A Review of Issues About Null Hypothesis Bayesian Testing

Tendeiro, Jorge; Kiers, H. A. L.

Published in:
Psychological Methods

DOI:
10.1037/met0000221

American Psychological Association

Psychological Methods

1082-989X/19/$12.00

2019, Vol. 24, No. 6, 774–795
http://dx.doi.org/10.1037/met0000221

A Review of Issues About Null Hypothesis Bayesian Testing

Jorge N. Tendeiro and Henk A. L. Kiers
University of Groningen

Abstract

Null hypothesis significance testing (NHST) has been under scrutiny for decades. The literature shows overwhelming evidence of a large range of problems affecting NHST. One of the proposed alternatives to NHST is using Bayes factors instead of *p* values. Here we denote the method of using Bayes factors to test point null models as "null hypothesis Bayesian testing" (NHBT). In this article we offer a wide overview of potential issues (limitations or sources of misinterpretation) with NHBT which is currently missing in the literature. We illustrate many of the shortcomings of NHBT by means of reproducible examples. The article concludes with a discussion of NHBT in particular and testing in general. In particular, we argue that posterior model probabilities should be given more emphasis than Bayes factors, because only the former provide *direct* answers to the most common research questions under consideration.

Translational Abstract

Null hypothesis significance testing (NHST) is the most common framework used by psychologists to test their research hypotheses. There are, however, several shortcomings associated with NHST, as has been shown in the scientific literature in the last decades. An alternative to NHST which is based on the Bayesian statistics paradigm uses the so-called Bayes factors. We denote the method of using Bayes factors to test point null models as "null hypothesis Bayesian testing" (NHBT). In this article we offer a wide overview of issues about NHBT which is currently missing in the literature. Our goal is to assist practitioners who are considering using Bayes factors to test their research hypotheses. We illustrate many of the shortcomings of NHBT by means of reproducible examples. The article concludes with a discussion of NHBT in particular and testing in general.

Keywords: *p* values, Bayes statistics, Bayes factors, null hypothesis significance testing, null hypothesis Bayesian testing

Supplemental materials: http://dx.doi.org/10.1037/met0000221.supp

The discussion regarding the current crisis in psychology is central in the scientific community (Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). Of the various actors that have been put forward as in part responsible for the current state of affairs, the overreliance on null hypothesis significance testing (NHST) is on the front stage. A large body of literature has dissected the various shortcomings of NHST. *p* values in particular have often been pointed out as problematic (Cohen, 1994; Edwards, Lindman, & Savage, 1963; Gigerenzer, Krauss, & Vitouch, 2004; Hubbard & Lindsay, 2008; Raftery, 1995; Sellke, Bayarri, & Berger, 2001; Wagenmakers, 2007; but see Harlow, Mulaik, & Steiger, 1997, for a balanced debate for and against significance testing).

The Bayesian paradigm is increasingly gaining traction in the social sciences as an alternative to the classical frequentist approach. Van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, and Depaoli (2017) performed a literature overview and concluded that over 1,500 Bayesian psychological articles were published between 1990 and 2015, and they showed a clear increase over time in that period. To see whether, specifically, the use of testing procedures using the Bayes factor in research applications, considered the Bayesian alternative to NHST, is also increasing, we conducted a small-scale search on Google Scholar using the terms "Bayesian test," "null hypothesis," and "psychology" from 2000 onward. After removing results that did not apply (e.g., articles from different fields that were selected because the keywords featured in the references, or repetitions), we were left with 272 references. Of these, 207 are statistical in nature (e.g., tutorials, methods development). We ended with a set of 65 references that consist of applications of Bayesian testing in psychology, which indeed show an increasing trend across years (see Figure 1). A quick read through the articles from 2018 (16 in total) showed that Bayes factors are being used either side-by-side with frequentist tests and confidence intervals (not always consistently), or only

This article was published Online First May 16, 2019.

Jorge N. Tendeiro and Henk A. L. Kiers, Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen.

Part of this work was presented at the IMPS 2018 meeting in New York, NY.

The data is available at https://osf.io/jmwk6/.

Correspondence concerning this article should be addressed to Jorge N. Tendeiro, Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, the Netherlands. E-mail: j.n.tendeiro@rug.nl

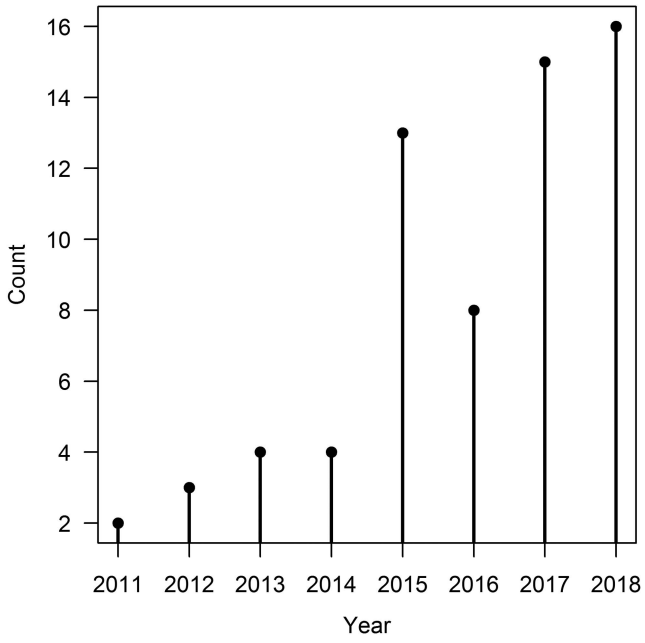

Figure 1. Count of results per year in Google scholar ("Bayesian test," "null hypothesis," "psychology" from 2000 onward). Only results that relate to *applications* of Bayesian testing were counted. See text for details.

when the frequentist test leads to a nonsignificant result. Thus, although the number of articles is still small, there is an increasing request for Bayesian alternatives to NHST.

Bayes factors are construed as a rational means to assess the evidence in the data supporting either of two competing hypotheses or models[1] (Kass & Raftery, 1995). The particular case where a point null model (e.g., M_0: $\theta = .5$) is compared to an alternative model (e.g., M_1: $\theta \neq .5$) will here be called Null Hypothesis Bayesian Testing (NHBT). NHBT is being advocated to have various advantages over NHST. For example, NHBT allows updating one's beliefs logically (by means of the Bayes' rule) and drawing support for either model under consideration (instead of rejecting vs. not rejecting the null model in NHST). Furthermore, NHBT does not depend on subjective data collection intentions, unlike NHST (Wagenmakers, 2007). Specifically, the computation of a p value depends not only on the data observed, but also on the data collection plan. For example (adapted from Lindley, 1993), in order to test whether a coin is biased upon observing six throws (five heads followed by one-tailed), it matters whether one intended to throw the coin six times from the offset (i.e., fixed sample size; $p = .109$) or whether one intended to throw the coin until the first tail appeared (i.e., fixed number of successes; $p = .031$). NHBT does not have this problem because only the observed data (and not the collection intentions) matter to decide upon the coin state. NHBT is also praised for automatically accommodating for multiple testing (e.g., because "by realizing that non-significant results may carry evidential value, Bayesian inference encourages to use all available data"; Dienes, 2016, p. 84) and allowing optional stopping (essentially because NHBT can gather evidence *for* either of the hypothesis being tested, unlike NHST); see Dienes (2016) and Rouder (2014) for details. Therefore, several researchers strongly defend that NHBT should replace NHST as the instrument of choice to test point null models. For example, Iverson, Lee, Zhang, and Wagenmakers (2009, p. 201) stated that "It should be the routine business of authors contributing to *Psychological Science* or any other Journal of scientific psychology to report Bayes Factors."

Given the shifting trend currently ongoing in psychology, we think it is of crucial importance that Bayes factors and NHBT are thoroughly scrutinized if they are to be seriously considered as a replacement of NHST. Any methodological approach has its advantages as well as its drawbacks and NHBT is no different. One year before this writing, the authors felt they did not know enough about NHBT to properly understand how it works, what its merits are, and what its potential limitations are. Therefore, we carried out an extensive literature study on NHBT and related topics. This article is the result of putting together a range of discussions on NHBT. While the merits of NHBT have been reviewed repeatedly (e.g., Dienes, 2014; Kass & Raftery, 1995; Morey, Romeijn, & Rouder, 2016; Wagenmakers et al., 2018), the present article aims to offer an overview of possible related issues, as can be found scattered across the literature. Our hope is that this article helps other methodologically well informed practitioners transitioning from the frequentist toward the Bayesian paradigm, or simply frequentists interested in the Bayesian alternative to the classic NHST paradigm.

The setup of the current article is such that, after an introductory section, each issue that we identified from the literature is described in detail and, whenever possible, illustrated with simple examples. The remaining of the article is organized as follows. First, the Bayes factor and NHBT are introduced. A list of 11 issues concerning various features of NHBT is presented; each topic is discussed in its own section. An "issue" can either be a limitation (according to us) or a feature that may (according to us) increase the risk of misuse or misinterpretation of a Bayes factor. We provide examples deemed helpful to clarifying a point being made. All examples can be reproduced by means of the accompanying R script available at the Open Science Framework (https://osf.io/jmwk6/). We conclude the article with an extensive discussion, where we summarize the main conclusions and offer our personal view on NHBT in particular and testing in general.

The Bayes Factor and NHBT

Suppose one wishes to compare two models (M_0 and M_1) specifying one or more population parameters, and suppose one has data drawn from the associated population. The research question is to what extent one can believe that M_0 or M_1 holds. This can be expressed by probabilities of a model being true, a priori or a posteriori (after having observed the data). Rather than inspecting probabilities, it is customary to consider ratios of probabilities, which are odds ratios as soon as models are jointly exhaustive (i.e., one believes that either M_0 or M_1 holds and there

[1] In this article we use the terms "hypothesis" and "model" interchangeably. Bayes factors can be used to test hypotheses in the classical sense (although composite hypotheses like $\theta \neq .5$ require the extra specification of a prior, as we discuss later), as well as to perform model comparison (e.g., compare nested regression models). Ultimately, different parameter values entail different models (as we discuss in section "Bayes factors test model classes"), thus the distinction between hypothesis and model is moot.

are no other possible models). The Bayes factor (Jeffreys, 1935, 1939; Kass & Raftery, 1995) quantifies the change in the prior odds ratio to the posterior odds ratio due to the data observed. Specifically, denoting the data by D, a direct application of Bayes' rule allows expressing the posterior probability of a model as the scaled product of its prior probability by the likelihood of the data under that model:

$$p(M_i|D) = \frac{p(M_i)p(D|M_i)}{p(M_0)p(D|M_0) + p(M_1)p(D|M_1)} \tag{1}$$

with $i = 0, 1$, where $p(M_i)$ denotes the prior probability for model M_i, $p(D|M_i)$ denotes the probability of observing the data given that M_i holds, and $p(M_i|D)$ denotes the posterior probability for M_i, given the data. The posterior odds for M_1 versus M_0 can then be expressed as follows:

$$\underbrace{\frac{p(M_1|D)}{p(M_0|D)}}_{\text{posterior odds}} = \underbrace{\frac{p(M_1)}{p(M_0)}}_{\text{prior odds}} \times \underbrace{\frac{p(D|M_1)}{p(D|M_0)}}_{\text{Bayes factor, } BF_{10}} \tag{2}$$

Equation 2 states that the Bayes factor is the amount by which the prior odds, which represent the relative belief for each model before observing the data, change to the posterior odds due to the observed data (Good, 1985). In other words, by observing the data we rationally update our relative beliefs about the models by the amount given by the Bayes factor. For example, if $BF_{10} = 5$ and the prior odds are 3 (implying that we believe that M_1 is three times as likely as M_0 before observing the data), then by Equation 2 we conclude that the posterior odds equal 15. This means that our belief about M_1 is 15 times as large as that about M_0 after looking at the data. Finally, by inverting all ratios in Equation 2 it is straightforward to conclude that $BF_{01} = 1/BF_{10}$.

Equation 2 can be used to express posterior model probabilities, $p(M_i|D)$, as functions of prior odds and Bayes factors. Upon observing that $p(M_1|D) = 1 - p(M_0|D)$, Equation 2 implies that

$$\begin{aligned} p(M_0|D) &= \frac{1}{1 + \text{Prior odds}_{10} \times BF_{10}}, \\ p(M_1|D) &= \frac{\text{Prior odds}_{10} \times BF_{10}}{1 + \text{Prior odds}_{10} \times BF_{10}} \end{aligned} \tag{3}$$

We will use this relation between Bayes factors and posterior model probabilities in the article.

Jeffreys (1935, 1939) proposed a Bayesian framework for null hypothesis testing, which we here denote as NHBT. NHBT is based on using the Bayes factor to compare a point null model M_0 specifying the parameter of interest to have a neutral value, to an alternative model M_1 specifying the parameter to have any other value, with a specific probability density function associated with it. This probability density function, often called a prior density, is associated only with M_1; we will refer to it as the *within-model prior*[2] (Kruschke & Liddell, 2018a) specifying the degree of belief in the parameter's value according to model M_1. The NHBT method has been advocated by many as a viable alternative to NHST. However, possible limitations or misinterpretations of NHBT need be well understood before one should consider using it. Ultimately, the goal of this article is to provide a critical overview of NHBT and hopefully helping researchers to avoid misusing this statistical tool.

Illustrative Example

We now consider one concrete example that highlights the core concepts associated to NHBT introduced above. We fall back to the classically simple case of ascertaining whether a coin is fair. That is, we want to test M_0: $\theta = .5$ versus M_1: $\theta \neq .5$, where θ is the true rate of heads. In the Bayesian framework, we must specify our uncertainty about the unknown parameter θ for each model. This is not a problem under M_0 because only one value is considered. However, as mentioned above, to adequately define M_1 a within-model prior must be chosen to specify the probability density of all values associated with $\theta \neq .5$. Here we choose to define the alternative model as "every value of θ is equally probable," hence defining the within-model prior to be the Uniform(0, 1) distribution, or equivalently, the Beta(1, 1) distribution: $p(\theta) = 1$ for $\theta \in [0, 1]$. Suppose that the coin was tossed five times and that two heads and three tails were observed; we refer to the observed outcome of our "experiment" as the data D. The ultimate question that NHBT tries to answer is what the *relative probability of the two models* is given the data. This is provided by the posterior odds as expressed in Equation 2. Clearly, given a prior odds ratio for the two models, the only ingredient needed to compute this posterior odds is the Bayes factor. For this reason, it has become common practice to calculate and report the Bayes factor, because it offers any reader the ability to calculate the posterior odds just by multiplying the Bayes factor with his or her own prior odds. The Bayes factor, being the ratio shown in Equation 2, indicates what the *relative predictive value of the two models* is. In other words, which of the two models predicts the observed data the best? This amounts to computing $p(D|M_0)$ and $p(D|M_1)$, that is, the probability of observing D under each model. If M_0 holds (i.e., if $\theta = .5$), then the probability of observing two heads in five tosses is $p(D|M_0) = \binom{5}{2} \times .5^5 = .3125$ (assuming all tosses are independent). Under M_1 the computations are more complex because now there is an infinity of parameter values to consider: We must consider $p(D|\theta = \theta_0) = \binom{5}{2}\theta_0^2(1 - \theta_0)^3$ for any real value θ_0 in the interval [0, 1]. The mathematical solution is based on averaging all such possible values $p(D|\theta = \theta_0)$ by the within-model prior (see Equation 4). In Appendix A in the online supplemental materials we work out the closed-form expression for this quantity; here we simply present the result: $p(D|M_1) = \binom{5}{2} \times .01667 = .1667$. We now have all ingredients to compute the Bayes factor: $BF_{10} = \frac{p(D|M_1)}{p(D|M_0)} = \frac{.1667}{.3125} = .5333$ or, equivalently, $BF_{01} = \frac{1}{.5333} = 1.875$.

As implicitly used above, there are two alternative ways to interpret the Bayes factor: (a) as the multiplicative factor that

[2] The term within-model prior could be seen as a misnomer, because the word "prior" suggests that there will also be interest in a "posterior." However, posterior probabilities for the parameter θ do not play a role here, and what is called "prior" here simply is the probability distribution for θ as *specified by* model M_1. After the analysis, one can make statements as to how likely or probable this *entire* data model (including its distribution specification) is, given the data. This thus entails a posterior probability for the entire model M_1 to be true, as it updates prior belief that M_1 is true, but the "within-model prior" is not being updated into a "within-model posterior" distribution. The use of the term prior actually can be seen as rather unfortunate, but we stick to it because it is so common in the literature.

transforms prior odds to posterior odds, and (b) as the ratio of the probabilities of observing the data under each of the competing models. The first interpretation in particular highlights that the Bayes factor is just the means necessary to update our relative belief between two models. The Bayes factor is not the main conclusion; that should be derived from the posterior odds instead (we further extend this idea in the article; see Point 4). This has been pointed out by others before, see for example Etz (2015): **"The conclusion is *not* represented by the Bayes factor, but by the posterior odds.** The Bayes factor is just one piece of the puzzle, namely the *evidence* contained in our sample. In order to come to a conclusion the Bayes factor has to be combined with the prior odds to obtain *posterior odds*. We have to take into account the information we had before we started sampling. I repeat: The posterior odds are where the conclusion resides. Not the Bayes factor." The second interpretation of the Bayes factor may implicitly lead to misinterpretations, because the Bayes factor is a ratio of probabilities of the *data* conditional on *models*, but for lay people this may easily be interpreted as a ratio of probabilities of the models to be true, given the data. In this respect, Bayes factors and p values are alike: Both are based on probabilities of data conditional on models. Thus, for the above example, one can only conclude that the data are slightly more likely to be observed under M_0 than under M_1.

In the above example, observe that prior belief concerning the truth of either model was not considered. That is, $p(M_0)$ and $p(M_1)$ were never invoked. Only upon specifying the prior odds $p(M_0)/p(M_1)$ we are in a position to answer the question how much more probable one model is than the other, given the data (and, obviously, given the prior odds). For example, Person A may strongly believe that the coin is fair before tossing it and therefore assuming $p(M_0) = .99$ and $p(M_1) = .01$ might describe her expectations well. This prior odds ratio of 99 in favor of M_0 would, given the data, lead to an even higher posterior odds ratio of $99 \times 1.875 = 185.6$ in favor of M_0, and posterior probabilities $p(M_0 \mid D) = \frac{1}{1+1/185.6} = .995$ versus $p(M_1 \mid D) = \frac{1/185.6}{1+1/185.6} = .005$. Person B may truly doubt the fairness of the coin and express his beliefs as $p(M_0) = .50$ and $p(M_1) = .50$ instead, thus assuming prior odds 1. On the basis of the data, his beliefs would change: The posterior odds are now $1 \times 1.875 = 1.875$ in favor of M_0, and the posterior probabilities are $p(M_0 \mid D) = \frac{1}{1+1/1.875} = .65$ and $p(M_1 \mid D) = \frac{1/1.875}{1+1/1.875} = .35$.

To avoid confusion (see also Footnote 2) it is crucial to distinguish between the within-model prior distribution, $p(\theta \mid M_i)$, and the model's prior probability, $p(M_i)$: The former concerns the researcher's specification of probability about the parameters within the model M_1 he or she wishes to compare to M_0, while the latter concerns the probability of the model holding as a whole. In theory, both types of prior should be carefully set up by the researcher; in practice, researchers often rely on defaults offered by prepackaged software (we will discuss this matter later on). Strictly speaking, Bayes factors do not depend on $p(M_i)$: These probabilities are part of the prior odds (Equation 2). We will further elaborate on this aspect of Bayes factors later in the article, but for now we want to stress that within-model priors are independent from prior model beliefs (Equation 2 is clear in this respect).

List of Issues About NHBT Studied in This Article

Below we summarize the various issues that we will cover. Each point will be treated in its own subsection.

1. Bayes factors can be hard to compute.
2. Bayes factors are sensitive to within-model priors.
3. Use of "default" Bayes factors.
4. Bayes factors are not posterior model probabilities.
5. Bayes factors do not imply a model is probably correct.
6. Qualitative interpretation of Bayes factors.
7. Bayes factors test model classes.
8. Mismatch between Bayes factors and parameter estimation.
9. Bayes factors favor the point null model.
10. Bayes factors favor the alternative.
11. Bayes factors often agree with p values.

Point 1: Bayes Factors Can Be Hard to Compute

To fully appreciate the mathematical expression of BF_{10} shown as the rightmost term in Equation 2, consider the common situation where model M_i ($i = 0, 1$) is expressed as a parameterized probability model, here generically denoted by $p(D \mid \theta_i, M_i)$. This model relates the observed data, D, to a vector of unknown model parameters, θ (for simplicity, in all our examples we treat θ as a single parameter). Each of the terms featuring in the expression of BF_{10} is computed by means of the following equation:

$$p(D \mid M_i) = \int_{\Theta_i} p(D \mid \theta_i, M_i) p(\theta_i \mid M_i) d\theta_i. \quad (4)$$

Equation 4 is a weighted likelihood (or *marginal* likelihood), that is, a weighted average of the likelihood of the observed data (the first term under the integral) across the entire parameter space Θ_i. The weights are provided by the within-model prior probability density for the model parameters, $p(\theta_i \mid M_i)$. The within-model prior distribution reflects uncertainty about the true parameter values before observing the data. Equation 4 is written under the assumption that θ_i is a vector of continuous random variables, but it also applies to discrete random variables (by replacing the prior density by a prior probability mass function and the integration by a summation).

As it happens, the integral in Equation 4 is difficult to solve analytically in all but a few instances. Hence, we need to resort to numerical procedures. Several numerical methods have been proposed (Berger & Pericchi, 2001; Carlin & Chib, 1995; Chen, Shao, & Ibrahim, 2000; Gamerman & Lopes, 2006; Gelman & Meng, 1998; Green, 1995; Gronau et al., 2017; Kass & Raftery, 1995) but are not easy to use in practice (Kamary, Mengersen, Robert, & Rousseau, 2014). Luckily, the recent years witnessed the surge of free and user-friendly software that allows computing Bayes factors in a wide range of settings for a (somehow) restricted range of

within-model priors (e.g., the R BayesFactor package and JASP, which relies on BayesFactor; JASP Team, 2018; Morey & Rouder, 2018). For instance, JASP is an open-source software that offers a very intuitive point-and-click graphical user interface that is reminiscent of SPSS. For the coin example introduced previously, we simply need to provide the data in a file (e.g., a CSV file with the five coin tosses in one column) and run the "Bayesian Binomial Test" procedure (a screenshot of JASP is available as online supplemental material).

Given the availability of Bayes factor-friendly software noted above, we may conclude that the difficulty of handling Equation 4 is a feature of Bayes factors that does not offer major problems for practitioners, at least for the most common types of tests used in the social sciences. For instance, the current version of the BayesFactor R package (0.9.12–4.2) allows computing Bayes factors *for predefined within-model priors and probability models*, in the following settings: One-sample, two independent samples, ANOVA (fixed and random effects), regression (continuous and/or categorical predictors), linear correlations, single proportions, and contingency tables. The extension of Bayes factors to more complex models is likely to happen in the coming years.

Point 2: Bayes Factors Are Sensitive to Within-Model Priors

The sensitivity of Bayes factors to within-model priors is well established in the literature (e.g., Gallistel, 2009; Kass, 1993; Kass & Raftery, 1995; Liu & Aitkin, 2008; Robert, 2016; Sinharay & Stern, 2002; Vampaemel, 2010; Withers, 2002). As argued before, Bayes factors are ratios of weighted likelihoods, the latter consisting of the likelihood function averaged over within-model prior distributions. Hence, different within-model priors imply different weighted likelihoods and, as a consequence, different Bayes factors. In particular, within-model priors which place a large weight on implausible parameter values will lead to lower weighted likelihoods and thus decrease the relative credibility of the corresponding model. In this respect, Bayes factors do not perform similarly as estimation of posterior distributions, for which poorly selected priors need not compromise the form of the posterior distribution. Below we illustrate this feature by means of an example from Liu and Aitkin (2008). The example is equivalent to the coin bias experiment previously introduced, that is, testing M_0: $\theta = .5$ versus M_1: $\theta \neq .5$, where θ is the success rate of each trial in a Bernoulli process. In general, the probability of observing r successes in n independent trials is given by the binomial likelihood function:

$$p(D|\theta, M_i) = \binom{n}{r}\theta^r(1-\theta)^{n-r}. \quad (5)$$

Under M_0, the weighted likelihood is simple because there is only one θ value to entertain: $p(D|M_0) = p(D|\theta = .5, M_0) = \binom{n}{r} \cdot 5^n$. The problem is more difficult under M_1 because a full range of θ values needs to be considered. We therefore need a within-model prior $p(\theta|M_1)$ so that the weighted likelihood $p(D|M_1)$ can be computed by means of Equation 4. Liu and Aitkin (2008) used the beta within-model prior since it is a so-called *conjugate* prior for the binomial likelihood: $\theta \sim Beta(a, b)$, for positive a and b. Conjugate priors are convenient because the corresponding posterior distribution can be expressed in closed form as a distribution in the same family as the prior (the beta family in this case). For this within-model prior, it is possible to compute $p(D|M_1)$ in closed form, which then allows computing the Bayes factor by definition (see Appendix A in the online supplemental materials for the details). Liu and Aitkin (2008) considered four within-model priors: The uniform prior ($a = b = 1$), Jeffreys' prior ($a = b = .5$), an approximation to Haldane's prior ($a = b = 0.05$), and an informative prior ($a = 3$, $b = 2$) (see Figure 2A). After having observed 60 successes in 100 trials, the Bayes factors equal $BF_{10} = .91, .60, .09$, and 1.55 for the uniform, Jeffreys, Haldane, and informative within-model priors, respectively. Thus, the Bayes factor ranged from $BF_{01} = 1/.09 = 11.1$ for Haldane's within-model prior (indicative of strong support for M_0) through $BF_{10} = 1.55$ for the informative within-model prior (indicative of weak evidence in favor of M_1), based on Jeffreys (1961) benchmarks.

Thus, a judicious choice of within-model priors is essential to use Bayes factors properly. The reason why the relative support for either model varies greatly across within-model priors is apparent by close inspection of the formula for the marginal likelihood shown in Equation 4. As explained previously, marginal likelihoods are weighted averages of the data likelihood, with weights provided by the within-model prior distribution. To illustrate this, the data likelihood, $p(D|\theta, M_1)$, multiplied by each of the four within-model priors considered in our example, $p_k(\theta|M_1)$ ($k = 1, \ldots, 4$), is depicted in Figure 2C. Because these functions represent the weighted likelihoods $p(D|\theta, M_1)p_k(\theta|M_1)$, the marginal likelihoods $p_k(D|M_1)$ are their integrals (see Equation 4), hence the marginal likelihood values equal the areas under the graphs of these functions. We will refer to these functions $p(D|\theta, M_1)p_k(\theta|M_1)$ as the "non-normalized posteriors." This is because they are non-normalized versions of the actual posterior distributions in Figure 2B, as can be readily seen from the Bayes formula:

$$\underbrace{p_k(\theta|D, M_1)}_{\substack{\text{posterior}\\ \text{(Figure 2B)}}} = \frac{\overbrace{p(D|\theta, M_1)p_k(\theta|M_1)}^{\substack{\text{non-normalized posterior}\\ \text{(Figure 2C)}}}}{\underbrace{p_k(D|M_1)}_{\substack{\text{normalizing constant}\\ \text{(Equation 4)}}}}.$$

It can now be seen that, whereas the actual posteriors (in Figure 2B) are very close to each other, the non-normalized posteriors in Figure 2C are quite different from each other, thus different within-model priors imply different non-normalized posteriors. As a consequence, while posterior distributions themselves hardly differ, Bayes factors do, because the areas under the curves, $p_k(D|M_1)$, differ a lot, and they actually are part of the Bayes factor:

$$BF_{10} = \frac{p_k(D|M_1)}{p(D|M_0)},$$

while $p(D|M_0)$ is the same for $k = 1, \ldots, 4$ ($p(D|M_0) = \binom{n}{r} \cdot 5^n$); this is shown in Figure 2C by the horizontal solid line. Therefore, we conclude that the marginal likelihood under M_1, $p_k(D|M_1)$, and inherently the Bayes factor, is strongly dependent on the within-model prior k. Within-model prior distributions that place a large weight on parameters that are unlikely to have generated the data end up lowering the weighted average. This is exactly what happened in the example. The data (60 successes in 100 trials) are largely inconsistent

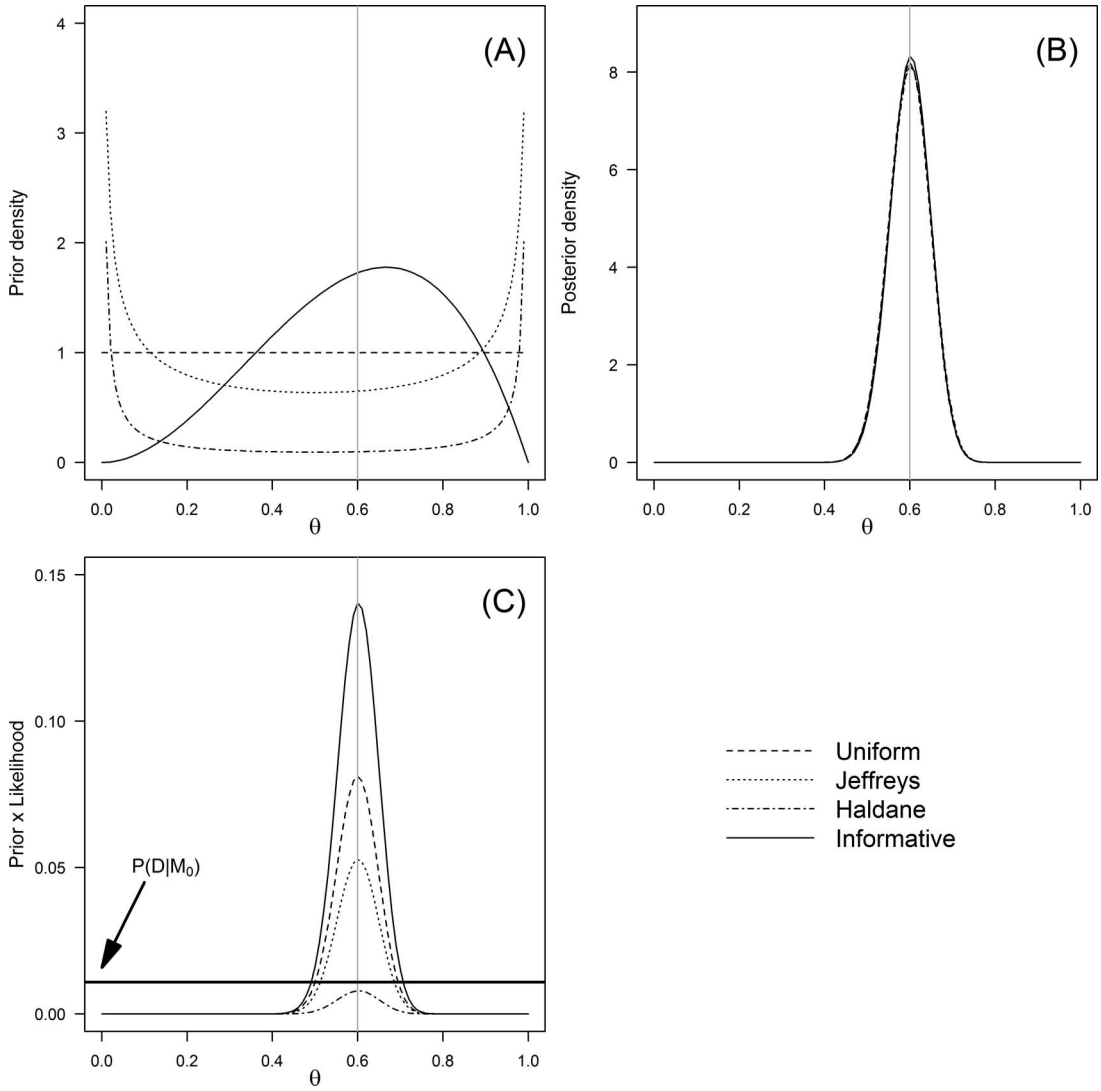

Figure 2. The four within-model prior distributions of Liu and Aitkin (2008) (panel A); the four corresponding posterior distributions (Panel B); and the non-normalized posteriors (Panel C). The posterior distributions (B) are indistinguishable in spite of the very different shapes of the within-model prior distributions (A). However, the non-normalized posteriors (C), on which the Bayes factor is based, are very different from each other. The vertical line concerns the observed data (60 successes in 100 trials).

with all but the informative within-model prior; as a consequence, the Bayes factor favors M_0 over M_1 for such within-model priors. On the contrary, the data are more consistent with the informative within-model prior (Figure 2A), which ultimately led to this Bayes factor displaying (some) support for M_1.

Other examples of the problem above have been given (e.g., Lavine & Schervish, 1999). Berger and Pericchi (2001) further discussed an example based on the normal data model that again warns users against using too vague within-model priors (i.e., priors that place too high probability on parameter regions that are inconsistent with the data). Suppose data are normally distributed, $Y_j \sim N(\mu, \sigma^2)$ for $j = 1, \ldots, n$, with known variance ($\sigma^2 = 1$). We want to test M_0: $\mu = 0$ against two alternative models, M_1: $\mu \sim N(0, \sigma_1^2)$ and M_2: $\mu \sim U(-\infty, \infty)$. The within-model prior for the mean parameter is increasingly vague under M_1 as σ_1 increases, whereas it is completely uninformative under M_2. Berger and Pericchi (2001; see also Berger & Delampady, 1987, or Rouder, Haaf, & Vandekerckhove, 2018, for BF_{10}) give the Bayes factors BF_{10} and BF_{20} (for completeness, Appendix B in the online supplemental materials provides the derivation of both formulas):

$$BF_{10} = \frac{1}{\sqrt{1 + n\sigma_1^2}} \exp\left[\frac{n\sigma_1^2}{2(1 + n\sigma_1^2)} z^2\right],$$
$$BF_{20} = \sqrt{\frac{2\pi}{n}} \exp\left(\frac{z^2}{2}\right), \tag{6}$$

where $z = \sqrt{n}(\bar{Y} - 0)/\sigma = \sqrt{n}\bar{Y}$ is the classical one-sample z test statistic. BF_{10} converges to 0 as σ_1 increases, thus the vaguer the

within-model prior the larger the support for M_0. For example, if $n = 10$ and $\bar{Y} = 1$ then $BF_{10} = 28.4, 9.2, 4.7, 0.9$, and 0.5 for $\sigma_1 = 1, 5, 10, 50$, and 100. The posterior distributions corresponding to the within-model priors, however, are practically indistinguishable from each other when σ_1 is larger than three (not shown; the supplementary R script includes the code necessary to produce the plot). Interestingly, the Bayes factor based on the *improper* within-model prior (i.e., a prior density which does not have a finite integral) under M_2 *is finite for fixed data*, because BF_{20} in Equation 6 depends only on the data (via n and $z = \sqrt{n}\bar{Y}$). Berger and Pericchi (2001, p. 143) summarize things by saying "*never use 'arbitrary' vague proper priors for model selection, but improper noninformative priors may give reasonable results.*"

In essence, application of Bayes factors crucially relies on the choice of within-model priors since the latter are used to weigh the likelihood (Equation 4). In particular, the use of too vague within-model priors (however "vague" is operationalized) for Bayes factors is ill-advised, because the null model will invariably end up being supported (Bayarri, Berger, Forte, & Garcia-Donato, 2012; Lindley, 1957; Morey & Rouder, 2011). As noted by Morey and Rouder (2011), it is striking that the inclusion of vague within-model priors, typically chosen in order to reduce the influence of the within-model prior on the posterior, has the deleterious effect of predetermining the result of a Bayesian test (i.e., support the null model). In spite of our example above being based on a small sample ($n = 10$), the problem stands for larger values of n (Bayarri et al., 2012; Berger & Pericchi, 2001; Kass & Raftery, 1995). This feature is remarkably distinct from what happens when estimating posterior distributions under the Bayesian framework (e.g., Gelman, Meng, & Stern, 1996; Kass, 1993). In general, under estimation, and except for completely "dogmatic" priors, the accumulation of evidence brought in by the data typically allows for a wide range of different prior beliefs to rationally converge to one common model.

As has been argued, models that employ too vague within-model priors often imply weighing regions of the parameter space that are highly inconsistent with the data. The corresponding weighted likelihoods will be lower than those for models which are less flexible but "closer" to the data. In other words, Bayes factors will naturally favor the model based on a less vague, "closer" to the data, within-model prior. This is not a bad feature per se. It has been considered as an ideal mechanism that favors "simpler" models (i.e., M_0) over unnecessary "complex" ones (i.e., M_1; Myung, 2000; Myung & Pitt, 1997). However, it is not clear to us why it is good to have a measure confounding appropriateness with simplicity. We take a more neutral stance here and agree with Vampaemel (2010, p. 491), who states "(. . .) if models are quantitatively instantiated theories, the prior can be used to capture theory and should therefore be considered as an integral part of the model. The viewpoint that the prior can be used as a vehicle for expressing theory implies that a model evaluation measure should be sensitive to the prior." In this sense, within-model priors should be specified to reflect our or other people's hypotheses as accurately as possible. However, this puts a considerable burden on the researcher because a researcher will most likely have difficulty formulating a hypothesis in terms of a full density function, and it can matter much what exact density function is chosen, as shown above.

We think that the dependence of Bayes factors on the choice of a within-model prior is a limitation of NHBT, especially because choosing such within-model priors and computing the related marginal likelihoods (Equation 4) are no easy tasks. How are then practitioners expected to choose within-model priors? Most important is that, when drawing conclusions in terms of evidence for one model over the other, both models should be specified precisely. One should never simplify the comparison to M_0: $\theta = .5$ versus M_1: $\theta \neq .5$ or M_0: $\mu = 0$ versus M_1: $\mu \neq 0$, simply because one does not have evidence on the value of the parameter in a general sense. The Bayes factor method does not just contrast statements of parameter values; it contrasts two distributionally specified models for the parameter. Indeed, if the method would give evidence on whether the parameter has a particular value, the choice of the within-prior would not matter. This obviously makes practical interpretation of results difficult, for the very reason that it depends on the within-prior specification. Statements like "We have found 6.1 times more support for the mean population effect size being 0 than for the mean population effect size being distributed as $N(0, 1)$" may be difficult to grasp, not the least because this distribution of population effect sizes will be hard to understand. Within the framework of NHBT, a partial remedy could be to study the sensitivity of the results to the choice of within-model priors for one's own data set. Sensitivity analysis consists of checking whether the main conclusions from the data analysis are robust to different prior specifications (Kass & Raftery, 1995; Myung & Pitt, 1997; Sinharay & Stern, 2002). If results are fairly stable under various priors then one can at least somewhat corroborate general statements in terms of support for there being or not being an effect (e.g., if widely different within-model priors all lead to strong support for or against M_0). If, on the other hand, the results are strongly dependent on the choice of within-model prior, then our best advice is to report the results from sensitivity analysis, explain why the chosen within-model prior makes M_1 a particularly interesting model to compare with M_0, and moderate the conclusions derived from the analysis. Although sensitivity analysis is not as commonly used in Bayesian analyses as it would be preferable (van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017), it is a valuable means of ascertaining the effect of the choice of the within-model prior on the Bayes factor and it should be routinely reported (e.g., Depaoli & van de Schoot, 2017). Sensitivity analysis is made available in JASP, but *only for the predefined within-model priors offered by the software* (see the next section for a discussion over default within-model priors). In general, sensitivity analysis for Bayes factors is a difficult endeavor because, as we learned in Point 1, it is not easy to compute Bayes factors under a wide range of families of within-model prior distributions.

Point 3: Use of "Default" Bayes Factors

As argued before, the choice of within-model priors is a delicate matter for Bayes factors. If researchers are to use Bayes factors, to what extent will they be able to choose an appropriate within-model prior? In our opinion, there is no easy answer. Various authors have, as an alternative, advocated the use of "default," "reference," or "objective" within-model priors (Bayarri et al., 2012; Jeffreys, 1961; Marden, 2000; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Zellner & Siow, 1980). Such priors are

carefully chosen so as to avoid the problem we discussed, for instance, by not allowing them to be too broad. As an example, the popular NHBT procedure from Rouder, Speckman, Sun, Morey, and Iverson (2009) for the Bayesian *t* test is based on the so-called JZS within-model prior, which consists of setting a Cauchy prior on the standardized effect size and an improper prior on the variance of the normal distribution. There are theoretical advantages to using the JZS within-model prior (Rouder et al., 2009). However, it is important to acknowledge that this choice of priors lacks clear empirical justification for any specific application. What researchers need to remember is that there are many Bayes factors, in fact, each within-model prior distribution implies a different Bayes factor. Default options offered by software packages are convenient and useful to the extent that they sufficiently adequately describe our tested hypotheses. A Bayes factor based on within-model prior distributions that poorly translate what we think about a phenomenon is misleading at best and egregiously wrong at times (Kruschke, 2011; Kruschke & Liddell, 2018a). Also, we may question how "objective" such within-model priors really are (Berger & Delampady, 1987). It is important to note that objective within-model priors were derived under a set of required desiderata (Bayarri et al., 2012; Berger & Pericchi, 2001). Bayarri et al. (2012) identified seven different desiderata divided among four classes: Basic, consistency criteria, predictive matching criteria, and invariance criteria. As an example, one consistency desideratum states that the posterior probability of the true model (i.e., the model that generated the data) should approach one as the sample size increases (Bayarri et al., 2012, Criterion 2). Such optimal criteria do impose restrictions on the within-model prior, which put the claimed objectivity into question (Berger & Delampady, 1987). The pressure for an "appearance of objectivity" (Berger & Pericchi, 2001, p. 141) instead of true objectivity might help explaining why objective within-model priors are often considered. For the sake of a balanced discussion, it needs to be said that formulation of appropriate within-model priors (prior elicitation) is by no means a simple task either and it may make it more difficult to compare results among studies.

The default JZS Bayesian *t* test models the uncertainty of the true effect size under the alternative hypothesis by means of a Cauchy distribution with scale γ. This implies that, under M_1, we allocate 50% of prior probability to effect sizes in the interval $(-\gamma, \gamma)$ and 50% of prior probability outside this interval. The scale parameter has been fixed at one (Rouder et al., 2009) and later on at $\sqrt{2}/2 = .707$ (e.g., Morey & Rouder, 2018). We may doubt whether specifying a model by means of such a prior makes sense, because high probability of large effect sizes may not be realistic in many social sciences contexts (but see Wagenmakers, Wetzels, Borsboom, Kievit, & Maas, 2011; Wetzels et al., 2011), and makes such models M_1 improbable to begin with. The previous discussion stressed the importance of not using too wide within-model priors so that the null model is not overly supported. Accordingly, in situations where we expect effect sizes of smaller magnitude, it is advisable to use a much smaller scale value that better suits the hypotheses of the analyst. This will provide a less conservative decision rule concerning the rejection of the null model, but it should be more realistic too.

All available software packages that allow computing Bayes factors do offer default within-model priors. In this sense, practitioners have little choice than to use one of the default options available, with the added feature of possibly tuning some of its parameters (like the scale parameter of the Cauchy distribution described above). Tuning parameters may provide an interesting flexibility in terms of changing the shape of the within-model prior toward distributions in line with the models we wish to compare. Also, this should be done as part of a sensitivity analysis, as discussed in Point 2. Unfortunately, the computation of Bayes factors using nondefault within-model priors is difficult to manage (recall Point 1) and is thus not a viable option to practitioners. There is therefore a limitation in what software currently offers, which is not a limitation of NHBT in general. Finally, we observe that one added risk of solely relying on default within-model priors is that practitioners may overly rely on how sensible those defaults are.

Point 4: Bayes Factors Are Not Posterior Model Probabilities

This section is a clarification concerning the relation between Bayes factors and posterior model probabilities, which was already referred to in the example in the introduction. Recall that Bayes factors indicate by how much our relative belief between two competing models should be updated in light of newly observed data. So, $BF_{01} = 15$ indicates that, after looking at the data, we revise our belief toward M_0 by 15 times, and we could say that the *data* are 15 times more probable under M_0 than under the alternative. The interesting question that follows is: *What does this imply concerning the probability of each model, given the observed data?* The answer to this question is, perhaps surprisingly: *On its own, nothing at all*. The posterior probability of each model depends on the corresponding prior probability; this information is entirely unrelated to the Bayes factor. Equation 2 is particularly illuminating in this regard: The Bayes factor is simply a multiplicative term that converts the prior model odds to the posterior model odds. To fully account for the posterior odds, we need to specify the prior odds (Stern, 2016; see Equation 2). Our own reading of the literature indicates that the last step seems to be often ignored in applications. That is, researchers seem content in deriving Bayes factors alone. But, as argued above, this brings no specific information concerning the plausibility of each model in light of the data, that is, of $p(M_i|D)$. This idea is not new (recall Etz, 2015), see for example Edwards, Lindman, and Savage (1963, p. 235): "It is uninteresting to learn that the odds in favor of the null hypothesis have increased or decreased a hundredfold if initially they were negligibly different from zero." Given the proliferation of reported Bayes factors as standalone pieces of evidence, we find it important to stress this particular point.

Equation 3 shows how Bayes factors and posterior model probabilities relate. As an example, suppose $BF_{01} = 32$. To make things more tangible, suppose there are two persons, Anna and Ben, giving some consideration to either model M_0 and M_1 *before looking at the data*. Anna is clueless as to what to expect and therefore decides to put equal faith on either model: $p(M_0) = p(M_1) = .50$. Ben, on the other hand, is convinced that the second model is much more likely: $p(M_0) = .01$, $p(M_1) = .99$. The posterior model probabilities for Anna and Ben shown in Table 1 are very instructive. For Anna, M_0 is more likely after looking at the data. This result seems in line with the Bayes factor ($BF_{01} = 32$). However, for Ben, model M_1 is (still) more likely, even

Table 1
Prior and Posterior Model Probabilities Updated by Means of a Bayes Factor, for Two Different Sets of Prior Model Probabilities

	Prior model probabilities			Posterior model probabilities	
	$p(M_0)$	$p(M_1)$	BF_{01}	$p(M_0 \mid D)$	$p(M_1 \mid D)$
Anna	.50	.50	32	.970	.030
Ben	.01	.99		.244	.756

though the Bayes factor indicates that the data favor M_0 over M_1 at odds of 32 to 1. This conflict can be explained by observing that Ben's initial position was in sharp contrast to the data observed. Further observe that the same Bayes factor of 32 applies to Anna and Ben equally:

$$\underbrace{(.970/.030)/(.5/.5)}_{\text{Anna}} = \overbrace{32}^{\text{BF}_{01}} = \underbrace{(.244/.756)/(.01/.99)}_{\text{Ben}}$$

This example highlights an important feature of Bayes factors: They indicate the rate of change of belief, not the belief itself. For the latter one needs to consider posterior probabilities under each model $p(M_i \mid D)$ instead. It is essential to understand this distinction in order to avoid erroneous interpretations of Bayes factors.

In some cases, researchers are willing to assume that both models are equally likely a priori (Marden, 2000). Under this assumption, the posterior odds equal the Bayes factor (see Equation 2) and derivation of the probabilities $p(M_i \mid D)$ from the Bayes factor is then straightforward (Equation 3). Thus, the assumption of equal a priori model probabilities does simplify somewhat the analysis. Naturally there are settings in which we have no prior preference for one of the models and such a rationale applies. But this need not always be the case (e.g., Hinkley, 1987; Kruschke & Liddell, 2018a). The Bayesian paradigm is particularly suitable to including existing information in the data analysis (this is at the core of using prior distributions). Not doing so by default seems ill-devised and a wasted opportunity for contributing to accumulating knowledge.

On the other hand, an advantage of Bayes factors is that they can be used to convert *any* prior odds ratio into a posterior odds ratio. This hence allows for a *modest* reporting style of just providing the Bayes factor and not drawing any conclusions relative to model probabilities. Instead, such considerations are left to the reader, based on his or her own prior and hence posterior odds ratio. Sometimes, it is stated that the Bayes factor pertains to "the evidence from the data." This is true in the sense that it changes the prior belief on the basis of evidence from the data. However, it could easily be confused for 'conclusive' evidence, which it only is if it takes prior beliefs into consideration. To avoid any confusion, we recommend to only draw conclusions concerning model preference on the basis of posterior odds, not on the basis of Bayes factors alone. We do note, however, that not all Bayesians favor reporting posterior probabilities over Bayes factors (e.g., Morey & Rouder, 2011; Wagenmakers et al., 2018). However, we believe that the risk of *misinterpreting* the Bayes factor is worrisome and we hope that practitioners can now better understand what is at stake based on the discussion above.

Point 5: Bayes Factors Do Not Imply a Model Is Probably Correct

The Bayes factor is a measure of relative plausibility among two models. Thus, a large Bayes factor indicates which of two models is more likely to have generated the observed data. This does not imply, however, that the favored model is likely to be *true* and users should refrain from this type of statements. Thus, Bayes factors provide no *absolute* evidence supporting either model under comparison (Gelman & Rubin, 1995). Simply, Bayes factors point at the model most likely to have generated the observed data, *regardless of that model being actually true or even approximately true*. Indeed, both models may be very improbable, and the Bayes factor may still indicate strong support for one over the other. Ly, Verhagen, and Wagenmakers (2016, p. 30) offer a discussion on this topic and include references that further clarify how Bayes factors are expected to perform under model misspecification. We observe that this property of Bayes factors is similar to that of other model selection criteria commonly used in the social sciences, including information-based criteria such as the AIC, BIC, and DIC (Burnham & Anderson, 2003; Spiegelhalter, Best, Carlin, & van der Linde, 2002).

One could argue that the fact that the evidence provided by Bayes factors is relative in nature is not a limitation of the Bayes factor per se (problems only arise if practitioners misinterpret the Bayes factor). Our alert to practitioners is to avoid misusing Bayes factors in this way by keeping their conclusions in perspective (after all, all that has been achieved is to perform one specific comparison). However, what this discussion makes salient is that Bayes factors concern the *comparison of two models only* (just like NHST, for that matter). In this sense, Bayes factors are limited. What one can do is compare various pairs of models (one Bayes factor per comparison) and, based on that, choose the better predicting model (do observe that Equation 2 implies that Bayes factors enjoy the following transitivity property: $BF_{21}BF_{10} = BF_{20}$). Even in this way, only a limited set of models for the parameter can be compared. Instead, analyzing the posterior distribution for the parameter being tested offers a much richer insight, because it specifies the probability density for all possible values of the parameter, and does so in one go. We will further elaborate on this point in the Discussion section.

Point 6: Qualitative Interpretation of Bayes Factors

As a ratio of weighted likelihoods (Equation 2), Bayes factors are a continuous measure of evidence on the non-negative real numbers set (Rouder et al., 2009). BF_{10} values larger than one imply that the data are more likely under M_1 than M_0, whereas BF_{10} values smaller than one imply that the data are more likely under M_0 than M_1. But, how "much more" likely? In other words, how can we qualify Bayes factors into grading strengths of evidence? Unfortunately there is no clear answer to this question: Qualitative interpretations of strength are subjective. Bayesians do not fully agree on this account and this helps understanding why several competing proposals have been introduced (e.g., Kass & Raftery, 1995; Jeffreys, 1961; Lee & Wagenmakers, 2013). We think that, similarly to using a fixed significance level of 5% or 1% under NHST, written-in-stone rules for Bayes factors are also not advisable because they give the misleading feeling of a "mecha-

nistic interpretation" of statistical results (van der Linden & Chryst, 2017). Bayes factors are a continuous measure of degree and are better viewed as such (e.g., Konijn, van de Schoot, Winter, & Ferguson, 2015).

It is not simple to advise practitioners about this particular aspect of Bayes factors. We do agree that labels like those introduced by Jeffreys (1961) feel rather arbitrary (e.g., BF_{10} between one and three and between three and 10 are labeled as "barely worth mentioning" and "substantial," respectively). On the other hand, simply arguing that Bayes factors should be interpreted as continuous evidential information falls short because practitioners not used to Bayes factors naturally lack intuition about the magnitude of their values. We strongly think that the best solution is to not report Bayes factors only, but to also report posterior model probabilities, as already discussed in Point 4.

Point 7: Bayes Factors Test Model Classes

A different interpretational difficulty concerning Bayes factors may be hard to grasp but we find it nevertheless crucial. Bayes factors are a measure of change from prior model odds to posterior model odds after considering the observed data. Thus, from $BF_{01} = 1/5$ we conclude that the data are five times more likely to have occurred under M_1 than under M_0. This interpretation is accurate when both models are point hypotheses of the type M_i: $\theta = \theta_i$; in such cases the Bayes factor is simply the classic likelihood ratio. However, for composite models of the type M_1: $\theta \neq \theta_0$, the Bayes factor requires computing the marginal likelihood (Equation 4), which is the weighted likelihood of the observed data across the entire parameter space. The marginal likelihood is, in fact, a weighted likelihood for a *model class*: Each parameter value θ defines one particular model in the class. In this sense, the Bayes factor is in fact a likelihood ratio of model classes (Liu & Aitkin, 2008). Under this light, $BF_{01} = 1/5$ means that the data are five times more likely to have occurred under the *model class* M_1, averaged over its within-model prior distribution (Liu & Aitkin, 2008).

Unfortunately, *the most likely model class need not include the true model that generated the data*. In other words, the Bayes factor can point at the "best" of two model classes, but not necessarily at the model class that contains the true model (in case the true model exists, of course). For example, consider BF_{10} from Equation 6. If $n = 100$ and the true mean μ is .1, and assuming that both σ and σ_1 are equal to 1, then $z = \sqrt{n}\mu = 1$ and therefore $BF_{01} = 1/BF_{10} = \sqrt{1+n\sigma_1^2}\exp\left[-\frac{n\sigma_1^2}{2(1+n\sigma_1^2)}z^2\right] = 6.1$, suggesting that the data are over six times more likely under M_0 than under M_1. Therefore, the Bayes factor points at M_0 even though $\mu = .1$ is one instance of the model class M_1.

In particular, poorly chosen within-model priors can distort the marginal likelihoods to the extent that the Bayes factor can indicate support for the wrong model class (i.e., the model class that does not include the true model, while the unsupported model class does contain it). One is therefore advised to moderate claims derived from Bayes factors in the general situation where weighted likelihoods are involved. After our own reading of the literature, we were surprised to realize that this perspective has been hardly noticed. Apparently, researchers interpret models of the type M_1: $\theta \neq \theta_0$ as one "model." We prefer the model class perspective, as it helps putting composite (i.e., nonpoint) models under a clearer light. Therefore, a very practical advice to overcome the model class issue is to explicitly mention it. At the very minimum, the role of the within-model prior should be made salient, because it is this prior that is used to weigh the likelihood, as explained before.

Point 8: Mismatch Between Bayes Factors and Parameter Estimation

Frequentist statistics is often blamed for suffering from serious flaws. However, one of its features that strikes us as ideal is a seemingly good match between estimation and testing. It is well known that the result of a two-sided NHST of level $100\alpha\%$ is directly related to the corresponding $100(1-\alpha)\%$ confidence interval: The test rejects the null model if and only if the null point is outside the confidence interval. This property does not hold under the Bayesian framework in general. Specifically, it is entirely possible that a $100(1-\alpha)\%$ credible interval excludes the null point (say, μ_0) but the Bayes factor shows (some) support for M_0: $\mu = \mu_0$ over M_1: $\mu \neq \mu_0$, or vice versa (Kruschke & Liddell, 2018b). To see an example, we extend the previous example of testing M_0: $\mu = 0$ against M_1: $\mu \sim N(0, \sigma_1^2)$ for normally distributed data ($Y_j \sim N(\mu, \sigma^2)$, $j = 1, \ldots, n$, with $\sigma^2 = 1$). The Bayes factor BF_{10} is given in Equation 6. We can draw support for M_0 when $BF_{10} < 1$. Solving this inequality with respect to $\bar{Y}$ implies that for observed sample means in the range

$$\bar{Y}_{BF} = \left[-\frac{\sqrt{2(1+n\sigma_1^2)}}{n\sigma_1}\left(\ln\sqrt{1+n\sigma_1^2}\right)^{1/2}, \frac{\sqrt{2(1+n\sigma_1^2)}}{n\sigma_1}\left(\ln\sqrt{1+n\sigma_1^2}\right)^{1/2}\right] \quad (7)$$

the Bayes factor BF_{10} is smaller than one and therefore there is evidence in the data supporting M_0.

We now need to derive an expression for the 95% credible interval for μ under M_1. The details can be found in Appendix C in the online supplemental materials; here we show the formula:

$$95\%\text{ credible interval} = \left[\mu_{post} - 1.96\sigma_{post}, \mu_{post} + 1.96\sigma_{post}\right], \quad (8)$$

with $\mu_{post} = \frac{n\bar{Y}\sigma_1^2}{1+n\sigma_1^2}$ and $\sigma_{post} = \frac{\sigma_1}{\sqrt{1+n\sigma_1^2}}$. The interval in Equation 8 indicates the range of μ values that are most likely based on the within-model prior and the observed data, so there is probability .95 that the true mean lies in this interval. In particular, it is of interest to see whether zero is in the interval, that is, $\mu_{post} - 1.96\sigma_{post} < 0 < \mu_{post} + 1.96\sigma_{post}$. This condition implies that $-1.96\sigma_{post} < \mu_{post} < 1.96\sigma_{post}$, hence the observed mean values $\bar{Y}$ in the range

$$\bar{Y}_{CI} = \left[-\frac{1.96\sqrt{1+n\sigma_1^2}}{n\sigma_1}, \frac{1.96\sqrt{1+n\sigma_1^2}}{n\sigma_1}\right] \quad (9)$$

are associated to credible intervals that include 0 (see Appendix C in the online supplemental materials for details).

Visual comparison of both $\bar{Y}_{BF}$ and $\bar{Y}_{CI}$ suggests that the two ranges of $\bar{Y}$ values shown in Equations 7 and 9 differ; Figure 3 makes the comparison concrete. The dashed areas relate to pairs of values $(n, \bar{Y})$ for which the 95% credible interval includes zero but the Bayes factor favors M_1 instead (i.e., $BF_{10} > 1$), or vice versa.

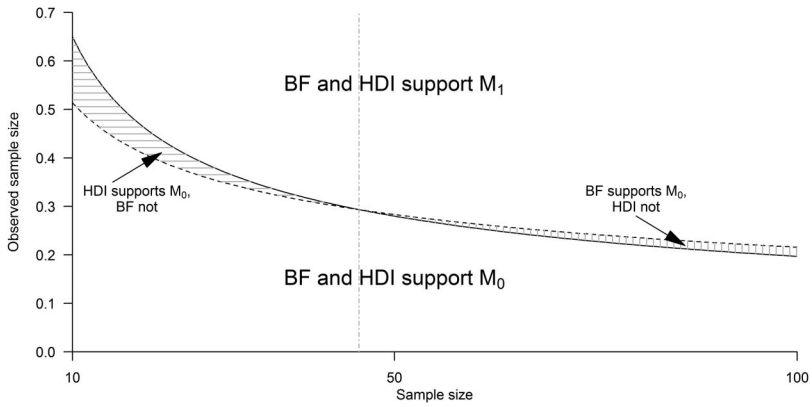

Figure 3. Agreement between Bayes factors and credible intervals (data $Y_j \sim N(\mu, \sigma^2 = 1)$; M_0: $\mu = 0$ vs. M_1: $\mu \sim N(0, \sigma_1^2 = 1)$). The solid curve is the upper bound of the 95% credible interval as a function of the sample size. The dashed curve is the upper bound of the rejection region specified by $BF_{10} = 1$. The dashed areas concern the $(n, \bar{Y})$ points for which there is disagreement between the Bayes factor and the credible interval. The plot is symmetric around the x-axis; here only the upper part of the plane is shown for convenience (i.e., when $\bar{Y} > 0$).

Thus, as can be seen, the outcome from credible intervals and Bayes factors need not coincide.

Our example is simply to illustrate that a common property from classical statistics need not hold under the Bayesian framework. By no means we encourage researchers to simply use credible intervals to decide rejecting a parameter value if it lies outside the credible interval and accepting it when it falls within it. First of all, this "rule" ignores the probabilistic nature of the interval; the parameter lies within the 95% credible interval with 95% probability, not absolute certainty. Furthermore, it ignores that the probability that the parameter has any specific value is zero anyway; we may only compute probabilities for ranges of values. The question is then what researchers should do concerning the mismatch between results from tests and credible intervals. First of all, this mismatch should be acknowledged in order to prevent misinterpretation of results. Thus, we should not be tempted to carryover this kind of intuition from the classical toward the Bayesian world. Furthermore, it is now apparent that, under the Bayesian paradigm, there is a stricter line separating estimation from testing. Both approaches seemingly answer different questions and therefore we should choose the approach that best suits the research question at hand (Kruschke, 2011). While some Bayesians argue that estimation and testing should indeed be kept apart (Jeffreys, 1939; Ly et al., 2016; Wagenmakers et al., 2018), others question this state of affairs (Robert, 2016). The problem is directly related to the use of the point null model, because such models prevent a specification of a proper within-model prior distribution. Concerning this problem, Kass (1993, p. 552) argued as follows: "When a sharp [point] hypothesis is involved, the prior must put mass on that hypothesis, e.g. $\psi = \psi_0$, whereas for 'estimating' ψ a continuous prior is used." The problem disappears when testing two point models against each other, as the Bayes factor reduces to the likelihood ratio (and credible intervals are of no direct concern in this setting).

For a unified treatment that attempts to bring point estimation, region estimation, and hypothesis testing under one common framework of decision theory, see Bernardo (2012) and the interesting discussions that follow therein.

Point 9: Bayes Factors Favor the Point Null Model

The nature of the point null model has an impact on the performance of the Bayes factor and its relation with classical inference. Berger and Sellke (1987) argued that posterior probabilities of M_0 are not aligned with classical p values when testing point null models, even though both types of probabilities are measures of evidence against M_0 (the smaller the probability the larger the evidence against M_0). The general pattern is that p values overstate the evidence against M_0, using Bayesian inference as the term of comparison. In other words, small p values (supporting evidence against M_0) are typically matched by large posterior probabilities of M_0 and hence of Bayes factors that support M_0 more strongly (assuming equal model probabilities a priori), for several families of "objective" within-model prior distributions. This finding is based on previous results from Edwards et al. (1963) and Dickey (1977). To illustrate what is at stake, we borrow from Example 1 in Berger and Sellke (1987) (alternatively, see Example 1 in Berger & Delampady, 1987) and consider again the Bayes factor B_{10} from Equation 6. This is the expression of the Bayes factor indicating by how much our prior belief favoring M_1: $\mu \sim N(0, \sigma_1^2)$ over M_0: $\mu = 0$ should shift after taking the data into account, for data assumed normally distributed with mean μ and known variance $\sigma^2 = 1$. The posterior model probabilities are given by Equation 3. Assuming that both models are equally likely a priori (as Berger & Sellke, 1987 did) the prior odds equal one and the following holds:

$$p(M_0|D) = (BF_{10} + 1)^{-1}.$$

We can replace BF_{10} by means of Equation 6, which leads to

$$p(M_0|D) = \left\{ \frac{1}{\sqrt{1 + n\sigma_1^2}} \exp\left[\frac{n\sigma_1^2}{2(1 + n\sigma_1^2)} z^2 \right] + 1 \right\}^{-1}. \tag{10}$$

The classical two-sided z test rejects M_0: $\mu = 0$ when $|z| \geq 1.96$ at 5% significance level. Because $z = \sqrt{n}\,(\bar{Y} - 0)/\sigma = \sqrt{n}\,\bar{Y}$, we infer that absolute sample means at least equal to $1.96/\sqrt{n}$ lead to

rejecting M_0 at 5% significance level. It is now straightforward to assess how much evidence such "extreme" data bring against M_0 either in terms of $p(M_0|D)$ (Equation 10) or BF_{10} (Equation 6). Figure 4 shows the values of $p(M_0|D)$ and $p(M_1|D)$ (left panel) and BF_{10} (right panel) as functions of the sample size. The variance of the within-model prior was fixed at 1, with no loss of generality. Note that, with increasing n, $\bar{Y}$ will decrease because we fix $z = \sqrt{n}\,\bar{Y}$ at 1.96. Thus, we study BF_{10} for data for which an NHST outcome would be exactly on the significance threshold (i.e., $z = 1.96$). As it can be easily verified from the left panel, there is strong disagreement between the p value (the horizontal dashed line) and the posterior probability of M_0: $p(M_0|D)$ is much larger. In fact, when $n \geq 42$ it is always the case that the classical test rejects M_0 whereas $p(M_0|D)$ *is larger than .50*, thus indicating support *for* M_0. The right panel shows the same pattern in terms of the Bayes factor. Therefore, we conclude that for small to moderate sample sizes ($n < 42$) both the z test and the Bayes factor indicated some level of support for M_1. As sample size increases, the NHBT is *much more conservative* than the classical null hypothesis test: The observed effect size should be much larger for the Bayes factor to favor M_1. Because in this case the sample means are also standardized effect sizes (since $d = (\bar{Y} - 0)/\sigma = \bar{Y}$), we conclude that Bayes factors regard effect sizes deemed "statistically significant" under the classical paradigm as simply being too small to warrant dismissing support for M_0, at least as the sample size increases. In sum: NHST is severely biased against M_0 (e.g., Sellke et al., 2001).

A natural question is then: What magnitude of effect sizes would make the Bayes factor shift toward M_1 instead? As explained in a previous section, we recommend drawing conclusions on the basis of posterior odds, and to get those, we need to specify prior odds. Here we choose that the prior odds are one for convenience, but any other prior odds could be chosen. Again rather arbitrarily, we suggest that a posterior odds ratio of 19 could be considered strong evidence for M_1. This pertains to $p(M_0|D) = .05$ and $p(M_1|D) = .95$. Alternatively, we could draw upon the guidelines from Jeffreys (1961) and use the following minimum reference values for BF_{10}: 1 (*barely worth mentioning*), 3 (*substantial evidence*), 10 (*strong*), 30 (*very strong*), and 100 (*decisive*), which implicitly also seem to take prior odds equal to one. These values are, according to Jeffreys' (1961) taxonomy, the minimum that allow qualifying the evidence supporting M_1 as indicated by the corresponding label. To find the value of $\bar{Y}$ that corresponds to a particular posterior odds (in this case equaling the Bayes factors), we need to solve Equation 6 with respect to $\bar{Y}$, for given values of BF_{10}. This leads to the solution

$$\bar{Y} = \frac{z^*}{\sqrt{n}}, \text{ where } z^* = \left[\frac{2(1+n\sigma_1^2)}{n\sigma_1^2}\log\left(BF_{10}\sqrt{1+n\sigma_1^2}\right)\right]^{1/2}. \tag{11}$$

Figure 5 displays values of the sample mean shown in Equation 11 for varying sample size n and a range of values of BF_{10}: 1, 3, 10, 19, 30, 100. The minimum evidence required to reject M_0 under NHST is matched with Bayes factors between 1 (*barely worth mentioning*) and 3 (*substantial evidence*). The problem decreases as the sample size increases, but for sample sizes below say, 100 it is relevant.

Berger and Sellke (1987) showed that $p(M_0|D)$ is commonly larger than the p value under a wide variety of circumstances by computing lower bounds on $p(M_0|D)$ for several classes of within-model priors under M_1. Berger and Delampady (1987) also concluded that p values tend to exacerbate the data evidence against the point null model in comparison to two Bayesian measures of evidence, namely, the Bayes factor and the Bayesian posterior probability of models. Thus, under a wide range of settings, we conclude that the evidence against M_0 indicated by p values is weak at best. Recent empirical assessments of published research

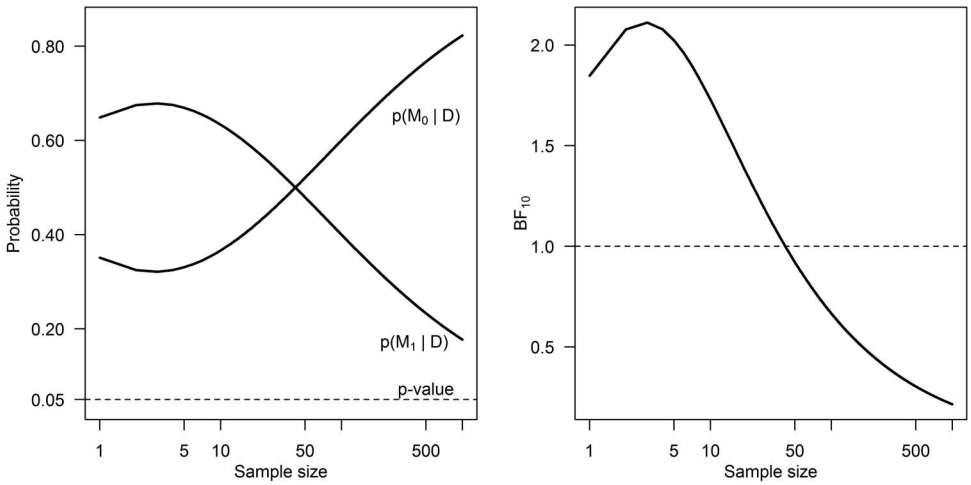

Figure 4. p values overstate the evidence against the null model (data $Y_j \sim N(\mu, \sigma^2 = 1)$; M_0: $\mu = 0$ vs. M_1: $\mu \sim N(0, \sigma_1^2 = 1)$). The left panel shows the posterior model probabilities when $|z| = 1.96$ (solid lines) in comparison to the p value (dashed line). The mismatch between $p(M_0|D)$ and the p value is clear. The right panel shows $BF_{10} = p(M_1|D)/p(M_0|D)$ (because the prior odds are equal to one; recall Equation 2), computed when $|z| = 1.96$ (solid line). The horizontal dashed line is the reference value (equal support for either model). As sample size increases, the Bayes factor decreases and eventually becomes smaller than one, showing increasing evidence *for* the null model.

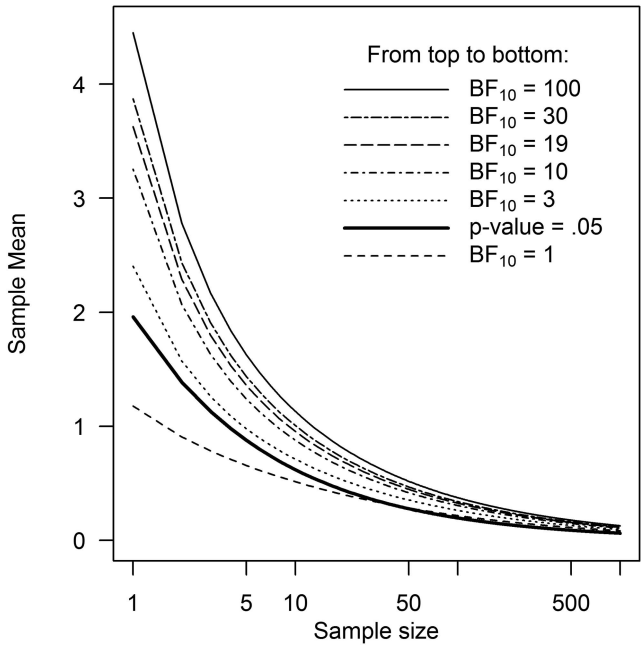

Figure 5. Evidence provided by p values against M_0, in terms of Bayes factors (data $Y_j \sim N(\mu, \sigma^2 = 1)$; M_0: $\mu = 0$ vs. M_1: $\mu \sim N(0, \sigma_1^2 = 1)$). The plot shows the minimum effect size (y-axis) required to meet either statistical significance (solid line) or particular Bayes factor levels (dashed lines), as functions of the sample size. The minimum evidence required to reject M_0 under NHST is matched with Bayes factors between 1 (*barely worth mentioning*) and 3 (*substantial evidence*).

made this point salient (Aczel, Palfi, & Szaszi, 2017; Hoekstra, Monden, van Ravenzwaaij, & Wagenmakers, 2018; Jeon & De Boeck, 2017; Wetzels et al., 2011).

Surprisingly, the previous results do not generalize to directional hypothesis testing problems (e.g., comparing $\mu > 0$ and $\mu < 0$; see Casella & Berger, 1987; Pratt, 1965), which is not strictly NHBT by our definition. In this case, Casella and Berger (1987) argued that $p(M_0|D)$ and p values can be very close to each other under a wide range of classes of within-model priors. In order to illustrate this close relation between Bayes factors and p values for directional hypotheses testing, we further extend the running example and consider testing M_2: $\mu > 0$ versus M_3: $\mu < 0$, with $\mu \sim N(0, \sigma_1^2)$. In this case the Bayes factor BF_{32} is a ratio of areas under the posterior distribution of μ, $p(\mu|D)$, as shown below:

$$BF_{32} = \frac{p(D|M_3)}{p(D|M_2)} = \frac{\int_{-\infty}^{0} p(D|\mu, M_3)p(\mu|M_3)d\mu}{\int_{0}^{\infty} p(D|\mu, M_2)p(\mu|M_2)d\mu} = \frac{\int_{-\infty}^{0} p(D|\mu, M_3)p(\mu|M_3)d\mu / p(D)}{\int_{0}^{\infty} p(D|\mu, M_2)p(\mu|M_2)d\mu / p(D)} = \frac{p(\mu < 0|D)}{p(\mu > 0|D)}. \quad (12)$$

The posterior distribution $p(\mu|D)$ has a closed form

$$\mu|D \sim N\left(\frac{\sqrt{n}z\sigma_1^2}{1 + n\sigma_1^2}, \frac{\sigma_1^2}{1 + n\sigma_1^2}\right) = N(\beta_1, \beta_2); \quad (13)$$

for a derivation see Equation B2 in Appendix B in the online supplemental materials with $\sigma = 1$ and $\bar{Y} = z/\sqrt{n}$. Hence, BF_{32} can be expressed in terms of the normal cumulative function:[3]

$$BF_{32} = \frac{\Phi\left(\frac{0 - \beta_1}{\sqrt{\beta_2}}\right)}{1 - \Phi\left(\frac{0 - \beta_1}{\sqrt{\beta_2}}\right)} = \frac{\Phi\left(-\frac{\beta_1}{\sqrt{\beta_2}}\right)}{\Phi\left(\frac{\beta_1}{\sqrt{\beta_2}}\right)} = \frac{\Phi(-\xi)}{\Phi(\xi)},$$

$$\text{with } \xi = \frac{\sqrt{n}z\sigma_1}{\sqrt{1 + n\sigma_1^2}}. \quad (14)$$

The classical one-sided z test rejects M_2 when $z < -1.64$ at 5% significance level, that is, when the sample mean is not larger than $\bar{Y} = z/\sqrt{n} = -1.64/\sqrt{n}$. The left panel of Figure 6 shows the posterior probabilities of both M_2 and M_3 for such effect sizes. It is clear that the posterior probability of M_2 converges quickly to the p value as the sample size increases, thus indicating a strong agreement between both measures of evidence against M_2. The right panel of Figure 6 shows the corresponding values of BF_{32} which are converging to .95/.05 = 19 (see Equation 14 and note that ξ converges to $z = -1.64$ as n increases). Figure 7 is analogous to Figure 5 as it tries to compare effect sizes at various Bayes factor thresholds against that required from the classical test to reject M_2. Now the minimum sample mean required to reject M_2 under the classical test is between the sample means that Bayes factors would label as *strong* ($B_{32} = 10$) and *very strong* ($B_{32} = 30$). In fact, the level of evidence required by the p value quickly converges to that of $BF_{32} = 19$ (i.e., $p(M_0|D) = .05$).

Casella and Berger (1987) showed that, in one-sided testing settings, the lower bounds of $p(M_0|D)$ can be equal *or even smaller* than the p value, for four classes of within-model priors (the example presented above belongs to the fourth class). In such cases the p values do not overstate the evidence against the null model (conversely, perhaps it is the Bayes factor who does). We do not see this point stressed in the literature of Bayes factors, possibly because tests of point null models are overly predominant in psychology.

Some caution should be exerted at this point. The p value, which is the probability of observing data at least as extreme as the data under consideration assuming that M_0 holds, is a probability statement of repeated sampling. The p value is *not* the probability that M_0 is true. The Bayesian posterior probability of M_0, $p(M_0|D)$, on the other hand, *is* a probability statement of M_0. The fact that p values and Bayes factors are numerically close in directional tests does not change this fact. Thus, researchers should not feel entitled to draw inferences concerning the "truth" of M_0 from p values simply because they are numerical close to Bayesian model probabilities, as this remains a logically invalid inferential step (e.g., Berger & Delampady, 1987). Nevertheless, one can gain some comfort from the agreement between both the classical and the Bayesian results.

[3] Equivalently, models M_2 and M_3 could have been defined using the truncated normal distribution: $\tilde{M}_2$: $\mu \sim N^+(0, \sigma_1^2)$ and $\tilde{M}_3$: $\mu \sim N^-(0, \sigma_1^2)$, where N^+ and N^- denote the truncated normal distribution on $(0, \infty)$ and $(-\infty, 0)$, respectively. It can be shown that $B_{23} = p(D|\tilde{M}_2)/p(D|\tilde{M}_3)$.

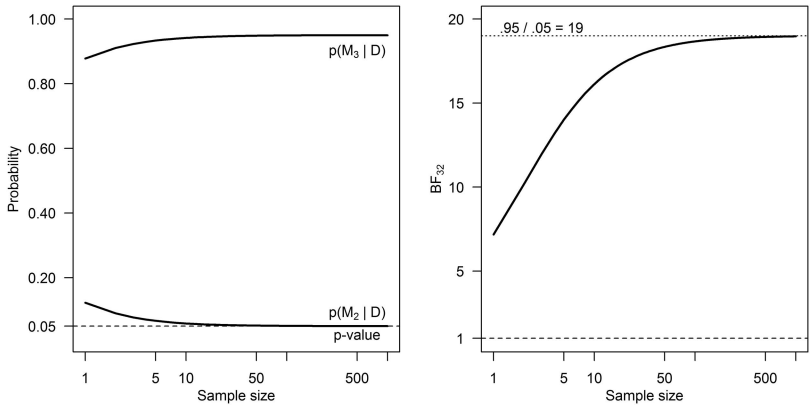

Figure 6. p values are well calibrated with posterior probabilities of the null model (data $Y_j \sim N(\mu, \sigma^2 = 1)$; M_2: $\mu \sim N^+(0, \sigma_1^2 = 1)$ versus M_3: $\mu \sim N^-(0, \sigma_1^2 = 1)$). The left panel shows the posterior model probabilities when $z = -1.64$ (solid lines) in comparison to the p value (dashed line). $p(M_2|D)$ and the p value quickly converge to each other as the sample size increases. The right panel shows $BF_{32} = p(M_3|D)/p(M_2|D)$ (because the prior odds are equal to 1; recall Equation 2), computed when $z = -1.64$ (solid line). The horizontal dashed line at $BF_{32} = 1$ is the reference value (equal support for either model). As sample size increases, the Bayes factor approaches the limiting value $BF_{32} = 19$, showing support for M_3.

One may wonder about the main reasons that differentiate directional and two-sided tests so distinctly. The "culprit" seems to be the point null model in two-sided tests: "It seems that if some prior mass is concentrated at a point (or in a small interval) and the remainder is allowed to vary over H_1, then discrepancies between Bayesian and frequentist measures will obtain" (Casella & Berger, 1987, p. 110). In empirical terms, it is difficult to defend that an interval of infinitesimal length is equally weighted as its complementary support, even in cases where a true null point may be meaningful (but Jeffreys, 1961 in particular disagreed completely with this opinion; see Etz & Wagenmakers, 2017 for a historical overview of point null models). We concur with the description by Vardeman (1987, p. 130): " I must say that I find the 'spike at θ_0' feature of the priors used . . . to be completely unappealing . . . The issue is simply that I do not believe that any scientist, when asked to sketch a distribution describing his belief about a physical constant like the speed of light, would produce anything like the priors used by Berger and Sellke. A unimodal distribution symmetric about the current best value? Probably. But with a spike or 'extra' mass concentrated at θ_0? No." We do agree that, theoretically, there could be some settings for which point null models are sensible, but we find it hard to believe these can be found to be true in actual social sciences research. Some authors suggest to view point null models as simplifying approximations (Berger & Delampady, 1987; Gallistel, 2009; Kass & Raftery, 1995; Konijn et al., 2015; Marden, 2000; Morey & Rouder, 2011; Rouder et al., 2009). We think indeed this makes sense, but then we would prefer to actually specify the hypothesis that is being approximated itself, rather than the approximating one. One could counter that the actual hypothesis might be hard to specify due to embarras du choix. However, smooth and possibly strongly peaked within-model priors in our view make more sense, because they are continuous and there are no qualitative differences between adjacent parameter values. Nevertheless, we realize that the point null model is so strongly entrenched in social sciences research that it is likely to stay at least for quite some time. It is then all the more important that researchers are aware of what is at stake when using NHBT. In particular, how well does the point null model approximate one's actual hypothesis (Berger & Delampady, 1987; Cohen, 1994; Kruschke & Liddell, 2018b; Morey & Rouder, 2011)?

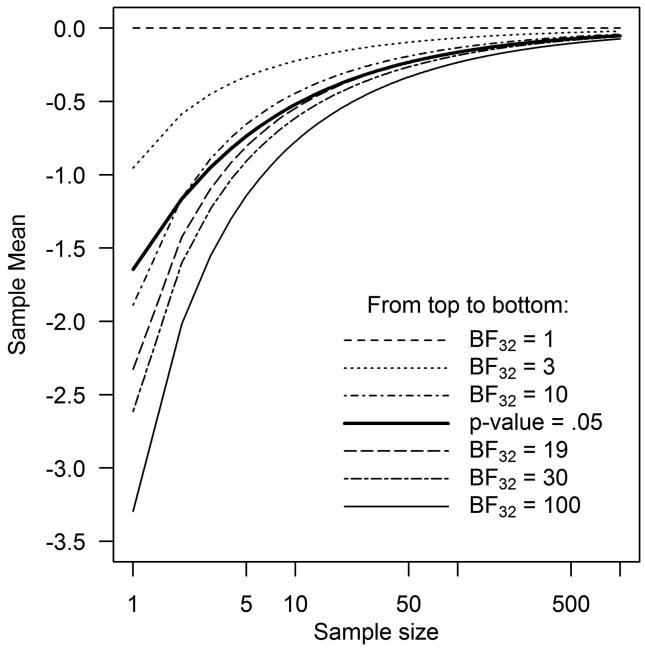

Figure 7. Evidence provided by p values against M_2, in terms of Bayes factors (data $Y_j \sim N(\mu, \sigma^2 = 1)$; M_2: $N^+(0, \sigma_1^2 = 1)$ versus M_3: $\mu \sim N^-(0, \sigma_1^2 = 1)$). The plot shows the minimum effect size (y-axis) required to meet either statistical significance (solid line) or particular Bayes factor levels (dashed lines), as functions of the sample size. The minimum evidence required to reject M_2 under NHST is typically matched with Bayes factors between 10 (*substantial evidence*) and 30 (*strong*).

Observe that the previous problem might be more of a theoretical than a practical one. The NHBT Bayes factor can be approximated as a ratio of areas under the posterior distribution defined by a small interval around the null point (see Berger & Delampady, 1987, Section 2.2), and conversely, tests based on point null models (e.g., M_0: $\mu = 0$ vs. M_1: $\mu \neq 0$) can be regarded as simplifications of tests of the type M_0: $|\mu| \leq \varepsilon$ versus M_1: $|\mu| > \varepsilon$.

What useful advice can practitioners draw from all the information above? First, it is important to keep in mind the relation between NHST and NHBT. We argued that *p* values overstate the evidence against the point null model, both in terms of the Bayes factor as in terms of posterior model probabilities. In this regard, NHBT does offer an advantage over NHST. However, we also highlighted that point null hypotheses are not necessarily natural specifications of theory in psychology. And quite remarkably, tests that do not rely on point null hypotheses seem to perform similarly to their frequentist counterparts (as shown by Casella & Berger, 1987). We think that the point null hypothesis, in the context of NHBT, is a strange entity that creates both conceptual as well as mathematical difficulties (in the sense that it forces considering both discrete and continuous measures of probability simultaneously). To this problem we can offer no easy solution other than suggesting alternative inferential types of analyses which are based on considering the full posterior distribution over the entire parameter space (thus, effectively, bypassing NHBT altogether; see the Discussion section for more details).

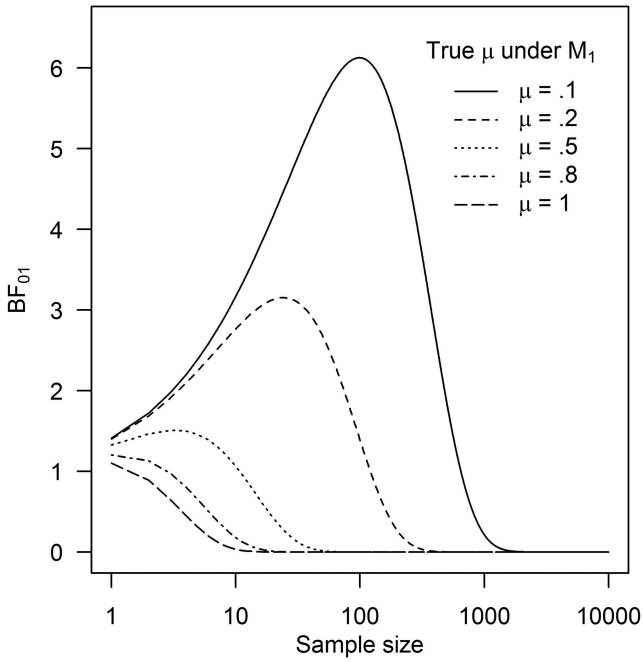

Figure 8. The Bayes factor favors the alternative model (data $Y_j \sim N(\mu, \sigma^2 = 1)$; M_0: $\mu = 0$ vs. M_1: $\mu \sim N(0, \sigma_1^2 = 1)$). The smaller the true effect under M_1, the larger the sample size required for the Bayes factor to display support for M_1. But BF_{01} will invariably converge to 0 as long as M_0 is not exactly true.

Point 10: Bayes Factors Favor the Alternative

We argued in the previous section how Bayes factors can favor point null hypotheses. Interestingly, we now observe that Bayes factors, similarly to NHST, have the property of leading to a rejection of M_0 as the sample size increases unless M_0 is exactly true. That is, if the true parameter is (ever so slightly) different from the exact value under M_0, both the *p* value and BF_{01} will approach zero as the sample size increases. Thus, even when the true effect is very small for practical purposes, with large enough sample size one is certain to reject the null model. This limiting property of the Bayes factor can be illustrated, for instance, by considering the limits of the Bayes factors in Equation 6. When M_0: $\mu = 0$ is false (even for arbitrarily small values of μ), both $BF_{01} = 1/BF_{10}$ and $BF_{02} = 1/BF_{20}$ converge to 0. This result is displayed in Figure 8 for various true values of μ (for BF_{01} only, the plot for BF_{02} is identical). For small sample sizes the Bayes factor favors the simpler null model because the data do not yet provide strong evidence supporting the alternative. As the sample size increases the evidence against M_0 accumulates and leads to the Bayes factor providing unbounded support for the alternative model.

Concerning this property, Morey and Rouder (2011, p. 411) stated: "In this regard, the JZS Bayes factor [and other Bayes factors] shares an unfortunate property of NHST: It provides no means of assessing whether rejections of the nil are due to trivial or unimportant effect sizes or are due to more substantial effect sizes." Thus the usual discussion of "statistical" versus "practical" significance in frequentist statistics also holds with NHBT. We should then be reminded to carefully look at the sizes of the effects being studied using tools alternative to NHBT, similarly as it is commonly advised in the frequentist framework (Wilkinson & Task Force on Statistical Inference, 1999). Furthermore, we should realize that Bayes factors do not provide information regarding the uncertainty of the observed effect sizes. For this purpose we could invoke a credible interval around the observed effect size. Hence, we will always need the posterior distribution of the parameter of interest to present the full picture.

Following Meehl (1967), Kruschke and Liddell (2018b) argued that null models of the type M_0: $\mu = 0$ are often false a priori (see our discussion in the previous section). That is, the real underlying process must surely imply that some effect, no matter how small, need exist (e.g., a small correlation, a small regression effect, a small proportion of explained variance). Assuming this is the case, it is then always possible to reject the null model using Bayes factors with a large enough sample. In other words, it is easy to *confirm* a theory for any non-null effect simply by collecting more data. But, as Kruschke and Liddell (2018b) argue, science should work the other way around: Data accumulation should make it easier to *disconfirm* a model simply because a model is almost surely missing the real true effect, even if only infinitesimally so. This is known as Meehl's paradox (Meehl, 1967). The conclusion is that Bayes factors do not solve Meehl's paradox, just like NHST. Actually, we argue that when null models are false a priori, there is no reason to use them to start with. In this case, the interest is to see whether the parameter is substantially larger or smaller than 0. After estimating the posterior distribution, this can be assessed directly.

On the positive side, it can be argued that this particular feature of Bayes factors is also adequate. The default Bayes factors ad-

vocated by Jeffreys (1961) are said to be *information consistent* in the sense that they display unbounded support for either model when data are overly informative either way (Ly et al., 2016). Per se, this is a good feature. It is the mismatch between the statistical versus practical significance that needs to be further considered in order to improve our inferences. Therefore, definitions of what counts as "practically worthwhile" (such as Kruschke's ROPE; Kruschke, 2013) are indispensable.

A different argument that highlights another way in which Bayes factors may favor the alternative has been put forward by Johnson and Rossell (2010). We apply this to NHBT comparing M_0: $\theta = \theta_0$ versus M_1: $\theta \neq \theta_0$.[4] Johnson and Rossell (2010) show the following: Bayes factors accumulate evidence in favor of true M_1 much faster than they do in favor of true M_0 as the sample size increases, for fixed sample-based estimates of θ. That is, although Bayes factors allow drawing support for either M_0 or M_1, they do so *asymmetrically*. This property is in contrast with the commonly praised feature of Bayes factors being symmetric (in the sense that they allow accumulating evidence for either model), unlike p values. Figure 9 illustrates what is happening with an example. Observe how $\log(BF_{01})$ increases at a much slower rate under M_0 than it decreases for various values under M_1, thus highlighting the asymmetry property reported above.

Summarizing, practitioners need to keep in mind that NHBT does not take effect size in consideration, particularly for large sample sizes (see Figure 8). It is therefore indispensable that estimation of the magnitude and uncertainty of the size of an effect is taken care of in addition to an NHBT analysis.

Point 11: Bayes Factors Often Agree With p Values

The point we want to stress here was already discussed before (when discussing two- vs. one-tailed tests): Results between Bayes factors and classical tests often agree with each other. This result may sound surprising given the widespread use of strong statements like "p-values are violently biased against the null hypothesis" (Edwards, 1965; Wagenmakers et al., 2018). But interestingly, Jeffreys (1961, p. 393, as cited in Ly et al., 2016) had already recognized it.

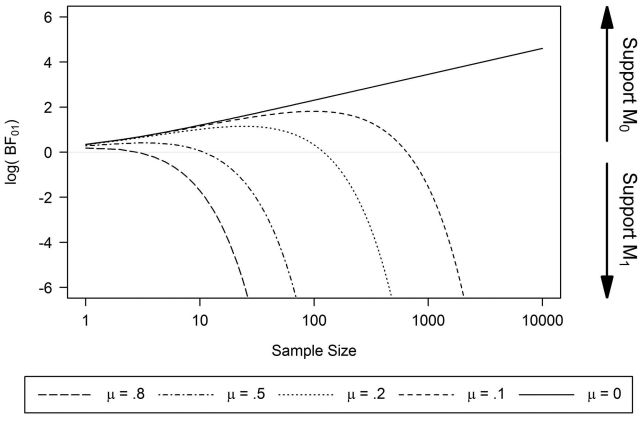

Figure 9. The Bayes factor favors the alternative model (data $Y_j \sim N(\mu, \sigma^2 = 1)$; M_0: $\mu = 0$ vs. M_1: $\mu \sim N(0, \sigma_1^2 = 1)$). The rate of change of $\log(BF_{01})$ as the sample size increases is notoriously smaller under M_0 (solid line) than under M_1 (the dashed lines).

To be clear: Yes, there are instances in which disagreement between Bayes factors and p values exists, with results even pointing in opposite directions (Edwards et al., 1963; Lindley, 1957; also the discussion by Luis Pericchi in Bernardo, 2012, p. 25). Some have argued that disagreements are prone to be found when the p value is just below the significance level (e.g., Wetzels et al., 2011). Thus, the issue may not be one of disagreement between p values and Bayes factors but, instead, one of calibration (Jeon & De Boeck, 2017; Wetzels et al., 2011). There have been recent proposals to redefine the typical statistical significance level of the classical test (Benjamin et al., 2018; please make note of arguments also against this proposal, for instance Crane, 2017; Ioannidis, 2018; Lakens et al., 2018). But this should not obscure the fact that also in many other situations (e.g., when the effect is strong) agreement between both approaches is to be expected.

Trafimow (2003) argued that the usual claim that p values are too liberal (i.e., too small in comparison to $p(M_0|D)$) is not founded in general. This can be shown as follows. First, consider Equation 1 slightly rewritten as follows:

$$p(M_0|D) = \frac{p(M_0)p(D|M_0)}{p(M_0)p(D|M_0) + [1 - p(M_0)]p(D|M_1)}. \quad (15)$$

In order to compute $p(M_0|D)$, we need to know three quantities: $p(D|M_0)$, $p(M_0)$, and $p(D|M_1)$. The first quantity is what the p value is based on, namely, the probability of the observed data under the null model. But the latter two quantities are essential to fully express $p(M_0|D)$ as shown in Equation 15. In particular, large values of $p(M_0)$ and, more importantly, *low values of* $p(D|M_1)$, will invariably lead to $p(M_0|D)$ being close to one, for $p(D|M_0)$ fixed. The difficulty under the classical paradigm is that the likelihood of the data under the alternative model is *not taken into account* when rejecting M_0. Suppose that the p value is below a predefined significance level (i.e., $p(D|M_0)$ is low). The null model will be rejected in spite of the fact that the data may even be more unlikely under M_1 (i.e., $p(D|M_1)$ may be even lower than $p(D|M_0)$). In this case, the posterior probability of M_0 and BF_{01} may be high, contrasting with the low p value. In this sense the p values are indeed too liberal. However, high values of $p(D|M_1)$, which indicate that the observed data are more likely under M_1 than under M_0, are associated with low posterior probabilities of M_0 (see Equation 15) and large values of BF_{10} (as long as $p(M_0)$ is not unacceptably large). In this case the p values are not liberal at all (Trafimow, 2003).

Let us try to understand the above argument by means of revisiting the success rate of a Bernoulli process from section 2. As before, suppose we want to test M_0: $\theta = .5$ versus M_1: $\theta \neq .5$, where now we consider the within-model prior *Beta* (*7, 2*). Suppose a series of $n = 100$ independent trials are conducted; r ($r = 0, \ldots, 100$) successes are recorded. We can express $p(D|M_0)$ and $p(D|M_1)$ as functions of r (using Equation 5 with $\theta = .5$, and Equation A3 in Appendix A in the online supple-

[4] The point being made here is general and applies to tests of the type M_0: $\theta \in \Theta_0$ versus M_1: $\theta \in \Theta_1$, where Θ_0 and Θ_1 are complementary of each other and cover the entire parameter range of interest, such that $p_0(\theta)$ and $p_1(\theta)$ are positive on Θ_1 and Θ_0, respectively. Johnson and Rossell (2010) dubbed priors verifying this property as being "local alternative prior densities." We focus on point null M_0 in the article for illustration purposes.

mental materials). We will assume $p(M_0) = p(M_1) = .5$. In Figure 10A we display the marginal likelihoods $p(D|M_0)$ and $p(D|M_1)$. The grayed area corresponds to the critical region of r values for which the frequentist binomial test rejects M_0 at 5% significance level ($r \leq 39$ and $r \geq 61$). The posterior probability of M_0 (Equation 15) and (the logarithm of) BF_{10} are shown in Figures 10B and 10C, respectively. First, observe that the r values in the critical region are associated to low probabilities under M_0 (i.e., $p(D|M_0)$ is low; Figure 10A), as we would expect. However, for r values close to 39, $p(D|M_1)$ *is even lower*. For example, $p(D|M_0) = .011$ and $p(D|M_1) = .002$ when $r = 39$, so $p(M_0|D) = .87$ and $BF_{10} = .15$ which indicates support for M_0. In this case a conflict between the p value and $p(M_0|D)$ (or BF_{10}) arose *because the observed data are largely inconsistent with the alternative model* M_1. The p value is insensitive to the likelihood of the data under M_1, whereas $p(M_0|D)$ and BF_{10} *are* sensitive to it, and this explains why the p value is biased against M_0. If we now focus on the critical region at or above 61, we conclude that $p(D|M_1)$ is always larger than $p(D|M_0)$. For example, $p(D|M_0) = .004$ and $p(D|M_1) = .012$ when $r = 61$, so $p(M_0|D) = .27$ and $BF_{10} = 2.7$ which indicates support for M_1. That is, the p value and $p(M_0|D)$ (or BF_{10}) are in agreement *because the observed data are largely consistent with the alternative model* M_1. So, in essence, p values are not biased against the null model when the data are likely under the alternative model (i.e., $p(D|M_1)$ is large). What NHBT allows, and clearly classical tests fail at, is to draw support *for* the null model. Thus, although large p values may be associated to large Bayes factors in favor of M_0, the former cannot logically be used to draw support for M_0. This is one crucial property that demarks Bayes factors as a better inferential tool than p values.

For practitioners, the take-away message is mostly conceptual: The outcomes from NHBT and NHST may agree more than anticipated, in particular if one takes the literature advocating Bayes factors as the reference. This is not to say, however, that NHBT and NHST are close to being equivalent: This is far from true. Recall that the p value is insensitive to the likelihood of the data under M_1, which is one of the main sources of disagreement between NHBT and NHST. Therefore, if one can decide on a sensible specification of M_1, NHBT does have the upper hand and is preferable over NHST.

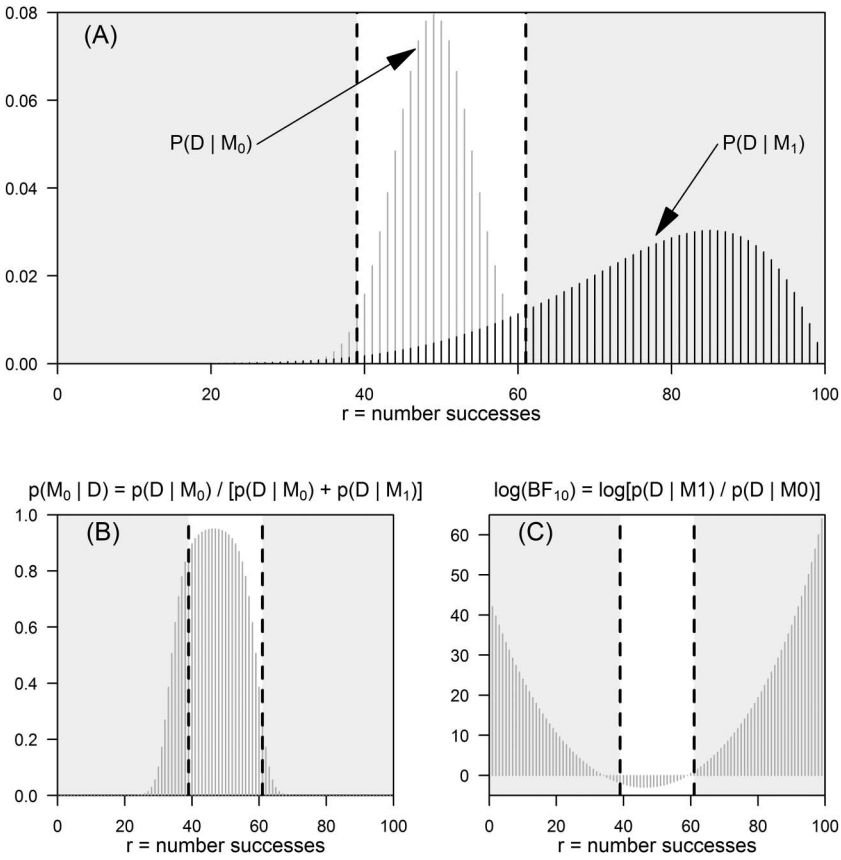

Figure 10. Distribution of $p(D|M_0)$ and $p(D|M_1)$ as functions of the number of successes in a Bernoulli process ($n = 100$; Panel A); posterior probability of M_0 (Panel B); and the logarithm of BF_{10} (Panel C). The grayed area corresponds to the critical region for which the frequentist two-sided binomial test rejects M_0: $\theta = .5$ at 5% significance level ($r \leq 39$ and $r \geq 61$). Values of $r \geq 61$ correspond to low posterior probabilities of M_0 and large BF_{10} values, thus corroborating the p value outcome. See text for details.

Bayes Factors and the Replication Crisis

In the current surge of research focusing on the replication crisis in the social sciences, NHBT is often being portrayed as part of the solution to the problem. However, NHBT has its own problems, as discussed throughout this article. Furthermore, what NHBT promises to offer might not be what researchers and journals are willing to use. Savalei and Dunn (2015) pointed out two specific issues that may limit the usefulness of NHBT in the context of the replication crisis. First, it is not yet established that the ability of NHBT to show support for null models may alleviate the widespread bias against publishing nonsignificant results. If journal editors keep insisting on only publishing extraordinary results then they will continue resisting publishing large Bayes factors supporting the null model, simply because the latter are associated to a "lack of effect." Second, Savalei and Dunn (2015) mentioned that "b-hacking" (see also Konijn et al., 2015) the results is very possible, suggesting that the widespread adoption of Bayes factors may open the way to the introduction of new types of questionable research practices. This is certainly true in the sense that no statistical tool (NHBT included) is immune to the unethical behavior of unscrupulous researchers. The risk is real and it is important to recognize it.

Discussion

The purpose of this article was to scan the literature and to learn as much as possible about NHBT. It is crucial that the scientific community gets acquainted with Bayes factors and NHBT, given its increased use in the literature. We find it ill-advised to suggest that researchers move from NHST toward NHBT without enduring the growing pains of learning the basics of this new statistical tool. Arguably, the failure of the NHST paradigm in current scientific practice is to be blamed by two main sorts of reasons. One is the technical pitfalls of NHST itself; this aspect has been exhaustively explored in the literature. The other reason is the misuse of NHST, for which the researchers (and not NHST) are to blame. For these researchers, NHBT should not be thought of as a panacea for all problems. First and foremost, they need to learn proper statistical thinking. Hopefully this article can aid toward this goal.

NHBT was shown to portray some idiosyncrasies. In a nutshell, we argued how the Bayes factors used in NHBT can be: Overly sensitive to within-model prior distributions, potentially misleading when so-called default within-model priors are invoked, interpreted without reference to posterior model probabilities, misleading in terms of selecting a "good" model, subjective to interpret qualitatively, sometimes biased toward and sometimes against the null model. Its role in improving the current state of affairs within the replication crisis is still to be determined. Altogether, we conclude that all of the above should be known to those who plan to use NHBT.

How to Use NHBT Properly

Throughout the article we pointed out various limitations and risks for misuse or misinterpretation of NHBT. For each point under discussion we offered suggestions on what we think is the best way to address the issue. Here we offer a small set of general guidelines that hopefully can assist researchers who want to use NHBT in their applied research. We summarize our ideas in three main points.

1. Set up the hypotheses. Carefully consider your research question. Are you really interested in assessing the plausibility of one particular parameter value? What are the alternative values of interest (one-sided or two-sided)? Do consider these questions *before* looking at the data.

2. (Within-model) priors. Software that includes Bayes factors in their toolbox is based on predefined within-model prior distributions. We advise practitioners to visualize (plot) these within-model priors. Try to change their parameters and adjust their shape so that they can approximate your own desired specification of M_1 as closely as possible. Keep in mind that the default within-model prior need not be the one that serves your purposes the best; this is not a matter of software, it is a matter of critical judgment. Do include all these analysis details in your report. Conduct a sensitivity analysis, report the results. Do observe that sensitivity to within-model prior distributions is not necessarily a problem; what is important is that: (a) the choice of the within-model prior distribution is sufficiently explained, and (b) the researcher covers several within-model prior distributions that allow catering to a wide audience (i.e., reflecting the views of people whose theoretical stance a priori is not necessarily aligned with our own). It is crucial that all these steps are clearly reported, for both transparency and reproducibility.

3. Interpretation. Distinguish between Bayes factors (which quantify the relative likelihood of the observed data under each model) and posterior model odds (which quantify the relative likelihood of each model given the observed data). A large BF_{10} value does not mean that the posterior probability of M_1 is very large; to know the latter we need to know the prior model probabilities. Also, always interpret a Bayes factor and posterior probabilities in relative terms. A large BF_{10} value simply implies that the data are much more likely under M_1 than under M_0; a large posterior probability for M_1 means that, given the data, M_1 is much more probable than M_0. It is still entirely possible that another model exists that has much more predictive value than M_1. Do not overly rely on qualitative labels for Bayes factors like the ones introduced by Jeffreys (1961); these are mere guidelines and should not be taken too strictly. Furthermore, remember that in NHBT composite hypotheses like $\mu \neq 0$ or $\mu > 0$ encompass a model class. So, NHBT compares the predictive value of the null point model with that of the alternative model weighted across the entire parameter range (with the within-model prior providing the weights).

But, Should We Use NHBT?

Given all that we gathered concerning NHBT, one interesting question is: "Should NHBT replace classical NHST?" The answer is imminently subjective, as it clearly depends on the person being

asked. Here we offer our personal take on this issue. First of all, we want to state very clearly that NHBT is an improvement over NHST, in spite of all the idiosyncrasies noted throughout this article. In particular, we appreciate that Bayes factors do not depend on data that were never observed, do not depend on the subjective stopping intentions of the researcher, do take the likelihood of the data under the alternative model into account, and do allow supporting either model under comparison. All the previous properties do not hold for NHST. Therefore, if testing hypotheses is the intended goal, then, clearly, choosing NHBT over NHST seems reasonable. On the other hand, the fact that the procedure is quite sensitive to the exact formulation of the alternative model in terms of a within-model prior density makes it less than ideal. If testing hypotheses is the intended goal, we suggest that comparing more balanced hypotheses, for instance comparing two point hypotheses (or two hypotheses specifying relatively tight density distributions around the points of interest), or comparing two mutually exclusive one-sided hypotheses (e.g., $\mu > 0$ vs. $\mu < 0$), may be a better alternative to NHST than the unbalanced comparison in NHBT.

In general, we think that testing is *very limitative*. NHBT is a model testing procedure that only compares two models. Bayes factors fall short in terms of quantifying the magnitude of an effect or the precision of the estimate of the magnitude (Kruschke & Liddell, 2018a). This goes against the guidelines that the American Psychological Association issued 20 years ago (Wilkinson & Task Force on Statistical Inference, 1999). Therefore, under no circumstance we conceive that reporting Bayes factors (or *p* values) *only* is a sound analytical strategy. In our opinion, any data analysis is incomplete without at least some indication of the magnitude and precision of the estimate of the effect of interest. Ideally, this is done by means of reporting full posterior distributions for the parameters of interest.

The above discussion is much more open to contention than it might seem at a first glance. Some researchers supporting Bayes factors argue that Bayes factors should be used as a *necessary* first step. Only after the data have shown a strong support for one of the models being considered (by means of a Bayes factor) should we attempt to estimate the magnitude and precision of the estimate of the effect under that model (e.g., Wagenmakers et al., 2018). In other words, one should not attempt estimating the magnitude of an effect or the precision of the estimate of the magnitude before establishing that the effect "exists." We find this procedure overly complex and unnecessary. We strongly believe that "true" point models are hard to conceive of in actual practice in social sciences research. A model of the type M_0: $\theta = \theta_0$ could be a sensible approximation to a model like M_0: $|\theta| < \varepsilon$, for some small value ε, and the only advantage of using such an approximation would seem not having to bother about how to choose ε. However, in return, one should assess how well the point null model actually approximates the hypothesis the researcher does wish to test.

While we do not object against using Bayes factors per se, we think that what we basically need for inference could also be based on the estimated full posterior distribution (see Williams, Bååth, & Philipp, 2017, for a very insightful discussion). We already indicated that we need posterior distributions for assessing probabilistic information on the effect size. Having the full posterior distribution can suffice for model comparisons too. For instance, comparing the model M_0: $|\theta| < \varepsilon$ stating that the parameter is close to zero, with its complement, M_1: $|\theta| > \varepsilon$, can be done by simply directly computing the posterior odds ratio for these two models on the basis of the full posterior distribution. In similar ways, one can compare M_a: $\theta < 0$ versus M_b: $\theta > 0$ or M_a: $\theta < \varepsilon$ versus M_b: $\theta > \varepsilon$.

Furthermore, we urge researchers to rethink the way they conceive their research questions and the corresponding data analysis. In particular, is testing really necessary for answering their research questions? We definitely do not endorse "mindless hypothesis testing" (Luce, 1988). Gigerenzer and Marewski (2015, p. 422) phrased it humorously: "One of us reviewed an article in which the number of subjects was reported as 57. The authors calculated that the 95% confidence interval was between 47.3 and 66.7 subjects. Every figure was scrutinized in the same way, resulting in three dozen statistical tests. The only numbers with no confidence intervals or *p* values attached were the page numbers." Our point is that what is called testing may have its place in inference (by means of NHBT), but it actually is just one way of describing one's belief with respect to the possible values of a parameter. Instead, we recommend estimation of the full posterior distribution, from which, for example, credible intervals for the unknown quantities of interest and posterior model probabilities can be derived (Kruschke, 2011; Kruschke & Liddell, 2018b; van der Linden & Chryst, 2017; Williams et al., 2017).

Final Conclusions

We want to emphasize that if a researcher wishes to do NHBT and compare models using Bayes factors, then (s)he should make sure that the compared models match well with the hypotheses under scrutiny, not just in terms of the range of parameter values specified, but also in terms of the specified or implied within-model prior density distributions. But to begin with, carefully consider whether NHBT answers your research questions while noting that estimation of the full posterior distribution offers a more complete picture. Finally, if NHBT is chosen, do use it properly and at least complement its results with an estimation of the full posterior.

References

Aczel, B., Palfi, B., & Szaszi, B. (2017). Estimating the evidential value of significant results in psychological science. *PLoS ONE, 12,* e0182651. http://dx.doi.org/10.1371/journal.pone.0182651

Bayarri, M. J., Berger, J. O., Forte, A., & Garcia-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics, 40,* 1550–1577. http://dx.doi.org/10.1214/12-AOS1013

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2,* 6–10. http://dx.doi.org/10.1038/s41562-017-0189-z

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science, 2,* 317–352. http://dx.doi.org/10.1214/ss/1177013238

Berger, J. O., & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In L. Parhasarathi (Ed.), *Model selection* (pp. 135–207). Beachwood, OH: Institute of Mathematical Statistics Lecture Notes - Monograph Series.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p* values and evidence. *Journal of the American Statistical Association, 82,* 112–122. http://dx.doi.org/10.2307/2289131

Bernardo, J. M. (2012). Integrated objective Bayesian estimation and hypothesis testing. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (Eds.), *Bayesian Statistics 9: Proceedings of the Ninth Valencia Meeting* (pp. 1–68). New York, NY: Oxford University Press.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach*. Berlin, Germany: Springer Science & Business Media.

Carlin, B., & Chib, S. (1995). Bayesian model choice via Markov-Chain Monte-Carlo Methods. *Journal of the Royal Statistical Society Series B. Methodological, 57,* 473–484. http://dx.doi.org/10.1111/j.2517-6161.1995.tb02042.x

Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association, 82,* 106–111. http://dx.doi.org/10.1080/01621459.1987.10478396

Chen, M.-H., Shao, Q.-M., & Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4612-1276-8

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

Crane, H. (2017). *Why "redefining statistical significance" will not improve reproducibility and could make the replication crisis worse* (SSRN Scholarly Paper No. ID 3074083). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=3074083

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods, 22,* 240–261. http://dx.doi.org/10.1037/met0000065

Dickey, J. M. (1977). Is the tail area useful as an approximate Bayes Factor? *Journal of the American Statistical Association, 72,* 138–142. http://dx.doi.org/10.1080/01621459.1977.10479922

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5,* 781. http://dx.doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72,* 78–89. http://dx.doi.org/10.1016/j.jmp.2015.10.003

Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin, 63,* 400–402. http://dx.doi.org/10.1037/h0021967

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

Etz, A. (2015). *Understanding Bayes: Evidence vs. conclusions*. Retrieved from https://alexanderetz.com/2015/11/01/evidence-vs-conclusions/

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science, 32,* 313–329. http://dx.doi.org/10.1214/16-STS599

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116,* 439–453. http://dx.doi.org/10.1037/a0015251

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science, 13,* 163–185. http://dx.doi.org/10.1214/ss/1028905934

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733–760.

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology, 25,* 165–173. http://dx.doi.org/10.2307/271064

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 392–409). Thousand Oaks, CA: SAGE Publications, Inc. http://dx.doi.org/10.4135/9781412986311.n21

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management, 41,* 421–440. http://dx.doi.org/10.1177/0149206314547522

Good, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 2* (pp. 249–269). New York, NY: Elsevier.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika, 82,* 711–732. http://dx.doi.org/10.1093/biomet/82.4.711

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., . . . Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology, 81,* 80–97. http://dx.doi.org/10.1016/j.jmp.2017.09.005

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What If There Were No Significance Tests?* Erlbaum Publishers.

Hinkley, D. V. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence [Comment]. *Journal of the American Statistical Association, 82,* 128–129. http://dx.doi.org/10.2307/2289134

Hoekstra, R., Monden, R., van Ravenzwaaij, D., & Wagenmakers, E.-J. (2018). Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. *PLoS ONE, 13,* e0195474. http://dx.doi.org/10.1371/journal.pone.0195474

Hubbard, R., & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18,* 69–88. http://dx.doi.org/10.1177/0959354307086923

Ioannidis, J. P. A. (2018). The proposal to lower p value thresholds to. 005. *Journal of the American Medical Association, 319,* 1429–1430. http://dx.doi.org/10.1001/jama.2018.1536

Iverson, G. J., Lee, M. D., Zhang, S., & Wagenmakers, E.-J. (2009). p(rep): An agony in five fits. *Journal of Mathematical Psychology, 53,* 195–202. http://dx.doi.org/10.1016/j.jmp.2008.09.004

JASP Team. (2018). JASP (Version 0.8.6)[Computer software]. Retrieved from https://jasp-stats.org/

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society, 31,* 203–222. http://dx.doi.org/10.1017/S030500410001330X

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, UK: The Clarendon Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Jeon, M., & De Boeck, P. (2017). Decision qualities of Bayes factor and p value-based hypothesis testing. *Psychological Methods, 22,* 340–360. http://dx.doi.org/10.1037/met0000140

Johnson, V. E., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society Series A, 72,* 143–170. http://dx.doi.org/10.1111/j.1467-9868.2009.00730.x

Kamary, K., Mengersen, K., Robert, C. P., & Rousseau, J. (2014). *Testing hypotheses via a mixture estimation model*. Retrieved from https://arxiv.org/abs/1412.2044

Kass, R. E. (1993). Bayes factors in practice. *The Statistician, 42,* 551–560. http://dx.doi.org/10.2307/2348679

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. http://dx.doi.org/10.1080/01621459.1995.10476572

Konijn, E. A., van de Schoot, R., Winter, S. D., & Ferguson, C. J. (2015). Possible solution to publication bias through Bayesian statistics, including proper null hypothesis testing. *Communication Methods and Measures, 9,* 280–302. http://dx.doi.org/10.1080/19312458.2015.1096332

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299–312. http://dx.doi.org/10.1177/1745691611406925

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142,* 573–603. http://dx.doi.org/10.1037/a0029146

Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review, 25,* 155–177. http://dx.doi.org/10.3758/s13423-017-1272-1

Kruschke, J. K., & Liddell, T. M. (2018b). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25,* 178–206. http://dx.doi.org/10.3758/s13423-016-1221-4

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., . . . Becker, R. B. (2018). Justify your alpha. *Nature Human Behaviour, 2,* 168–171. http://dx.doi.org/10.1038/s41562-018-0311-x

Lavine, M., & Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician, 53,* 119–122.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139087759

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44,* 187–192. http://dx.doi.org/10.1093/biomet/44.1-2.187

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics, 15,* 22–25. http://dx.doi.org/10.1111/j.1467-9639.1993.tb00252.x

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology, 52,* 362–375. http://dx.doi.org/10.1016/j.jmp.2008.03.002

Luce, R. D. (1988). The tools-to-theory hypothesis: Review of G. Gigerenzer and D. J. Murray, "Cognition as intuitive statistics." *Contemporary Psychology, 33,* 582–583. http://dx.doi.org/10.1037/030460

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72,* 19–32. http://dx.doi.org/10.1016/j.jmp.2015.06.004

Marden, J. I. (2000). Hypothesis testing: From *p* values to Bayes factors. *Journal of the American Statistical Association, 95,* 1316–1320. http://dx.doi.org/10.2307/2669779

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115. http://dx.doi.org/10.1086/288135

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72,* 6–18. http://dx.doi.org/10.1016/j.jmp.2015.11.001

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16,* 406–419. http://dx.doi.org/10.1037/a0024377

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204. http://dx.doi.org/10.1006/jmps.1999.1283

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4,* 79–95. http://dx.doi.org/10.3758/BF03210778

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716. http://dx.doi.org/10.1126/science.aac4716

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7,* 528–530. http://dx.doi.org/10.1177/1745691612465253

Pratt, J. W. (1965). Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society Series A, 27,* 169–203.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* 1995 (Vol. 25, pp. 111–163). Malden, MA: Blackwell Publishing.

Robert, C. P. (2016). The expected demise of the Bayes factor. *Journal of Mathematical Psychology, 72,* 33–37. http://dx.doi.org/10.1016/j.jmp.2015.08.002

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21,* 301–308. http://dx.doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, Part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review, 25,* 102–113. http://dx.doi.org/10.3758/s13423-017-1420-7

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

Savalei, V., & Dunn, E. (2015). Is the call to abandon *p*-values the red herring of the replicability crisis? *Frontiers in Psychology, 6,* 245. http://dx.doi.org/10.3389/fpsyg.2015.00245

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

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician, 56,* 196–201. http://dx.doi.org/10.1198/000313002137

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B, Methodological, 64,* 583–639. http://dx.doi.org/10.1111/1467-9868.00353

Stern, H. S. (2016). A test by any other name: *p* values, Bayes factors, and statistical inference. *Multivariate Behavioral Research, 51,* 23–29. http://dx.doi.org/10.1080/00273171.2015.1099032

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review, 110,* 526–535. http://dx.doi.org/10.1037/0033-295X.110.3.526

Vampaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54,* 491–498. http://dx.doi.org/10.1016/j.jmp.2010.07.003

van der Linden, S., & Chryst, B. (2017). No need for Bayes factors: A fully Bayesian evidence synthesis. *Frontiers in Applied Mathematics and Statistics, 3,* 1–3. http://dx.doi.org/10.3389/fams.2017.00012

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22,* 217–239. http://dx.doi.org/10.1037/met0000100

Vardeman, S. B. (1987). Testing a point null hypothesis: The irreconcilability of *p* values and evidence [Comment]. *Journal of the American Statistical Association, 82,* 130–131. http://dx.doi.org/10.2307/2289136

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779–804. http://dx.doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25,* 35–57. http://dx.doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Kievit, R., & Maas, H. (2011). *Yes, psychologists must change the way they analyze their data: Clarifications for Bem, Utts, and Johnson (2011)*. Retrieved from https://www.researchgate.net/publication/221705056_Yes_

psychologists_must_change_the_way_they_analyze_their_data_Clarifications_for_Bem_Utts_and_Johnson_2011

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6,* 291–298. http://dx.doi.org/10.1177/1745691611406923

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. http://dx.doi.org/10.1037/0003-066X.54.8.594

Williams, M. N., Bååth, R. A., & Philipp, M. C. (2017). Using Bayes factors to test hypotheses in developmental research. *Research in Human Development, 14,* 321–337. http://dx.doi.org/10.1080/15427609.2017.1370964

Withers, S. D. (2002). Quantitative methods: Bayesian inference, Bayesian thinking. *Progress in Human Geography, 26,* 553–566. http://dx.doi.org/10.1191/0309132502ph386pr

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 585–603). Valencia, Spain: University Press.

Received August 8, 2018
Revision received February 28, 2019
Accepted March 19, 2019 ■