BAYESIAN ESTIMATION AND INFERENCE: A USER'S GUIDE

Michael J. Zyphur

University of Melbourne

Frederick L. Oswald

Rice University

Acknowledgements: *For their assistance with previous drafts of this manuscript, the authors would like to acknowledge Bengt Muthén, Kristopher Preacher, Zhen Zhang, and Raymond Zammuto*

Corresponding author: Michael J. Zyphur, Department of Management & Marketing, Business & Economics, University of Melbourne, Parkville, VIC 3010

Email: mzyphur@unimelb.edu.au

ABSTRACT

This paper introduces the “Bayesian revolution” that is sweeping across multiple disciplines but has yet to gain a foothold in organizational research. The foundations of Bayesian estimation and inference are first reviewed. Then, two empirical examples are provided to show how Bayesian methods can overcome limitations of frequentist methods: (1) a structural equation model of testosterone's effect on status in teams, where a Bayesian approach allows directly testing a traditional null hypothesis as a research hypothesis, and allows estimating all possible residual covariances in a measurement model, neither of which are possible with frequentist methods; and (2) an ANOVA-style model from a true experiment of ego-depletion's effects on performance, where Bayesian estimation with informative priors allows results from all previous research (via a meta-analysis and other previous studies) to be combined with estimates of study effects in a principled manner, yielding support for hypotheses that is not obtained with frequentist methods. Data are available from the first author, code for the program Mplus is provided, and tables illustrate how to present Bayesian results. In conclusion, the many benefits and few hindrances of Bayesian methods are discussed, where the major hindrance has been an easily solvable lack of familiarity by organizational researchers.

Keywords: Bayes; frequentist; structural equation modeling; null hypothesis significance testing

BAYESIAN ESTIMATION AND INFERENCE: A USER'S GUIDE

A Bayesian revolution is underway in social science methods (Kruschke, Aguinis, & Joo, 2012), bringing many benefits. First, Bayesian methods expand the range of testable hypotheses, and results can be interpreted in intuitive ways that do not rely on null hypothesis significance testing (NHST; Orlitzky, 2012). Second, Bayesian estimation can combine prior findings with new data, yielding results that are automatic meta-analyses. Third, because Bayesian estimation can use prior findings, data from small-sample studies is less problematic. Fourth, Bayesian methods allow estimating models when traditional estimation fails because of model complexity.

Although new to organizational research, Bayesian methods exist in many disciplines, such as physics (Dose, 2003; von Toussaint, 2011), atmospheric science (Wilks, 2011), geology (Curtis & Wood, 2004), biomedicine (Ashby, 2006; Greenland, 2006, 2007), and economics (Chib, Griffiths, Koop, & Terrell, 2008). Philosophers of science note that Bayesian methods are ideal for scientific inference (Bovens & Hartmann, 2003; Howson, 1997, 2001), and decision scientists model rational decision makers as Bayesian (e.g., Tversky & Kahneman, 1974).

However, quantitative researchers in the social and organizational sciences have historically relied on estimation and inference methods such as maximum likelihood, *p*-values, confidence intervals, and NHST, which all have non-Bayesian origins (e.g., Fisher, 1922, 1925, 1935; Neyman, 1937, Neyman & Pearson, 1933). The drawback is that these methods all rely on a theory of probability called *frequentist* (and its associated inferential methods). Where frequentist probability references the probability of observed data, Bayesian probability references the probability of parameters that are of true interest to researchers (Hacking, 2001; Jaynes, 2003). Indeed, frequentist probability hinders organizational research in many ways: it creates confusion when making inferences with *p*-values and confidence intervals; it inhibits small-sample research; and it makes estimating many statistical models difficult or impossible.

To overcome these issues, we advocate a Bayesian approach to organizational research. We do this by addressing three fundamental questions that researchers may be asking when considering a Bayesian approach: “How does Bayesian estimation and inference work?” “Why should I bother with it?” and, “Can I publish with it?” We begin by addressing the How Does it Work question, explaining Bayesian fundamentals to frame its advantages. With two empirical examples, we then address the Why Should I Bother and Can I Publish questions, demonstrating Bayesian advantages and illustrating how results can be interpreted and presented.

Our first example is a structural equation model, where Bayesian methods show evidence for one of two competing theories by directly *supporting* a null hypothesis (versus NHST, which only “fails to reject the null”). This example also shows how to analyze relationships that are impossible using frequentist methods because of model identification issues. The second example shows an ANOVA-style analysis with a true experiment and a small sample, where prior information derived from meta-analysis and other work is incorporated into analyses, reducing noise in small-sample data and supporting hypotheses that are consistent with the data but frequentist methods fail to support. All data are available from the authors and Appendices give Mplus program code (although AMOS and R also allow estimating these models).

PROBABILITY, ESTIMATION, AND INFERENCE

We explain the difference between Bayesians and frequentists on three dimensions: probability, estimation, and inference. We treat these as they relate to Bayesian and frequentist logic, with added emphasis on Bayesian logic to describe it for organizational researchers. We italicize terms that may be new to readers, but which we define and explain along the way.

Probability

No data or results are definitive; they all carry uncertainty. Probability quantifies this uncertainty. Bayesian or frequentist probability can be used for this, but they have important differences. Although debate on probability's meaning has been heated for 150 years (Daston, 1994, 1995; Galavotti, 2005), Bayesian probability is usually associated with *degrees of belief* or *degrees of knowledge*. Alternatively, frequentist probability is usually based on *features of hypothetically observable systems* and is associated with the relative frequency of an event "in the long run" (e.g., how often something happens in an infinite series of observations).

Consider the statement, "There is a .50 probability that a fair coin will land heads." From a Bayesian perspective, this means that someone's belief is evenly divided between the coin landing heads or tails, or alternatively, that there is no amount of knowledge favoring heads over tails. From a frequentist perspective, however, this statement means that if the coin were flipped a hypothetical infinite number of times, heads would be seen half of the time (Hájek, 1997).

With this fundamental difference in the meaning of probability, Bayesian and frequentist probabilities can be applied to different things. Bayesians can apply probability to anything that can be the subject of belief or knowledge, such as hypotheses, statistical parameters, and entire statistical models. Bayesians can also apply probability to worldly states of affairs, such as past or future single events or long-running relative frequencies, which means Bayesian approaches can subsume frequentist approaches (see Lewis, 1980). Alternatively, frequentists are limited to inference about the relative frequency of events or observations in the long run, even though this long run is always unobservable and is disconnected from researchers' actual interest in single events, theories, hypotheses, or statistical parameters and models (Eagle, 2004; Hájek, 2009).

These two theories of probability affect how statistical models are interpreted. For example, assume linear regression is used to determine the effect of high performance work systems (data along x) on organizational performance (along y). The linear regression model can be represented by the familiar formula $y = \alpha + \beta x + \varepsilon$, where the coefficient β captures the effect of work systems. Using p-values, the frequentist effectively asks, "assuming the null hypothesis that $\beta = 0$ is true, how improbable is the observed effect of work systems or any effect more different from 0)?” To do this, frequentists estimate a sampling distribution for the estimated effect. This distribution simulates the effects that would be observed if the same study were done an infinite number of times, with only sampling error affecting results (Barnard, 1947).

Thus, for frequentists, the data carry uncertainty (in the form of a sampling distribution) while parameters have fixed population values (such as the null that $\beta = 0$; see Fisher, 1921, 1922; Neyman & Pearson, 1933). For statistical inference, NHST is used, even though everyone knows the null hypothesis is false before conducting the analysis (the probability that an unknown parameter is any single value is always equal to zero). This leaves researchers in the strange position of testing a null they never believed in and, regardless of the test’s outcome, they cannot say how probable the null (or any effect) is. Indeed, *p*-values only reflect the probability of the estimated effect, assuming the null is true; to get this *p*-value, frequentists rely on hypothetical infinite replications of the study (via sampling distributions) that never occurred.

Statistical inference would be improved if researchers could forget about the null and make probability statements directly about parameters like β. Conveniently, Bayesians do exactly this by *inverting* the assignment of probability. Bayesian probability refers directly to the parameter β itself (instead of treating parameters as fixed null hypotheses), and observed data in y are treated as fixed (instead of being considered random; Jeffreys, 1939/1998; Savage, 1954). As we now describe, uncertainty in the effect β is quantified by estimating its probability across a range of values, allowing direct probabilistic statements about β based on observed data.

To introduce the logic of estimation and how researchers can make inferences with Bayesian probability, we now discuss Bayesian estimation and inference. We use notation that might be somewhat unfamiliar, but is helpful and is common in the statistical literature.

Bayesian Estimation and Inference

Going beyond linear regression, the symbol θ is often used to represent a parameter or parameters for *any* statistical model (covariances or correlations, the βs in multiple regression or structural equation modeling). The notation $P(\theta|z)$ symbolizes the probability of parameters in θ for any model, *given* all the observed data, represented by z, which would contain both x and y in our regression example. Thus, $P(\theta|z)$ is a very useful piece of information that is not provided in traditional NHST (Howson & Urbach, 1993). $P(\theta|z)$ is also called a *posterior probability* (or *Bayesian posterior*) because it indicates parameters' probability *after* observing the data.

For example, to describe the probability that high performance work systems have a practical impact on organizational performance, Bayesian estimates of θ would contain the regression slope β, allowing a statement such as, "given the observed data, there is a high probability that high performance work systems have an effect around $\beta = 0.20$." Recalling that Bayesian probability is knowledge or belief, this statement indicates that most knowledge or belief is centered at $\beta = 0.20$, with little variance around this value. This statement is intuitive because of its direct focus on the probability of parameters, whereas rejecting the null in NHST does not say anything about the probability of the parameter, and only indirectly relates to the hypothesis of interest (i.e., the data likely did not come from a world where $\beta = 0$, and therefore the estimated β value is preferred). By using Bayesian posterior probabilities, Bayesian analysis frees researchers to make direct statements of the sort they actually want to make (Gigerenzer, 1993; Schwab, Abrahamson, Starbuck, & Fidler, 2011; Wagenmakers, Lee, Lodewyckx, & Iverson, 2011). These are the kinds of statements that a CEO or PhD student would intuitively understand, and they are exactly the kind of statements that are mistakenly used when interpreting the traditional frequentist p-value (Gigerenzer, 1993).

In Bayesian analysis, probabilities are assigned to the range of values that parameters can take (Gill, 2008). In our regression example, θ reflects the full range of values possible for β (a theoretically infinite range). Therefore, $P(\theta|z)$ is a continuous distribution or *probability density function*, where "density" describes the probability associated with the range of all possible β values. In our regression example, *any* value is a possible value for β, but which values are most probable *given* the data? The probability density function answers this question, where Figure 1 shows that the probability associated with the effect of work systems on performance is highest at $\beta = .20$, with 95% of the probability distribution falling inside a *credibility interval* between $\beta = .10$ and $\beta = .30$. The credibility interval shows the most probable range of values for the effect, allowing statements that are impossible with frequentist methods, such as "there is a 95% chance that the effect of work systems on performance ranges between .10 and .30."

INSERT FIGURE 1 ABOUT HERE

As Figure 1 implies, Bayesians treat parameters in θ as *random* or *variable* across a range of possible values, while the observed data in y are treated as *fixed* or *constant*. This is the "inverse" of the treatment found in traditional NHST, where the data in z are treated as random and the population parameters are assumed to be fixed, so that the null hypothesis must be treated as a single value (e.g., $\beta = 0$) to compute a p-value for the observed data.

In summary, Bayesian probability is belief or knowledge. Frequentist probability is the relative frequency of an event in a hypothetical infinite series of events. Bayesian probability allows making intuitive inferences because it references the parameters of interest to researchers, rather than irrelevant null hypotheses. Moving on from probability, we now treat the How Does it Work question about Bayesian analysis by describing how posterior probabilities are derived.

This is done by combining observed data with prior information using Bayes' rule—named after Thomas Bayes (1763). Although our discussion requires introducing some technical terms, we keep it broad and offer a range of accessible references for readers interested in more detail.

Bayes' Rule: The Prior, Likelihood, and Posterior. *Bayes' rule* estimates $P(\theta|z)$, the probability of parameters in θ given observed data in z. Bayes' rule is a way to update initial hypotheses with data to arrive at conclusions by combining initial information about parameters prior to a study being conducted with the probability of observed data given the parameters (Howson & Urbach, 1993). Formally, Bayes' rule is

$$P(\theta|z) \propto P(z|\theta)P(\theta). \qquad (1)$$

This equation is very important. Describing it technically: Results of a Bayesian analysis are the *posterior probabilities* of parameters given the data, $P(\theta|z)$, and these are proportional to (i.e., $\propto$) the probability of the data as informed by the parameters, which is the *likelihood*, $P(z|\theta)$, multiplied (i.e., weighted) by the *prior probability* of the parameters, $P(\theta)$, where prior probabilities are the probability of parameters *before* any data are collected (see Figures 2a-2c).

INSERT FIGURES 2A – 2C ABOUT HERE

To explain, posterior probabilities indicate how probable *all* parameter values are in θ, given the data. These probabilities of parameters are proportional to two pieces of information: 1) the *prior probability* distribution $P(\theta)$, which is the probability of all parameters prior to any data being collected and 2) the probability of the data *given* the range of possible parameters $P(z|\theta)$, or the *likelihood*, which is information that observed data contribute during estimation. When drawing conclusions, a Bayesian analyst takes the posterior probability distribution and considers the median or the peak to be the best (most probable) estimate of parameters (see Figure 1). The more pronounced the peak, the better it serves as an estimate, because the single value is much more probable than other values; when this is the case, the credibility interval indicating a 95% range of probable parameter estimates is small; conversely, when this peak is shallow, then other estimates are also reasonable, meaning that the credibility interval indicating a 95% range of probable parameter estimates is large. In other words, better inferences about a specific parameter value are possible when there is a tall peak in the posterior distribution.

Which Prior Distribution to Specify? Bayesian analysis is liberating because, unlike NHST, it allows making direct statements about the probability of parameters of interest. However, being able to do this requires that researchers choose a prior probability distribution for the parameters before an analysis, and this can be challenging. There has been substantial debate on this issue for over two centuries (see Berger, 2006; Browne & Draper, 2006; Casella, 1985; Efron & Morris, 1972; Howson, 1997; Jeffreys, 1939/1998; Kass & Wasserman, 1996; Lewis, 1980; Suppes, 2007). Simplifying these debates, we note that prior distributions can be 1) *informative priors* based on previous findings and theoretical predictions, 2) *empirical priors* based on observed data, or 3) *diffuse*, *non-informative*, or *uninformative priors* based on no prior knowledge or belief, or a desire to eliminate the influence of a prior distribution during estimation. We discuss each of these three types of priors in turn.

Informative Priors. Informative priors map previous findings or theory onto parameters in θ before conducting a new study (Gill, 2008, Ch. 5; Howson, 1997; Suppes, 2007). For example, findings from a meta-analysis of the effect of feedback on performance could be used to specify a prior distribution for a standardized regression coefficient. Based on Kluger and DeNisi's (1996) finding, a Cohen's *d* of $\mu_d = .20$ and variance $\sigma_d^2 = .97$, the *d* to *r* transformation leads to a Bayesian prior with a mean effect $\mu_\beta = .20$ and a variance $\sigma_\beta^2 = .44$.

Informative priors have many advantages. First, researchers do not have to estimate parameters from scratch each time they conduct a study; past research can inform the current research, which is aligned with the desire of scientists to synthesize past findings and update them with new data (e.g., meta-analysis has been viewed as Bayesian since Schmidt & Hunter, 1977). Using Bayes' rule, researchers can use of past findings for informative priors, as is done in fields like physics and medicine (see Ashby, 2006; Dose, 2003). This analysis yields results that combine all old results as well as the data from a current study—as if all the existing data sets were analyzed simultaneously. As such, Bayesian analysis with informed priors from past research allows continuously updating findings in organizational research, rather than waiting every few years to go through the arduous process of conducting a new meta-analysis.

Second, informed priors facilitate small-sample research, which involves a lot of uncertainty due to sampling error variance. The uncertainty (i.e., variance) in posterior distributions that are used for inferences is reduced when informed priors are consistent with the likelihood derived from data (see Figure 2b; Gelman & Shalizi, 2012), helping hypothesis tests.

Third, informed priors allow estimating parameters with information from both observed data (likelihoods) that can be *supplemented* or *augmented* with prior distributions. Frequentist approaches, by contrast, only rely on observed data (likelihoods). Thus, Bayesian analyses with informed priors allow estimating parameters even when likelihoods do not offer enough information to do so (Garrett & Zeger, 2000) as long as priors offer enough information for estimation (Muthén & Asparouhov, 2012). For instance, in our factor analysis example, all factor loadings *and* residual covariances are estimated among all scale items, with "small variance" priors that allow the estimation of residual covariances that frequentist estimation does not.

At this point, readers may be wondering if informative priors are useful, given that editors and reviewers seem to demand a constant stream of new theory and models. Although informative prior distributions can be based on previously studied effects, they do not have to be. Informative priors can formalize different educated guesses about the prior distribution of possible parameter values, to be subjected to revision by Bayes' rule as the research data come in. Also, researchers often extend existing findings with analyses of moderation and mediation. Using previous findings for informed priors is useful in such cases because finding effects with high probability (or rejecting nulls in NHST) is notoriously difficult.

Empirical Priors. As the name suggests, empirical priors come from *empirical Bayes estimation,* where prior distributions are estimated from a dataset itself (e.g., with maximum likelihood; see Casella, 1985; Efron & Morris, 1972). Empirical priors can have the advantage of using all observed data to estimate parameters that are associated with only a subset of the data, as in multilevel modeling where all of the data are used to estimate a group's mean (Raudenbush & Bryk, 2002). However, in single-level models, there is the caution that empirical priors use the same observed data to estimate priors *and* likelihoods, thereby making an "undesirable double use of the data" (Berger, 2006, p. 399). This is similar to pretending that a sample size is twice as large as it actually is, which results in posterior probability distributions that are too narrow (see Figures 2a-2c, where priors that are similar to likelihoods lead to narrower posteriors). Though arguments can be made for empirical priors in some cases (e.g., Efron, 2010), they are often at odds with the point of Bayesian estimation, where priors are to be *updated* with *new* data.

Uninformative Priors. When a study is exploratory, there may be little to no prior knowledge that can be used for estimation. Similarly, prior knowledge may be diffuse because of contradictory findings or competing theories, leading to prior distributions that are also diffuse. Alternatively, researchers may decide to eliminate the importance of priors in the estimation process to rely as much as possible on the likelihood (i.e., rely primarily on the data). In these cases, it is common to specify prior probabilities that allow the data (the likelihood) to dominate, such as by specifying a uniform (flat) prior distribution. This distribution specifies that, before any data are collected, no parameter values are more probable than others. A distribution that serves a similar purpose is a prior distribution with a huge variance, such as a normal distribution for an effect β with a mean $\mu_\beta = 0$ and variance $\sigma_\beta^2 = 10^{10}$. This variance makes the prior probability distribution of the parameter values nearly flat, which is the default setting in some statistics programs (see Asparouhov & Muthén, 2010a; Muthén, 2010). Such distributions are so uninformative that they allow the data to dominate the estimation of posteriors through the likelihood (Gill, 2008; Kass & Wasserman, 1996). This is shown in Figure 2c, where the prior is less informative (i.e., more flat) compared with the priors from the other figures. Using uninformative priors in this manner has a long history (e.g., Bayes, 1763; Keynes, 1921/2008; Laplace, 1825/1995). Recent approaches with formalized rules for specifying uninformative priors for different types of distributions have been termed *objective priors* (see Ghosh, 2011). However, the word *objective* here simply means "with standardized rules."

As we describe below, uninformative priors allow Bayesian estimation to mimic frequentist estimation, because prior information does not influence results (see Gill, 2008). However, *it remains fundamentally different* from frequentist estimation because the "inverse" Bayesian interpretation is still expressed in terms of the probability of the parameters given the data, thus still facilitating intuitive inferences about model parameters.

In conclusion, although prior distributions can be categorized into three types, and although the choice of a prior distribution can be debated, the impact of the prior in determining posteriors is important for smaller sample sizes and diminished as sample sizes increase. Thus, "in problems with large sample sizes we need not work especially hard to formulate a prior distribution that accurately reflects all available information." (Gelman, Carlin, Stern, & Rubin, 2004, p. 108). Therefore, even in cases where empirical or informative priors are used, results from Bayesian analysis will converge with frequentist estimation as sample sizes increase (see Wasserman, 2004, p. 181), yet still provide the advantage of the probabilistic interpretation of parameters. In any case, researchers can study the sensitivity of posterior distribution results to priors by using different priors for the same analysis and examining differences in posteriors—a study of *prior dependence* (Asparouhov & Muthén, 2010b; Muthén, 2010).

Bayesian Estimation. Assuming a prior has been specified, Bayesian methods for computing posterior distributions has generated much discussion (see initial work by Bayes, 1763; Laplace, 1774/1986; see modern work by Gelman et al., 2004; Gill, 2008). Historically, computing the posterior probability of parameters has been very difficult. This severely limited the usefulness of Bayesian methods when they were popularized in the late 1700s and 1800s (Stigler, 1986). However, in recent decades, conceptual and computational advances now allow researchers to estimate posterior probability distributions for many models.

We offer a simple example of a coin being flipped to introduce the general statistical logic behind Bayesian estimation. Consider two hypotheses that are in θ, H1: "a coin is unbiased", and H2: "a coin is biased and will always show heads." This can be shown as the set:

$$\theta = \left\{ \begin{array}{l} \text{H1: A coin is unbiased} \\ \text{H2: A coin is biased and will always show heads} \end{array} \right\}$$

Before observing the data of observed flips, a person might presume that the probabilities that H1 and H2 are true are equal to .99 and .01, respectively. These priors in $P(\theta)$ reflect experience and knowledge, namely that coins almost always seem fair and are not designed to be unfair. Using Bayes' rule, it is possible to update the probability of the hypotheses in θ with data that are collected in z, where z is a series of flips, recording whether the coin showed heads or tails. This updating requires working with the likelihood $P(z|\theta)$, which is the probability of the data *given* the hypotheses. To derive posteriors, priors are multiplied by likelihoods, as in Equation 1.

For example, if the coin is flipped twice, each time landing heads, the difference in the prior probabilities for H1 and H2 will not change much, because if the coin is fair, the probability of two heads is relatively high, with the likelihood: *P*(2 heads|H1) = $.50^2$ = .25. Given the prior is *P*(H1) = .99, then the posterior is *P*(H1|2 heads) ∝ .99×.25, which is .2475. Likewise, the posterior probability of H2 is *P*(2 heads|H2) = 1^2 = 1, because H2 stipulates that the coin will always land heads. Because the prior is *P*(H2) = .01, the posterior is *P*(H2|2 heads) ∝ .01 × 1, which is .01 and therefore still exceptionally small. However, to interpret these posteriors for H1 and H2 as probability, this requires that we *normalize* them. To do this, all posteriors must sum to 1, because total probability is always equal to 1. To do this, we sum .2475 and .01, and divide each by the sum. This results in the following posterior probabilities:

P(H1|2 heads) = .2475/.2575 = .961

P(H2|2 heads) = .01/.2575 = .039

Thus, observing two heads leads to revising the prior probabilities towards H2, the hypothesis that the coin always shows heads, but the hypothesis of a fair coin dominates because observing only two heads still has a high probability if the coin is fair (as does observing any outcome).

However, assume the coin lands heads 10 times in a row… then 20 times… and then 50 times! These new data would substantially revise the prior probability of the hypotheses because the probability of observing 50 heads in a row with a fair coin is minuscule, with the likelihood: *P*(50 heads|H1) = $.50^{50}$ = .000000000000000888. However, the probability of observing 50 heads under the assumption that the coin always shows heads is always 1, with likelihood: *P*(50 heads|H2) = 1^{50} = 1. Using Bayes' rule, the posterior probabilities would become *P*(H1|50 heads) ∝ .99 × .000000000000000888 = .000000000000000879. However, *P*(H2|50 heads) ∝ .01 × 1 = .01. To normalize these values and obtain posterior probability, again, requires dividing them by their sum, which is .010000000000000879. The resulting posterior probabilities, then, are

P(H1|2 heads) = .000000000000000879/.010000000000000879 = .000000000000879 (nearly 0)

P(H2|2 heads) = .01/.010000000000000879 = 0.999999999999920 (nearly 1)

As additional heads are observed, the probability of the coin being biased (H2) would increase (even though the prior probability absent any data was very small), whereas the probability that the coin was fair would decrease (even though its prior probability absent any data was very large). Eventually, with more data, the likelihood dominates the posterior, and prior probabilities become almost entirely irrelevant when considering the probability of H1 and H2. Further, we can categorically falsify H2 by observing just one outcome of tails. If this occurs, then the posterior probability of H1 will be 1, because all of the probability will be associated with it.

But what if the coin is biased in a way not captured by H1 and H2, such as if it produced 60% heads and 40% tails? In this case, neither hypothesis would have a high probability, indicating that additional hypotheses should be included. This highlights the need for good statistical model specification in θ, which is still an issue for Bayesian methods.

Our coin-flipping example indicates that the prior probability distribution $P(\theta)$ is where researchers can formally incorporate theory and previous findings into analyses (Howson & Urbach, 1993). Looking at Figures 2a-2c, note that the final parameter estimate (which could be the peak of the posterior probability distribution) lies between the peak of the prior distribution and the peak of the likelihood. This demonstrates how a posterior distribution is a weighted combination of the prior distribution and likelihood, and how Bayesian results relying on priors from previous research would combine them with new data to obtain updated results.

Markov Chain Monte Carlo. The most common method for Bayesian estimation in applied problems involves *Markov Chain Monte Carlo* (MCMC) estimation (Gelman et al., 2004; Gill, 2008), which is “the great success story of modern-day Bayesian statistics” (Efron, 2011, p. 1052). The point of MCMC is that it allows specifying many types of priors. For MCMC, there are multiple methods (e.g., Casella & George, 1992; Chib & Greenberg, 1995) that can be implemented in popular software (e.g., Asparouhov & Muthén, 2010a, 2010b).

Because multiple introductory discussions of MCMC exist (e.g., Muthén, 2010; Muthén & Asparouhov, 2012), we direct the reader to these discussions and offer only a few brief points.

MCMC is an iterative process, where a prior distribution is specified, and posterior values for each parameter are estimated in many iterations (estimates form a “chain”). Posterior values are estimated to build up and define the posterior distribution. MCMC is carried out from at least two starting points (i.e., at least two “chains”), to ensure the convergence of the iteration process on a stable estimate of posteriors. Convergence can be evaluated by calculating the *potential scale reduction* (*PSR*; Asparouhov & Muthén, 2010a; see also Gill, 2008). *PSR* is complementary (opposite and inverse) to an intraclass correlation. Formally the *PSR* is the ratio of (Total Variance across chains/Pooled variance within chain); once this ratio reflects very little variance between chains when compared to within-chain variance (i.e., a $PSR < 1.05$), estimation is halted because different iterative processes (i.e., different chains) yield equivalent results. The Kolomogorov-Smirnov test is a recent test for equal posterior distributions across chains; it is stricter than *PSR* in supporting convergence, and it helps identify where a model is not properly identified (this test can be conducted in Mplus 7; Muthén & Muthén, 1998-2012).

Conveniently, MCMC generates many independent posterior estimates, gracefully handling non-normal and skewed posterior distributions, such as those that arise when testing indirect or mediated effects (Yuan & MacKinnon, 2009). Thus, in relation to traditional methods, MCMC may sound like other forms of simulation called bootstrapping, which may be a helpful analogy to draw (see Efron, 2011; Newton & Raftery, 1994; Rubin, 1981).

Bayesian Inference. Examining posteriors allows drawing statistical conclusions in many powerful and intuitive ways (e.g., Casella & Berger, 2002; Edwards, Lindman, & Savage, 1963; Gill, 2008; Hoijtink, Klugkist, & Boelen, 2010; Kadane, 2011; see foundations in Edwards, 1954; Jeffreys, 1939/1998; Savage, 1954; Wald, 1950). Two useful ways for making

Bayesian inferences are described here, with a focus on the distribution of posterior probabilities for parameters of interest (see Gill, 2008).

One way to make inferences is to examine the range of parameter estimates that captures 95% or 99% of the posterior probability distribution. This is called a *credible* or *credibility interval* (Gelman et al., 2004; Gill, 2008), which allows testing hypotheses in multiple ways (Muthén, 2010). For example, Bayesians can perform a type of NHST by seeing if a 95% or 99% credibility interval contains a null value; if not, the null can be rejected as improbable (Berger, 2003; for initial work see Jeffreys, 1939/1998). This is in contrast with traditional confidence intervals, which are based on the idea of an infinite number of replications of a study, and do not directly reference the probability of the null. Further, a Bayesian *p*-value can be reported by taking the peak of the posterior distribution as the Bayesian estimate of a parameter, then reporting the proportion of the distribution that exists on the other side of a null value (Muthén, 2010). For example, the Bayesian *p*-value associated with an effect $\beta = 0.50$ is the proportion of the posterior at or below $\beta = 0$ (Jeffreys, 1939/1998). With a low probability (e.g., $p < .05$), researchers might reject the idea that the population effect is at or below zero.

However, with the ability to reference the probability of parameters directly, focusing on an irrelevant null value is not very helpful. Bayesians can seek support for hypotheses in a more confirmatory manner (Bovens & Hartmann, 2003; Edwards et al., 1963; Howson, 2001; Howson & Urbach, 1993). With information on the posterior probability of a parameter, researchers can focus on where a parameter can be credibly described to exist. Based on credibility intervals, this approach is stronger than the many recommendations for using confidence intervals (e.g., APA Publication Manual, 6th Ed., 2010; Cumming & Finch, 2005). For example, if a null effect $\beta = 0$ was a legitimate research hypothesis, the proportion of the posterior distribution that existed between standardized values of -.10 and .10 could be determined. This proportion indicates the probability that the effect is within a range of values associated with what might be considered practically irrelevant, and if this proportion were large, it would support the hypothesis of no practically meaningful effect (similar to Example 2 below). Being able to make direct inferences regarding the probability of no practically meaningful effect is a powerful new addition to the repertoire of organizational research methods, one that is difficult in a frequentist framework.

Another way to make Bayesian inferences is by evaluating an entire model or by comparing different models (for methods alternative to ours, see work on Bayes factors in Gill, 2008, p. 243; Kass & Raftery, 1995; Raftery, 1992). This can be done with what is called *posterior predictive checking* (Asparouhov & Muthén, 2010a, 2010b; Gill, 2008; Muthén, 2010; Muthén & Asparouhov, 2012), which answers the question, “Do the estimated parameters in the model produce data that look like the observed data?” Posterior predictive checking samples posterior estimates of model parameters and uses them to generate a dataset that is the same size as the observed dataset. The probability of the observed data and the probability of the generated data are then each estimated with χ^2 values, and the latter χ^2 value is subtracted from the former. This is done over many iterations, which creates a distribution of the χ^2 differences. Positive average differences in χ^2 values indicate poor model fit, because it means the observed data tend to have larger χ^2 values than generated data. However, when the model conforms to the data, the observed and generated data are equally likely; when this is the case, the average χ^2 difference between observed and generated datasets equals zero. Thus, posterior predictive checking involves two sources of uncertainty: uncertainty in sample data (by comparing observed versus generated data), and uncertainty in parameters themselves (by sampling parameters from the posterior distribution). Frequentist approaches to assessing model fit assume fixed parameters and do not incorporate this latter source of uncertainty, giving Bayesian analysis an advantage.

In posterior predictive checking, the χ^2 difference values allow computing *posterior predictive* p-*values* (*PPP*), which reflect the proportion of times that the observed data are more probable than the generated data (i.e., the proportion of times observed data have a smaller χ^2 than the generated data). Thus, *PPP*-values of .50 indicate that, on average, the observed data are just as probable as the generated data, implying good model fit (Asparouhov & Muthén, 2010b). However, small *PPP*-values (e.g., < .05) indicate poor model fit because this means that the observed data fit better than generated data very infrequently (e.g., less than 5% of the time), prompting model respecification (Lynch & Western, 2004).

Another index for comparing models uses the *deviance information criterion* (*DIC*; Muthén, 2010). As with Akaike's and Bayesian information criteria (AIC and BIC) that are often reported in structural equation modeling output, the *DIC* rewards models that strike a balance between parsimony and fit (see Gill, 2008, p. 260). Models can be compared using the *DIC* even when they are not nested, and smaller *DIC* values indicate better models.

Summary. Bayes' rule allows working with priors and likelihoods to derive posterior probabilities. Priors can be empirical, uninformative, or informed by previous findings. A range of posterior distributions are estimated with MCMC, and checks for convergence of posteriors done with *PSR* statistica. Hypotheses associated with parameters can be tested with credibility intervals and Bayesian *p*-values, and model fit and model comparisons can be made by comparing observed versus generated data (where positive χ^2 values indicate poor fit), *PPP*-values (where values < .05 indicate poor fit), and *DIC* (for model comparisons).

Frequentist Estimation and Inference

We make three points about frequentist methods to remind the reader of the stark contrast between frequentist and Bayesian methods. First, frequentist probability is applied to random events or observations *z* and not to parameters in θ (Fisher, 1921). Populations are defined by fixed parameters in θ (Fisher, 1922), so that parameters are unknowable but take on singular values with a probability of 1.0 (e.g., the null hypothesis $\beta = 0$). Whenever frequentist model comparisons are made, probabilities refer to the data *given* the fixed parameters under one model versus another. Models with higher likelihoods receive greater support, but *without* considering the prior probability of the model parameters (all parameters are equally likely a priori).

Indeed, frequentists have no priors (Howson & Urbach, 1993). Frequentist probability refers to the probability of the data given a parameter (e.g., the null) and computing this requires assuming that all data are sampled from the same fixed-but-unknown underlying population distribution (i.e., they are identically distributed). This is why Bayesian and frequentist estimates converge when priors are uninformative: likelihoods dominate posteriors (Berger, 2006).

Third, because probability is accorded to data and not parameters, uncertainty in sample estimates is reduced as a function of increasing sample sizes. With small sample sizes, this means that there can be so much uncertainty that even massive effects result in *p*-values and confidence intervals that are too large to reject null hypotheses (Neyman, 1937). This means that even if researchers found *exactly* what they predicted they could not support their hypotheses. In contrast, Bayesian analysis incorporates priors to assist or qualify the informativeness of small samples, so that the data do not have to stand on their own every time a new study is conducted.

EMPIRICAL EXAMPLES

We now provide two empirical examples that demonstrate Bayesian estimation and inference. In the first, Bayesian estimation gives evidence against one of two competing theories in the form of support for a null effect, which cannot be accomplished with frequentist methods. The second example is a true experiment with a small sample and informative priors, where the results support hypotheses that frequentist methods fail to support. Each example is a study from published research that originally relied on a frequentist approach.

Example 1: Structural Equation Modeling

This example is based on Zyphur, Narayanan, Koh, and Koh's (2009) investigation of testosterone's effect on social status. In brief, Mazur's dominance theory proposes a positive testosterone-status link because of testosterone's effect on engagement and persistence in dominance contests (Mazur, 1985). Alternatively, status characteristics theory indicates that status is conferred to individuals based on their ability to engage in and complete socially relevant tasks rather than dominance contests (Berger, Cohen, & Zelditch, 1972). If dominance theory is correct the testosterone-status effect should be positive, but if status characteristics theory is correct then testosterone will probably be unrelated to status.

Zyphur et al. (2009) addressed this question by randomly assigning 579 students to 92 project teams and measured their testosterone by salivary assay (in pg/ml). At the conclusion of the semester, each member of the team rated other members along a 5-item status scale, and responses were aggregated across raters. To remove an effect of gender, both status and testosterone were centered around gender. Also, both testosterone and status (peer rating) were within-group centered to test the hypothesis of a testosterone-status effect within groups.

Frequentist Analysis. In Zyphur et al. (2009) a structural equation model was specified to estimate a latent status factor. The five observed status variables indicated a single latent variable, and no residual covariances were specified. To examine the testosterone-status effect, the status factor was regressed on levels of testosterone (see results in Table 1 and Figure 3a). This model showed a poor fit to the data in NHST terms ($\chi^2_{(14)} = 114.85$, $p < .01$), and modification indices indicated this misfit came from the non-zero residual covariances among the observed items. However, standardized fit indices were acceptable enough to interpret the testosterone effect (CFI = .96, TLI = .95, SRMR = .01, RMSEA = .11). The testosterone-status effect was effectively 0 and not statistically significant, where $\beta = -0.001$ ($SE = .002$, $p = .71$).

INSERT TABLE 1 AND FIGURE 3 ABOUT HERE

Zyphur et al. (2009) note that this result lends more support for status characteristics theory than dominance theory. Yet, the *p*-value of .71 does not directly support the null; it merely fails to indicate that the observed data are unlikely if the parameter value in the null hypothesis is true. Thus, *p*-values do not directly inform the probability of a null or a research hypothesis (Berger & Sellke, 1987; Cohen, 1994). In any case, because of the focus on rejecting nulls with frequentist methods, theorizing null effects is almost never done, even though hypothesizing nulls can be useful (Greenwald, 1975; Kruschke, 2011; Schwab et al., 2011).

Also, the frequentist analysis has a shortcoming that causes misfit. The measurement model specifies independence of the observed items, meaning no residual correlations among items. Thus, the latent variable must account for all observed covariance among the items. While this is a typical specification for a general factor model, the χ^2 value indicates that this is not justified and, indeed, under the strict NHST approach, the model is rejected. Unfortunately, using frequentist methods, it is not possible to estimate factor loadings and all possible residual covariances among observed variables simultaneously because there is not enough information in the data alone to estimate all of these parameters (i.e., the model is not identified; Bollen, 1989). This is unfortunate because even with a unidimensional scale there is no reason to assume exactly zero covariance among residuals because the content between items might covary to some small degree beyond the trait being measured. In other words, measurement models are too strict. Conveniently, a Bayesian analysis can deal with this by relying on information data *and* priors that reflect a strong assumption of no residual correlation.

Bayesian Analysis. The data from Zyphur et al. (2009) were re-analyzed with a Bayesian estimator. Given the possibility of observing different testosterone-status effects, a normally distributed uninformative prior was used for the effect. Default priors were used in Mplus, where variances and covariances and regression effects have very uninformative priors. Priors for regression effects are normal with $\mu_\beta = 0$ and a very flat prior distribution by specifying a huge variance of $\sigma_\beta^2 = 10^{10}$ (see Muthén & Asparouhov, 2012). Two chains were estimated and in 100 iterations reached an appropriate convergence criterion of *PSR* = 1.05 (near 1.0).

The first step in examining Bayesian estimation results is to evaluate the quality of the model. Checking the quality of a model is done with posterior predictive checking (recall the *PPP*-value is how often the observed data are more probable than the model-generated data). Consistent with the large χ^2 from the frequentist analysis, the *PPP*-value is less than .01, which is not desirable because it means the observed data are improbable given the model. Also, the 95% confidence interval of the difference between observed and generated data χ^2 values ranged between 69.12 and 102.82, indicating a large χ^2 and, therefore, poor model fit—desirable values would have a range centered at zero. Given the model specification, one of the causes for model misfit is partly due to unmodeled residual covariances among the observed variables.

These covariances often exist, but in factor analysis they cannot be estimated without creating identification problems because there is not enough information in the data to estimate them. A solution to this is possible in Bayesian analysis by specifying normally-distributed priors for these covariances with a mean of 0.0 and a small variance of .01 to allow estimating a posterior distribution for these parameters even though the mean estimate is set to 0.0. This creative solution is described in Muthén and Asparouhov (2012) and can be applied to underidentified cross-loadings in models with multiple latent factors (see Muthén, 2010) as well as underidentified direct effects. In the case of factor analysis, such a prior specification is a principled way to indicate that, a priori, observed variables are assumed independent conditional on a latent variable, but the possibility exists of small residual covariances.

The model from Zyphur et al. (2009) was re-estimated by setting priors for all residual covariances with a mean of 0.0 and a variance of .01 (see Method 3 in Muthén & Asparouhov, 2012; see Appendix A for program code). The model converged after 27,100 iterations when a *PSR* < 1.05 was reached. Model fit was substantially improved through this respecification, with a *PPP*-value of .70 and a 95% confidence interval of the difference between observed and generated data χ^2 between -24.21 and 12.91, indicating good fit. Also, the *DIC* was 3723.73, while the *DIC* for the above model was 3811.13, meaning a reduction of 87.4, indicating improved fit. As Table 1 shows, the estimated standardized residual covariances are small, and their credibility intervals cover zero, but they are clearly large enough to impact model fit—standardized values are reported to assist in interpreting their magnitude.

Given these results, we proceed and interpret model output, focusing on the testosterone-status effect (for results see Table 1 and Figure 3b). The median of the posterior distribution for this effect is < .001, and the 95% credibility interval ranges between -.003 and .002, meaning there is a 95% chance that the testosterone-status effect is between -.003 and .002 (see Figure 4).

Notice that this last statement is a *direct* probabilistic interpretation of the parameter, and the finding directly supports a hypothesized null (not possible with frequentist *p*-values). We conclude from this analysis that there is virtually no empirical evidence favoring a positive testosterone-status effect, thus providing more support for status characteristics theory than for a dominance theory. This highlights a benefit of Bayesian estimation and inference: Hypothesizing a traditional null as a research hypothesis is straightforward, and posterior probability estimates can provide evidence *in support of* or against such a hypothesis (Kruschke, 2011).

Turning back to the Bayesian analysis, despite the support for no effect, we can go further by examining the probability that the parameter lies in a region supporting a dominance theory, where the testosterone-status effect should be positive. To do this, the probability that the standardized testosterone-status effect (i.e., the correlation) ranges between +.10 and +1.0 was examined by saving the posterior estimates and computing the proportion of values in this range (see Appendix A, which shows the SAVEDATA option in the Mplus language to save posterior estimates). Approximately 0.35% of the parameter's distribution fell between +.10 and +1.0, meaning the probability that the testosterone-status effect is positive and in a range supporting a dominance theory is very low, at .0035. This speaks strongly against a testosterone-oriented theory of dominance. This analysis is helpful because it allows for an intuitive statement about the probability of the parameter existing in a region that is mapped directly onto hypotheses and theory (Berger & Delampady, 1987). Indeed, the finding speaks against a dominance theory much more strongly than did the original frequentist (NHST) results from Zyphur et al. (2009).

Summary. A Bayesian method of estimation has multiple advantages. First, parameters such as residual covariances that are not identified in a frequentist approach can be estimated to represent a more realistic expectation of the measurement characteristics of a scale, improving model fit (Muthén & Asparouhov, 2012, Muthén, 2010). Second, the probability of parameters taking on null values at and around the traditional null value of $\beta = 0$ can be directly investigated to test theories of weak or nil effects, which cannot done with a frequentist (NHST) approach. Further, our Table 1 shows how to present Bayesian results, including priors, median posterior estimates for parameters, and credibility intervals (as well as *PSR*, *PPP*, and *DIC* values).

Example 2: ANOVA Model and a True Experiment

This experiment comes from Zyphur, Warren, Landis, and Thoresen (2007, see Study 2), who investigated the effect of prior self-regulation (i.e., self control) on future performance.

Theory on ego depletion characterizes self-regulatory capacity as a limited resource that takes time to regenerate after use, like a muscle (see Baumeister, 2002). To test this, participants were 80 undergraduate students randomly assigned to a self-regulation condition (n = 40) or a control condition (n = 40). In both conditions, half of the participants watched a movie clip that was either funny (n = 38) or sad (n = 42), and those in the ego depletion condition were instructed to self-regulate by suppressing their emotional reactions to whichever movie they were assigned (see Baumeister, Bratslavsky, Muraven, & Tice, 1998; Muraven, Tice, & Baumeister, 1998).

Subsequently, students participated in a naval combat simulation (see Hollenbeck et al., 1995), where they made decisions about how to respond to incoming targets. There was a training scenario and then 15 additional scenarios. Dependent variables were an overall score representing the number of correct decisions (on a 7-point scale), the amount of time spent on the task (in seconds), and the average number of target attributes measured (between 0 and 9).

Frequentist Analysis. Data from Zyphur et al. (2007) were re-analyzed using OLS regression, which has good small-sample properties for estimating standard errors (Fisher, 1922). The dependent variables (overall score, time taken, and number of attributes measured) were regressed on variables that dummy coded for condition, video type, and their interaction.

Results are in Table 2. There was no main effect of video type or the interaction on the three dependent variables. However, there was a statistically significant effect of condition on score (β = .12, p < .01), with lower scores for those engaging in self-regulation. The effect of condition on the number of attributes measured neared statistical significance, with fewer attributes measured for those engaging in self-regulation (β = .01, p = .08). Because score was the most important outcome variable, these results offer support for the research hypothesis.

However, these findings highlight a shortcoming of frequentist analyses: theory and prior findings exist, but they are not formally incorporated into the analyses themselves. Although

Zyphur et al. (2007) hypothesized that those who engaged in self-regulation in the happy or sad film would tend to have shorter times (i.e., less persistence) and measure fewer attributes in the simulation, these hypotheses are not incorporated into the analyses conducted except by rejecting the null hypothesis; yet the small sample size makes rejecting the null difficult, even though all effects are in the predicted direction. This makes a point about frequentist methods: because they cannot incorporate information from past studies, small-sample research is made difficult.

Also, OLS estimates were computed instead of model-wide maximum likelihood estimates because the asymptotic properties and benefits of maximum likelihood are not present in small samples. The question of what sample size is “large enough” to provide the benefits of such estimation is a topic in many papers (e.g., Anderson & Gerbing, 1984; MacCallum, Browne, & Sugawara, 1996). Relying on OLS estimation here results in little more than having to estimate three regression equations instead of a single multivariate path analysis. However, in other cases, a small sample presents difficulties when researchers want to estimate complex models. Alternatively, Bayesian estimation generally does not suffer the same dilemmas when working with informed prior distributions (Gill, 2008; Howson & Urbach, 1993).

Bayesian Analysis. A meta-analysis by Hagger, Wood, Stiff, and Chatzisarantis (2010) shows that ego depletion has a negative effect on performance. They found an effect size of ego-depletion on effort-related dependent variables of $d = 0.64$, with $\sigma_d^2 = .0256$. Also, this finding was consistent, with little variance in the effect of self-regulation across different types of tasks (e.g, persistence in menial tasks as well as difficult exams). Thus, meta-analytic evidence from prior work offers strong empirical justification that past self-regulation has a consistent negative effect on the future ability to self-regulate—enough evidence to create an informative prior.

To illustrate using informative priors and examine experimental effects with a Bayesian estimator, data from Zyphur et al. (2007) are re-analyzed with a path model in Mplus. Prior distributions for the residual variances of the dependent variables were chosen from previous literature. (See Appendix B for how to parameterize priors for variances.)

Next, the prior distribution of effects of the independent variables and the interaction were specified. Normal distributions for these effects were specified as $\beta \sim N(\mu_\beta, \sigma_\beta^2)$ where "~" means "distributed as" and "*N*" means "normal." Prior work with the same manipulations as Zyphur et al. (2007) shows no effect of video or the interaction between video and condition (see Baumeister et al., 1998; Muraven et al., 1998). Therefore, the priors for video and the interaction term were estimated with a mean of 0, but with a variance of .01 to express some uncertainty in the effects. An interesting feature of this prior is that it is a principled way for researchers to specify and test for a null relationship when they believe that an effect is actually zero (see Edwards et al., 1963). If researchers truly believe in a null, then the observed data (i.e., likelihood) can fight this belief and pull the effect's posterior distribution away from the prior.

Priors for the effect of condition were taken from Hagger et al.'s meta-analysis (2010) and were converted from $d = 0.64$ and $\sigma_d^2 = .0256$—of course, the results from Zyphur et al. (2007) are included in this meta-analysis, but it is still useful as an example for our discussion. These values translate into an unconditional standardized regression coefficient (i.e., correlation) of $\beta = 0.30$ with $\sigma_\beta^2 = .0128$ (for *d*-to-*r* transformations see Rosenthal, Rosnow, & Rubin, 2000). The estimates are correlations because with random assignment to condition and no expected effect of video or the interaction on the dependent variable this *d* is expected to translate directly to a standardized coefficient. To work with the standardized effect of .30 and its variance .0256 in Mplus requires that these be unstandardized, such as with

$$\beta_{\text{unstandardized}} = \beta_{\text{standardized}} \frac{SD_y}{SD_x}, \tag{2}$$

where variables' variances are described in Appendix B. This conversion leads to specifying the prior distribution as normal with a mean effect of condition on score of $\mu_\beta = .06$ with $\sigma_\beta^2 = .005$, a mean effect on time of $\mu_\beta = -4.5$ with $\sigma_\beta^2 = .384$, and a mean effect on number of attributes measured of $\mu_\beta = .03$ with $\sigma_\beta^2 = .00256$ (see MODEL PRIORS in Appendix C).

The model was estimated with 10,000 iterations and two chains. *PSR* values below 1.05 indicated good convergence. The high *PPP*-value of .48 indicates adequate model fit (expected because the model is saturated, with covariance allowed among dependent variables). Recalling this *PPP* value indicates the proportion of χ^2 values indicating better fit to observed versus generated data, it is not surprising that the 95% confidence interval for the difference in these χ^2 values ranged between -17.00 and 16.36 (centered almost exactly at zero, indicating good fit).

Parameter estimates and credibility intervals are in Table 2. Contrasting the Bayes estimates with those from OLS shows the benefits of incorporating previous findings as priors. Specifically, results show that the 95% credibility intervals for the effects of condition on score, time taken, and the number of attributes measured do not contain zero. This is a substantially different outcome than NHST based on OLS estimation, where the null hypothesis was retained. Also, by relying on a meta-analysis for priors, this Bayesian analysis reflects the combination of all prior research on ego depletion with the results from Zyphur et al. (2007). Future research in any field can update existing meta-analytic results in a similar manner using Bayesian analysis.

Summary. The frequentist procedures show one statistically significant effect, even though *all* effects were consistent with theory and hypotheses. By incorporating prior meta-analytic findings into analyses, the accuracy of the estimates increased and credibility intervals and point estimates from the Bayesian analyses directly support proposed hypotheses.

DISCUSSION

Bayesian methods allow for a straightforward test of the probability that parameters lie in a particular region, such as a range associated with a null hypothesis or any other values. This goes beyond frequentist methods using NHST, which limit researchers to descriptions of the probability of data given fixed parameters. Also, Bayesian methods provide a principled way to incorporate prior information into analyses to limit uncertainty in estimates, thus potentially supporting hypotheses that are not supported with frequentist methods—even though data may be consistent with theory and hypotheses. We now discuss Bayesian methods more broadly.

Expertise and the Growth of Knowledge

Organizational scholars have accrued a vast collection of theories and findings that describe and predict many important micro- and macro-level phenomena. From decision making (e.g., Simon, 1967) and sense making (e.g., Weick, 1979) to team functioning (e.g., Thompson, 1967) and institutional processes (e.g., DiMaggio & Powell, 1983), organizational scholars are masters of their content—often when they are most critical (e.g., March, 2003).

With such a large body of expertise, it is unfortunate that researchers have been unable to incorporate this knowledge formally as priors in the frequentist statistical estimates dominating organizational research. Perhaps this is one reason that meta-analysis is so immensely popular: meta-analysis follows the Bayesian spirit of combining prior information for the benefit of current and future research. Indeed, some meta-analytic methods in organizational research even take a Bayesian approach (see Brannick, 2001; Newman, Jacobs, Bartram, 2007; Steel & Kammeyer-Mueller, 2008). Conveniently, meta-analytic results could be updated in real time, if past findings are used as priors for Bayesian analyses, which would allow a straightforward way for organizational scholars to combine past studies with current findings. Further consideration of this Bayesian idea would be needed to make it operational, but if the idea gained traction, it could revolutionize how data across studies are accumulated and reported.

In such cases, the weight of evidence in a new study (the likelihood based on current data) often has to be radically different from the weight of past studies (the prior distribution of parameter estimates) in order to substantially revise knowledge in a domain (create a posterior distribution that meaningfully differs from the prior). As shown in our second example, small-sample research that formally incorporates informative priors can provide better information about the nature of the parameters that symbolize phenomena of interest; with Bayesian analysis, we do not have to start from scratch every time a new study is conducted. This can open the door to small sample research that currently finds a barrier to publication because of low statistical power. Extensive additional work on Bayesian methods is needed to explore how to map theory and findings onto prior distributions to solidify this practice as a mainstream method.

To this end, our example with informative priors shows one method for choosing priors: past findings grounded in theory, as is done in fields like physics and medicine (e.g., Ashby, 2006; Dose, 2003). However, there are ways of choosing priors that represent bad-faith research practices. For example, researchers could investigate their data and then choose priors that agree with the data, so that posterior distributions support their hypotheses; similarly, researchers could choose theory-driven priors that “drive” the posteriors in a way not consistent with observed data. Such cases show two ways for Bayesian analysis to be conducted in bad faith.

Therefore, researchers should be honest about where they get their priors, and without good justifications for a prior, researchers may report results from different priors, including a non-informative (diffuse) prior. Honesty about describing priors in the Bayesian framework mirrors honesty required for good-faith research in a frequentist framework, such as avoiding “hypothesizing after results are known,” or HARKing (Kerr, 1998). In HARKing, frequentists describe what appear to be a priori hypotheses, but they are actually formed after data have been analyzed and examined for their conclusions, raising the danger of capitalizing on chance and biased parameter estimates. In other words, Bayesian priors offer a clear way to act in bad faith, but this already exists in the frequentist approach. Indeed, in any framework involving data, statistics, and probabilistic inference, there is a possibility of practicing science in bad faith.

To guard against dishonesty, however, Bayesian methods offer a unique opportunity: Authors can be transparent about where they have derived priors, and the standards of a profession and its journals can regulate these derivations to fit a researcher's particular context. Future work should address priors in light of the conflict that characterizes empirical scientists' call to be honest *and* to publish or perish. As an example, to derive priors, organizational researchers can consult not only published work, but also their expert colleagues who can provide priors without being biased by any data that have already been collected by a researcher.

Why Bayes and Why Now?

As Edwards et al. (1963) lamented almost 50 years ago in a pro-Bayesian article from *Psychological Review*, "The textbook that would make all the Bayesian procedures mentioned in this paper readily available to experimental psychologists does not yet exist, and perhaps it cannot exist soon" (p. 193). What Edwards et al. could have added here is that difficulties deriving posteriors would limit the application of Bayesian techniques until modern computers and MCMC methods were developed. Alternatively, frequentist estimation methods with known asymptotic (i.e., large-sample) properties made statistical estimation and probabilistic inference easy by comparison. Indeed, as Box and Tiao (1973) note when describing frequentist methods, "the nice solutions… have been popular for another reason—they were easy to compute" (p. 1).

Today, however, Bayesian solutions are equally attainable given modern computer power and software implementing MCMC methods. For example, with problems involving highdimensional numerical integration (as is required with maximum likelihood estimation with categorical observed variables and latent variables), Bayesian estimation can be easily accomplished in cases that are completely intractable using traditional methods (see Muthén & Asparouhov, 2012). Also, as demonstrated previously, Bayesian models and methods are flexible, allowing parameter estimation where frequentist approaches are impossible due to model identification issues. With this flexibility in Bayesian estimation and inference, it is safe to say that the majority of future complex data analytic methods may be Bayesian—indeed, they may *have* to be Bayesian to cope with the increasingly complicated model specifications that correspond with increasingly sophisticated theories that are multilevel and dynamic in nature.

Organizational researchers can take heart in the fact that Edwards et al. (1963) are no longer correct; there is now a wealth of useful Bayesian literature. Discussions that are focused around implementation in Mplus are by Asparouhov and Muthén (2010a, 2010b), Muthén (2010), and Muthén and Asparouhov (2012). An exemplary book is by Gill (2008), but many others are also very good (e.g., Bolstad, 2007; Christensen, Johnson, & Branscum, 2011; Gelman et al., 2004; Hoff, 2009; Jackman, 2009; Kruschke, 2011; Lee, 2004; Lynch, 2010; for concise pro-Bayesian arguments see Robert, 2007, Chapter 11). Further, a truly exceptional introduction to Bayesian scientific reasoning is by Howson and Urbach (1993).

Beyond these texts, there are many accessible treatments of Bayesian methods (e.g., Efron, 2005; Little, 2006) and special issues on Bayesian methods can be found in *Statistical Science* (see Christian & Casella, 2004; Lahiri & Slud, 2011); the *British Journal of Mathematical and Statistical Psychology* (2013, Vol 66, Issue 1); and the journal *Bayesian Analysis* is free of charge (see www.ba.stat.cmu.edu). In other words, organizational researchers can bolster their knowledge with ready access to a healthy and growing Bayesian literature, along with an active and supportive community of researchers and methodologists.

Conclusion

Our paper introduces organizational science to Bayesian methods by contrasting them with the usual logic of NHST and the frequentist methods that have dominated the field since its inception. To be clear, we are not saying that Bayesian analysis is a cure-all for issues in data analysis, but it has literally been called “revolutionary” in many other disciplines due to its flexibility and the range of possible analyses it allows, all combined with the benefit of the intuitive interpretations and inferences that result from such analyses. The field of management should not shy away from promising new methods that offer improvements or complements to traditional frequentist methods that have taken organizational researchers far, but not without limitations. At the very least, diverse qualitative and quantitative methods should be explored when they facilitate addressing organizational issues. To this end, Bayesian analysis now allows researchers to fulfill goals for estimation and inference that simply cannot be accomplished by existing approaches. As the Bayesian revolution unfolds in organization science, we hope it substantially benefits organizational researchers and organizational stakeholders alike.

References

Anderson, J. C., & Gerbing, D. W. 1984. The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika,* 49: 155-173.

Ashby, D. 2006. Bayesian statistics in medicine: A 25 year review. *Statistics in Medicine,* 25: 3589-3631.

Asparouhov, T., & Muthén, B. (2010a). Bayesian analysis using Mplus: Technical implementation. Retrieved from: www.statmodel.com/download/Bayes3.pdf

Asparouhov, T., & Muthén, B. (2010b). Bayesian analysis of latent variable models using Mplus. Retrieved from www.statmodel.com/download/BayesAdvantages18.pdf

Barnard, G. A. (1947). The meaning of a significance level. *Biometrika,* 34: 179-182.

Baumeister, R. F. (2002). Ego depletion and self-control failure: An energy model of the self's executive function. *Self and Identity,* 1: 129-136.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology,* 74: 1252-1265.

Bayes, T. P. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* 53: 370-418.

Berger, J., Cohen, B. P., & Zelditch, M. Jr., (1972). Status characteristics and social interaction. *American Sociological Review,* 37: 241-255.

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis,* 3: 385-402.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science,* 2: 317-335.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association,* 82: 112-122.

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

Bolstad, W. M. 2007. *Introduction to Bayesian statistics* (2nd ed.). New York: Wiley.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford: Oxford University Press.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. New York: Wiley.

Brannick, M. T. (2001). Implications of empirical Bayes meta-analysis for test validation. *Journal of Applied Psychology,* 86: 468-480.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis,* 3: 473-514.

Bunderson, J. S. (2003). Recognizing and utilizing expertise in work groups: A status characteristics perspective. *Administrative Science Quarterly,* 48: 557-591.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician,* 39: 83-87.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Pacific Grove: Wadsworth.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *Journal of the American Statistical Association,* 46: 167-174.

Chib, S., Griffiths, W., Koop, G., & Terrell, D. (2008). *Advances in econometrics: Bayesian econometrics* (Vol. 23). Bingley: JAI Press.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician,* 49: 327-335.

Christensen, R., Johnson, W., & Branscum, A. (2011). *Bayesian ideas and data analysis: An introduction for scientists and statisticians*. Boca Raton: Chapman Hall.

Christian, P. R., & Casella, G. (2004). Introduction to the special issue: Bayes then and now. *Statistical Science,* 19: 1-2.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cumming, G., & Finch, S. (2005). Inference by Eye: Confidence intervals and how to read pictures of data. *American Psychologist,* 60: 170-180.

Curtis, A., & Wood, R. (2004). *Geological prior information*. Bath: Geological Society.

Dabbs, J. M. (1992). Testosterone and occupational achievement. *Social Forces,* 70: 813-824.

Daston, L. (1994). How probabilities came to be subjective and objective. *Historia Mathematica,* 21: 330-344.

Daston, L. (1995). *Classical probability in the enlightenment*. Princeton: Princeton University Press.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review* 48: 147-160.

Dose, V. (2003). Bayesian inference in physics: Case studies. *Reports on Progress in Physics,* 66: 1421-1461.

Eagle, A. (2004). Twenty-one arguments against propensity analyses of probability. *Erkenntnis,* 60: 371-416.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin,* 51: 380-417.

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

Efron, B. (2005). Bayesians, frequentists, and statisticians. *Journal of the American Statistical Association,* 100: 1-5.

Efron, B. (2010). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction*. Cambridge: Cambridge University Press.

Efron, B. (2011). The bootstrap and Markov-chain Monte Carlo. *Journal of Biopharmaceutical Statistics,* 21: 1052-1062.

Efron, B., & Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators—Part II: Empirical Bayes. *Journal of the American Statistical Association*, 67: 130-139.

Fisher, R. A. (1921). On the "probable error" of a correlation coefficient deduced from a small sample. *Metron,* 1: 3-32.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London,* A, 222: 309-368.

Fisher, R. A. (1925). Applications of "Student's" distribution. *Metron,* 5: 90-104.

Fisher, R. A., (1935). *The design of experiments*. Edinburgh: Oliver & Boyd, Ltd.

Galavotti, M. C. (2005). *Philosophical introduction to probability*. Palo Alto: CSLI Publications.

Garrett, E. S., & Zeger, S. L. 2000. Latent class model diagnosis. *Biometrics,* 56: 1055-1067.

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

Gelman, A., & Shalizi, C. R. (2012). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*.

Ghosh, M. (2011). Objective priors: An introduction for frequentists. *Statistical Science,* 26: 187-202.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gill, J. (2008). *Bayesian methods: A social and behavioral science approach* (2nd ed.). Boca Raton: Chapman & Hall.

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology,* 35: 765-775.

Greenland, S. (2007). Bayesian perspectives for epidemiological research. II. Regression analysis. *International Journal of Epidemiology,* 36: 195-202.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin,* 82: 1-20.

Hacking, I. (2001). *An introduction to probability and inductive logic*. Cambridge: Cambridge University Press.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin,* 136: 495-525.

Hájek, A. (1997). 'Mises redux'—redux: Fifteen arguments against finite frequentism. *Erkenntis,* 45: 209-227.

Hájek, A. (2009). Fifteen arguments against hypothetical frequentism. *Erkenntis,* 70: 211-235.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York: Springer.

Hoijtink, H., Klugkist, I., & Boelen, P. A. (Eds.) (2010). *Bayesian evaluation of informative hypotheses*. New York: Springer.

Hollenbeck, J. R., Ilgen, D. R., Douglas, S. J., Hedlund, J., Major, D. A., & Phillips, J. (1995). Multilevel theory of team decision making: Decision performance in teams incorporating distributed expertise. *Journal of Applied Psychology,* 80: 292-316.

Howson, C. (1997). A logic of induction. *Philosophy of Science,* 64: 268-290.

Howson, C. (2001). *Hume's problem: Induction and the justification of belief*. Oxford: Oxford University Press.

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach* (2nd ed.). Peru, Il: Open Court Publishing.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. West Sussex: Wiley.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Jeffreys, H. (1939/1998). *Theory of probability* (3rd ed.). Oxford: Clarendon Press.

Kadane, J. B. (2011). *Principles of uncertainty*. London: Chapman & Hall.

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

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association,* 91: 1343-1370.

Kerr, N. L. (1998). HARKing: Hypothesizing after results are known. *Personality and Social Psychology Review,* 2: 196-217.

Keynes, J. M. (1921/2008). *A treatise on probability*. Rough Draft Printers.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin,* 119: 254-284.

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

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods,* 15: 722-752.

Lahiri, P., & Slud, E. (2011). Special issue on Bayesian methods that frequentists should know. *Statistical Science,* 26: 161-162.

Lancaster, (2004). *Introduction to modern Bayesian econometrics*. New York: Wiley.

Laplace, P.-S. (1774). Mémoire sur la probabilité des causes par les événements. *Mémoires de Mathemhatique et de Physique Presentés à L'academie Royale des Sciences, par Divers Savants, & lûs dans ses Assemblés,* 6: 621-656. [see Laplace, P. S. (1986). Memoir on the probability of the causes of events. *Statistical Science,* 1, 364-378.].

Laplace, P.-S. (1825/1995). Philosophical essay on probabilities (Trans. A. I Dale). New York: Springer-Verlag.

Lee, P. M. (2004). *Bayesian statistics: An introduction* (3rd ed.). London: Wiley.

Lewis, D. A. (1980). A subjectivists guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. II: pp. 263-294). Berkeley: University of California.

Little, R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician,* 60: 213-223.

Lynch, S. M. (2010). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer.

Lynch, S. M., & Western, B. (2004). Bayesian posterior predictive checks for complex models. *Sociological Methods & Research,* 32: 301-335.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power and determination of sample size for covariance structure modeling. *Psychological Methods,* 1: 130-149.

March, J. G. (2003). A scholar's quest. *Journal of Management Inquiry,* 12: 205-207.

Mazur, A. (1985). A biosocial model of status in face-to-face primate groups. *Social Forces,* 64: 377-402.

Muraven, M., Tice, D. M., & Baumeister, R. F. (1998). Self-control as limited resource: Regulatory depletion. *Journal of Personality and Social Psychology,* 74: 774-789.

Muthén, B. (2010). Bayesian analysis in Mplus: A brief introduction. www.statmodel.com/download/IntroBayesVersion%203.pdf

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods,* 17: 313-335.

Muthén, L. K. & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Newman, D. A., Jacobs, R. R., & Bartram, D. (2007). Choosing the best method for local validity estimation: Relative accuracy of meta-analysis versus a local study versus Bayes analysis. *Journal of Applied Psychology,* 92: 1394-1413.

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistics Society,* B: 56, 3-48.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London,* A: 236, 333-380.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society,* A: 231, 289-337.

Orlitzky, M. (2012). How can significance tests be deinstitutionalized. *Organizational Research Methods,* 15: 199-228.

Raftery, A. E. (1992). Bayesian model selection in structural equation models. www.stat.washington.edu/~raftery/Research/PDF/bollen1993.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park: Sage.

Robert, C. P. (2007). *The Bayesian choice: From decision-theoretic foundations to computational implementation* (2nd ed.). New York: Springer.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavior research*. Cambridge: Cambridge University Press.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics,* 9, 130-134.

Savage, L. J. (1954). *Foundations of statistics*. New York: Wiley.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology,* 62: 529-540.

Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. (2011). Researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science,* 22: 1105-1120.

Simon, H. (1967). Motivational and emotional controls of cognition. *Psychological Review,* 74: 29-39.

Steel, P. D. G., & Kammeyer-Mueller, J. D. (2008). Bayesian variance estimation for meta-analysis: Quantifying our uncertainty. *Organizational Research Methods,* 11: 54-78.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge: Belknap Press.

Suppes, P. (2007). Where do Bayesian priors come from? *Synthese,* 156: 441-471.

Thompson, J. D. (1967). *Organizations in action: Social science bases of administrative theory*. New Brunswick: Transaction Publishers.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185: 1124-1131.

von Toussaint, U. (2011). Bayesian inference in physics. *Review of Modern Physics,* 83: 943-999.

Wald, A. (1950). *Statistical decision functions*. New York: Wiley.

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York: Springer.

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. New York: Springer.

Weick, K. E. (1979). *The social psychology of organizing* (2nd ed.). New York: McGraw-Hill.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences*. Oxford: Academic Press.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods,* 14: 301-322.

Zyphur, M. J. (2003). *The effects of prior self-regulation on ability to respond to feedback: An empirical investigation*. Unpublished manuscript.

Zyphur, M. J., Narayanan, J., Koh, G., & Koh, D. (2009). Testosterone-status mismatch lowers collective efficacy in groups: Evidence from a slope-as-predictor multilevel structural equation model. *Organizational Behavior and Human Decision Processes,* 110: 70-79.

Zyphur, M. J., Warren, C. R., Landis, R. S., & Thoresen, C. J. (2007). Self regulation and performance in high-fidelity simulations: An extension of ego-depletion research. *Human Performance,* 20: 103-118.

Table 1

Quasi-Experiment with Null as a Research Hypothesis

Frequentist Analysis with Maximum Likelihood Estimator

Effect	-2.5%	β	+2.5%	*p*-value
Factor Loading				
Y1	-----	1.00	-----	-----
Y2	0.90	0.94	0.97	<.01
Y3	0.72	0.76	0.79	<.01
Y4	0.80	0.85	0.88	<.01
Y5	0.64	0.68	0.72	<.01
Regression Effects				
T→Status	-0.004	-0.001	0.002	.71

Bayesian Analysis with Uninformative Priors

Effect	*Prior*		*Posterior*	
	$\mu\ (\sigma^2)$	-2.5%	μ_β	+2.5%
Factor Loading				
Y1	$0.00(10^{10})$	-----	1.00	-----
Y2	$0.00(10^{10})$	0.85	0.94	1.03
Y3	$0.00(10^{10})$	0.67	0.76	0.84
Y4	$0.00(10^{10})$	0.75	0.84	0.93
Y5	$0.00(10^{10})$	0.61	0.68	0.75
Regression Effects				
T→Status	$0.00(10^{10})$	-0.003	-0.001	0.002
Standardized Residual Covariances				
Y1←→Y2	0.00(.01)	-0.02	0.04	0.12
Y1←→Y3	0.00(.01)	-0.04	0.03	0.11
Y1←→Y4	0.00(.01)	-0.04	0.03	0.10
Y1←→Y5	0.00(.01)	-0.02	0.03	0.09
Y2←→Y3	0.00(.01)	-0.03	0.02	0.08
Y2←→Y4	0.00(.01)	-0.02	0.05	0.13
Y2←→Y5	0.00(.01)	-0.04	0.03	0.06
Y3←→Y4	0.00(.01)	-0.04	0.04	0.12
Y3←→Y5	0.00(.01)	-0.02	0.04	0.10
Y4←→Y5	0.00(.01)	-0.06	0.01	0.09

Table 2

Experiment with Random Assignment

Frequentist Analysis with OLS Estimator

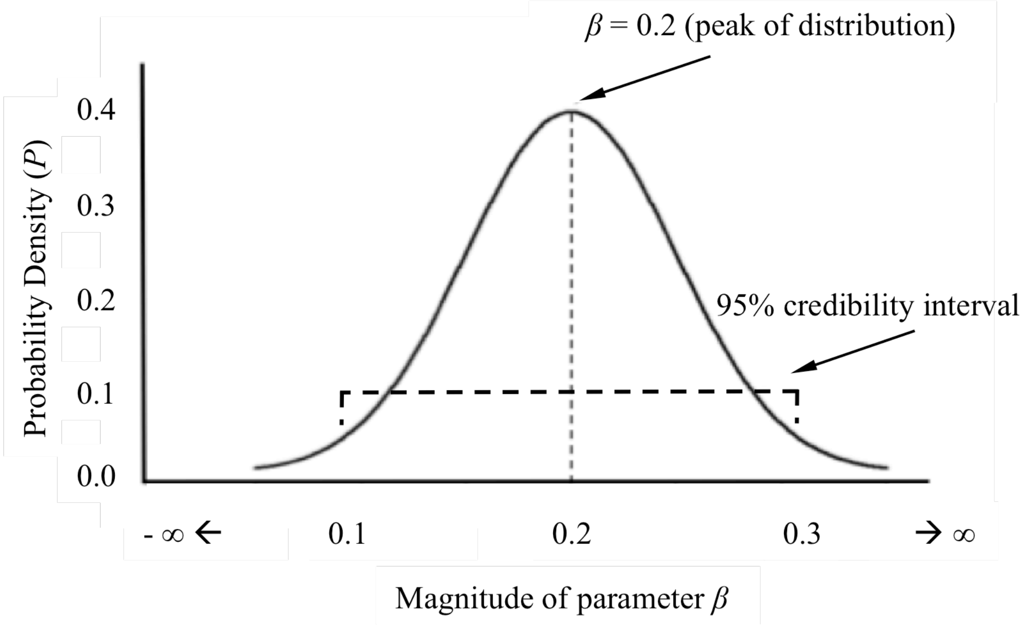
Dependent variable	*Predictor*	-2.5%	β	+2.5%	*p*-value
Overall score	Condition	0.03	0.12	0.22	.01
	Video	-0.04	0.05	0.15	.29
	Interaction	-0.12	-0.03	0.07	.56
Time remaining	Condition	-4.89	-1.57	1.75	.35
	Video	-1.76	1.56	4.87	.35
	Interaction	-1.88	1.44	4.75	.39
Attributes measured	Condition	-0.01	0.01	0.02	.08
	Video	-0.01	0.01	0.01	.64
	Interaction	-0.01	0.01	0.02	.46

Bayesian Analysis with Informative Priors

Dependent variable	*Predictor*	*Prior*		*Posterior*	
		$\mu\ (\sigma^2)$	-2.5%	μ_β	+2.5%
Overall score	Condition	0.06(0.005)	0.02	0.10	0.18
	Video	0.00(.01)	-0.04	0.04	0.13
	Interaction	0.00(.01)	-0.11	-0.03	0.06
Time remaining					
	Condition	-4.50(.384)	-5.25	-4.16	-2.98
	Video	0.00(.01)	-0.20	0.01	0.20
	Interaction	0.00(.01)	-0.19	0.01	0.20
Attributes measured	Condition	0.03(.003)	0.01	0.01	0.02
	Video	0.00(.01)	-0.01	0.01	0.02
	Interaction	0.00(.01)	-0.01	0.01	0.02

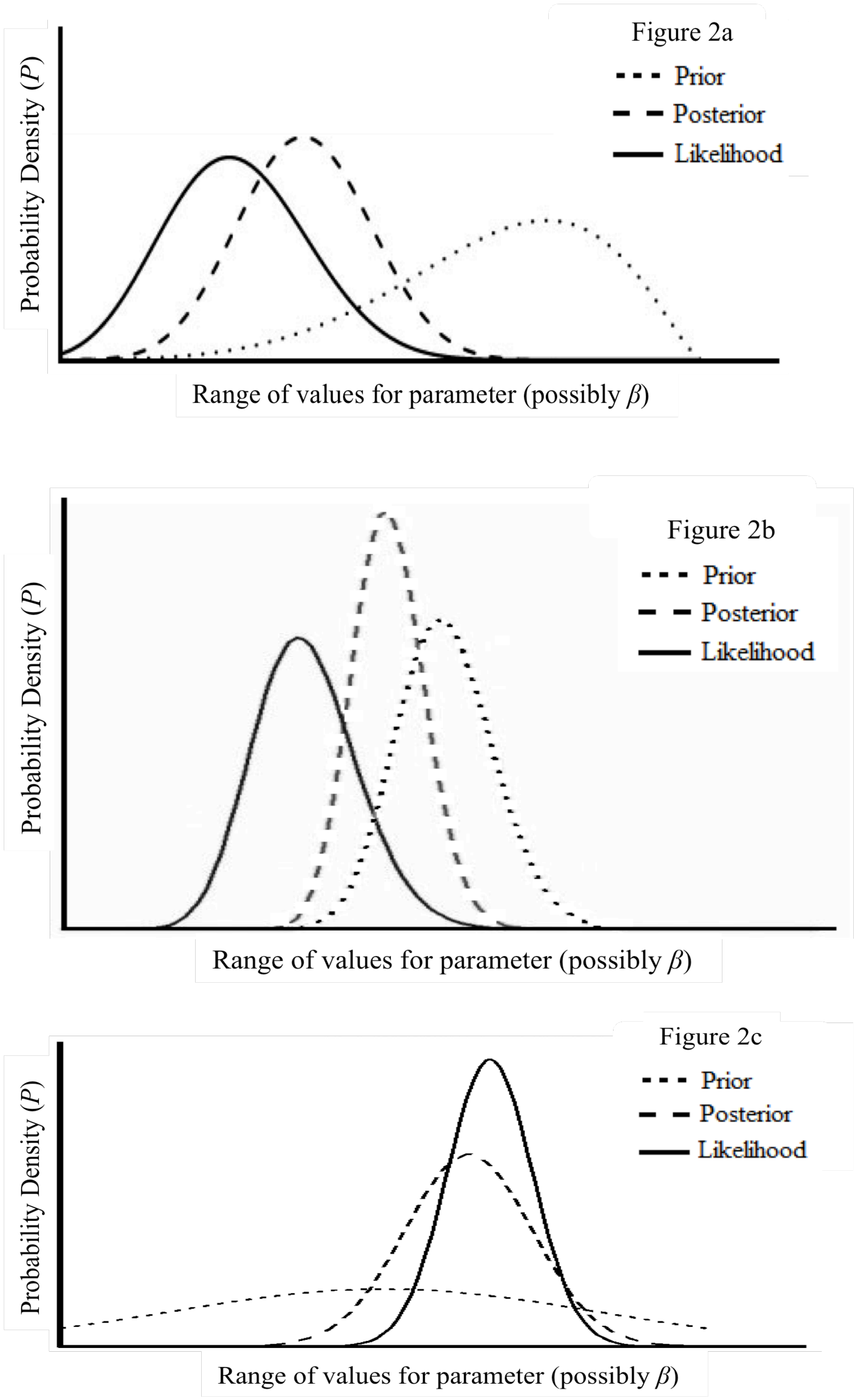
Figure 1

Hypothetical Probability Distribution for a Regression Coefficient *β*



Figures 2a-2c

Hypothetical Probability Density Distributions of Priors, Likelihoods, and Posteriors



Figures 3a-3b

Results from Frequentist Analysis (Top) and Bayesian Analysis (Bottom)

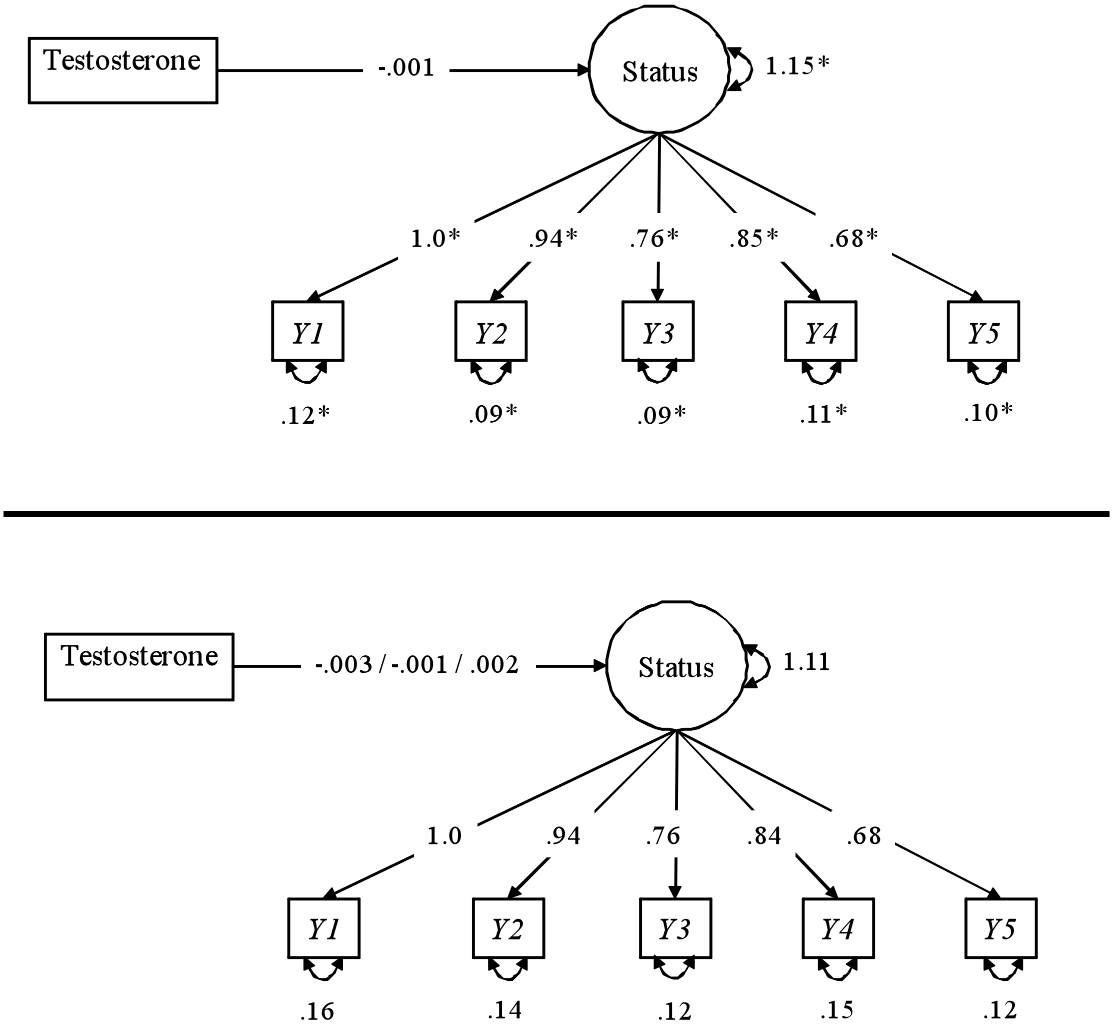
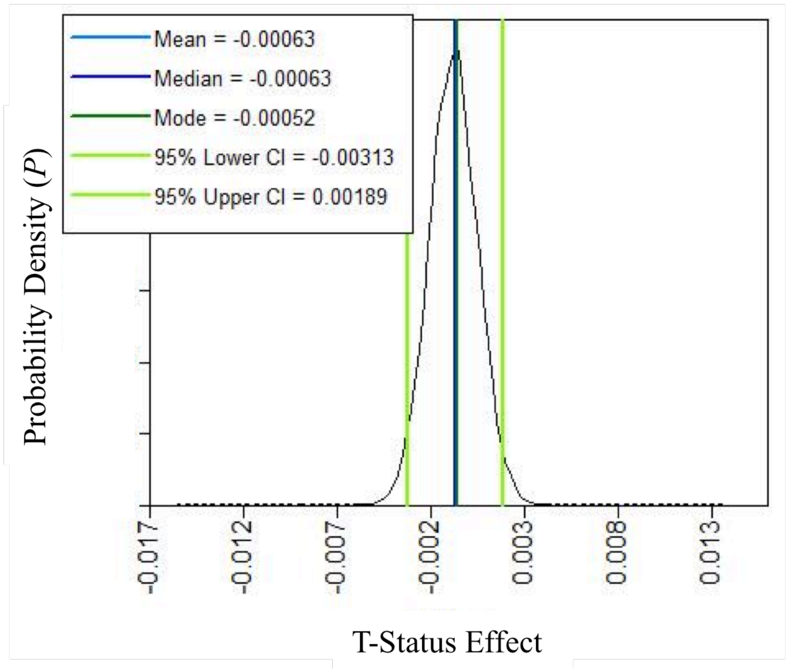


Figure 4

Credibility Interval for Testosterone-Status Effect



APPENDIX A

```
DATA:
File = data.dat; ! Names data file

VARIABLE:
Names = Y1-Y5 T; ! Variable names, Y1-Y5 are status indicators, T is testosterone

ANALYSIS:
Estimator = Bayes;! Requests Bayesian estimation

MODEL:
Status BY Y1-Y5; ! Defines status factor
Status ON T;     ! Regresses latent status variable on testosterone

leadt1_y WITH convt1_y (RESCOV1);     ! Names covariances RESCOV1-10
leadt1_y WITH inft1_y (RESCOV2);
leadt1_y WITH domt1_y (RESCOV3);
leadt1_y WITH powt1_y (RESCOV4);
convt1_y WITH inft1_y (RESCOV5);
convt1_y WITH domt1_y (RESCOV6);
convt1_y WITH powt1_y (RESCOV7);
inft1_y WITH domt1_y (RESCOV8);
inft1_y WITH powt1_y (RESCOV9);
domt1_y WITH powt1_y (RESCOV10);

MODEL PRIORS:
RESCOV1 ~ N(0,.01); ! Specifies normal distribution with mean 0, variance .01
RESCOV2 ~ N(0,.01);
RESCOV3 ~ N(0,.01);!
RESCOV4 ~ N(0,.01);
RESCOV5 ~ N(0,.01);
RESCOV6 ~ N(0,.01);
RESCOV7 ~ N(0,.01);
RESCOV8 ~ N(0,.01);
RESCOV9 ~ N(0,.01);
RESCOV10 ~ N(0,.01);

OUTPUT:
TECH1           ! Requests parameter matrices
CINTERVAL(hpd); ! Requests credibility intervals with highest posterior density

SAVEDATA:
BPARAMETERS = savedata.dat; ! Saves posterior estimates to file savedata.dat

PLOT:
Type = Plot2; ! Requests plots to assess estimation and posterior distributions
```

APPENDIX B

Prior distributions of variances are parameterized with inverse gamma (*IG*) distributions (e.g., Asparouhov & Muthén, 2010a; Muthén, 2010), which have positive density from $0.0 \rightarrow \infty$ and are characterized by two parameters: a shape parameter α and a scale parameter β, such that $\sigma_Y^2 \sim IG(\alpha,\ \beta)$. In their Appendix A, Asparouhov and Muthén (2010b) note that

$$\alpha = 2 + \frac{m^2}{v} \quad \text{and} \quad \beta = m + \frac{m^3}{v}$$

where *m* is the mean of the prior ($\mu_{\sigma_Y^2}$) and v is the variance ($\sigma_{\sigma_Y^2}^2$).

Previous literature reports variances for participants' scores on the naval simulation of around $\mu_{\sigma_Y^2} = .04$, with virtually no variance in the estimates across studies (e.g., Hollenbeck et al., 1995). Yet, as noted by Zyphur et al. (2007), the simulation was designed specifically for their purposes. Therefore, priors come from a mixture of past findings, with variation in the prior added to reflect possible differences between the procedures used in previous research and those used by Zyphur et al. (2007), such that $\sigma_{\sigma_Y^2}^2 = .02$. These translate to *IG*(2.08, .043).

In terms of the time it took participants, past research does not report this. However, in an unpublished master's thesis, Zyphur (2003) notes that the scenarios were designed to take approximately 90 seconds +/- 30 seconds. As such, an *SD* of 15 seconds was chosen to approximate a 30 second interval, translating to $\mu_{\sigma_Y^2} = 225$, with variation around this estimate $\sigma_{\sigma_Y^2}^2 = 100$ to represent uncertainty in the estimate. These translate to *IG*(508.25, 114141.3).

The number of attributes measured is crucial (Zyphur et al., 2007). Therefore, very little variation in this variable is expected, with a point estimate chosen of $\mu_{\sigma_Y^2} = .01$ and a variance $\sigma_{\sigma_Y^2}^2 = .001$. These translate into *IG*(2.1, .011).

APPENDIX C

```
DATA:
File is data.dat;     ! Names data file

VARIABLE:
Names = video condition X score time measure;     ! Names independent variables
! video, condition, their interaction X, and dependent variables score, time taken, and
! the number of target attributes measured

ANALYSIS:
Estimator = Bayes;     ! Requests Bayesian estimation
FBIterations = 10000;     ! Requests 10,000 iterations

MODEL:
score (sVAR);     ! Names variance of score sVAR
time (tVAR);     ! Names variance of time tVAR
measure (mVAR);     ! Names variance of measure mVAR
score ON condition (sONc);     ! Names effect of condition on score sONc
time ON condition (tONc);     ! Names effect of condition on time tONc
measure ON condition (mONc);     ! Names effect of condition on measure mONc
score ON video (CONTROL1);     ! Names effect of video on score CONTROL1
time ON video (CONTROL2);     ! Names effect of video on time CONTROL2
measure ON video (CONTROL3);     ! Names video effect on measure CONTROL3
score ON X (CONTROL4);     ! Names effect of interaction on score CONTROL4
time ON X (CONTROL5);     ! Names effect of interaction on time CONTROL5
measure ON X (CONTROL6);     ! Names interaction effect on measure CONTROL6

MODEL PRIORS:
sVAR ~ IG(2.08, .043);     ! Sets inverse gamma parameters for shape, scale
tVAR ~ IG(508.25, 114141.3);     ! Sets inverse gamma parameters for shape, scale
mVAR ~ IG(2.1, .011);     ! Sets inverse gamma parameters for shape, scale
CONTROL1-CONTROL6 ~ N(0, .01);     ! Sets mean, variance effect of controls
sONc ~ N(.06, .005);     ! Sets effect of condition on score mean, variance
tONc ~ N(-4.5, .384);     ! Sets effect of condition on time mean, variance
mONc ~ N(.03, .00256);     ! Sets effect of condition on measure mean, variance

OUTPUT:
TECH1     ! Requests parameter matrices
CINTERVAL(hpd);     ! Requests credibility intervals with highest posterior density

PLOT:
Type = Plot2;     ! Requests plots to assess estimation and posterior distributions
```

Author/s:
Zyphur, MJ; Oswald, FL

Title:
Bayesian Estimation and Inference: A User's Guide

Date:
2015-02-01

Citation:
Zyphur, M. J. & Oswald, F. L. (2015). Bayesian Estimation and Inference: A User's Guide. JOURNAL OF MANAGEMENT, 41 (2), pp.390-420. https://doi.org/10.1177/0149206313501200.

Persistent Link:
http://hdl.handle.net/11343/247891