Organizational Research Methods
15(4) 722-752
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1094428112457829
http://orm.sagepub.com

The Time Has Come: Bayesian Methods for Data Analysis in the Organizational Sciences

John K. Kruschke[1], Herman Aguinis[2], and Harry Joo[2]

Abstract

The use of Bayesian methods for data analysis is creating a revolution in fields ranging from genetics to marketing. Yet, results of our literature review, including more than 10,000 articles published in 15 journals from January 2001 and December 2010, indicate that Bayesian approaches are essentially absent from the organizational sciences. Our article introduces organizational science researchers to Bayesian methods and describes why and how they should be used. We use multiple linear regression as the framework to offer a step-by-step demonstration, including the use of software, regarding how to implement Bayesian methods. We explain and illustrate how to determine the prior distribution, compute the posterior distribution, possibly accept the null value, and produce a write-up describing the entire Bayesian process, including graphs, results, and their interpretation. We also offer a summary of the advantages of using Bayesian analysis and examples of how specific published research based on frequentist analysis-based approaches failed to benefit from the advantages offered by a Bayesian approach and how using Bayesian analyses would have led to richer and, in some cases, different substantive conclusions. We hope that our article will serve as a catalyst for the adoption of Bayesian methods in organizational science research.

Keywords

quantitative research, computer simulation procedures (e.g., Monte Carlo, bootstrapping), quantitative research, multilevel research

The use of Bayesian methods for data analysis is creating a revolution in fields ranging from genetics to marketing. The magnitude of this change from traditional methods is suggested by the following article titles: "Bayesian Computation: A Statistical Revolution" (Brooks, 2003), "The Bayesian

[1]Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA
[2]Department of Management and Entrepreneurship, Kelley School of Business, Indiana University, Bloomington, IN, USA

Corresponding Author:
John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington, IN 47405-7007, USA
Email: JohnKruschke@gmail.com

Revolution in Genetics'' (Beaumont & Rannala, 2004), ''A Bayesian Revolution in Spectral Analysis'' (Gregory, 2001), and ''The Hierarchical Bayesian Revolution: How Bayesian Methods Have Changed the Face of Marketing Research'' (Allenby, Bakken, & Rossi, 2004). A primary reason for this Bayesian revolution is that traditional data analysis methods (e.g., maximum likelihood estimation, MLE) that rely on null hypothesis significance testing (NHST) have several known problems (Cashen & Geiger, 2004; Cortina & Folger, 1998; Cortina & Landis, 2011; Lance, 2011). Specifically, what researchers want to know is the parameter values that are credible, given the observed data. In particular, researchers may want to know the viability of a null hypothesis (e.g., zero correlation between two variables or zero difference in mean scores across groups) given the data, $p(Ho|D)$. However, traditional methods based on NHST, also labeled *frequentist* statistics to distinguish them from *Bayesian* statistics, tell us the probability of obtaining the data in hand, or more extreme unobserved data, if the null hypothesis were true, $p(D|Ho)$ (Aguinis, Werner, et al., 2010). Unfortunately, $p(Ho|D) \neq p(D|Ho)$. As noted by Cohen (1994), a test of statistical significance ''does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!'' (p. 997). In short, what we want to know is the credibility of candidate parameter values given the data that we actually observed.

Bayesian methods provide an alternative to the widely criticized traditional NHST-based methods because they provide the information we seek. Moreover, Bayesian methods are available for numerous data-analytic applications, including those that have been documented to be the most frequently used in organizational science research (Aguinis, Pierce, Bosco, & Muslin, 2009; Scandura & Williams, 2000). In addition, Bayesian methods are particularly useful for estimating parameter values in nonlinear hierarchical models that are flexibly adapted to the specifics of research and application contexts (e.g., Pierce & Aguinis, in press). Bayesian methods thereby open the door to extensive new realms of modeling possibilities that were previously inaccessible. As such, if embraced by organizational researchers, we believe that the use of Bayesian methods is likely to lead to important theoretical advancements and subsequent practical applications.

As noted earlier, Bayesian methods are taking hold across a broad range of scientific disciplines. Before we began writing our article, we suspected that the organizational sciences have not kept abreast of the Bayesian revolution. To formally assess the veracity of our presumption, we conducted a literature review to examine the extent to which Bayesian methods have been used in published articles in the organizational sciences between January 2001 and December 2010. Our review focused on 15 organizational science journals based on their visibility and impact. We used the database Business Source Premier and the keywords *Bayes* and *Bayesian* using the full-text search option.

Results of our review are summarized in Table 1 and indicate that Bayesian methods are essentially absent from the organizational science literature. Out of the more than 10,000 articles published in the 15 journals included in our 10-year review, only 42 used Bayesian methods. In other words, fewer than half of 1% of articles in the review period used Bayesian approaches. Only two journals published more than 1% of articles using Bayesian methods: *Management Science* (about 2%) and *Organizational Research Methods* (*ORM*) (about 1.4%). The publication of articles on newer methodological approaches such as Bayesian methods in *ORM* is not entirely surprising given findings based on a content analysis of all *ORM* articles published from 1998 through 2007 (Aguinis et al., 2009). Aguinis et al. (2009) concluded that although more traditional methodological approaches are still popular, several novel approaches have become at least as popular in *ORM* in a very short time. Stated differently, researchers focusing on methodological issues and publishing in *ORM* are more likely to become interested in and investigate newer methodological approaches compared with substantive researchers who publish their work in nonmethodological journals. Nevertheless, Table 1 shows that, for all practical purposes, Bayesian approaches are not currently in use in the organizational behavior/human resource management (e.g., *Journal of Applied Psychology*),

Table 1. Number of Articles Using Bayesian Methods in Selected Organizational Science Journals (2001-2010)

Journal	Total Number of Articles Published	Number (%) of Articles Using Bayesian Methods
Academy of Management Journal	736	1 (0.14)
Administrative Science Quarterly	587	1 (0.17)
Industrial & Labor Relations Review	651	0
Journal of Applied Psychology	1,082	3 (0.28)
Journal of Business Venturing	429	0
Journal of International Business Studies	673	1 (0.15)
Journal of Management	486	0
Journal of Organizational Behavior	589	0
Leadership Quarterly	642	0
Management Science	1,407	29 (2.06)
Organization Science	544	0
Organizational Behavior & Human Decision Processes	518	0
Organizational Research Methods	364	5 (1.37)
Personnel Psychology	933	0
Strategic Management Journal	731	2 (0.27)
Total	10,372	42 (0.40)

strategy (e.g., *Strategic Management Journal*), and entrepreneurship (e.g., *Journal of Business Venturing*) literatures, to mention a few research domains. In short, the organizational sciences do not seem to be taking advantage of the Bayesian revolution taking place in many other scientific fields. We believe that the time has come for the organizational sciences to consider the use of Bayesian techniques, which yield richer inferences than do traditional methods that rely on NHST, *p* values, and confidence intervals.

The goal of our article is to introduce organizational researchers to Bayesian methods and describe why and how they should be used. Our article is organized as follows. First, we offer a brief description of Bayesian data analysis in general. Second, we use multiple linear regression as the framework to offer a step-by-step demonstration, including the use of software, regarding how to implement Bayesian methods. This demonstration includes complete details regarding how to execute and report the Bayesian analysis, including how to determine the prior distribution, how to compute the posterior distribution, and how to produce a write-up describing the entire Bayesian process, including graphs, results, and their interpretation. Third, we offer a summary of the advantages of using Bayesian analysis and examples of how specific published research based on frequentist analysis-based approaches failed to benefit from the advantages offered by a Bayesian approach and how using Bayesian analyses would have led to richer and, in some cases, different substantive conclusions. We hope that our article will serve as a catalyst for the adoption of Bayesian methods in organizational science research.

Bayesian Data Analysis: Rationale and Brief Overview

Bayesian analysis determines what can be inferred about parameter values given the actually observed data. Bayesian analysis is the mathematically normative way to reallocate credibility across parameter values as new data arrive.

Reallocation of credibility across possible causes is common in everyday reasoning. For example, suppose there are two unaffiliated suspects for a crime, and strong evidence implicates one suspect. We infer that the other suspect is exonerated. Thus, data that increase culpability of one suspect produce a reallocation of culpability away from the other suspect. The complementary reallocation is

also common, as can be paraphrased from the fictional detective Sherlock Holmes: When you have eliminated all other possibilities, then whatever remains, no matter how improbable, must be the truth (Doyle, 1890). In this case, data that reduce the credibility of some options increase the credibility of the remaining options, even if the prior credibility of those options was small. When the reallocation of credibility is done in the mathematically correct way, then it is Bayesian. Dienes (2011) showed that scientific intuitions about the interpretation of data match the results from Bayesian analysis.

Formal Bayesian data analysis begins with a descriptive model, just as in classical statistics. The descriptive model has meaningful parameters that describe trends in the data. Unlike classical methods, Bayesian analysis yields a complete distribution over the joint parameter space, revealing the relative credibility of all possible combinations of parameter values. Decisions about parameter values of special interest, such as zero, can be made directly from the derived distribution.

Bayesian analysis is named after Thomas Bayes, who discovered a simple mathematical relation among conditional probabilities that is now known as Bayes' rule (Bayes & Price, 1763). When the rule is applied to parameters and data, it can be written as follows:

$$\underbrace{p(\theta|D)}_{\text{posterior}} = \underbrace{p(D|\theta)}_{\text{likelihood}}\ \underbrace{p(\theta)}_{\text{prior}} \Big/ \underbrace{p(D)}_{\text{evidence}}, \tag{1}$$

where D is the observed data and θ is a vector of parameters in the descriptive model. The posterior distribution, $p(\theta|D)$, specifies the relative credibility of every combination of parameters given the data. Because the range of parameter values defines the complete space of possible descriptions, the distribution of credibility sums to 1 and is tantamount to a probability distribution. The posterior distribution provides the most complete information that is mathematically possible about the parameter values given the data (unlike the point estimate and confidence interval in classical statistics, which provide no distributional information, as will be explained later). To make this abstraction concrete, we proceed now to the most common data-analytic application in organizational science research over the past three decades (Aguinis et al., 2009), multiple linear regression.

Bayesian Multiple Linear Regression

Consider a common situation in organizational research, where we are interested in predicting an outcome based on three predictors. For concreteness, the outcome is job performance and the predictors are general mental ability (GMA), conscientiousness, and biodata (i.e., collected using a biographical inventory). To use a realistic illustration, we generated data regarding each of these variables for 346 individual workers from meta-analytically derived population correlations reported by Roth, Switzer, Van Iddekinge, and Oh (2011) (the appendix provides details for accessing the data file and conducting the analysis). All four variables were generated from a multivariate normal distribution and were then rounded to the nearest Likert scale value from 1 to 7.

Job performance, denoted y, is randomly distributed around a linear function of the other three variables, denoted x_1 (GMA), x_2 (conscientiousness), and x_3 (biodata). Formally, the relation is

$$p(y_i|\hat{y}_i, \sigma) = N(\hat{y}_i, \sigma) \text{ and } \hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \tag{2}$$

where $\hat{y}_i$ is the predicted value of y_i and where the subscript i denotes values for the ith individual.

Notice that five parameters must be estimated from the data: β_0 (i.e., the intercept), β_1 (i.e., the regression coefficient for GMA), β_2 (i.e., the regression coefficient for conscientiousness), β_3 (i.e., the regression coefficient for biodata), and σ (i.e., the standard deviation for job performance

scores). Equation 2 expresses the probability of observed data values given any candidate values of the parameters and thereby constitutes the likelihood function in Bayes' rule shown in Equation 1.

Establishing the Prior Distribution

Recall that Bayesian analysis reallocates credibility across candidate parameter values. Therefore, we must establish a prior distribution of credibility on the parameter values, without the newly considered data. When there is little publicly agreed prior knowledge about the parameters, then the prior distribution can be very broad so that no parameter value is given much more credence than any other parameter value. If, however, previous research provides clear guidance regarding plausible parameter values, then the prior distribution can be specified to place more credibility on the plausible parameter values than on the implausible parameter values. Crucially, the prior distribution cannot be set capriciously to favor a researcher's subjective and idiosyncratic opinion. The prior distribution is explicitly specified and justified for a skeptical scientific audience. When skeptics disagree about the appropriateness of a prior distribution, then a noncommittal broad prior distribution can be used. Another useful procedure is to conduct the analysis with more than one prior distribution to demonstrate that the posterior distribution is essentially invariant under reasonable changes in the prior. In typical analyses, a noncommittal broad prior is used, and therefore, the posterior distribution is very robust.

Even though the prior distribution is often selected to be noncommittal, this does not imply that the prior distribution is an inconvenient nuisance for which a researcher must apologize. To the contrary, in many applications, a well-informed prior distribution can provide inferential leverage. As a simplistic example, consider the goal of estimating the possible bias of a coin. Suppose we flip the coin once and it comes up heads. If we have little prior knowledge about the coin, then the single flip tells us only very *un*certainly that the coin might be somewhat head biased. If, however, we have prior knowledge that the coin came from a magic shop and must be either an extremely head-biased coin or an extremely tail-biased coin, then the single flip tells us with high certainty that the coin is of the extremely head-biased type, because the extremely tail-biased coin would almost never exhibit even a single head. Because Bayesian analysis is able to incorporate prior knowledge when it is appropriate, Bayesian analysis is consistent with the epistemological position that organizational science theories will advance to the extent that we engage in empirical research that relies on the accumulation of knowledge (Schmidt, 2008).

Not only can prior knowledge be useful, but ignoring it can cause a derailment in the accumulation of knowledge as well as ineffective practical applications. Consider the use of drug screening as part of the preemployment testing process. Suppose we have a drug test that correctly detects drug use 95% of the time and gives false alarms only 5% of the time. Suppose we select a job applicant at random and the test result is positive. What is the probability that the person uses the drug? An answer that ignores the base rate of drug use might be near the detection rate of 95%. But the correct answer takes into account the base rate of drug usage. If the base rate in the population is 5%, then the actual probability of drug use in the randomly tested person is only 50%, not 95%. This follows directly from Bayes' rule:

$$
\begin{aligned}
p(\theta = \text{user}|D = +) &= \frac{p(D = +|\theta = \text{user})p(\theta = \text{user})}{p(D = +)} \\
&= \frac{p(D = +|\theta = \text{user})p(\theta = \text{user})}{p(D = +|\theta = \text{user})p(\theta = \text{user}) + p(D = +|\theta = \text{nonuser})p(\theta = \text{nonuser})} \\
&= \frac{0.95 \times 0.05}{0.95 \times 0.05 + 0.05 \times 0.95} = 0.50.
\end{aligned}
$$

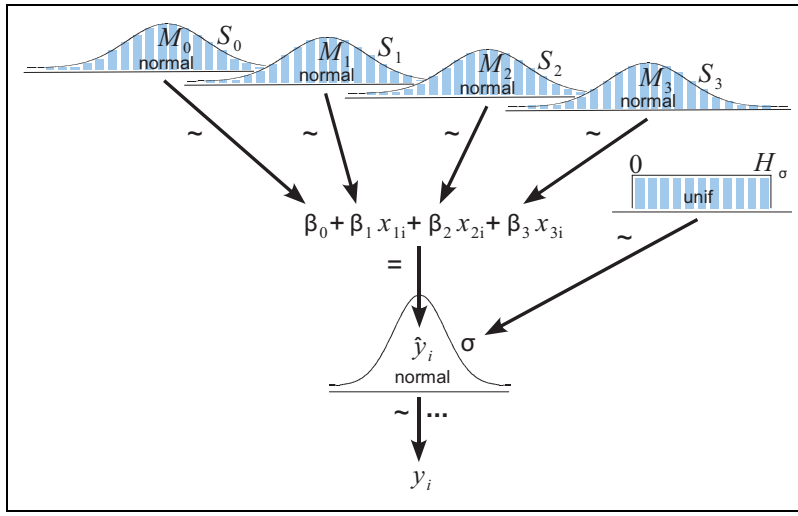

Figure 1. Hierarchical diagram for multiple linear regression. The prior distribution has histogram bars superimposed to indicate correspondence with posterior distributions shown in subsequent figures.

In this situation, the parameter θ being estimated is the person's drug use, which has two nominal values, θ = user or θ = nonuser. The base rate is the prior distribution over the two values of the parameter—namely, $p(\theta = \text{user}) = .05$ and $p(\theta = \text{nonuser}) = .95$. The prior distribution in this scenario has been established from extensive previous research. Not to use the well-informed prior would be a mistake because it would unfairly deny employment opportunities to job applicants. In any other domain of scientific inference, if we have a well-informed prior distribution, not to use it could also be a costly mistake.

For our specific application to multiple linear regression, although we generated data from an established substantive research domain (i.e., human resource selection), our goal is to provide a generic example without specific prior commitments, and therefore we use a noncommittal, vague prior. For each regression coefficient, the prior distribution is a very broad normal distribution, with a mean of zero and a standard deviation that is extremely large relative to the scale of the data. The same assumption is made for the prior on the intercept. Finally, the prior on the standard deviation of the predicted value is merely a uniform distribution extending from zero to an extremely large value far beyond any realistic value for the scale of the data. In the specific analyses demonstrated in this section of our article, the data were standardized so that the prior would be broad regardless of the original scale of the data. The analysis results were then simply algebraically transformed back to the original scale. For the standardized data, the prior on the intercept and regression coefficients was a normal distribution with mean at zero and standard deviation of 100. This normal distribution is virtually flat over the range of possible intercepts and regression coefficients for standardized data. The prior on the standard deviation parameter (σ in Equation 2) was a uniform distribution from zero to 10, which again far exceeds any reasonable value for σ in standardized data. Thus, the prior places essentially no bias on the posterior distribution.

To facilitate our presentation, the full model is illustrated graphically in Figure 1. The lower part of the diagram illustrates the likelihood function of Equation 2: The arrow to y_i indicates that the data are normally distributed with mean $\hat{y}_i$ and standard deviation σ. The arrow pointing to $\hat{y}_i$ indicates that the predicted value equals a linear function of the predictors. The upper part of the diagram illustrates the prior distribution. For example, in the upper left, it is shown that the intercept β_0 has a

normal prior with mean M_0 and standard deviation S_0. The prior distribution is a joint distribution across the five-dimensional parameter space, defined here for convenience as the product of five independent distributions on the five parameters.

The prior distribution is a continuous mathematical function indicated by the black curves, but it is illustrated with superimposed histograms because the parameter distribution will be represented by using a very large (e.g., 250,000) representative random sample from the parameter space. Thus, the Bayesian analysis will reallocate the very large set of representative parameter values from the prior distribution to a posterior distribution, illustrated by histograms of the representative values. For any fixed set of data and prior distribution, there is one true posterior distribution, represented by a very large representative sample of parameter values. The larger the representative sample, the higher resolution picture we have of the true posterior.

In summary, Bayesian inference involves a reallocation of credibility across the space of parameter values. The reallocation is based on the actually observed data, not on imagined data that might have been obtained from a null hypothesis if the intended sampling were repeated. The reallocation starts from a prior distribution. As noted earlier, the prior distribution is not capricious and must be explicitly reasonable to a skeptical scientific audience.

Computing the Posterior Distribution

In Bayes' rule shown in Equation 1, the likelihood function and prior distribution have mathematical forms. The two mathematical forms are multiplied together in the numerator. We use the term *evidence* to refer to the denominator of Bayes' rule, $p(D)$, which is also known as the marginal likelihood. Computing the value of $p(D)$ can be difficult because it is actually an integral over the parameter space: $p(D) = \int p(D|\theta)p(\theta)d\theta$. For many years, Bayesian analysis was confined to mathematical forms that could be analytically integrated or analytically approximated. Fortunately, new computer-based methods have emerged that allow Bayesian analysis to be computed flexibly and for very complex models.

The new method is called Markov chain Monte Carlo (MCMC). The idea is to accurately approximate the posterior distribution by a very large representative random sample of parameter values drawn from the posterior distribution. From this very large sample of representative parameter values, we can determine the mean or modal parameter value, the quantiles of the parameter distribution, its detailed shape, and the forms of trade-offs between values of different parameters. What makes this approach possible is that MCMC methods generate the random sample without needing to compute the difficult integral for $p(D)$. Moreover, recent advances in algorithms and software have made it possible for a researcher merely to specify the form of the likelihood function and prior distribution, and the software is able to apply MCMC methods.

The hierarchical diagram in Figure 1 contains all the information that must be communicated to a computer program so that MCMC sampling can be conducted. Every arrow in the hierarchical diagram has a corresponding declaration in the software for Bayesian analysis, which is explained in more detail in the appendix. The program has been packaged such that all the user needs to do is type three simple commands. One command loads the data, a second command creates the MCMC chain, and a third command creates graphs of the posterior distribution.

The MCMC chain provides a large sample of credible *combinations* of parameter values in the five-dimensional parameter space. We examine various one-dimensional projections of the posterior. Figure 2 shows the posterior distribution for data that have been standardized, and Figure 3 shows the posterior distribution on the original data scale. The middle row of Figure 2 shows the three regression coefficients. Notice that the posterior provides an explicit distribution of the credibility (i.e., probability) of each possible value for the regression coefficient. The upper right panel of Figure 2 shows the posterior for the intercept, and the upper middle panel shows the posterior for

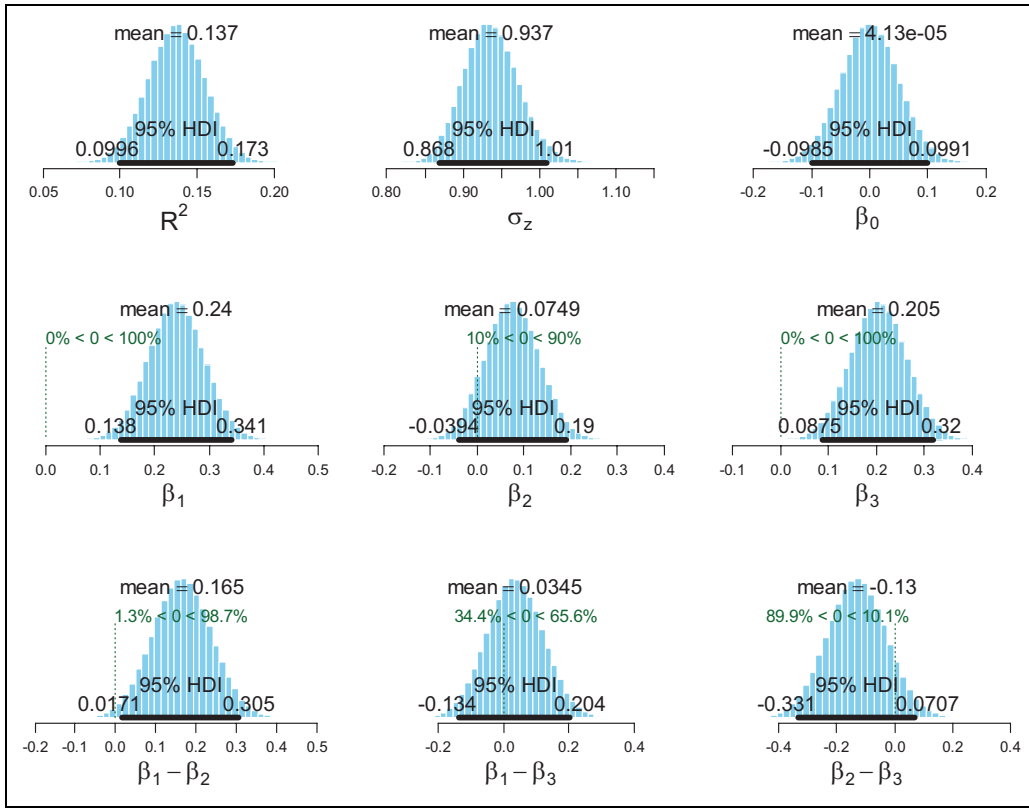

Figure 2. Posterior distribution for the multiple linear regression example, showing parameters for standardized data
Note: HDI = highest density interval.

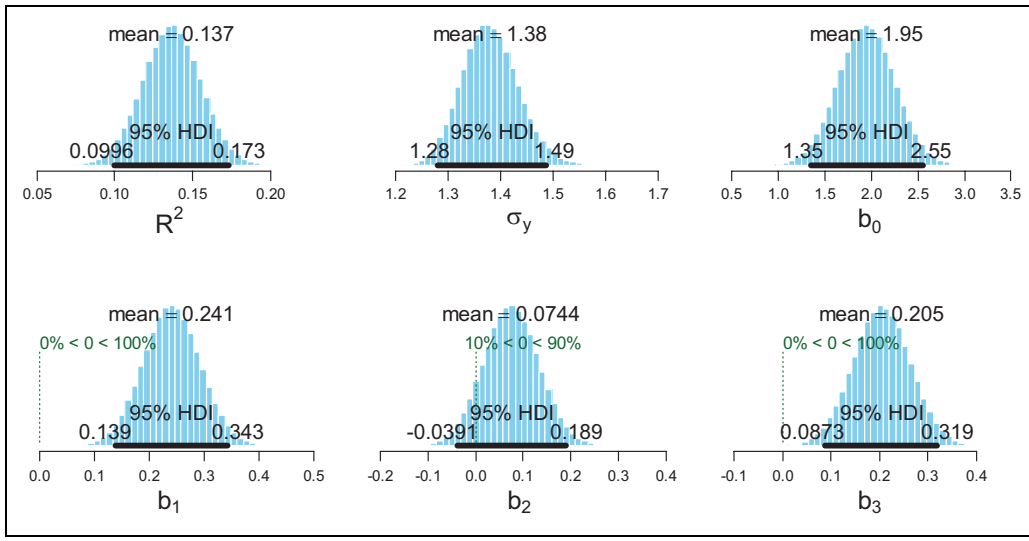

Figure 3. Posterior distribution for the multiple linear regression example, showing original-scale parameters
Note: HDI = highest density interval.

the standard deviation of the data around the linear prediction. By contrast, results from traditional analysis provide no distribution over parameter values.

Figure 2 shows additional information about the regression coefficients. The lower row shows the differences between standardized regression coefficients and, as we describe later in more detail, this information is useful in terms of understanding practical decisions regarding the use of each of the predictors in the model. Although it is well known that standardized regression coefficients must be interpreted with caution because the scales are brought into alignment only in terms of the sample-based standard deviations of the predictors (e.g., King, 1986), a comparison of standardized coefficients can nevertheless be useful. For example, if it is equally costly to test a job applicant for GMA or for conscientiousness, but we would prefer to avoid the double cost of testing both, then we may want to know which test yields higher predictiveness for job performance. The credible differences are determined by computing the difference between regression coefficients at every step in the MCMC chain and plotting the result.

Finally, the upper left panel of Figure 2 shows an entire distribution of credible values for the proportion of variance accounted for, denoted R^2. At each step in the MCMC chain, a credible value of R^2 is computed as simply a reexpression of the credible regression coefficients at that step: $R^2 = \sum_j \beta_j r_{y,x_j}$, where β_j is the standardized regression coefficient for the jth predictor at that step in the MCMC chain, and r_{y,x_j} is the sample correlation of the criterion values, y, with the jth predictor values, x_j. The equation for expressing R^2 in terms of the regression coefficients is used by analogy to least squares regression, in which the equation is exactly true (e.g., Hays, 1994, Eq. 15.14.2, p. 697). The mean value in the distribution of R^2 is essentially the maximum likelihood estimate when using vague priors and the least squares estimate when using a normal likelihood function. The posterior distribution reveals the entire distribution of credible R^2 values. (The posterior distribution of R^2, defined this way, can exceed 1.0 or fall below 0.0, because R^2 here is a linear combination of credible regression coefficients, not the singular value that minimizes the squared deviations between predictions and data.)

The panels in Figure 2 show an interval marked as HDI, which stands for *highest density interval*. Points inside an HDI have higher probability density (credibility) than points outside the HDI, and the points inside the 95% HDI include 95% of the distribution. Thus, the 95% HDI includes the most credible values of the parameter. The 95% HDI is useful both as a summary of the distribution and as a decision tool. Specifically, the 95% HDI can be used to help decide which parameter values should be deemed not credible, that is, rejected. This decision process goes beyond probabilistic Bayesian inference per se, which generates the complete posterior distribution, not a discrete decision regarding which values can be accepted or rejected. One simple decision rule is that any value outside the 95% HDI is rejected. In particular, if we want to decide whether the regression coefficients are nonzero, we consider whether zero is included in the 95% HDI. For example, Figure 2 shows that zero is excluded from the 95% HDI for both Predictor 1 and Predictor 3. The lower left panel indicates that the regression coefficients on Predictors 1 and 2 are credibly different (i.e., a difference of zero is not among the 95% most credible values). The upper left panel indicates that R^2 is well above zero.

Accepting the Null Value

A more sophisticated decision rule also has a way of accepting a (null) value, not merely rejecting it. This extended decision rule involves establishing a region of practical equivalence (ROPE) around the value of interest. For example, if we are interested in the null value (i.e., zero) for a particular regression coefficient, we establish slope values that are equivalent to zero for practical purposes in the particular application. Suppose we specify that slopes between –0.05 and +0.05 are practically equivalent to zero. We would decide to *reject* the null value if the 95% HDI falls completely *outside*

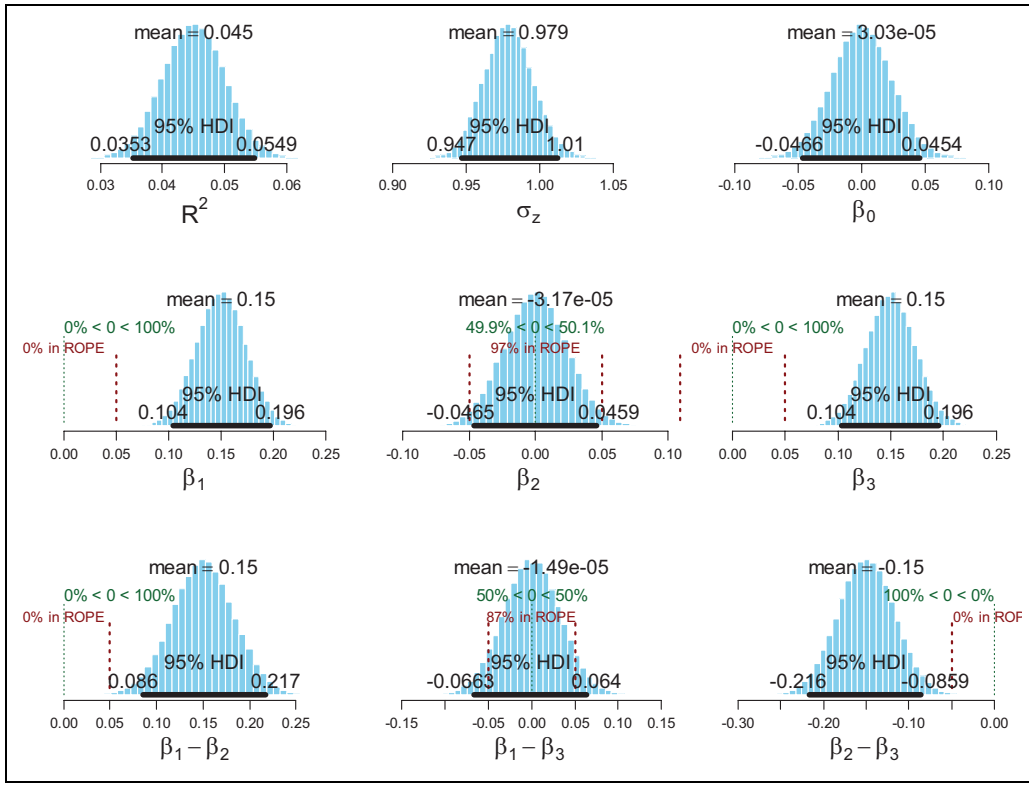

Figure 4. Posterior distribution for a large illustrative data set ($N = 1{,}730$), showing standardized parameters. Notice that the 95% highest density interval (HDI) for the second regression coefficient (middle) falls within the region of practical equivalence (ROPE).

the ROPE, because none of the 95% most credible values is practically equivalent to the null value. Moreover, we would decide to *accept* the null value if the 95% HDI falls completely *inside* the ROPE, because all of the 95% most credible values are practically equivalent to the null value. The 95% HDI gets narrower as the sample size gets larger.

To illustrate a case of accepting the null value, we simulated a larger sample ($N = 1{,}730$) of data from the linear regression model using true regression coefficients of .15, .00, and .15. The resulting posterior distribution is shown in Figure 4 for standardized parameters and in Figure 5 for original-scale parameters. Regarding Figure 4, notice in the middle panel that the posterior distribution for the regression coefficient on the second predictor has its 95% HDI entirely within the ROPE. Because the bulk of the credible values are practically equivalent to zero, we decide to accept the null value. The lower middle panel in Figure 4 shows that the difference between the first and third regression coefficients is centered on zero, but the 95% HDI of the difference does not quite fall entirely within the ROPE.

Classical NHST-based analysis has no way of accepting the null hypothesis, which typically consists of specifying the absence of an effect or relationship (cf. Cortina & Folger, 1998). Indeed, in NHST, because the null hypothesis can only be rejected, a researcher is guaranteed to reject the null hypothesis, even when it is true, if the sample size is allowed to grow indefinitely and the researcher tests with every additional datum collected (e.g., Aguinis & Harden, 2009; Anscombe, 1954; Cornfield, 1966; Kruschke, in press). This "sampling to reach a foregone conclusion" does not happen in

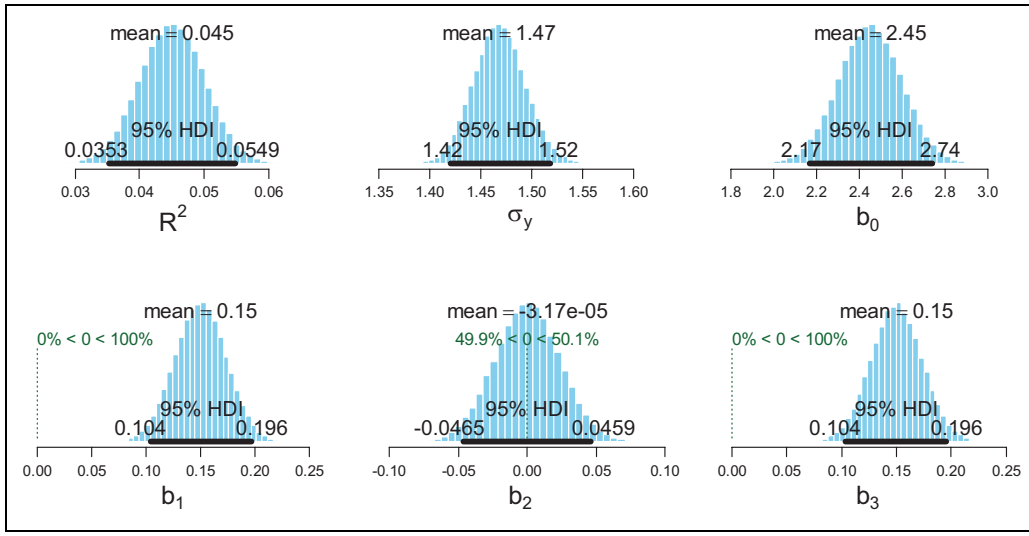

Figure 5. Posterior distribution for a large illustrative data set ($N = 1{,}730$), showing original-scale parameters
Note: HDI = highest density interval.

a Bayesian approach. Instead, because the HDI narrows as sample size increases, and therefore the null has greater probability of being accepted when it is true, it is the case that the probability of false alarm asymptotes at a relatively small value (depending on the specific choice of ROPE). For an illustration of rates of false alarms and acceptances in sequential testing, see Kruschke (in press). Independently of its use as a decision tool for Bayesian analysis, the use of a ROPE has also been suggested as a way to increase the predictive precision of theories in the organizational sciences (Edwards & Berry, 2010).

Summary of Rich Information Provided by Bayesian Analysis

As shown in Figures 2 through 5, the posterior distribution reveals complete distributional information about which values of the parameters are more or less credible, including the stardard deviation of the noise in the data. The posterior distribution simultaneously reveals the differences between standardized regression coefficients. Moreover, a complete distribution of credible R^2 values is provided.

We emphasize that a frequentist sampling distribution of parameter estimates, from an assumed fixed parameter value, is not the same thing as the posterior distribution on the parameter. To create a sampling distribution of estimates, one starts with an assumed fixed parameter value, and then, either analytically or through bootstrapping, creates a sampling distribution of parameter estimates. The process is as follows: (1) Assume a fixed parameter value. (2) From a hypothetical population with that parameter value, repeatedly generate random samples according to an intended sampling design, that is, stopping rule. (3) For each randomly generated sample, compute a sample statistic that estimates the parameter value. (4) From the repeated random samples, create a sampling distribution of the estimator. The result is a distribution that specifies the probability of each estimator value given the assumed parameter value and the sampling design. Notice that the distribution of sample estimates is not about the probability of a parameter value, given the data; indeed, the sampling process starts with a single assumed parameter value. Notice also that the estimator and the parameter are apples and oranges. The distinction between estimator and parameter is especially evident when the parameter of interest is the bias of a coin (e.g., estimating the proportion of left-

handers in a population). The sample estimate is the number of "heads" in the sample divided by the sample size, N. The estimate can therefore take on any of the discrete values $0/N$, $1/N$, $2/N$, . . . , N/N (assuming that fixed N was the intended design). The sampling distribution of the estimator is a distribution on those discrete proportions (specifically, a binomial distribution if we assume N was fixed by design). In contrast, the posterior distribution on the underlying bias is a continuous distribution on the interval [0,1]. The posterior distribution is explicitly a probability distribution on parameter values, given the actually observed data. A major advantage of the Bayesian approach is that the posterior distribution can be directly understood and interpreted: It reveals the relative credibility of every possible combination of parameter values, given the data. A sampling distribution of parameter estimates, from some assumed parameter value, has only a convoluted interpretation and little direct inferential relevance. (And, as we describe later in this article, there is no unique sampling distribution because it depends on the design intention—a.k.a. the stopping rule.)

From the posterior distribution, decisions can be made about landmark values of interest, such as null values. By using the HDI and ROPE, a reseacher can decide whether to reject or accept a candidate parameter value. And the decision is made without reference to p values and sampling distributions as in NHST. Note that one needs a Bayesian approach to implement the decision rule involving the ROPE. In particular, consider the decision to accept a (null) value if the 95% HDI falls entirely within the ROPE around the value. This decision rule only makes sense because the 95% HDI represents the bulk of the credible parameter values. Crucially, the frequentist 95% confidence interval (CI) does not indicate the 95% most credible values of the parameter. The CI is by definition different from the HDI, as we will describe later. One clear illustration that the CI is not an HDI comes from multiple tests: The 95% CI grows much wider when a researcher intends to conduct more tests, but the HDI is unchanged (because the posterior distribution is unchanged).

Although not illustrated by our multiple linear regression example, Bayesian analysis is exceptionally well suited for more complex applications. When the data are skewed or have outliers, it is easy to change the program (see the appendix) to use a nonnormal distribution, and interpretation of the posterior distribution proceeds seamlessly as before. By contrast, generating p values for NHST can be challenging, especially for nonnormal distributions in nonlinear hierarchical models. When a researcher is interested in nonlinear trends in data, it is easy to change the program to model the trend of interest, whether it is polynomial, exponential, a power law, or sinusoidal, to name a few. Importantly, a hierarchical structure is easily implemented in Bayesian software. For example, regression curves can be fit to each of many individuals, with higher level parameters describing trends across groups or conditions, even for nonlinear and nonnormal models. Another example of a hierarchical structure is the inclusion of a higher level distribution to describe the distribution of the regression coefficients across predictors. This sort of hierarchical structure allows each regression coefficient to inform the estimates of the other regression coefficients, while all simultaneously informing the higher level estimates of their dispersion across predictors. The resulting shrinkage of estimated regression coefficients helps attenuate artificially extreme estimates caused by noisy data.

Reasons Why a Bayesian Approach Overcomes Deficiencies in the Frequentist Approach

The information provided by the classical frequentist approach is quite different from the information provided by Bayesian analysis. Whereas Bayesian analysis provides the complete posterior distribution of credibility on the joint parameter space, from which multiple decisions can be made, the frequentist approach provides only a point estimate (from least squares or maximum likelihood) without any distributional information. To make decisions in the frequentist approach, the analyst must create sampling distributions to determine p values and limits on CIs. Unfortunately, the p values and CI limits can change dramatically depending on the intended tests, comparisons, and stopping rules for

sampling of data. Moreover, the CIs provide no distributional information about the parameter values. We address these issues in more detail next.

In multiple linear regression, the statistical significance of a regression coefficient is assessed by computing how much the fit of the full model worsens when the coefficient is removed (i.e., set to zero) and then by computing the probability of that magnitude of worsening if the null hypothesis were true. For example, to assess the importance of the regression coefficient β_1, we consider the residual sum-squared error of the full model, $E_F = \sum_i (y_i - [\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}])^2$, versus the residual sum-squared error of the restricted model in which β_1 is set to zero: $E_R = \sum_i (y_i - [\beta_0 + 0x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}])^2$. Note that the best-fitting values of β_0, β_2, and β_3 may differ in the full and restricted models. The full model will always fit at least as well as the restricted model, even when the true value of β_1 is zero, because the full model can better fit random sampling noise and the restricted model is nested within the full model. Therefore, the crux is the probability of that amount of worsening if we were to imagine randomly sampling data from the null hypothesis.

Adopting classical frequentist statistics involves using a descriptive statistic, denoted F, that measures the average increase in error per parameter restricted, relative to the baseline error per data point in the full model (Maxwell & Delaney, 2004):

$$F = \frac{(E_R - E_F)/(df_R - df_F)}{E_F/df_F}, \tag{3}$$

where df is the degrees of freedom, which is the number of data values N minus the number of parameters in the model (other than the "nuisance" parameter, σ). Thus, for testing a single regression coefficient from the set of three predictors, the full model has $df_F = N - 4$, and the restricted model has $df_R = N - 3$.

To determine a sampling distribution for F, we repeatedly generate random data from the null hypothesis, in which all the (nonnuisance) parameters are zero, and we examine the distribution of F when the null hypothesis is true. In the null hypothesis, the simulated data are normally distributed. For each set of randomly generated data, we use the stopping rule intended by the researcher. The stopping rule plays an important role because it determines the space of possible imaginary outcomes when the null hypothesis is true, relative to which the actual outcome is judged. Conventionally, data are assumed to be sampled until N reaches a threshold size. Other stopping rules are possible, as will be explained subsequently. When sampling from the null hypothesis, usually F will be around 1.0, because the random worsening of fit by a restricted parameter (the numerator of F) will be about the same as the background noise per data point (the denominator of F). But, by chance, the F value will sometimes be larger than 1.0, because the random sample of data will happen to have an accidental arrangement that is much better fit by the full model than by the restricted model. Our goal is to determine exactly the probability that the null hypothesis generates F_{samp} values as large as or larger than the particular F_{act} value observed in the actual data. This probability is called the p value and can be formally specified as

$$p(\text{any } F_{samp} \geq F_{act} | \beta = \text{null}), \tag{4}$$

which means, in words, the probability that any sampled F value equals or exceeds the actually observed F value, given that the parameter values are set to the null hypothesis values. When the p value of Equation 4 is thought of as a function of F_{act} (with β fixed), then we can solve for the value of F_{act} that yields $p = .05$ and call that value the critical value, F_{crit}, because when $F_{act} > F_{crit}$, then $p < .05$. When the p value of Equation 4 is thought of as a function of β (with F_{act} fixed), then we can solve for the values of β that yield $p = .05$ and call those values the limits of the confidence

interval on β, because values within those limits yield $p \geq .05$ and hence are not rejected, but values beyond those limits yield $p < .05$ and hence are rejected.

When the frequentist approach is applied to our job performance data introduced earlier in the article, the point estimate and the limits of the conventional 95% CI closely match the mean and 95% HDI of the Bayesian posterior distribution in Figure 3. It is crucial to understand, however, that the frequentist result does not provide a continuous distribution over the parameters as the Bayesian analysis does. The frequentist result provides no test or CI on the noise parameter, σ. The frequentist result provides no probability distribution on the proportion of variance accounted for, R^2, whereas the Bayesian approach does. A frequentist CI can be constructed for R^2 (e.g., Steiger & Fouladi, 1992), but it provides no distributional information and its limits depend on the sampling intention of the analyst, as explained below. The basic frequentist result also provides no comparative tests of standardized regression coefficients, as is revealed by the posterior distribution in Figure 2. There are frequentist comparisons of standardized regression coefficients, but the conclusions from the tests depend on which tests are intended to be conducted.

Indeed, in the frequentist approach, every test incurs an additional opportunity for false alarm, and the primary concern for NHST decisions is keeping the false alarm rate capped at, for example, 5% (Aguinis, Werner, et al., 2010). When additional tests are conducted, the space of possible F_{samp} enlarges because every test contributes F_{samp} values to the sampling distribution. Therefore, the p value in Equation 4 increases, even though the actual value F_{act} is unchanged. In other words, the p value measures how much of the space of imaginary unobserved outcomes exceeds the single actual outcome, and when the analyst's imagination changes to include additional tests, the p value changes.

In particular, the default behavior of most frequentist software packages is to test each regression coefficient as if it were the only test being conducted. But when testing multiple regression coefficients, it is sensible to "correct" the p value of each test to take into account the larger sampling space of multiple tests (e.g., Mundfrom, Perrett, Schaffer, Piccone, & Roozeboom, 2006). Moreover, if additional comparisons of (standardized) regression coefficients are conducted, then the sampling space is again enlarged, and stricter corrections must be applied to the p values of each test. Thus, a single observed estimate of a regression coefficient can have many different p values and confidence intervals, depending on the testing intentions of the analyst.

A p value depends not only on the space of intended tests but also on the intended stopping rule for data sampling, for the same reason: The stopping intention, like the testing intention, determines the space of imaginary F_{samp} values relative to which the actual F_{act} is judged. The conventional stopping rule is to imagine sampling data from the null hypothesis until the sample size, N, meets some fixed threshold. But the data themselves bear no signature of why data collection stopped, because researchers are careful to collect data so that every datum is completely insulated from all other data collected before or after. Therefore, many researchers stop collecting data for reasons other than reaching a threshold sample size. In particular, a common stopping rule is sampling until a temporal threshold is reached (e.g., collect data for two weeks, or until 5 o'clock Friday afternoon). Under this stopping rule, the sampling distribution from the null hypothesis involves samples of various sizes, all of which could be collected in the same temporal window. Some imaginary F_{samp} values use sample sizes that are larger or smaller than the sample size in F_{act}, but all imaginary F_{samp} values use the same temporal threshold as F_{act}. Because there are different imaginary F_{samp} values, the sampling distribution for stop-at-threshold-sample-size is different from the sampling distribution for stop-at-threshold-time, and therefore the p values are different (for specific examples, see Kruschke, 2010a, in press). Unfortunately, because the data bear no signature to distinguish whether sampling was stopped because of reaching threshold sample size or reaching threshold time, we do not know which p value is appropriate for judging the data. The dilemma is not solved by asking the data collector about his or her stopping intention, because identical data could be collected under

different intentions, implying that two people with identical data could come to different conclusions about the statistical significance of the data.

There are other plausible stopping rules for data collection, which again yield different p values for any one set of data. Consider collecting data until a summary description of the data exceeds a preset threshold, and the variable being measured is how many data values had to be collected before that threshold was exceeded. For example, suppose we set a stopping threshold of $F_{stop} = 2.0$. We collect data until $F_{act} > F_{stop}$, and we count how many data values were sampled. If there is a nonnull effect in the data, then it should not take much data to exceed the threshold, but if there is no effect in the data, it will take a long time to exceed the threshold. For the observed count of data values, we compute the probability that so few data would be collected if the null hypothesis were true, and this is the p value according to this stopping intention. Importantly, this p value differs from the p value computed from the same data under the intention of stopping until threshold sample size is reached. For more information about this sort of stopping rule, in the context of dichotomous-scale data instead of metric-scale data, see, for example, Berger and Berry (1988). For an example in the context of the traditional t test, see Kruschke (in press). The situation was succinctly summarized by Berry (2006) as follows: ''The p-value depends on the design of the trial and the intentions of the investigator'' (p. 31).

The dependency of the p value on the stopping intention is distinct from concerns about repeatedly testing data as they are collected. The stopping rules discussed earlier did not repeatedly test data; they merely considered stopping at threshold sample size, at threshold time, or at threshold data-summary value. Once the threshold was reached, a single test was conducted. If, to the contrary, an analyst tests repeatedly as successive data are collected, then the p value is inflated because of the repeated tests, just as in the case of multiple comparisons (but the magnitude of inflation is different because of different structural overlap in the tests).

The problem of ill-defined p values is not based only on the arguments from the intended stopping rule and intended multiple tests of parameters for a single sample. An analogous argument applies when there are multiple groups. Just as different stopping intentions change the space of possible outcomes from the null hypothesis, different intended comparisons of groups change the space of possible outcomes from the null hypothesis. Effectively, the stopping rule for observing tests affects the p value just as the stopping rule for observing data.

Please note that the same logic described previously also applies to bootstrapping and resampling, which is, by definition, the creation of sampling distributions. So, although different from analytically derived sampling distributions in their mechanics and assumptions about the null hypothesis, bootstrapping actually inherits all the problems of p values and confidence intervals that apply to analytically derived sampling distributions. In particular, when the sampling intention changes, a bootstrapped or resampled p value changes, and the confidence interval changes.

Another problem with the frequentist approach is that it cannot accept a null hypothesis. Consider the large-sample-size example discussed in conjunction with Figures 4 and 5. Recall that the Bayesian analysis concluded that the second regression coefficient was equivalent to the null value for practical purposes. When conventional frequentism is applied to the data, the test of the second regression coefficient results in a p value close to 1. But this result does not let us accept the null value; it means merely that the null hypothesis would almost always produce F_{samp} values greater than $F_{act} \cong 0$. In fact, a frequentist would *reject* the null hypothesis because the probability of getting F_{samp} *less than* F_{act} is extremely small, and therefore, something about the null hypothesis is probably wrong, such as the assumption of independence in the data.

Importantly, we cannot import a ROPE into frequentist decision making and use the CI like the HDI because of two reasons (although some have explored this approach, e.g., Rogers, Howard, & Vessey, 1993; Westlake, 1976, 1981). First, unlike the HDI, the limits of the CI change when the intended stopping rule or comparisons change. Recall that the limits of the 95% CI are the parameter

values at which the p value in Equation 4 is .05 (or .05/2 for two-tailed tests). When the p value changes, the CI changes. Thus, we simply cannot tell whether "the" 95% CI is within a ROPE because the 95% CI itself is ill-defined. Second, unlike the HDI, the CI tells us nothing about the credibility of the parameters in its bounds. It is tempting to psychologically confer distributional qualities upon the CI that are not actually there. For example, we could plot the p value in Equation 4 as a function of the parameter (e.g., Poole, 1987; Sullivan & Foster, 1990). This graph shows the probability of extreme F values as a function of the hypothetical parameter value; it does not show the posterior probability of the parameter value, and it is not even a probability distribution (e.g., it does not integrate to 1). More sophisticated variations of the approach build probability distributions on the parameter, such that different intervals under the distribution correspond to different levels of confidence (e.g., Schweder & Hjort, 2002; Singh, Xie, & Strawderman, 2007). But these definitions of confidence distributions still assume that sampling proceeds until threshold sample size has been reached, and the confidence distributions change when the sampling intention changes, just as p values and confidence intervals change. Another way to impose a distributional interpretation on a confidence interval is by superimposing the sampling distribution of the sample statistic onto the parameter axis centered at the best point estimate of the parameter (Cumming, 2007; Cumming & Fidler, 2009). Unfortunately, this approach has limited meaningfulness and applicability. The distribution it produces is not a distribution of parameter values; instead, it is a distribution of sample statistics assuming a fixed parameter value. The distinction between a sample statistic and a parameter value is especially stark when the approach is applied to dichotomous data, for which the sample statistic is a discrete proportion but the parameter is a continuous parameter. The distribution of sample statistics is also subject to the vicissitudes of different sampling intentions and is therefore just as ill-defined.

Some readers may wonder whether the prior distribution in a Bayesian analysis has the same slippery status as a sampling-space intention in the frequentist approach. To the contrary, a prior distribution and a sampling intention are quite different. First, the prior distribution is explicit, and its influence on the posterior distribution is easily checked. With the prior established, the posterior distribution is a fixed entity, unchanging whether or not various comparisons are considered and uninfluenced by the stopping intention of the data collector. On the other hand, stopping and testing intentions in frequentism are usually not explicit and usually rationalized post hoc. Second, and more importantly, the prior distribution *should* influence our interpretation of data, but the sampling-space intentions should *not*. Bayesian analysis indicates exactly how credibility should be reallocated across parameter values, starting from the prior distribution. If there is strong prior knowledge, it can be a blunder not to use it, as was previously explained, for example, in the case of random drug or disease testing. If there is only vague prior knowledge, Bayesian analysis nevertheless provides a complete posterior distribution over parameter values. On the other hand, researchers carefully insulate the data collection process from the intentions of sampling and testing, and therefore it is antithetical to base interpretive conclusions on p values and CIs that depend on such intentions. Because research design and scientific reasoning make the sampling and testing intentions irrelevant, a researcher's intuitions about data analysis match Bayesian interpretation, not frequentism (Dienes, 2011).

We are not the first to point out flaws of the frequentist approach, but the particular examples presented here are novel. Some previous critiques include those by Berger and Berry (1988), Cohen (1994), Kline (2004), McCloskey (1995), Nickerson (2000), and Wagenmakers (2007). Many of those critiques point out difficulties in correctly understanding NHST (e.g., Aguinis, Werner, et al., 2010). Our argument is that especially when traditional frequentist analysis is *correctly* understood, including the use of confidence intervals, it is seen to be fundamentally flawed in its foundational logic.

Recommendations and Illustration of How to Report Bayesian Analysis Results

Guidelines for reporting a Bayesian data analysis are provided by Kruschke (2011a, chap. 23). We offer five recommendations regarding what information to include in a manuscript reporting Bayesian analysis. First, motivate the use of Bayesian analysis if the targeted readership is not familiar with such an approach. Second, describe the model and its parameters because the parameter values bear the meaning of the analysis. Third, describe and justify the prior distribution for a potentially skeptical readership. Fourth, mention the MCMC details, including evidence that the chains are fully representative of the underlying posterior distribution. Finally, interpret the posterior distribution. (Additional optional points are discussed by Kruschke, 2011a, chap. 23.) Next, we offer the following paragraphs as an illustration of writing up a realistic data analysis using our job performance example and implementing the aforementioned five recommendations. Before analyzing data, however, we need to describe the data themselves.

We were interested in predicting job performance based on GMA, conscientiousness, and biodata. We initially targeted a random sample of 485 employees of a firm. Thirty-five employees did not have current contact information. Of those we could attempt to contact, 362 responded and, of those, 346 provided complete information about the three predictors, and we also had criterion data available (i.e., supervisory ratings of performance). The predictors and criterion were all measured on 1-to-7 Likert-type scales.

We analyzed the data with Bayesian linear regression. Unlike traditional NHST-based statistics, Bayesian analysis yields complete distributional information regarding the parameters in the regression model. Bayesian analysis uses only the observed data and does not use p values and confidence intervals that are based on hypothetical unobserved data that might have been obtained assuming a particular stopping intention about sample size of the researcher. Moreover, traditional analysis assumes that the stopping intention was to cease data collection at a preset sample size—an assumption that does not apply to our data. (Presumably, as Bayesian analyses become routine, this sort of explicit justification will not be needed.)

We used the standard linear regression model, in which the criterion value is described as normally distributed around a linear combination of predictor values:

$$p(y_i|x_{1i}, x_{2i}, x_{3i}, \beta_0, \beta_1, \beta_2, \beta_3, \sigma) = N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \sigma).$$

The model has five parameters: β_j for $j \geq 1$ indicates how much the criterion increases when the jth predictor increases one unit and the other predictors are held constant, β_0 indicates the value of the criterion when all three predictors are zero, and σ indicates the standard deviation of the residual scores in the criterion around the linearly predicted value. Our primary interest is in the magnitudes of the regression coefficients on the three predictors.

We used a noncommittal broad prior on the parameters so that the prior had minimal influence on the posterior. For the analysis, the data were standardized, and the intercept and slope parameters had normal priors with mean zero and standard deviation of 10, which is very large relative to the scale of the standardized data (because standardized regression coefficients will tend to fall between -1 and $+1$). The residual-noise parameter had a uniform prior extending from zero to 10 (which is extremely broad and inclusive relative to the standardized noise of 1). The estimated parameters were linearly transformed back to the original scale (see Eq. 17.1, p. 459, of Kruschke, 2011a).

The posterior was generated as an MCMC sample using the free software R, rjags, and JAGS (Plummer, 2003, 2011; R Development Core Team, 2012). Three chains were initialized at the maximum likelihood values of the parameters and well burned in (for 1,000 steps), and a total of 250,000 steps were saved. There was very little autocorrelation in the well-mixed chains. The resulting MCMC sample is therefore highly representative of the underlying posterior distribution.

As shown in Figure 3, the marginal posterior on b_1 had a mean of 0.241 and a 95% HDI that extended from 0.139 to 0.343. Indeed, all of the 250,000 representative values in the posterior were well above zero, so zero was deemed to be not credible. (As these reports become routine in the literature, the language could be compressed; e.g., marginal posterior mean b_1 [GMA] = 0.241, 95% HDI = 0.139, 0.343; zero deemed not credible). Marginal posterior mean b_2 (conscientiousness) = 0.0743, 95% HDI = –0.0385, 0.189; zero among the credible values. Thus, although conscientiousness (x_2) might have a nonzero predictiveness for job performance, the uncertainty in its estimated influence (b_2) is large relative to the magnitude of the influence. Marginal posterior mean b_3 (biodata) = 0.205, 95% HDI = 0.0885, 0.321; zero deemed not credible (with 100% of the posterior greater than zero). Marginal posterior mean intercept b_0 (job performance) = 1.94, 95% HDI = 1.36, 2.56. Marginal posterior modal residual-noise σ = 1.37, 95% HDI = 1.28, 1.48. The posterior on R^2 had 95% HDI from 0.122 to 0.137.

We were interested in assessing whether the predictors were differentially predictive of job performance, and therefore we examined the posterior distributions of the differences of standardized regression coefficients, keeping in mind that the comparison is based on standardization using the sample only (e.g., King, 1986). Figure 2 shows that for GMA versus conscientiousness, the posterior mean difference of $\beta_1 - \beta_2 = 0.165$, 95% HDI = 0.0183, 0.307, with 98.7% of differences being greater than zero. Therefore, if collecting GMA and conscientiousness data is equally costly, and if we wish to avoid the double cost of measuring both, then it is probably more effective to assess GMA than conscientiousness.

Notice that the preceding summary of the Bayesian analysis never mentioned p values, confidence intervals, or corrections for multiple tests. If we had conducted a frequentist analysis, we would have had to correct the p values and confidence intervals for the six or more tests we conducted (corresponding to the panels of Figure 2), taking into account the structural relations of the tests (e.g., Mundfrom et al., 2006). With each test, the p values increase, and the CIs widen. By contrast, the posterior distribution from the Bayesian analysis is a fixed entity based on the observed data. The posterior distribution highlights explicitly the uncertainty in the estimation of each parameter.

Synthesis of Advantages of Bayesian Data Analysis

In this section, we summarize a number of key advantages of Bayesian methods for data analysis over frequentist analysis. Table 2 includes a summary of these advantages. This section also includes several examples of how specific published research that is based on frequentist analysis-based approaches failed to benefit from the advantages offered by a Bayesian approach—and how using Bayesian analyses would have led to richer and, in some cases, different substantive conclusions. In the very few instances when available, we actually illustrate the advantages of Bayesian analysis using published research that did adopt this approach.

Use of Prior Knowledge

The first advantage included in Table 2 is that Bayesian data analysis incorporates not only the data at hand but also prior accumulated knowledge to generate parameter estimates that tend to be both different and more accurate than those derived under frequentist methods. This is especially true when the prior accumulated knowledge corresponds to a distribution that is substantially different from a noncommittal broad prior distribution.

As an illustration of this advantage in a specific organizational science context, Brannick (2001) described a situation involving 10 local validation studies of a situation judgment test. The 10 corresponding correlations were calculated using frequentist analysis. Because of the nature of

Table 2. Brief Summary of Selected Advantages of Bayesian Analysis Over Frequentist Analysis

Issue	Frequentist	Bayesian
1. Use of prior knowledge	Does not incorporate prior knowledge into estimation of parameters. Instead, only uses the data at hand. As a result, the estimated parameters can be less precise.	Explicitly incorporates prior knowledge into estimation of parameters by expressing the prior knowledge in terms of a prior distribution, which is then combined with the data to produce the posterior.
2. Joint distribution of parameters	Creates only a local approximation.	Generates accurate joint distribution of parameter estimates.
3. Assessment of null hypotheses	Null hypothesis cannot be accepted.	Null values can be accepted as well as rejected.
4. Ability to test complex models	Often difficult to derive *p* values, confidence intervals on parameters, and confidence intervals on predictions for new data.	Flexibility in adapting a model to diverse data structures and distributions (e.g., nonnormal) with no change in inference method.
5. Unbalanced or small sample sizes	In unbalanced designs in analysis of variance (ANOVA), user must choose "Type I" or "Type III" error terms. In chi-square tests, small sample violates assumptions for obtaining an accurate *p* value.	Posterior distribution reveals greater uncertainty in parameter estimates for cells with smaller sample sizes. Sample size does not affect inference method.
6. Multiple comparisons	Requires corrections for inflated Type I error rates (i.e., false positives), and correction depends on the comparisons that are intended by the researcher.	No reference to intended or post hoc comparisons. Ease of hierarchical modeling allows rational sharing of information across groups to "shrink" estimates.
7. Power analysis and replication probability	Virtually unknowable because of lack of distributional information in confidence interval.	Precisely estimated from complete distribution of parameter values.

frequentist analysis, the calculation of each frequentist analysis-derived correlation did not involve the use of prior knowledge, which in this example consists of correlations calculated from the other nine studies. As a result, one of the studies yielded a statistically nonsignificant correlation of .15 ($p > .05$). In contrast, when Bayesian analysis was employed such that prior knowledge regarding the other correlations was taken into account, the same study yielded a Bayesian-based correlation of .18. Moreover, the corresponding 95% HDI did not contain the correlation value of .00. In short, the substantive conclusion based on a frequentist approach was that the correlation was not different from zero, but using a Bayesian approach led to the conclusion that the correlation was different from zero.

A difference of less than .05 correlation units—about 33% change— in the estimated magnitude of a correlation might not seem impressive in a number of contexts. At the same time, in many other contexts such as in the case of Brannick's 10 local validation studies, such a difference was practically significant enough for two local validation studies' correlations to be judged as different from zero (under Bayesian analysis) but statistically nonsignificant (under frequentist analysis). This is precisely the type of apparently small difference that actually has important practical consequences (Cortina & Landis, 2009). From a substantive perspective, the situation judgment test is now deemed

valid in the local context, whereas the frequentist analysis may have reached the conclusion that the test should not be used because the relationship between predictor and criterion scores was not statistically different from zero. The great potential of Bayesian methods applied specifically to meta-analysis led Schmidt and Raju (2007) to conclude that "the medical model [i.e., adding new studies to the existing database and recalculating the meta-analysis] and the alternative Bayesian procedure proposed here should be the methods of choice for updating meta-analytic findings in most research areas" (p. 305).

In summary, in practical applications, we benefit from using both prior knowledge and the data at hand to derive accurate parameter estimates. Bayesian data analysis does just that. By translating the prior knowledge into a prior distribution, and then using Bayes' rule to combine the prior with the current data, the resulting posterior distribution gives the researcher parameter estimates that are the best available lacking further information.

Joint Distribution of Parameters

Mentioned in Issue 2 in Table 2, the second advantage of Bayesian data analysis is that the method produces a *joint* (i.e., simultaneous) distribution of credible parameter estimates across multiple predictors in a multidimensional parameter space, thereby allowing the researcher to examine the trade-offs among values of different parameter estimates across the multiple predictors. The computer program described in the appendix produces scatter plots of jointly credible values for all pairs of parameters. For linear regression with normally distributed noise, the scatter plots tend to be simple oval shapes, but for other models, the scatter plots can be "banana" or S-shaped, indicating interesting trade-offs in credible parameter values. Frequentist analysis can use an asymptotic approximation to the likelihood function, but this approach fails to describe the posterior distribution in general. Although we did not include an illustration of this issue in our article due to space constraints, the programs provide graphical displays of the posterior for all pairs of parameters.

Being able to know the accurate trade-offs among credible parameter estimates across multiple predictors is not a mere technical refinement. Consider our example from Figures 2 and 3, where Bayesian data analysis is used to fit a linear regression model to data, and job performance is predicted by GMA, conscientiousness, and biodata. Previous research indicates that conscientiousness and biodata are fairly strongly correlated ($r = .51$; Roth et al., 2011). If we also have prior knowledge about how strongly either conscientiousness or biodata predicts job performance, then we can use that prior knowledge to leverage more precise estimation of the other predictor (Western & Jackman, 1994). The inferential leverage derives from the trade-off in parameter estimates: The prior knowledge about one parameter narrows the estimate of the other parameter because of the trade-off.

Assessment of Null Hypothesis

The third advantage summarized in Table 2 refers to the assessment of the null hypothesis. Within a Bayesian approach, as described earlier, accepting a null value involves establishing a ROPE around the value of interest. For example, if we are interested in the null value (i.e., zero) for a regression coefficient, we establish slope values that are equivalent to zero for practical purposes in the particular application. On the other hand, the classical frequentist framework has no way of accepting the null hypothesis (cf. Cortina & Folger, 1998). The frequentist approach cannot incorporate an analogous decision rule involving a ROPE and CI (as opposed to HDI) because the CI, unlike the HDI, does not indicate which parameter values are credible, and the CI changes its size when the sampling and testing intentions change.

Furthermore, in NHST, a researcher is mathematically guaranteed to reject the null hypothesis even when it is true, if the sample size is allowed to grow indefinitely and a test is conducted with

every additional datum (e.g., Aguinis & Harden, 2009; Anscombe, 1954; Cornfield, 1966; Kruschke, in press). This "sampling to reach a foregone conclusion" does not happen in a Bayesian approach. Instead, because the HDI narrows as sample size increases, and therefore the null has greater probability of being accepted when it is true, it is the case that the probability of false alarm asymptotes at a relatively small value (depending on the specific choice of ROPE).

Ability to Test Complex Models

Issue 4 in Table 2 refers to another important advantage of Bayesian methods, which is the ease of analyzing complex data structures and models. The model specification language is quite flexible in the Bayesian software JAGS (see the appendix). All the usual forms of the generalized linear model can be specified, including multiple linear regression, analysis of variance (ANOVA), analysis of covariance, logistic regression, ordinal regression, log-linear models for contingency tables, and many other data-analytic approaches that are the most frequently used by organizational science researchers (e.g., Aguinis et al., 2009; Gelman & Hill, 2007; Jackman, 2009; Kruschke, 2011a; Scandura & Williams, 2000). A Bayesian approach can also be used to implement mixed models and nesting of variables in hierarchical models. Bayesian models are especially useful also for nonlinear models. Moreover, a researcher is not limited to normal distributions for describing metric data; other distributions can be used instead to accommodate outliers or skew in robust regression (O'Boyle & Aguinis, 2012; Kruschke, 2011a). Mixture models are handled by Bayesian approaches because the assignment of data to mixture components is probabilistic, and the estimated parameters are jointly distributed with the assignments of data. In other words, instead of only a single assignment of data to mixture components, multiple credible assignments are assessed simultaneously.

In general, a researcher can flexibly create a hierarchical nonlinear model that reflects a structure appropriate to the data. The inference of credible parameter values takes place regardless of the type of model. In all cases, a complete joint distribution of credible parameter values is created, even for dozens or hundreds of parameters, given the single set of actually observed data. Nonlinear models are becoming more important in the organizational sciences, as we gain precision in specifying the functional forms of relations between variables (Edwards & Berry, 2010; Pierce & Aguinis, in press). Moreover, hierarchical models are becoming increasingly pervasive in many organizational science domains (Aguinis et al., 2009). For example, in many applications, researchers measure aspects of individuals within different groups, and the researchers want to estimate the effects of both individual-level variables (e.g., job satisfaction) and group-level variables (e.g., team cohesion) on individual-level outcomes (e.g., performance). Bayesian software allows specification of any number of levels. Within each level, there can be complex models. For example, we might have a multiple regression model for each individual and a higher level model of how individual-level regression coefficients are distributed across groups (for a complete working example, see section 16.3 of Kruschke, 2011a, or Lykou & Ntzoufras, 2011). Hierarchical modeling is also useful for meta-analysis, in which particular studies play the role of individuals, and higher level model structure has parameters that describe overall tendencies across studies (e.g., Kruschke, 2011b). Frequentist applications can be exceedingly difficult for complex applications, especially for nonlinear or nonnormal models, because generating sampling distributions and confidence intervals is often intractable (O'Boyle & Aguinis, 2012). Moreover, even in complex hierarchical applications, Bayesian methods straightforwardly generate predictions that accurately incorporate the full distribution of credible parameter values instead of just a single point estimate.

As an example of how the flexibility of Bayesian methods might have affected recent work in the organizational sciences, consider the extensive interest in how creativity and organizational innovation relate to competitiveness (Zhou, 2003). Research by Hirst, Van Knippenberg, and Zhou (2009) revealed a complex, polynomial-trend relationship between employee learning orientation and

employee creativity that was moderated by a third factor, team learning behavior. The researchers used frequentist hierarchical linear models. Despite the sophistication of their analysis, had the researchers used a Bayesian approach, they would have discovered similar trends but with complete distributional information about the trade-offs between coefficients on different trend components. Moreover, the researchers could have implemented nonnormal distributions at any level in the hierarchical model, to accommodate outliers, skewed data, or alternative assumptions about the distributions of trend parameters across different teams. Bayesian analyses such as these have great potential to be used to critically evaluate the robustness of recent conceptualizations involving nonnormal (O'Boyle & Aguinis, 2012) and nonlinear (Pierce & Aguinis, in press) relationships.

Unbalanced or Small Sample Sizes

Regarding Issue 5 in Table 2, Bayesian methods can be used regardless of the overall sample size or relative sample sizes across conditions or groups. Essentially, in Bayesian analysis, there is an updating of the parameter estimates for every individual datum, so it does not matter, computationally, where in the design each datum appears. By contrast, in frequentist ANOVA, when different cells of the design include different sample sizes, a researcher must decide between using Type I or Type III error terms in computing the best estimates and corresponding p value. Similarly, in frequentist approaches to moderated multiple regression models with categorical moderator variables, an unequal number of individuals across groups (e.g., more men than women or more Whites than African Americans) leads to important errors in prediction (Aguinis, Culpepper, & Pierce, 2010). These problems are not relevant in the Bayesian analogues of multiple regression and ANOVA. As another example, in the frequentist chi-square tests of independence, the p value is determined by approximating the sampling distribution of the discrete Pearson chi-square value with the continuous chi-square sampling distribution, but that approximation is reasonably good only when the expected cell frequency is about 5 or larger. Therefore, researchers must acknowledge that results and substantive conclusions might be incorrect. There is no such issue in Bayesian analogues of chi-square tests because the analysis estimates parameter values without relying on large-sample approximations.

Dynamic analyses of social networks in organizations are becoming an increasingly important research area in the organizational sciences (Newman, Barabasi, & Watts, 2006). Recently, Sasovova, Mehra, Borgatti, and Schippers (2010) used longitudinal data on friendship relations in the department of an organization to find that "high self-monitors were more likely than low self-monitors to attract new friends and to occupy new bridging positions over time" (p. 639). In doing so, Sasovova et al. (2010) used a frequentist procedure called the quadratic assignment procedure (QAP). However, as explained by Casciaro and Lobo (2008), QAP assumes completely balanced sample sizes across distinct respondents, even when it is usually the case that individuals have different numbers of social network ties with other individuals such that the assumption of completely balanced sample sizes is untrue. Thus, it is likely that QAP's assumption of equal sample sizes biased the parameter estimates and the standard errors reported in Sasovova et al. (2010). Had Sasovova et al. (2010) used a parametric descriptive model and Bayesian estimation, results would not have been susceptible to these problems.

Multiple Comparisons

Issue 6 in Table 2 refers to another advantage of Bayesian methods that results from obviating p values and confidence intervals. In frequentist analysis, the essential decision criterion is that α (i.e., Type I error rate) $= .05$ (or some other value such as .01), which means simply that the probability of falsely rejecting the null value is capped at the nominal level (i.e., typically 5%). A new problem arises when applying this decision criterion to situations in which there are multiple tests. The

problem is that every additional comparison presents an opportunity for a Type I error (i.e., false positive). Traditional analysts want to keep the overall probability of false positives, across the whole family of tests, at 5%. To do so, each individual comparison must be "corrected" to require a more stringent (i.e., smaller) *p* value for a difference to be deemed statistically significant. This is a scientifically dubious practice because the difference between two groups can be declared to be significant or not entirely on the basis of whether a researcher is inquisitive and decides to conduct many comparisons or feigns disinterest and runs only a few comparisons.

In Bayesian analysis, there is no use of *p* values and confidence intervals and no influence from which comparisons, and how many, a researcher might or might not want to make. Instead, the distribution of credible parameter values is determined purely by the data and the structure of the model. Bayesian methods do not escape the possibility of false positives, of course, because they are caused by coincidences of rogue data that can arise in any random sample. But Bayesian methods do not try to control for inflation in Type I error rates on the basis of a researcher's explicit or implicit intentions. Bayesian methods can incorporate rational constraints in a model's structure so that the data from different groups mutually inform each other's estimates and thereby reduce the extremity of outlying estimates. For example, Bayesian approaches to ANOVA can use a hierarchical model structure such that the estimates of group means shrink toward the overall mean. In other words, the data themselves dictate how much the estimates of outlying groups should be shrunken. This shrinkage in the estimates of group means attenuates false positives (Gelman, Hill, & Yajima, 2009; Kruschke 2010a, 2010b). The same approach applies to shrinking estimates of regression coefficients across multiple predictors (e.g., Kruschke, 2011a; Lykou & Ntzoufras, 2011). Hierarchical models are not unique to Bayesian estimation, but they are especially straightforward to implement and evaluate in Bayesian software.

Power Analysis and Replication Probability

The seventh issue summarized in Table 2 contrasts traditional and Bayesian approaches regarding statistical power analysis and replication probability. In traditional NHST, statistical power is the probability that a null hypothesis would be rejected if a particular alternative nonnull effect were true and sampling proceeded in a particular way. More generally, statistical power is the probability that an existing population effect will be detected in a sample of observed data. The key prerequisite in any power analysis is to determine a targeted nonzero effect size. In traditional power analysis, there is only one targeted effect size with no distribution of other reasonable values (e.g., Scherbaum & Ferreter, 2009). Because we have no sense of how uncertain the power estimate is, traditional analysis yields estimates that have little stability and are, according to a recent analysis, "virtually unknowable" (Miller, 2009; see also Gerard, Smith, & Weerakkody, 1998; Thomas, 1997). This may be one of the reasons that published research is still underpowered (Maxwell, 2004) despite repeated calls that researchers conduct power analyses (e.g., Aguinis, Beaty, Boik, & Pierce, 2005).

In a Bayesian framework, the estimate of power is robust because the hypothetical model uses the complete distribution of credible parameter values. Power is estimated by using a large number of credible parameter-value combinations and from each generating simulated data. The simulated data are analyzed by Bayesian methods and tallied with respect to whether the effect was detected. Across many simulations, power is thereby estimated while taking into account the full uncertainty of the posterior parameter distribution. This method works for any model, regardless of hierarchical complexity or distributional assumptions. For an example of Bayesian power analysis, see Kruschke (in press), and for complete technical details, see Kruschke (2011a).

As an example of the use of power analysis in organizational science research, consider a recent study of situational influences on personality states through time (Huang & Ryan, 2011). Because some of the observed effects trended in the predicted directions but did not reach statistical

significance based on a traditional approach, the researchers conducted a power analysis to estimate the sample size that would have been needed to obtain a value of .80. The power and sample size estimates were based on the point estimate of an effect size in the frequentist analysis without considering the uncertainty of that point estimate. Therefore, we do not know whether the estimated power and needed sample size are stable or not. A Bayesian analysis, on the other hand, would take into account the uncertainty in the effect size estimate.

Concluding Comments

Although we believe that adopting a Bayesian approach has many benefits and advantages compared with frequentist analysis, we readily acknowledge several potential challenges. First, as is the case with any new methodological approach, researchers interested in adopting Bayesian methods will need to invest the necessary time and effort in the learning process. Given the several textbooks available, as well as introductory courses offered at many universities and professional organization meetings, we believe that the learning curve should not be too steep. However, we also recognize that understanding the benefits of a new approach and adopting it are best accomplished by actually trying it. The program described in the appendix is packaged for easy use with the illustrative data file we used in our article. Second, regarding the issue of implementation, researchers interested in adopting a Bayesian approach will also need to invest time and effort into learning new software tools. Although initially this may seem like a daunting process, packages are becoming increasingly user-friendly, and several textbooks provide step-by-step illustrations as well as data sets to facilitate the learning process (see the appendix; Culpepper & Aguinis, 2011; Kruschke, 2011a, in press).

Another issue to consider is that in actual empirical research, the sample of data comes from an unknown underlying process in the world. The researcher chooses a descriptive model based on theoretical motivations prior to the computational analysis. This choice of descriptive model confronts both Bayesian and non-Bayesian analysis, and it is not magically solved by a Bayesian approach. Once candidate models are selected, however, Bayesian methods are excellent for model comparison because they automatically take into account differences in the number of parameters and structural complexity. Bayesian methods therefore provide a formal mechanism for ''strong inference'' via competitive testing that has recently been advocated for the organizational sciences (Gray & Cooper, 2010; Leavitt, Mitchell, & Peterson, 2010).

Finally, it has been argued that Bayesian analysis is an almost magical inferential process that has been ''oversold as an all-purpose statistical solution to genuinely hard problems'' (Gelman, 2008, p. 446). To address this concern, it is important to understand that Bayesian methods are a type of data-analytic and inferential procedure and are affected by research design and measurement issues much like any other inferential data-analytic procedure. For example, if the data in hand were collected using a nonexperimental design, then we will not be able to draw conclusions about causality. Also, if the data were collected using unreliable (i.e., ''noisy'') measures, the resulting estimates will be imprecise, but the Bayesian analysis will explicitly reveal the uncertainty in the parameter values. In short, Bayesian analysis is an important improvement over existing frequentist methods because it provides information that researchers want to know: the credible parameter values, given the data that have been obtained thus far. However, using Bayesian statistics does not diminish the importance of research design and measurement choices, and it certainly is no substitute for good conceptual thinking.

In closing, Bayesian methods are essentially absent from the organizational science literature. There may be many reasons for why this may be the case, including a lack of clear understanding of advantages as well as a lack of clear guidelines on how to actually conduct and report results of a Bayesian analysis. As recently noted by Orlitzky (2012), ''Methods training should clearly and explicitly unmask as illusory the belief that NHST is some deus ex machina for instilling objectivity

and factuality . . . in the long run, this would involve a shift toward a Bayesian view of probability" (p. 209). We hope that our article will serve as a catalyst for the adoption of Bayesian methods in the organizational sciences, and we look forward to future Bayesian applications leading to important theoretical advances and organizational practices.

Appendix

Software for Bayesian Multiple Linear Regression

Complete information for installing the software is provided at http://www.indiana.edu/~kruschke/BMLR/. "BMLR" stands for Bayesian multiple linear regression. The programs linked in the website provide a complete working example, along with the data used for Figures 2 through 5 in the main text of the article.

After the software has been installed, please do the following. Open the file, BMLRexample.R, in the editor RStudio. Make sure that R's working directory is the folder with the BMLR programs. The program can be run "as is" to produce graphs such as Figures 2 through 5. The program also produces scatter plots of the posterior for all pairs of parameters, thereby showing the trade-offs in parameter estimates.

An initial step executed by the program BMLRexample.R is loading of the relevant data. The data are formatted in a matrix, such that there is one row per measurement unit, with the first column containing the predicted value (y), and the remaining columns containing the predictor values (x). The following programming code assumes that the data matrix is loaded into R in a matrix called `dataMat`. Then, the multiple linear regression functions are loaded into R's active memory, using the command:

```
source("BMLR.R")
```

The MCMC chain for the posterior distribution is generated with the command:

```
mcmcChain = BMLRmcmc(dataMat)
```

The posterior distribution is plotted using the command:

```
BMLRplot(mcmcChain)
```

The resulting plots can be saved as desired using the interactive menu in R, and the MCMC chain can then be saved, if desired, using R's save command. The BMLRplot command also outputs a numerical summary of the posterior distribution. The comments in the program, BMLRexample.R, provide more details.

The program is packaged such that the user does not need to delve any deeper into the inner workings of the software. But, one of the benefits of the software is that models can be modified or created for various applications, using, for example, nonnormal distributions or hierarchical structures. Here, we briefly explain how multiple linear regression is implemented, as an illustration of the model specification language. The key to understanding the model specification is that every arrow in the hierarchical diagram of Figure 1 has a corresponding statement in the model specification. For example, the lowest arrow in Figure 1, which points to the data from a normal distribution, is stated in the model specification as follows:

```
y[i] ~ dnorm(y.hat[i], tau)
```

The programming code says that y_i is distributed as a normal density with mean $\hat{y}_i$ and "precision" τ, where precision is the reciprocal of variance. (Bayesian specifications of normal distributions often use precision instead of standard deviation for historical reasons.) The next arrow up in Figure 1 indicates that the predicted value, $\hat{y}_i$, is a linear combination of the predictors. This is expressed in the model specification as follows:

```
y.hat[i] <- b0 + inprod(b[1:nPred], x[i,1:nPred])
```

where "inprod" denotes the inner product of the vector of regression coefficients and the vector of predictors. The complete model specification is as follows:

```
model {
  # Likelihood:
  for(i in 1:N) {
    y[i] ~ dnorm(y.hat[i], tau)
    y.hat[i] <- b0 + inprod(b[1:nPred], x[i,1:nPred])
  }
  # Prior (assumes standardized data):
  tau <- 1/pow(sigma,2)
  sigma ~ dunif(0, 10)
  b0 ~ dnorm(0, 1.0E-2)
  for (j in 1:nPred) {
    b[j] ~ dnorm(0, 1.0E-2)
  }
}
```

The model specification language can be modified to express a variety of other assumptions. For example, if the user wants to accommodate outliers in the data, then a *t* distribution can be used instead of a normal distribution. Polynomial trends or interaction terms can also be added to the prediction equation. Moreover, different prior assumptions can be implemented. For example, there might be strong prior knowledge about one of the regression coefficients, and this can be incorporated as well.

The model is built for software called JAGS (Just Another Gibbs Sampler; Plummer, 2003, 2011), which is written in C++ and usable on most computer platforms. JAGS is an alternative to its predecessor, called BUGS, which stands for Bayesian inference Using Gibbs Sampling. BUGS has a stand-alone version for Windows called WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), with a subsequent version called OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009). The JAGS modeling language is virtually the same as BUGS. Because BUGS got an earlier start, it is presently more extensively used than JAGS, but JAGS is gaining in popularity because it can be used more easily on non-Windows operating systems and is more robust in operation. Both BUGS and JAGS can be accessed from R and other software environments. All the software is available free of charge.

After the model is specified, there are only four more steps in programming a Bayesian analysis: loading the data, specifying initial values for the MCMC chain, running the MCMC chain, and examining the results. An important aspect of using MCMC methods for representing a distribution is making sure that the random chain is truly representative of the distribution, which means that the chain has smoothly explored the entire posterior distribution and does not have lingering influence from its starting position. To prevent a randomly outlying starting point from influencing the results, we chose the starting point to be in the midst of credible parameter values (via least squares estimation), and the initial 1,000 steps of the chain are excluded from consideration. This is called the *burn-in period*. Thanks to the speed of modern personal computers, the chain can be run to great lengths so that a very large sample of representative points is collected. We use 250,000 steps for illustration, but more can be used in research applications as desired. The MCMC sample therefore is robustly representative of the true underlying posterior distribution. In typical applications of linear regression, the Markov chains are sufficiently well behaved that a run of 250,000 steps provides a precise representation of the posterior distribution. The program allows a sophisticated user to check for autocorrelation and convergence of chains in cases with strongly correlated predictors. This issue

merely affects the efficiency of obtaining an MCMC representation of the posterior distribution, not the validity of Bayesian analysis.

Acknowledgment

We thank Jose M. Cortina and three *Organizational Research Methods* anonymous reviewers for highly constructive and detailed feedback that allowed us to improve our manuscript substantially.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *90*, 94-107.

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology*, *95*, 648-680.

Aguinis, H., & Harden, E. E. (2009). Cautionary note on conveniently dismissing χ^2 goodness-of-fit test results: Implications for strategic management research. In D. D. Bergh & D. J. Ketchen (Eds.), *Research methodology in strategy and management* (vol. 5, pp. 111-120). Howard House, England: Emerald Group Publishing.

Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. (2009). First decade of *Organizational Research Methods:* Trends in design, measurement, and data-analysis topics. *Organizational Research Methods*, *12*, 69-112.

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, *13*, 515-539.

Allenby, G. M., Bakken, D. G., & Rossi, P. E. (2004). The hierarchical Bayesian revolution: How Bayesian methods have changed the face of marketing research. *Marketing Research*, *16*, 20-25.

Anscombe, F. (1954). Fixed sample size analysis of sequential observations. *Biometrics*, *10*, 89-100.

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S. Communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S. *Philosophical Transactions*, *53*, 370-418. doi:10.1098/rstl.1763.0053

Beaumont, M. A., & Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics*, *5*, 251-261.

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

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews: Drug Discovery*, *5*, 27-36. doi:10.1038/nrd1927

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

Brooks, S. P. (2003). Bayesian computation: A statistical revolution. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *361*, 2681-2697.

Casciaro, T., & Lobo, M. (2008). When competence is irrelevant: The role of interpersonal affect in task-related ties. *Administrative Science Quarterly*, *53*, 655-684.

Cashen, L. H., & Geiger, S. W. (2004). Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies. *Organizational Research Methods*, *7*, 151-167.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, *49*, 997-1003.

Cornfield, J. (1966). A Bayesian test of some classical hypotheses, with applications to sequential clinical trials. *Journal of the American Statistical Association*, *61*, 577-594.

Cortina, J. M., & Folger, R. G. (1998). When is it acceptable to accept a null hypothesis: No way, Jose? *Organizational Research Methods*, *1*, 334-350.

Cortina, J. M., & Landis, R. S. (2009). When small effect sizes tell a big story, and when large effect sizes don't. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 287-308). New York, NY: Routledge.

Cortina, J. M., & Landis, R. S. (2011). The earth is not round ($p = .00$). *Organizational Research Methods*, *14*, 332-349.

Culpepper, S. A., & Aguinis, H. (2011). R is for revolution: A cutting-edge, free, open source statistical package. *Organizational Research Methods*, *14*, 735-740.

Cumming, G. (2007). Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics*, *29*, 89-93.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift fur Psychologie*, *217*, 15-26.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274-290.

Doyle, A. C. (1890). *The sign of four*. London, England: Spencer Blackett.

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods*, *13*, 668-689.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, *3*, 445-450.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gelman, A., Hill, J., & Yajima, M. (2009). *Why we (usually) don't have to worry about multiple comparisons* (Technical report). New York: Department of Statistics, Columbia University.

Gerard, P. D., Smith, D. R., & Weerakkody, G. (1998). Limits of retrospective power analysis. *Journal of Wildlife Management*, *62*, 801-807.

Gray, P. H., & Cooper, W. H. (2010). Pursuing failure. *Organizational Research Methods*, *13*, 620-643.

Gregory, P. C. (2001). A Bayesian revolution in spectral analysis. In A. Mohammad-Djafari (Ed.), *AIP conference proceedings* (pp. 557-568). Berlin, Germany: Springer.

Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.

Hirst, G., Van Knippenberg, D., & Zhou, J. (2009). A cross-level perspective on employee creativity: Goal orientation, team learning behavior, and individual creativity. *Academy of Management Journal*, *52*, 280-293.

Huang, J. L., & Ryan, A. M. (2011). Beyond personality traits: A study of personality states and situational contingencies in customer service jobs. *Personnel Psychology*, *64*, 451-488.

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

King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, *30*, 666-687.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kruschke, J. K. (2010a). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 658-676.

Kruschke, J. K. (2010b). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*, 293-300.

Kruschke, J. K. (2011a). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press/Elsevier.

Kruschke, J. K. (2011b, August 1). Extrasensory perception (ESP): Bayesian estimation approach to meta-analysis [Blog post]. Retrieved from http://doingbayesiandataanalysis.blogspot.com/2011/08/extrasensory-perception-esp-bayesian.html

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

Lance, C. E. (2011). More statistical and methodological myths and urban legends. *Organizational Research Methods*, *14*, 279-286.

Leavitt, K., Mitchell, T. R., & Peterson, J. (2010). Theory pruning: Strategies to reduce our dense theoretical landscape. *Organizational Research Methods*, *13*, 644-667.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions (with discussion). *Statistics in Medicine*, *28*, 3049-3082.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325-337.

Lykou, A., & Ntzoufras, I. (2011). WinBUGS: A tutorial. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*, 385-396.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147-163.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, NJ: Erlbaum.

McCloskey, D. N. (1995). The insignificance of statistical significance. *Scientific American*, *272*, 104-105.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, *16*, 617-640.

Mundfrom, D. J., Perrett, J. J., Schaffer, J., Piccone, A., & Roozeboom, M. (2006). Bonferroni adjustments in tests for regression coefficients. *Multiple Linear Regression Viewpoints*, *32*, 1-6.

Newman, M. E. J., Barabasi, A. L., & Watts, D. J. (2006). *The structure and function of dynamic networks*. Princeton, NJ: Princeton University Press.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301.

O'Boyle, E. H., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology*, *65*, 79-119.

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

Pierce, J. R., & Aguinis, H. (in press). The too-much-of-a-good-thing effect in management. *Journal of Management*.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3ssrd International Workshop on Distributed Statistical Computing, March 20-22*. Vienna, Austria: DSC. Retrieved from http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/

Plummer, M. (2011). rjags: Bayesian graphical models using MCMC. R package version 3-5 [Computer software]. Retrieved from http://CRAN.R-project.org/package=rjags

Poole, C. (1987). Beyond the confidence interval. *American Journal of Public Health*, *77*, 195-199.

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

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

Roth, P. L., Switzer, F. S., Van Iddekinge, C. H., & Oh, I. S. (2011). Toward better meta-analytic input matrices: How matrix values change conclusions in human resource management simulations. *Personnel Psychology*, *64*, 899-935.

Sasovova, Z., Mehra, A., Borgatti, S. P., & Schippers, M. C. (2010). Network churn: The effects of self-monitoring personality on brokerage dynamics. *Administrative Science Quarterly*, *55*, 639-670.

Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal*, *43*, 1248-1264.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, *12*, 347-367.

Schmidt, F. L. (2008). Meta-analysis: A constantly evolving research integration tool. *Organizational Research Methods*, *11*, 96-113.

Schmidt, F. L., & Raju, N. S. (2007). Updating meta-analytic research findings: Bayesian approaches versus the medical model. *Journal of Applied Psychology*, *92*, 297-308.

Schweder, T., & Hjort, N. L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics*, *29*, 309-332.

Singh, K., Xie, M., & Strawderman, W. E. (2007). Confidence distribution (CD)—distribution estimator of a parameter. In R. Y. Liu, W. E. Strawderman, & C.-H. Zhang (Eds.), *Complex datasets and inverse problems* (Vol. 54, pp. 132-150). Beachwood, OH: Institute of Mathematical Statistics.

Steiger, J. H., & Fouladi, R. T. (1992). R2: A computer program for interval estimation, power calculations, sample size estimation, and hypothesis testing in multiple regression. *Behavior Research Methods*, *24*, 581-582.

Sullivan, K. M., & Foster, D. A. (1990). Use of the confidence interval function. *Epidemiology*, *1*, 39-42.

Thomas, L. (1997). Retrospective power analysis. *Conservation Biology*, *11*, 276-280.

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

Western, B., & Jackman, S. (1994). Bayesian inference for comparative research. *American Political Science Review*, *88*, 412-423.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, *32*, 741-744.

Westlake, W. J. (1981). Response to bioequivalence testing—a need to rethink. *Biometrics*, *37*, 591-593.

Zhou, J. (2003). When the presence of creative coworkers is related to creativity: Role of supervisor close monitoring, developmental feedback, and creative personality. *Journal of Applied Psychology*, *88*, 413-422.

Bios

John K. Kruschke is Professor of Psychological and Brain Sciences and Adjunct Professor of Statistics at Indiana University. He is a seven-time winner of Teaching Excellence Recognition Awards from Indiana University, has authored an acclaimed textbook on Bayesian data analysis, and has presented numerous workshops. He guest-edited a section on Bayesian analysis for the journal *Perspectives on Psychological Science*, and is an action editor for the *Journal of Mathematical Psychology*. His current research interests include the science of moral judgment and the education of Bayesian methods. He received the Troland Research Award from the National Academy of Sciences, and is an elected Fellow of the Society of Experimental Psychologists and the Association for Psychological Science and other professional groups.

Herman Aguinis is the Dean's Research Professor, a professor of organizational behavior and human resources, and the founding director of the Institute for Global Organizational Effectiveness in the Kelley School of Business, Indiana University. He has been a visiting scholar at universities in the People's Republic of China (Beijing and Hong Kong), Malaysia, Singapore, Argentina, France, Spain, Puerto Rico, Australia, and South Africa. His research interests span several human resource management, organizational behavior, and research methods and analysis topics. He has published five books and more than 100 articles in refereed journals. He is the recipient of the 2012 Academy of Management Research Methods Division Distinguished Career Award and a former editor-in-chief of *Organizational Research Methods*.

Harry Joo is a doctoral student in organizational behavior and human resource management in the Kelley School of Business, Indiana University. His research interests include performance management and research

methods and analysis. His work has appeared in several refereed journals including *Organizational Research Methods, Academy of Management Perspectives*, and *Business Horizons*. He has delivered presentations at the meetings of the Academy of Management and elsewhere.