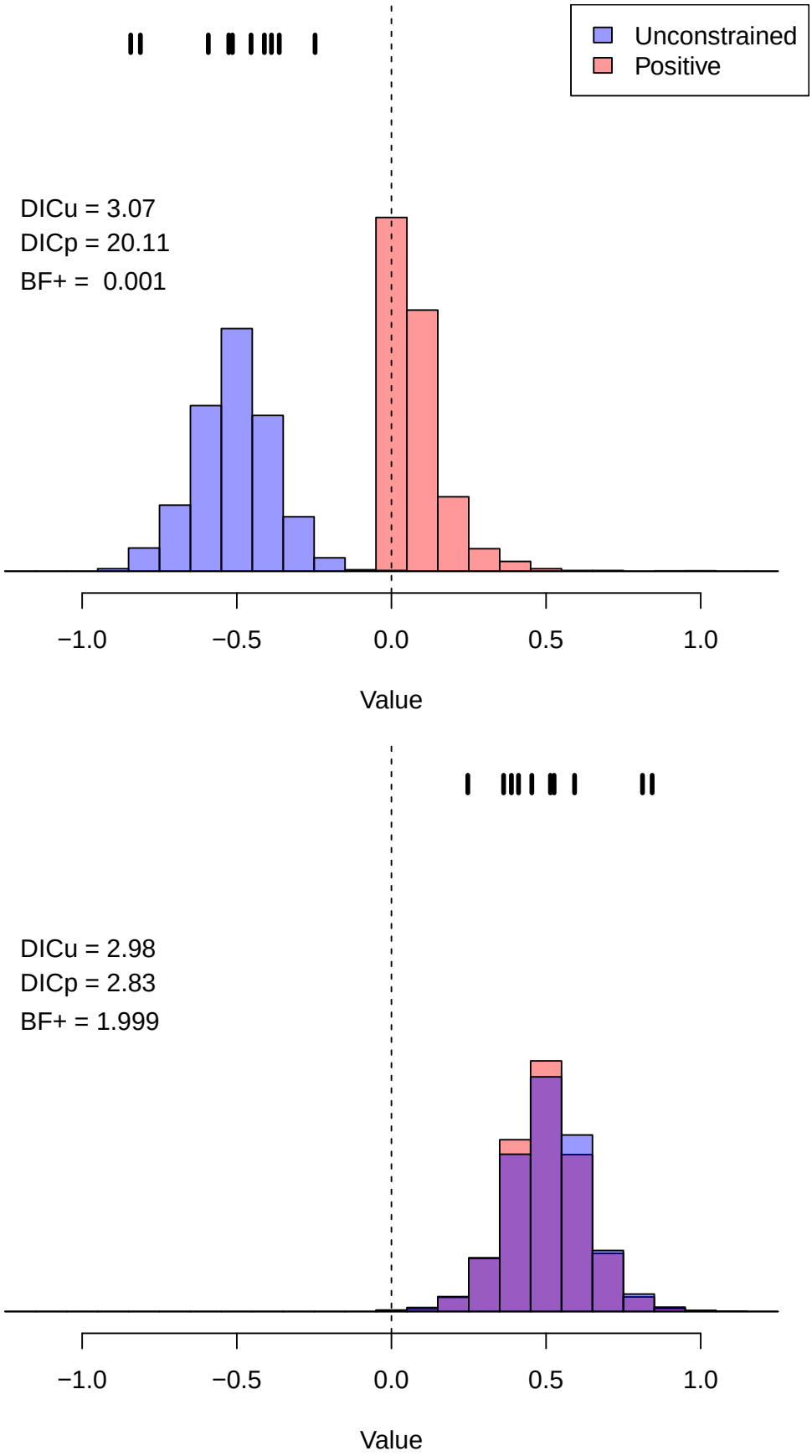

Figure 10. Behavior of posteriors, DIC, and Bayes factors for the comparison of the unconstrained and positive model for two different data sets. The critical panel is the bottom one. The data are broadly compatible with the positive constraint. Even so, the posteriors are largely equivalent leading to equivocal DIC values. The Bayes factor, in contrast, evaluates