American Journal of Evaluation
2018, Vol. 39(2) 278-289
© The Author(s) 2017
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1098214017704302
journals.sagepub.com/home/aje
SAGE

Bayesian Posterior Odds Ratios: Statistical Tools for Collaborative Evaluations

Tyler Hicks[1], Liliana Rodríguez-Campos[2], and Jeong Hoon Choi[1]

Abstract

To begin statistical analysis, Bayesians quantify their confidence in modeling hypotheses with priors. A prior describes the probability of a certain modeling hypothesis apart from the data. Bayesians should be able to defend their choice of prior to a skeptical audience. Collaboration between evaluators and stakeholders could make their choices more defensible. This article describes how evaluators and stakeholders could combine their expertise to select rigorous priors for analysis. The article first introduces Bayesian testing, then situates it within a collaborative framework, and finally illustrates the method with a real example.

Keywords

Bayesian statistics, collaborative evaluation, Bayes's factor, posterior odds ratios

Bayesian methods are promising statistical tools for evaluators (Pollard, 1986). Yet a lesser known fact is that they could facilitate collaboration. Evaluators have multiple reasons for collaborating with stakeholders (Azzam, 2010; Brandon, 1989; Cousins & Earl, 1992; Fetterman, 2001; House & Howe, 1998; Orr, 2010; Patton, 1978; Smith, 1999). For example, collaboration could improve relevance, shared ownership, and accuracy of evaluations (Rodríguez-Campos, 2012). This article clarifies how authentic collaboration could increase the rigor of data analysis with Bayesian methods.

Bayesian methods, in particular Bayesian testing, have a different logic than conventional statistical tests, such as t tests, F tests, and χ^2 tests. Because Bayesian logic may be unfamiliar, we begin with an overview to foreground the methods described in this article in their theoretical context. After we present an overview, we provide an example illustrating the focal procedure in an actual collaborative evaluation.

The Bayesian Approach to Statistical Testing

This section introduces the logic of Bayesian methods to newcomers, and those already familiar with it can skip this section. Suppose a fortune-teller claimed to be a reliable predictor of the future.

[1] University of Kansas, Lawrence, KS, USA
[2] University of South Florida, Tampa, FL, USA

Corresponding Author:
Tyler Hicks, Department of Special Education, University of Kansas, 1450 Jayhawk Blvd, Lawrence, KS 66045, USA.
Email: tahicks@ku.edu

Skeptical, you test her claim. You discover that in 10 of 12 attempts, she predicted the outcome of a fair coin toss. Perhaps she got lucky. You decide to formally test this model (i.e., she was guessing).

Traditional tests only "reject" or "fail to reject" models depicting null hypotheses. This binary is not optimal. It makes accepting the null model impossible—a "failure to reject" is not the same as an "accept." Although you prefer the null model, you proceed with the conventional test. Much to your frustration, the p value comes out very low ($p < .05$). Thus, the test rejects the null model at the 5% level.

Despite the unwelcome outcome, you persist in clinging to the null model. You still find it to be defensible in comparison to the alternative (i.e., the fortune-teller predicted the future). You argue to yourself that, despite the p value, the best science supports the null model. Albeit less certain, you are not yet ready to give up the null model.

Your reaction in the above scenario strikes Bayesians as prudent (Berger, 1985). One problem with traditional tests, according to Bayesians, is that they may exclude knowledge that should stay in play (i.e., background knowledge). Although background knowledge can have a subjective component, its integration could improve a test's accuracy (Bolstad, 2007). When tests consider all information, including contextual evidence, they reach better conclusions (Howson & Urbach, 2006).

Recall, when conventional p value procedures reject the null, we knock the null out of the competition. Yet p values say nothing about the probability of the null model given data. Properly interpreted, a p value only tells us how frequently random sampling would yield data more surprising than the obtained data, if the null was true. Thus, a null with a low p value may still end up being more probable than an alternate model, given our contextual knowledge (Hoff, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009). In this light, testers may prefer a different approach to testing.

Bayesian Testing

The foremost consideration in Bayesian testing is the probability of a model given sample data, $P(M|Y)$, where M is the model and Y is the obtained data. In Bayesian vocabulary, $P(M|Y)$ is called a "posterior." If M_0 and M_1 denote a null and alternate model, respectively, then testers compare their posteriors. If $P(M_0|Y) = 0.80$ and $P(M_1|Y) = 0.20$, then the null is 4 times more probable than the alternative given data.

Bayesian methods derive their name from the formula for posteriors, Bayes's rule:

$$P(M|Y) \propto P(M)P(M|Y), \tag{1}$$

where $P(M)$ is the prior and $P(M|Y)$ is the likelihood.

Simply put, the posterior is proportional to the product of the "prior" and the "likelihood." The prior is the credibility of the model apart from data. The likelihood is the chance that we obtained the data given the model.

Posterior odds ratios (PO_{01}) rather than p values facilitate Bayesian testing, briefly:

$$PO_{01} = \frac{P(M_0|Y)}{P(M_1|Y)}, \tag{2}$$

where M_0 is the null model and M_1 is the alternate model.

The subscripts in PO_{01} indicate that the null model is in the numerator position (PO_{10} indicates that the alternate is in the numerator).

Posterior odds ratios (PO_{01}) are the product of two terms, odds prior ratios and Bayes's factor:

$$\frac{P(M_0|Y)}{P(M_1|Y)} = \frac{P(M_0)}{P(M_1)}\frac{P(Y|M_0)}{P(Y|M_1)}, \tag{3}$$

where $\frac{P(M_0)}{P(M_1)}$ is the ratio of priors and $\frac{P(Y|M_0)}{P(Y|M_1)}$ is the ratio of likelihoods.

The ratio of likelihoods is known as Bayes's factor (B_{01}). It quantifies the relative disparity in the likelihoods of data under each model. A B_{01} of 10, for example, shows that the data are 10 times likelier under the null than alternate model. The ratio of prior probabilities quantifies the relative disparity in priors between models. A prior odds ratio of 1, for example, indicates no disparity in priors.

Priors introduce contextual knowledge into the analysis. Returning to the fortune-teller example, you strongly expected M_0 to be accurate. Setting $P(M_0) = 0.99$ is one way you could incorporate your scientific expectation into the test. Because only two models were considered in the example, the prior for the alternate must be $0.01[0.01 = 1 - P(M_0)]$. Priors, like any model component, need only to approximate reality to work.

In the alternate model, modelers need to depict the fortune-teller as a reliable predictor of the future, yet they may disagree over how to quantify "reliable." Obviously, reliability should exceed 50% (i.e., better than a coin flip). However, reliable could be far less than 100%. Perhaps it makes sense to operationalize reliability as 75% (i.e., halfway between 50% and 100%). Later, we will show how Bayesians could define reliability with a range of values.

=If analysts model the number of correct predictions as binomially distributed, then they could test two models:

$$M_0 : 10 \sim \text{Binomial}(\theta = 0.50,\ N = 12), \tag{4}$$
$$M_1 : 10 \sim \text{Binomial}(\theta = 0.75,\ N = 12), \tag{5}$$

where θ is the probability of a successful prediction and N is the number trials.

Because a fair coin toss only has two outcomes, random guessing has a 50% chance of being accurate. Consequently, θ is set to 0.50 in the null model (M_0) to express the "lucky guess" scenario. In the alternate model (M_1), θ is set to 0.75 to depict a reliable prediction rate.

The prior odds ratio favors the null model at 99:1 [99 = 0.99/0.01]. The null is thus 99 times more probable than the alternate apart from data. The Bayes's factor (B_{10}), obtained by hand calculations favors the alternate at 14.416:1. It indicates that the data are 14.416 times likelier if she reliably predicted them rather than guessed. Balancing these considerations with Bayes's rule, the alternate will win (i.e., $PO_{01} = 6.87$). Yet because the odds for your null shrank from 99 to 6.87 times, you might collect more data to resolve the matter.

Bayesian Modeling

Recall the value of θ in M_1 was set to 0.75. But why not 0.76 or 0.74? In this light, 0.75 seems arbitrary. Uncertainty about how to set θ can be made transparent in Bayesian modeling through priors. Setting a prior on θ would indicate the uncertainty modelers had about the best value to plug into the alternate model before they examined data.

A sensible choice for a prior on θ in M_1 could be a uniform distribution from 0.60 to 1. This prior affords flexibility. It assumes our fortuneteller's accuracy rate could be as low as 0.60 or as high as 1. This new alternate model ($M_1{}'$) can be formally stated as follows:

$$M_1{}' :\ 10 \sim \text{Binomial}(\theta,\ N = 12)\ P(\theta) \sim \text{Uniform}(0.60, 1). \tag{6}$$

Comparison of M_1 and $M_1{}'$ reveals that they only differ in the value of θ. In M_1, the value of θ is certain 0.75. However, in $M_1{}'$ the value of θ is uncertain. It could be anywhere from, say, 0.60 to 1, and any value in that range is as credible as another.

To perform a Bayesian test comparing M_0 (null) to $M_1{}'$ (new alternative), testers would again obtain the requisite Bayes's factor (B_{10}). This time it requires taking an integral and can be accomplished by hand:

$$\frac{L(Y|M_1{}')}{L(Y|\,M_0)} = \int_{P=0.6}^{P=1} \frac{p^{10}(1-p)^2}{(0.5)^{12}}\,dp. \tag{7}$$

The calculus derivation (not shown) intimates the data are 4.497 likelier under $M_1{}'$ than M_0. If the previous arrangement for priors is employed, that is, $P(M_1{}') = 0.01$ and $P(M_0) = 99$, then $PO_{01} =\ 22.01$. The null model is (about) 22 times more probable.

Bayesian and Traditional Testers

The abovementioned example shows that the conclusions of Bayesian and traditional testing can diverge. Defenders of traditional tests, such as t tests, χ^2 tests, and F tests are frequentists. This name derives from their concern about a statistical procedure's frequency properties (Hacking, 2001). For example, frequentists designed p value procedures to control error rates in the limits (e.g., a t test may have a 5% chance of making a Type I error in repeated applications). Bayesian tests, in contrast, serve a different purpose.

Instead of balancing error rates in the limits, Bayesians conform tests to the formal logic of induction (Howson & Urbach, 2006). Bayesian tests subscribe to Bayes's rule—a rule showing how a rational agent would modify uncertain beliefs with evidence (Jaynes, 1968). Should tests balance error rates or conform to Bayes's rule? It depends on the goals of the tester (Gelman & Shalizi, 2013). Frequentists care about what happens in the limits. However, Bayesians such as Keynes (1923) have argued that controlling what happened in the limits was moot because "we are all dead in the long run" (p. 80).

To recap, the fortune-telling example illustrates the usefulness of priors for data analysis. With a frequentist test, we would reject the hypothesis that the fortune-teller was guessing because of a low p value. However, this conclusion ignores contextual knowledge about fortune telling. When we appropriate outside knowledge into the analysis through a prior, we would conclude the null was more probable than the alternative. Posterior odds ratios, unlike p values, consider the data as well as contextual knowledge. The remainder of this article explores how to pick defensible priors for analysis.

A Test for Collaborative Evaluators

The Model of Collaborative Evaluation

Several collaborative methodologies exist (Fetterman, Rodríguez-Campos, Wandersman, & O'Sullivan, 2014), each has advantages and disadvantages. We prefer the one articulated in the model for collaborative evaluations (MCE; Rodríguez-Campos & Rincones-Gómez, 2012). The MCE has six interdependent components: (a) identify the situation, (b) clarify the expectations, (c) establish a collective commitment, (d) ensure communication is open, (e) encourage effective practices, and (f) follow specific guidelines. Within an MCE approach, evaluators retain control of production while collaborating with stakeholders. This arrangement helps safeguard the credibility of evaluation products, such as models, while integrating collaboration into the design.

Capitalizing on the MCE, we propose testers blend insights from both evaluators and stakeholders to create models depicting rival hypotheses. When people invest in the models, they are more likely to care about the models. Another advantage of teamwork is that it increases the buy-in for the overall evaluation finding. There are three steps to constructing an insightful prior in partnerships with stakeholders (Gill & Walker, 2005): (a) select an appropriate distribution for the prior (e.g., a normal-shaped, a *t*-shaped, or customized), (b) specify the parameters that will define it, if any (e.g., a normally shaped prior can be defined with a mean and a standard deviation), and then (c) crossvalidate the prior. This last step typically involves verifying that the prior is a passable representation of an expert's opinion—that is, one that could survive critical scrutiny.

An Example of Collaboratively Setting Priors

A good partner in prior setting would be a stakeholder who (a) had limited exposure to the data (i.e., the sample being tested), (b) was well positioned to articulate hypotheses (e.g., participated in past evaluations, a qualified expert in the content area, some understanding of statistics), and (c) would be a credible expert to others.

To illustrate, suppose a prior must be set on an effect (μ_d). First, you specify a shape for the distribution. You either select a standard distribution for the prior shape, such as normal, or ask the stakeholder to help you stretch out its shape (Kruschke, 2011). Let us suppose we picked a normal shape. Next, ask a stakeholder to provide an upper and lower boundary (i.e., What is the smallest credible effect? What is the largest credible effect?). These two guesses mark the points three *SD*s away from the mean.

To be more concrete, suppose that a stakeholder thinks a phonics-based reading program works well for fifth graders struggling to decode words. The stakeholder frames her guesses about the effect in a standardized metric (e.g., Cohen's *d*). She posits that the lowest credible effect is $d = 0.2$ while the highest passable effect is $d = 0.8$. She thinks any estimate outside of that range is unlikely. Combining your chosen shape for the prior (i.e., normal) and her best guess of the effect, a mean and *SD* for the prior can be defined, as shown below.

The requisite prior can be given a mean of 0.5 (i.e., the midpoint between 0.2 and 0.8) and a *SD* of 0.1 (i.e., $0.1 = (0.8 - 0.5)/3$). Let μ' denote this mean and σ' denote this *SD*, then the normally shaped prior for the effect can be precisely defined as follows: $P(\mu_d) \sim N(\mu' = \ 0.5, \sigma' = \ 0.1)$. This notation is read as the prior for the effect is normal with a mean (μ') of 0.5 and an *SD* (σ') of 0.1. Testers can use this prior to build an alternate model for testing against a null hypothesis.

An Example of a Bayesian Test for Collaborative Evaluation

This example comes from the education sector and is focused on a new school reform model—the Schoolwide Integrated Framework for Transformation (SWIFT). SWIFT was the product of the SWIFT Center, a national K–8 technical assistance center at the University of Kansas. The core of SWIFT reform is a Multi-Tiered System of Support (MTSS; McCart, Sailor, Bezdek, & Satter, 2014). A fuller description of MTSS can be found online (www.swiftschools.org). For present purposes, it suffices to know that schools outfitted with MTSS can educate students, with and without disabilities, in integrated environments. In MTSS, all school resources are accessible to all students rather than to just the select few taught in entitlement programs.

In this evaluation, evaluators sought to quantify the typical effect of MTSS on reading outcomes. Two elementary schools participated. In the first year of implementation, evaluators visited each school and rated fidelity of implementation using a validated tool, SWIFT-FIT (Algozzine et al., 2014). The fidelity of implementation was low at one school and moderate at the other.

The evaluators matched schoolchildren between schools using propensity score matching. The evaluators matched schoolchildren on baseline covariates linked by past research to reading outcomes, such as previous reading achievement, grade, and disability. For the main analysis, differences in annual reading gains among matched pairs ($n = 20$), as indexed by state reading tests, was computed using the following formula:

$$Y_D = Y_M - Y_L, \tag{8}$$

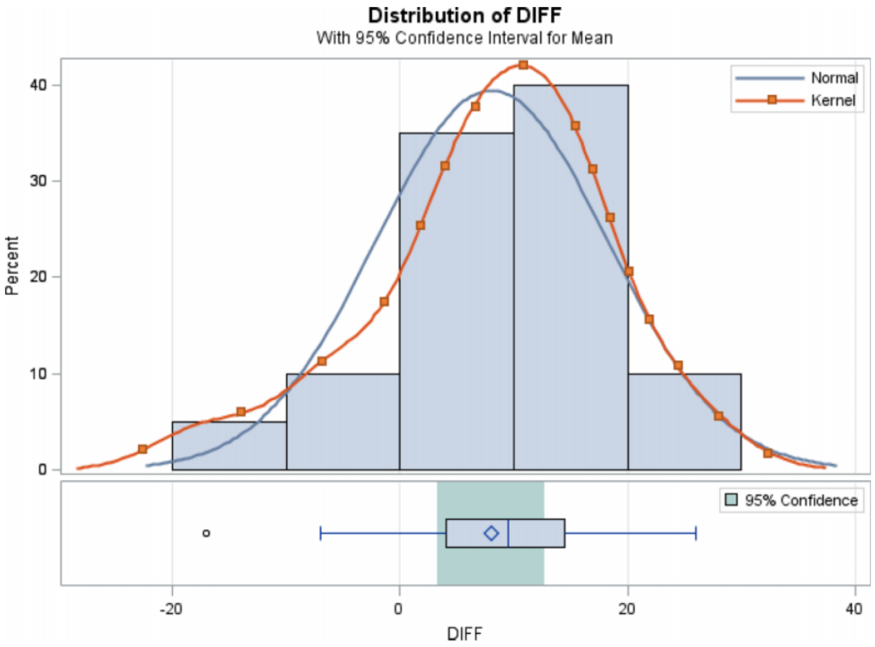

Figure 1. Visual depiction of the difference in scores among matched pairs of schoolchildren at low and moderate fidelity schools, $Y_D = Y_{\text{treated}} - Y_{\text{control}}$. As a reference, we imposed a "normal" and "kernel" density to reconstruct the population density. The first curve assumes normalcy and the second curve makes no such assumptions. We obtained the kernel curve by "smoothing" the obtained sample distribution.

where Y_D represents the difference in annual reading gain between paired students and Y_M and Y_L the annual reading gain of matched students from the moderate- and low-fidelity school, respectively. Figure 1 depicts the distribution of difference scores. The typical difference in annual gain scores in reading between matched pairs in the sample $(\bar{Y}_D \approx\ 6; SD \approx 10)$ was found to be of moderate size, $d \approx 0.6$.

To proceed with a Bayesian test, evaluators first needed to identify a model structure. The data were ultimately modeled as being normally distributed:

$$\begin{aligned} &Y_D \sim N(\mu_D, \sigma_{Y_D}), \\ &\text{where } \mu_D = \Delta(\sigma_{Y_D}). \end{aligned} \tag{9}$$

In this model, Δ is the standardized effect, μ_D the raw effect, and σ_{Y_D} the *SD*. This parameterization is convenient because a prior can be set on Δ rather than μ_D.

Next, evaluators needed to set up a null and alternate model. The value of the effect (Δ) in the null model (M_0) can be set to 0. However, the claim "MTSS has an effect" (i.e., $\Delta \neq 0$) is too vague for an alternate model in a valid Bayesian test. To build an adequate model depicting an effective MTSS scenario, evaluators had to collaborate with stakeholders. In particular, evaluators needed help formulating a meaningful prior on Δ in the alternate model.

To move forward in Bayesian modeling, evaluators partnered with a stakeholder who had extensive experience with both MTSS and evaluations. She participated in a related pilot study on MTSS in another school district. In that study, her team had not seen dramatic increases in outcomes until full fidelity of implementation was reached. She thus did not expect to see a large effect in this case but considered a modest effect likely.

A prior for Δ in the alternate model was coconstructed using the above information. Evaluators first asked the stakeholder to guess the smallest and largest credible effect. She responded with $d = 0.4$ and $d = 0.7$, respectively. She did not think estimates outside of these boundaries were believable. Evaluators decided to use the procedure described in the last section to translate these guesses into a normal prior. The result was a normal prior with a mean of 0.55, that is, $0.55 = (0.4 + 0.7)/2$, and an *SD* of 0.05, that is, $0.05 = (0.55 - 0.40)/3$.

Having settled on adequate values for Δ in both models, evaluators subsequently considered the second parameter, σ_{Y_D}. They treated it as a nuisance parameter rather than the focus of the test. Consequently, they set a "neutral" prior on it, Jeffreys (1946) prior. Jeffreys priors can be determined by formulas. For an *SD*, the formula for a Jeffreys prior is:

$$P(\sigma_{Y_D}) \propto \frac{1}{\sigma_{Y_D}}. \tag{10}$$

A Jeffreys prior exerts little influence on posteriors, and posteriors make convenient defaults for nuisance parameters needing priors.

The null and alternate models to be tested in this evaluation can be formally defined as:

$$M_0 : Y_D \sim N(0, \sigma_{Y_D})\ P(\sigma_{Y_D}) \propto \frac{1}{\sigma_{Y_D}}, \tag{11}$$

$$M_1 : Y \sim N(\Delta\sigma_{Y_D}, \sigma_{Y_D})\ P(\sigma_{Y_D}) \propto \frac{1}{\sigma_{Y_D}}, P(\Delta) \sim N(\mu' = 0.55, \sigma' = 0.05), \tag{12}$$

where μ' and σ' are the prior's mean and *SD*.

Having specified these models, the next step for the evaluators was to obtain the Bayes's factor. Evaluators used a procedure in SAS 9.4 (PROC MCMC) designed for Bayesian modeling to obtain Bayes's factor. Annotated SAS code is available with the article online at http://journals.sagepub.com/doi/suppl/10.1177/1098214017704302.

Obtaining Bayes's Factors With Simulation Methods

To estimate B_{01} with PROC MCMC, testers needed to set up a complex model, wherein M_0 and M_1 were included as subcomponents (Kruschke, 2011). In this umbrella model, a new parameter, say M_i, is introduced to index the two models. This new parameter takes on two possible values, 1 and 0. The conditional probability of Δ given M_i is:

$$P(\Delta|M_i)\begin{cases} M_0(\Delta) \text{ if } M_i = 0 \\ M_1(\Delta) \text{ if } M_i = 1 \end{cases}. \tag{13}$$

The notation $M_0(\Delta)$ is interpreted as Δ defined in accordance with M_0. Thus, if, in the big model, $M_i = 0$, then $\Delta = 0$, as M_0 states. (Otherwise, M_1 wins.) The parameter M_i needs a prior. If a binary prior is set on M_i, that is, $M_i \sim \mathrm{B}(p_0)$, then p_0 can be given an impartial value, such as 0.5, to obtain the Bayes's Factor (Christensen, Johnson, Branscum, & Hanson, 2011).

Testers then used Markov Chain Monte Carlo (MCMC) simulation to derive the relevant posterior. Recall that statisticians are fond of using samples to reconstruct unknown populations. Astonishingly, statisticians can reconstruct abstract populations, such as posteriors, through sampling. Posteriors are populations of possible parameter values and, hence, recoverable through sampling. In particular, MCMC methods simulate sampling from posteriors. The logic of MCMC methods has been amply described elsewhere (Kruschke, 2011). In short, MCMC involves randomly jumping back and forth between values in the target population. At each jump, we will sample either the new or the last value, depending on their relative frequencies in the population. This random walk then creates a "chain" of simulated draws from the target density.

Table 1. Selected Output of Records From a Posterior Sample Drawn Generated Using Markov Chain Monte Carlo Simulation Methods for the Bayesian Test of the Schoolwide Integrated Framework for Transformation Program.

Records	Δ	σ	M_I	LogPrior	LogLike	LogPost
32	.9640	14.8114	0	−5.1905	−79.6690	−84.8594
78	.7987	12.5666	0	−4.8624	−79.2548	−84.1172
303	.7384	10.3979	1	−4.6146	−74.2091	−78.8238
500	.8792	9.1228	1	−4.6213	−74.2672	−78.8885

Note. The posterior sample had 100,000 records, and M_I obtained the value 1 nearly 95% of the time. The selected output suggests that when the random walk stumbled upon relatively high values for σ (say 15), then the posterior for the null model might win on that occasion. Because σ rarely happened to draw a value that high, the alternate model won the vast majority of time. The ratio of wins for the alternate model estimates its posterior probability.

In this example, when M_i is included, at each step in an MCMC simulation it will be assigned either a 1 or 0 depending on which model (i.e., M_0 or M_1) yielded the higher posterior for the set of parameter values at that iteration (i.e., Δ, σ). For example, if $M_i = 1$ in the third step of the chain, then M_1 must have had a higher posterior at whatever spot it landed. As the values of M_i fluctuate back and forth between 1 and 0, the consequent posterior sample will eventually approximate Bayes's factor.

Table 1 records a few draws from an MCMC simulation to illustrate M_i. The table shows that at each iteration, M_i indexed which model yielded the highest posterior on that particular draw. Because M_i was given a neutral prior and 1s counted in favor of the alternate, its posterior will be equivalent to an estimated likelihood of the data under the alternate, $L(Y|M_1)$. To transform this posterior into the estimated Bayes's factor, plug the obtained likelihood value into the equation for $B_{10}\left(\text{i.e.,} B_{10} = \frac{L(Y|M_1)}{1-L(Y|M_1)}\right)$.

The Results and Interpretation of Bayesian Tests

To perform a Bayesian statistical test, evaluators requested that PROC MCMC produce a posterior sample with 10,010,000 draws to estimate B_{10}. To reduce autocorrelation, they discarded the first 10,000 (the "burn-in period") and retained every 1,000th iteration thereafter ("thinning"). The reduced MCMC chain had 100,000 iterations. Because convergence to the posterior is not a guaranteed proposition, diagnostics were performed to evaluate the chain's convergence. Figure 2 presents a summary of diagnostics information outputted from PROC MCMC.

The estimate for the model index parameter (M_i) drawn from a posterior sample obtained through MCMC methods strongly indicated that, in comparison to the null model (M_0), the alternate model (M_1) is more probable than the null given data ($M_i > 0.94$). If a default prior rather than a coconstructed one had been used in the alternate model then the estimated for B_{10} would be 18.6971 (i.e., $18.6971 = \frac{0.9492}{1-0.9492}$). This suggests data would be almost 19 times likelier on the alternate than the null. But B_{10} rose to 110.1 when a coconstructed prior was used rather than a neutral prior. This increase shows why constructed priors can be real assets. Kass and Raftery (1995) reported a reference scale to help interpret B_{10}. On their scale, a $B_{10} > 3$ is positive evidence, a $B_{10} > 20$ is strong evidence, and a $B_{10} > 150$ is very strong evidence.

Frequentist testers might conclude "There was a statistically significant difference ($M = 6$, $SD = 10$) between the groups, $t(19) = 3.56$, $p = .0021$." Yet, in this case, their test presumes the fiction of randomization (i.e., they randomly selected 20 matched pairs from the population to perform the test). Without such a fiction, there is no matter of fact about the correct p value (Howson & Urbach, 2006). Simply put, we cannot predict what will happen in an infinite series of

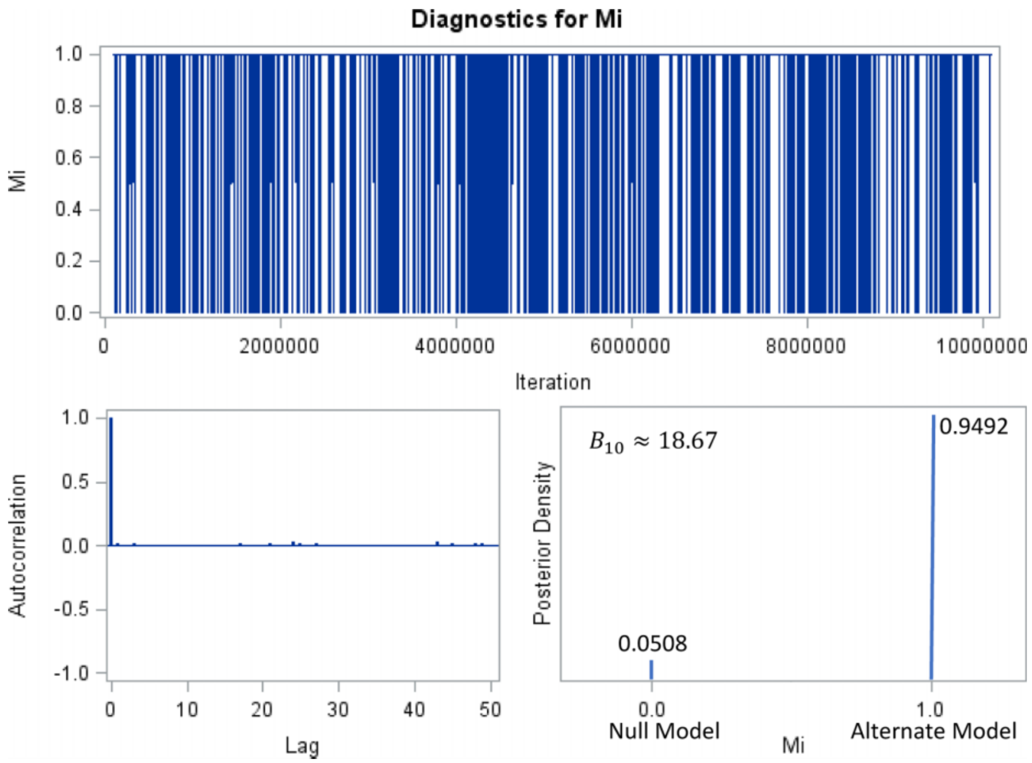

Figure 2. Summary of diagnostic plots on the posterior sample yielded by Markov Chain Monte Carlo (MCMC) simulation to check for convergence. Recall that M_i can assume any value between 0 and 1, and if $M_i = 0.5$ then $B_{10} = 1$. The trace plot (top) shows values for M_i returned at steps in the "random walk." This plot indicates adequate exploration. The plot of lag (bottom right corner) indicates minimal autocorrelation in the sample. Thus, the MCMC simulated sample approximates a simple random sample of the posterior. The posterior plot (bottom left corner) depicts the posterior for the model index parameter (M_i). We can transform this posterior into an approximate Bayes's factor.

replication studies, even if we assume the null was true, unless we first imagine testers used a random sampling plan. Bayesian tests do not require such a fiction for their warrant. The data, as the Bayesian test showed, were likelier to occur under the alternate than null. The point is that Bayesian tests are logically valid even in cases when samples are not truly random, a common occurrence in evaluations (Hicks, 2015).

In the official write-up, testers could nicely summarize Bayesian test results as follows:

Bayes's factor analysis with a coconstructed prior on the effect in reference to the null indicates the obtained data was almost 110 times likelier to occur given a moderate effect of MTSS than a null effect, as indexed by state tests ($\hat{\Delta} = 0.55;\ \beta_{01} = 110.1$).

Discussion

The integration of prior probabilities into statistical tests requires a change in how we evaluate test performance. Performance standards must align with statistical paradigms. In a conventional paradigm, a test's abstract performance in the limits is what matters. Frequentists, for example, insist a valid *t* test should balance Type I error rate and Type II error rate. Bayesians do not normally design their tests to meet frequentist standards, like balanced error rates.

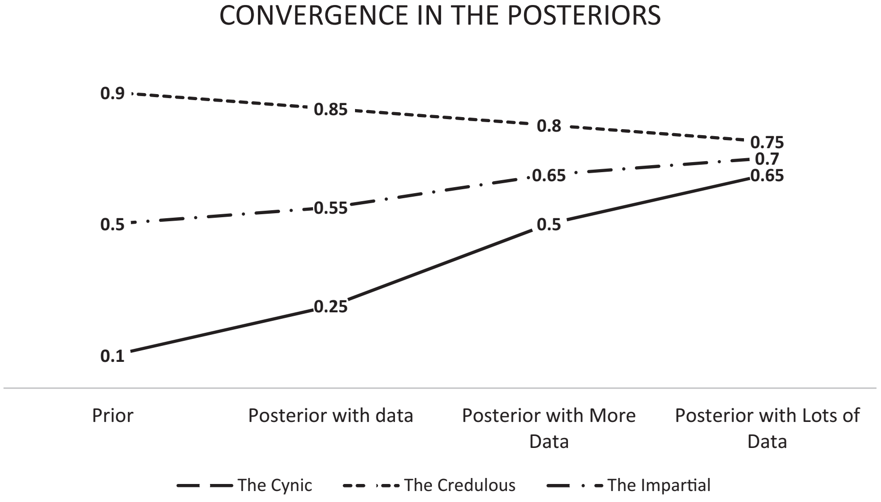

Figure 3. Illustration of Bayesian convergence. The cynic, credulous, and impartial tester start with different priors in the beginning but converge in their posteriors as evidence accumulates so long as they all use Bayes's rule to modify their beliefs as evidence arrives.

Recall, Bayesian tests conform to Bayes's rule. Given two rival models and the evidence, a posterior odds ratio shows which model is more probable given evidence and by how much. As long as the test conforms to inductive logic (i.e., Bayes's rule), Bayesians will not care how the tests performs in an infinite series of repeated applications. Testers who use posterior odds ratios have switched from a frequentist to a Bayesian statistical paradigm.

The subjectivity of Bayesian modeling is worth some discussion. Whether in a Bayesian or frequentist context, testing is controlled by subjective choices (e.g., in a frequentist framework testers must set an α-level). Because priors are a component of models in a Bayesian scheme, Bayesian tests are sensitive to them. It is thus helpful to see how the choice of priors affects the test by running the test a few times with different priors. One may find that results were not very sensitive. However, if they were sensitive to the chosen prior, then the choice of prior is critical. In such situations, collaboration in the process of setting up the prior could increase buy-in for the test outcome.

Moreover, testers always have the option of gathering more data to make conclusions insensitive to priors. Bayesian convergence theorems guarantee that, under sensible conditions, large samples cause likelihoods to dominate priors in the formation of posteriors (Hawthorne, 1994). Priors count most when data are scarce. This strikes us as exactly how it should be. Before evidence reaches a critical mass, reasonable people will disagree.

To illustrate Bayesian convergence theorems, consider three scientists. One thinks a hypothesis is highly probable, another thinks it highly improbable, while the last is unsure. Naturally, if they use different priors, their posteriors will differ. However, if more data arrive, and they recycle their old posteriors as new priors, they will reach consensus. Figure 3 illustrates such convergence, regardless of prior, as evidence accumulates. Note that the theorem holds regardless posteriors were obtained (e.g., analytically, MCMC simulation, etc.).

Final Remarks

Collaboration rather than formulas may be the more successful route for prior setting in many evaluations. This recognition affords room for innovation in mixed methods. Perhaps testers could deploy qualitative methods to attain insights from evaluators and stakeholders about modeling hypotheses. This mixed methods approach could enhance the rigor of data analysis. Such a novel realignment of qualitative and quantities methods is long overdue.

Bayesian methods permit testers a way to leverage previous knowledge, make evaluators and stakeholders partners in analysis, and accept null models. Yet Bayesian tests are only as meaningful as the models they compare (Kruschke, 2011). Consequently, statisticians have sought to automate the process of setting up priors to build models (Kass & Wasserman, 1996; Rouder et al., 2009). Yet, in all the hype over formulas, one can forget about the rewards of using collaboration rather than a textbook equation to specify priors.

Authors' Note

The views expressed herein do not necessarily represent the positions or policies of the Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service, or enterprise mentioned in this publication is intended or should be inferred.

Acknowledgments

OSEP Project Officers Grace Zamora Durán and Tina Diamond served as the project officers.

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) disclosed receipt of the following financial support for the research and/or authorship of this article: SWIFT Center produced the data for this article under U.S. Department of Education, Office of Special Education Programs Grant No. H326Y120005.

Supplemental Material

The annotated SAS code is available with the article online at http://journals.sagepub.com/doi/suppl/10.1177/1098214017704302.

References

Algozzine, B., Morsbach Sweeney, H., Choi, H., Horner, R., Sailor, W., McCart, A., . . . Lane, K. (2014). *SWIFT fidelity implementation tool: Development and preliminary technical adequacy*. Lawrence: University of Kansas, SWIFT Center.

Azzam, T. (2010). Evaluator responsiveness to stakeholders. *American Journal of Evaluation*, *31*, 45–65.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York, NY: Springer.

Bolstad, W. M. (2007). *Introduction to Bayesian statistics*. Hoboken, NJ: John Wiley.

Brandon, P. R. (1989). Stakeholder participation for the purpose of helping ensure evaluation validity: Bridging the gap between collaborative and non-collaborative evaluations. *American Journal of Evaluation*, *19*, 325–337.

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). *Bayesian ideas and data analysis*. New York, NY: CRC Press.

Cousins, J. B., & Earl, L. M. (1992). The case for participatory evaluation. *Educational Evaluation and Policy Analysis*, *14*, 397–418.

Fetterman, D. M. (2001). The transformation of evaluation into a collaboration: A vision of evaluation in the 21st century. *American Journal of Evaluation*, *22*, 381–385.

Fetterman, D. M., Rodriguez-Campos, L., Wandersman, A., & O'Sullivan, R. (2014). Collaborative, participatory and empowerment evaluation: Building a strong conceptual foundation for stakeholder involvement approaches to evaluation. (A response to Cousins, Whitmore and Shulha, 2013) [Letters to the Editor]. *American Journal of Evaluation*, *35*, 144–148.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 8–38.

Gill, J., & Walker, L. D. (2005). Elicited priors for Bayesian model specifications in political science research. *Journal of Politics*, *67*, 841–872.

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

Hawthorne, J. (1994). On the nature of Bayesian convergence. *Philosophy of Science*, *1*, 241–249.

Hicks, T. (2015). *What you know counts: How to elicit priors from experts to improve quantitative analysis with qualitative analysis* (Doctoral dissertation). Retrieved from http://scholarcommons.usf.edu/etd/5493

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

House, E. R., & Howe, K. R. (1998). The issue of advocacy in evaluations. *American Journal of Evaluation*, *19*, 233–236.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). Chicago, IL: Open Court.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions of System Science and Cybernetics*, *4*, 227–241.

Jeffreys, H. (1946, September). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *186*, 453–461. doi:10.1098/rspa.1946.0056

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

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

Keynes, J. (1923). *A track on monetary reform*. Toronto, Canada: MacMillan.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York, NY: Elsevier.

McCart, A. B., Sailor, W. S., Bezdek, J. M., & Satter, A. L. (2014). A framework for inclusive educational delivery systems. *Inclusion*, *2*, 252–264. doi:10.1352/2326-6988-2.4.252

Orr, S. T. (2010). Exploring stakeholder values and interests in evaluation. *American Journal of Evaluation*, *31*, 557–569.

Patton, M. Q. (1978). *Utilization-focused evaluation*. Thousand Oaks, CA: Sage.

Pollard, W. E. (1986). *Bayesian statistics for evaluation research*. Thousand Oaks, CA: Sage.

Rodríguez-Campos, L. (2012). Stakeholder involvement in evaluation: Three decades of the American journal of evaluation. *Journal of Multidisciplinary Evaluation*, *8*, 57–59.

Rodríguez-Campos, L., & Rincones-Gomez, R. (2012). *Collaborative evaluations: Step-by-step* (2nd ed.). Stanford, CA: Stanford University Press.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychological Bulletin Review*, *16*, 225–237.

Smith, M. F. (1999). Participatory evaluation: Not working or not tested? *American Journal of Evaluation*, *20*, 295–308.