Diagnosing the Misuse of the Bayes Factor in Applied Research

Jorge N. Tendeiro[1] , Henk A. L. Kiers[2] , Rink Hoekstra[2] , Tsz Keung Wong[3] , and Richard D. Morey[4]

[1]Education and Research Center for Artificial Intelligence and Data Innovation, Hiroshima University, Hiroshima, Japan; [2]Faculty of Behavioral and Social Sciences, University of Groningen, Groningen, Netherlands; [3]School of Social and Behavioral Sciences, Tilburg University, Tilburg, Netherlands; and [4]School of Psychology, Cardiff University, Cardiff, Wales

Abstract

Hypothesis testing is often used for inference in the social sciences. In particular, null hypothesis significance testing (NHST) and its p value have been ubiquitous in published research for decades. Much more recently, null hypothesis Bayesian testing (NHBT) and its Bayes factor have also started to become more commonplace in applied research. Following preliminary work by Wong and colleagues, we investigated how, and to what extent, researchers misapply the Bayes factor in applied psychological research by means of a literature study. Based on a final sample of 167 articles, our results indicate that, not unlike NHST and the p value, the use of NHBT and the Bayes factor also shows signs of misconceptions. We consider the root causes of the identified problems and provide suggestions to improve the current state of affairs. This article is aimed to assist researchers in drawing the best inferences possible while using NHBT and the Bayes factor in applied research.

Keywords

null hypothesis Bayesian testing, Bayes factor, questionable reporting and interpreting practice

The use of frequentist statistics to perform inference in applied research is riddled with difficulties. There is strong evidence suggesting that p values and confidence intervals are often misinterpreted in practice (Belia et al., 2005; Falk & Greenbaum, 1995; Haller & Kraus, 2002; Hoekstra et al., 2014; Oakes, 1986), and the numerous types of misinterpretations have been often reiterated (e.g., Goodman, 2008; Greenland et al., 2016). Suggestions to improve the current state of affairs are various. Researchers have advocated for improving statistical education (e.g., Guo & Ma, 2022; Lakens, 2021), strengthening the bounds for decision-ruling (Benjamin et al., 2017), "retiring" the categorical flavor inherent to statistical significance (Amrhein et al., 2019), or even banning null hypothesis significance testing (NHST) altogether (Trafimow & Marks, 2015). The apparent mismatch between what practitioners wish to accomplish and what frequentist inference permits may be at the core of the many problems that have been identified. In this respect, Bayesian statistics is being

advocated by some as a better alternative (Kruschke & Liddell, 2018; Wagenmakers, 2007).

The last 10 years have witnessed an increase in published materials aiming at promoting the Bayesian paradigm to researchers in the social sciences (Etz & Vandekerckhove, 2018; Świątkowski & Carrier, 2020; van de Schoot et al., 2014). But Bayesian statistics is still relatively unknown and novel among social scientists. Hence, it would not be surprising if researchers would be making interpretation mistakes when using some of the newly learned Bayesian inferential tools. In this article, we mostly focus on null hypothesis Bayesian testing (NHBT) and the Bayes factor, that is, the Bayesian counterparts to NHST and the p value, respectively. A

Corresponding Author:
Jorge N. Tendeiro, Education and Research Center for Artificial Intelligence and Data Innovation, Hiroshima University
Email: tendeiro@hiroshima-u.ac.jp

first study by Wong et al. (2022) suggested that there are indeed misunderstandings related to the practical use of Bayesian hypothesis testing and the Bayes factor.

This article has been written for applied social scientists for whom the Bayes factor is still a relatively new tool. The article has two main objectives. The first is to provide a full account of what a correct use of the Bayes factor entails. To this effect, we offer a commented reanalysis of a published result, carefully explaining how the Bayes factor can be adequately used to draw inferences. At the same time, we refer to some pitfalls that are important to avoid. We intend this part of the article to be used as a template of good practices for those wishing to use the Bayes factor in their work. The second objective of this article is to provide an overview of how the Bayes factor has been suboptimally handled by practitioners in published research. We offer an extension to the work of Wong and colleagues by covering a wider range of articles and assessment criteria. Furthermore, Wong et al. (2022) did not elaborate in detail on the main factors behind the identified problems. In this article, we offer an extended discussion that aims at going to the root of each problem. Specifically, we identified various reasons that may help understanding the occurrence of such idiosyncrasies. This discussion is of great value because we can aim at improving matters only after the source of the problems has been clearly identified. On the basis of the results of our discussions, we suggest possible future avenues for improvement.

The article is organized as follows. We start by offering a short introduction to the Bayes factor and how it can be used to test hypotheses (or perform model comparison in general). Next, we showcase the Bayes factor by analyzing data from a real example and discussing both good and less ideal approaches. We then summarize the main findings from the work from Wong and colleagues and present the details of the current study. After presenting the main results, we elaborate on the reasons that may help illuminate why these problems seem to occur more or less consistently. The article ends with a short summary of the previous discussion and with some constructive suggestions for the future.

The Concept of the Bayes Factor

The Bayes factor offers a means of comparing the predictive ability of two models (say, $\mathcal{M}_0$ and $\mathcal{M}_1$). These models encompass two competing explanations for the real-world phenomenon under study. The "best" of the two models is the one that better predicts the data that were observed. In mathematical terms, the Bayes factor is the ratio of two *marginal likelihoods*,

$$BF_{10} = \frac{p(D \mid \mathcal{M}_1)}{p(D \mid \mathcal{M}_0)}, \qquad (1)$$

where D = observed data, and

$$p(D \mid \mathcal{M}_i) = \int_{\Theta_i} \underbrace{p(D \mid \theta_i, \mathcal{M}_i)}_{\text{likelihood}} \underbrace{p(\theta_i \mid \mathcal{M}_i)}_{\text{prior}} d\theta_i \qquad (2)$$

for $i = 0, 1$. In words, $p(D \mid \mathcal{M}_i)$ is the probability (or probability density for continuous data) of the observed data under $\mathcal{M}_i$. This probability is actually a *weighted average* of $p(D \mid \theta_i, \mathcal{M}_i)$, which is the likelihood of the observed data under $\mathcal{M}_i$ at a particular value of the model parameter θ_i.[1] The set of all possible values of θ_i is denoted by Θ_i. The weights of the weighted average are provided by $p(\theta_i \mid \mathcal{M}_i)$, the prior probability associated with θ_i. The prior probability of θ_i is typically chosen before looking at the observed data. The idea is then that the marginal likelihood, $p(D \mid \mathcal{M}_i)$, is a *value* based on the probability of the observed data at various parameter values and the prior probabilities of each such parameter value.

The Bayes factor in Equation 1 offers a relative assessment of the probability of the observed data under the two competing models. For example, $BF_{10} = 5$ means that the observed data are 5 times as likely in case $\mathcal{M}_1$ were true than if $\mathcal{M}_0$ were true. Conversely, $BF_{10} = 0.2$, which can be rewritten as $BF_{01} = \frac{1}{0.2} = 5$ (notice the updated subscript), means that the observed data are 5 times as likely in case $\mathcal{M}_0$ were true than if $\mathcal{M}_1$ were true.

An alternative means of portraying the Bayes factor is based on assuming that $\mathcal{M}_0$ and $\mathcal{M}_1$ are the only possible models of interest. In this sense, we act as if these are the only two models that could have generated the data that were observed. This is of course rather limited and contrived, which may incidentally be a source of confusion for users of the Bayes factor, as we discuss later. Thus, conditional on either $\mathcal{M}_0$ or $\mathcal{M}_1$ being true, both before and after observing the data, the probabilities of the two models are complementary, that is, they sum to one:

$$p(\mathcal{M}_1) = 1 - p(\mathcal{M}_0) \text{ and } p(\mathcal{M}_1 \mid D) = 1 - p(\mathcal{M}_0 \mid D). \quad (3)$$

Ratios of complementary probabilities are known as *odds*. In the current context we have two odds: the *prior* odds of $\mathcal{M}_1$ against $\mathcal{M}_0$, $\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}$, and the *posterior* odds of $\mathcal{M}_1$ against $\mathcal{M}_0$, $\frac{p(\mathcal{M}_1 \mid D)}{p(\mathcal{M}_0 \mid D)}$. Odds offer a different means of thinking about relative probabilities. For example, prior odds of $\mathcal{M}_1$ against $\mathcal{M}_0$ equal to 4 to 1 means that $\mathcal{M}_1$ is 4 times as likely as $\mathcal{M}_0$, which implies that

$$p(\mathcal{M}_0) = \frac{1}{1+4} = .20 \text{ and } p(\mathcal{M}_1) = \frac{4}{1+4} = .80 = 4 \times p(\mathcal{M}_0).$$

It is easy to show that the following equation holds:

$$\underbrace{\frac{p(\mathcal{M}_1 \mid D)}{p(\mathcal{M}_0 \mid D)}}_{\text{posterior odds}} = BF_{10} \times \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}}_{\text{prior odds}}. \qquad (4)$$

Written this way, BF_{10} quantifies the change in the relative likelihood of either model from before to after observing the data D. The prior odds represent our relative belief in either model before looking at the data. If $BF_{10} = 5$ then, *regardless of the prior odds*, one's initial relative belief should be revised by a factor of 5 to 1 in favor of $\mathcal{M}_1$ over $\mathcal{M}_0$. The relative revised belief is given by the posterior odds.

By rewriting the posterior odds as $\dfrac{1 - p(\mathcal{M}_0 \mid D)}{p(\mathcal{M}_0 \mid D)}$ we can derive expressions for both posterior-model probabilities from Equation 4:

$$p(\mathcal{M}_1 \mid D) = \frac{BF_{10} \times \dfrac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}}{1 + BF_{10} \times \dfrac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}}, \; p(\mathcal{M}_0 \mid D) = 1 - p(\mathcal{M}_1 \mid D).$$

(5)

From Equation 4 we see that the Bayes factor is equal to the ratio of the posterior odds to the prior odds. The Bayes factor is therefore a ratio of two odds, or an *odds ratio*. Regarded this way, $BF_{10} = 5$ means that the posterior odds of $\mathcal{M}_1$ over $\mathcal{M}_0$ are 5 times as large as the prior odds of $\mathcal{M}_1$ over $\mathcal{M}_0$. Following the same example as above, if the prior odds of $\mathcal{M}_1$ against $\mathcal{M}_0$ equal 4, and thus $p(\mathcal{M}_0) = .20$ and $p(\mathcal{M}_1) = .80$, then the posterior odds equal $5 \times 4 = 20$, which implies that

$$p(\mathcal{M}_1 \mid D) = \frac{20}{1 + 20} = .952 \text{ and } p(\mathcal{M}_0 \mid D) = 1 - .952 = .048$$

by Equation 5. The observed data thus allowed us to reinforce our belief in model $\mathcal{M}_1$ (its probability increased from .80 to .952), whereas model $\mathcal{M}_0$ loses some credibility (its probability decreased from .20 to .048). In sum, the Bayes factor indicates how a rational agent should reallocate probability among two competing models by taking into account the information in the observed data, provided that one indicates what the prior probabilities of the models are.

The Bayes factor offers a rather general framework for model comparison. In the Bayesian framework, a "model" consists of two elements: a likelihood function (seen as a function of the data given one or more model parameters) and a set of prior distributions for the model parameters. A likelihood and a prior together yield a predictive distribution for the data. Using this predictive distribution, any two such models may be compared via the Bayes factor. In the social sciences, however, the Bayes factor is primarily used via NHBT (Tendeiro & Kiers, 2019). One of the models, the null model, stipulates that the model parameters of interest are equal to a constant (e.g., a true mean is exactly 0), or that several parameters are equal to one another (e.g., all true means are the same). Such hypotheses operationalize the concept of an "absence" of an effect or "invariance" of parameters (Rouder et al., 2009). An alternative model, then, is *one possible* operationalization of "existence" or "variance." Null hypothesis testing is common in social-sciences research, and in fact it is in this particular setting that most introductions to Bayesian-model comparison and the Bayes factor are portrayed.

A Worked-Out Example

Haeffel et al. (2023) conducted a series of studies to learn about cognitive vulnerability to depression (original data available at https://osf.io/umg9p). Their research focused on five different groups (Honduran young adults, Nepali adults, Western adults, Black U.S. adults, and U.S. undergraduates). Cognitive vulnerability was measured by means of the Cognitive Style Questionnaire (CSQ; Haeffel et al., 2008). We performed a reanalysis[2] of two-tailed independent-samples t-tests reported in the article, which compares the CSQ scores between U.S. undergraduates (USugrad group; $n = 110$, $\bar{x} = 4.25$, $SD = .84$) and Western adults (Western group; $n = 104$, $\bar{x} = 4.12$, $sd = .92$). In what follows, we highlight both correct and also incorrect (or at least not ideal) takes on the Bayes factor. These *questionable reporting or interpreting practices* (QRIPs) are the primary focus of our main study and are introduced after this section. Readers should refer to Table 2, in which we define the various QRIPs that we analyzed in our study. We identify various QRIPs already in this worked example.

The test's null hypothesis $\mathcal{H}_0$ stipulates that there is no difference in mean CSQ score between the two groups in the population. The alternative hypothesis $\mathcal{H}_1$ indicates that there is a difference, either positive or negative. The result of the classical t-test is as follows: $t = 1.11$, $df = 207.6$, $p = .27$.[3] By most levels of significance in use in the social sciences, we would "fail to reject" $\mathcal{H}_0$. We next carry out a Bayesian t-test for the same group comparison. In the materials we shared at the OSF we show in detail how this can be done either in R by means of the BayesFactor package (Morey & Rouder, 2021) or in JASP, an easy-to-use graphical user interface reminiscent of SPSS (JASP Team, 2023). Here we explain the most important steps that need to be considered to optimally perform a Bayesian test. Readers may want to refer to our suggested checklist on how to perform a Bayesian hypothesis test (see Appendix). We also elaborate on a few ideas that are important to keep in mind while interpreting Bayes factors in practice.

Null hypothesis, alternative hypothesis, and prior assumptions

We assume that the CSQ scores are normally distributed in either group, with potentially different mean parameters (USugrad group: μ_U; Western group: μ_W) and with a common standard deviation σ. The null hypothesis for the Bayesian test is the same as for the classical t-test; it stipulates that there is no difference in mean CSQ

scores between the Western group and the USugrad group in the population, and hence the *t*-test statistic will have a Student's *t* distribution. Defining $\mu = \mu_W - \mu_U$, the null hypothesis is then that $\mu = 0$. The alternative hypothesis, on the other hand, is not exactly the same as that from the classical *t*-test. In Bayesian hypothesis testing, it does not suffice to specify the possible values of the parameter being tested (i.e., $\mu \neq 0$). One must also choose a prior distribution. In simple terms, this prior distribution assigns probability to each possible value of the parameter. Importantly, the prior should not be informed by the data. That is, the prior should reflect information that is independent from the observed data. The prior distribution may come about in various ways, for instance, to reflect current knowledge, differing scientific perspectives (e.g., skeptical, liberal, or mainstream), or known constraints of the parameter (e.g., priors for variances should be truncated below 0). Often, default priors have been set in place in commonly available software. Such priors, although not incorrect on their own, rely on mathematical idealized desiderata and may lack an empirical foundation.

For the Bayesian independent-samples *t*-test, the BayesFactor R package offers the Cauchy prior for the standardized grouped difference given by $\delta = \mu / \sigma$. The Cauchy distribution is the *t* distribution with one degree of freedom, and it resembles the normal distribution but with heavier tails. By adjusting the Cauchy distribution's scale parameter *r* we can determine how concentrated around 0 the prior should be. Parameter *r* implies that, a priori, there is 50% probability that the true standardized difference between the two groups means is at most *r* in magnitude. By default, $r = \sqrt{2} / 2$—about 0.707—but this value can be changed at will. Priors with non-zero means can easily be used as well. But other priors (say, asymmetric), although possible to use in theory, may require some changes to the analysis parameters or extra programming to implement them in practice.

We, the authors, lack a deep insight on the topic of cognitive vulnerability to depression. It is therefore difficult to choose a prior that is well-informed. Experts may be able to argue that standardized differences larger than 0.1, or perhaps 0.3 or 0.5, are quite unlikely. Such information could be used to specify a prior. In our case, we settle by using the default scale value of 0.707, but we also run a sensitivity analysis. This means that we consider the test result at various competing values of the scale parameter. Furthermore, priors with different values of the location parameter can also be explored. Do observe that priors symmetric around 0 allocate equal prior credence to symmetric values around 0. This may not be reasonable or properly reflect the current state of affairs (e.g., whether it is sensible that both $d = 0.5$ and $d = -0.5$ are a priori equally likely). In such cases it may be best to entertain varying prior location

values while also studying how sensitive the Bayes factor is to such variations. All in all, sensitivity analyses help us determine whether the test result is not too dependent on the chosen prior. Strong prior dependence should be acknowledged, and one needs to exert caution in drawing conclusions from the results. Sensitivity analyses are nearly always a good thing to try, even when we have given very careful thought into choosing our priors.[4] Concerning parameter σ, we note that this is a so-called *nuisance* parameter: It occurs in both models being compared, and it is not the main focus of the test. In such cases, its prior is assumed not to be very influential (Rouder et al., 2009). Following the default implementation in the BayesFactor R package, we assume Jeffreys's improper prior on the variance: $p(\sigma^2) = 1 / \sigma^2$. As a general rule, we note that it is good practice to always report which priors were used (see QRIPS 3b and 3c) and whenever possible to also provide a justification for the choice made (see QRIP 3a).

Interpretation

The result of the test—the Bayes factor—is $BF_{10} = 0.27$ or, equivalently, $BF_{01} = \dfrac{1}{BF_{10}} = 3.7$. This can be interpreted as follows (recall Equation 1): The observed data are 3.7 more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$. Alternatively, and recalling Equation 4, we can also conclude that the observed data tell us that we should revise our relative initial belief by a factor of 3.7 to 1 in favor of $\mathcal{H}_0$. Thus, someone with no prior preference for either hypothesis (i.e., prior odds = 1) should now believe that the null model is 3.7 times more probable than this alternative model (i.e., posterior odds = $3.7 \times 1 = 3.7$). In terms of posterior-model probabilities (Equation 5), this implies that $p(\mathcal{H}_1 | D) = \dfrac{0.27 \times 1}{1 + 0.27 \times 1} = .21$ and $p(\mathcal{H}_0 | D) = 1 - p(\mathcal{H}_1) = .79$. Another person, say someone who truly believed originally that $\mathcal{H}_1$ has probability .80 (and therefore $p(\mathcal{H}_0) = .20$ and prior odds = $\dfrac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)} = \dfrac{.8}{.2} = 4$), must now revise their beliefs and conclude that the data are $0.27 \times 4 = 1.1$ times more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$. In terms of posterior-model probabilities we have that $p(\mathcal{H}_1 | D) = \dfrac{0.27 \times 4}{1 + 0.27 \times 4} = .52$ and $p(\mathcal{H}_0 | D) = .48$. After having observed the data, this person is now more uncertain about the relative merit of either hypothesis than initially. It is important to reiterate the fact that the interpretation of the Bayes factor and probabilities reported above is contingent on the two specific chosen hypotheses only. Only in this sense can the prior and posterior-model probabilities be complementary to each other (Equation 3).
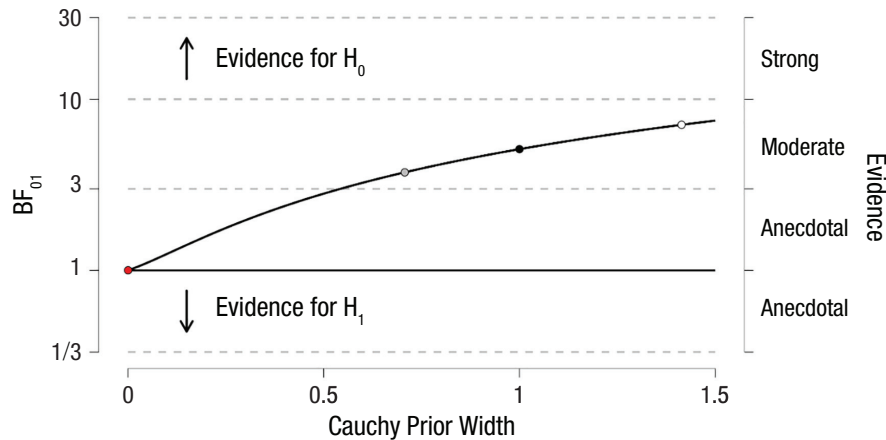
Fig. 1. Analysis of sensitivity to prior width for the independent *t*-test Bayes factor contrasting the mean Cognitive Style Questionnaire scores between the U.S. undergraduate and Western groups. The *x*-axis is the value of the scale parameter of the Cauchy prior for the standardized difference between the two group means under the alternative hypothesis. For the Cauchy-prior scale values between 0 and 1.5, the Bayes factor ranges between 1 and about 7 in favor of the null hypothesis against an alternative hypothesis with this particular prior distribution. This level of evidence brought about by the data is labeled as at most moderate on the basis of the classification of Lee and Wagenmakers (2014).

Supporting the null hypothesis

The running example is one interesting case of the Bayes factor providing relative support in favor of the null hypothesis compared with the particular alternative hypothesis used. It is well known that classic frequentist procedures do not allow supporting the null (although equivalence tests do exist; Wellek, 2003). From the frequentist *t*-test result and at customary significance levels, we may claim only that there was not enough evidence allowing us to reject the null hypothesis. Faced with such an outcome, it is likely that researchers may expect to find some added value in the Bayes factor.

Bayes factor versus posterior odds

Observe that the Bayes factor is a statement about the relative probability of the *data* under the two competing hypotheses or models (Equation 1). The posterior odds, on the other hand, do offer a relative assessment of the probability of the *hypotheses* after observing the data. The Bayes factor and the posterior odds are different, and this is important to recognize (see QRIPs 1 and 6). It is commonly observed in the literature that the Bayes factor gauges the predictive ability of both models under comparison. This observation may tempt unwary practitioners to "forget" about the data and project their reasoning on the models only. To be clear: If the interest is in looking at the relative likelihood of both hypotheses after observing the data, then one needs to look at the posterior odds instead of the Bayes factor. Specifically, and for the running example, it is incorrect to state that "$\mathcal{H}_0$ is 3.7 more likely than $\mathcal{H}_1$ after observing the data"

under all prior odds except unity. In such cases and to avoid ambiguity, practitioners are advised to explicitly state that their prior odds equal 1, so the Bayes factor and the posterior odds are equal to each other.

Relative evidence, priors, and labels

Hypothesis testing, or model comparison more generally, is an inherently relative endeavor. The merits of any one hypothesis are dependent on what other hypothesis we choose for the test. This is true regardless of the inferential paradigm of choice (frequentist or Bayesian), but it is perhaps more exacerbated in Bayesian testing because of the role played by prior distributions. Avoiding making absolute statements favoring one hypothesis (while disregarding its testing counterpart) is better avoided (see QRIP 4). Furthermore, sensitivity analyses showing the sensitivity of the Bayes factor to varying priors are important. Figure 1 shows how the Bayes factor for our test varies as a function of the scale of the Cauchy prior under the alternative hypothesis. It can be seen that there is relative evidence in favor of $\mathcal{H}_0$ for varying Cauchy priors under $\mathcal{H}_1$, with the value of BF_{01} ranging between about 3 and about 7 for a range of scale-parameter values between 0.5 and 1.5. We can conclude that the relative evidence in favor of the null hypothesis is at most moderate for a broad range of prior distributions under the alternative model (we did not explore the sensitivity to priors under varying location parameters here but, of course, could have). Qualitative labels such as "anecdotal" or "moderate" are alien to the Bayes factor (Kass & Raftery, 1995; Lee & Wagenmakers, 2014) and have been introduced merely to assist

researchers in their interpretations. It is best at all times to report the numerical value of the Bayes factor because it conveys a more complete picture of the amount of evidence encapsulated in the Bayes factor (see QRIP 10). Furthermore, practitioners are advised not to capitalize too much on the specific wording of the labels. Evidence labeled "moderate," for example, may (and should) be perceived differently among different researchers or even among different research fields. If labels are to be used, it is best to describe what they represent by taking the entire context into account (e.g., research field, specifics of the experiment, research design, models being compared).

Bayes factor and effect size

The Bayes factor is not a valid measure of the effect size (see QRIP 7). For example, for the Bayesian *t*-test above, increasing the sample size with no bound will lead to an increase of BF_{10} also with no bound, provided that the true difference between both groups is different from zero. In this sense, a Bayes factor of $BF_{01} = 3.7$ does not necessarily reflect a smaller standardized group difference than the value $BF_{01} = 20$ would. For such queries, one needs to rely on valid effect-size indices (e.g., d, r, ω^2). Our advice is for applied researchers to report a measure of the size of the effect being tested together with either confidence or credible intervals and also the Bayes factor as a measure of the evidence in the data. For the running example, we found Cohen's $d = 0.15$, an effect that may be considered of small magnitude—95% confidence interval = $[-0.12, 0.42]$.

Presence versus absence

Simplistic phrasing of research hypotheses such as "$\mathcal{H}_0$: There is no difference" and "$\mathcal{H}_1$: There is a difference" can arguably lead researchers to use the Bayes factor with the goal of establishing either the absence ($\mathcal{H}_0$) or presence ($\mathcal{H}_1$) of an effect. Logically speaking, the Bayes factor in isolation cannot establish either theory (see QRIP 5). The Bayes factor may be used to gather relative evidence in the data supporting either hypothesis. Hypotheses receiving strong support, especially after a sequence of multiple well-established experiments, will naturally lead researchers to update their theories. But Bayes factors are just a stochastic expression of our knowledge and should not be used as if they were a proof of a theorem. For the running example, the relative evidence weakly favors the null hypothesis compared with the particular alternative hypothesis used. Depending on our prior odds, our relative belief between both hypotheses is now slightly shifted toward the null hypothesis of an absence of an effect. However, as argued before, using

other alternative hypotheses may change the outcome quite drastically. More evidence and more experiments are probably needed before the scientific community can reach a consensus. The important point we want to make is that such a decision lies beyond the conceptual boundaries of the Bayes factor itself and requires extraneous information and agreed-on norms (akin to the particle-physics community accepting the existence of the Higgs boson on the basis of the five-sigma rule).

Inconclusive evidence

Bayes factors of (about) 1 imply that the observed data are equally likely under either hypothesis under comparison (Equation 1). In other words, there is lack of evidence either way. This should not be confused with evidence of absence, that is, that it is likely that there is no effect (see QRIP 9). A simple analogy is that of a nonsignificant frequentist test result. For the running example, BF_{01} approaches 1 as the Cauchy prior under $\mathcal{H}_0$ concentrates around 0 (see Fig. 1). At the convergence limit, null and alternative hypotheses coincide, and thus the data are perfectly uninformative.

The Bayes Factor in Applied Research

In the previous section we provided a detailed account of how to use the Bayes factor by means of an example. Although the origins of the Bayes factor go back about 100 years (Etz & Wagenmakers, 2017), the interest on its use in applied work only increased since the 1990s with the seminal article by Robert Kass and Adrian Raftery (Kass & Raftery, 1995). In addition, the availability of faster computers and dedicated software such as JASP (JASP Team, 2023) and BayesFactor (Morey & Rouder, 2021) facilitated a wider adoption of this tool in practice in the last, say, 10 years. It is therefore natural to question how well practitioners have been dealing with the Bayes factor in applied research. However, there is not a lot of literature on this topic. To the best of our knowledge, Wong et al. (2022) is the only article of the kind. Because the current article builds on Study 1 in Wong et al. (2022), here we present a brief summary of the main findings in Wong et al. (2022). We then present the details of our extension to Wong et al. (2022).

Wong et al. (2022)

Study 1 of Wong et al. (2022) is a small peer-reviewed literature study of 73 published applied articles. The study focused exclusively on how researchers used NHBT. Each article was inspected, and the occurence of any of the eight QRIPs was marked down. Table 1 identifies and provides a brief description of each QRIP,

Table 1. QRIPs for Null Hypothesis Bayesian Testing

QRIP (incidence)	Brief description
1—Describing the BF as posterior odds (4.1%)	Defining or elaborating on BFs as posterior odds ratios
2—Not specifying null and alternative hypotheses (24.7%)	Unclear which models are being tested by the BF
3—Incomplete reporting of prior distributions (69.9%)	Omitting the prior distribution from the alternative hypothesis
4—Not referring to the comparison of models (60.3%)	Presenting BFs as absolute evidence for one of the two models
5—Making absolute statements (56.2%)	Concluding that there is (not) an effect on the basis of the BF
6—Using the BF as posterior odds (17.8%)	Interpreting BFs as ratios of posterior model probabilities
7—Considering the BF as effect size (12.3%)	Associating the size of the BF to the size of the effect
8—Mismatch between statistical and research hypotheses (2.7%)	BF applied to incorrect operationalizations of research hypotheses

Note: Retrieved from Wong et al. (2022). QRIP = questionable reporting or interpreting practice. BF = Bayes factor.

together with the corresponding incidence in the sampled articles. As can be seen, the three most common QRIPs were the third (incomplete reporting of prior distributions), fourth (not referring to the comparison of models), and fifth (making absolute statements). Wong et al. (2022) also recorded other occurences of QRIPs beyond those mentioned in Table 1. One in particular was found often (21.9%): Bayes factors close to 1—which should imply that the models under comparison were relatively equally predictive of the observed data—were instead interpreted as supporting the null model of absence.

The current study

This article used the setup and findings from Wong et al. (2022) as a template, and both conceptually replicated and significantly extended their study design. We describe the details of our study in the Method section. Here we just list the main additions of our study to that by Wong et al. (2022):

1. *Extended literature search*. We performed a larger search for applications of the Bayes factor in the social-sciences literature. After filtering we were left with a set of 167 articles, which more than doubled the original study. All articles from the original study are also included in the new sample.
2. *Extended criteria*. We included new criteria for assessment (see Table 2). We also labeled some of these criteria QRIPs because they reflect inappropriate applications of the Bayes factor. While updating the list of criteria, we decided to discontinue QRIPs 2 and 8 by Wong et al. (2022), as we explain later. Furthermore, we included some criteria that are simply descriptive in nature and

partly reflect the Bayes factor usage intentions from the researchers.

3. *Abstracts*. In our study we distinguished between the abstract and the rest of the article. The reason is that an abstract, by nature, is more condensed than the body of the article. This fact may have consequences in how results including the Bayes factor may be reported. In our study we first present the results excluding all abstracts. Results specifically from the abstracts are reported separately.
4. *Extended discussion*. Importantly, we included an extended discussion of our results. Our goal is to go beyond reporting the results and to try to understand the rationale supporting our findings. Simply put: Why do these inconsistencies come about as often as they do? This discussion is of value if one wants to take the next step forward, which is to propose measures aiming at curtailing the prevalence of these problems.
5. *Recommendations*. Based on our findings, we offer concrete suggestions for improvement. Among all our suggestions, we highlight the checklist that we developed (see Appendix). This checklist aims at aiding both authors as well as journal reviewers and editors in using the Bayes factor in practice.

Method

Article selection

The first author (J. N. Tendeiro) performed the article selection. An advanced search for research articles was conducted on Google Scholar on December 22, 2021, using the key (“bayes factor” AND “bayesian test” AND psychol) from 2010 on. This led to 508 hits. From these hits, 399 were dropped. The dropped

hits were either repetitions (e.g., preprints of also selected final published articles) or false hits (e.g., theses, non-English-language sources, no PDF available, book/book chapters, no empirical applications of the Bayes factor, or not a research article). Articles that used only the Bayes factor through the Bayesian information criterion (Raftery, 1995; Schwarz, 1978; Wagenmakers, 2007) were also removed because not all the required criteria could be inspected. Thus, $508 - 399 = 109$ articles from this search entered our study.

We further complemented our sample with the result from an advanced search on the Web of Science on November 29, 2021, using the following key: `(TI=((bayes factor OR bayes* selection OR bayes* test*) AND psychol*) OR AB=((bayes factor OR bayes* selection OR bayes* test* OR bf*) AND psychol*) OR AK=((bayes factor OR bayes* selection OR bayes* test* OR bf*) AND psychol*)) AND PY=(2010-2022)`. This led to 730 hits. Of these, 27 overlapped with the Google search, so 703 unique hits remained. However, only 58 new articles survived the removal of repetitions or false hits. The main problem was the inclusion of the "bf" acronym, which led to selecting a wide range of false positives (e.g., body fat, big five). In summary, our study includes $109(\text{GS}) + 58(\text{WoS}) = 167$ articles.

Article grading

We independently graded 10 articles randomly selected from the sample of 167 articles. Each of us graded the same 10 articles. The purpose of this pilot study was to calibrate the grading procedure to be used in the entire sample. Prior to the pilot study, we decided to cover the eight criteria listed in Table 1 plus two more:

- #9: When faced with an inconclusive Bayes factor (e.g., $1/3 < \text{BF} < 3$), conclude that there is no effect.
  Explanation: A Bayes factor around 1 implies that both models under comparison are equally predictive for the data observed. Concluding that there is no effect amounts to claiming support for the null hypothesis in NHBT. This is a clear mistake.

- #10: Interpret the Bayes factor simply using cutoffs (e.g., 1–3, 3–10).
  Explanation: The Bayes factor encapsulates the evidence in the data (see Morey et al., 2016). Evidence through the Bayes factor is best interpreted as a ratio-scaled value on the continuum between 0 and infinity. Discretizing the Bayes factor value implies losing valuable information, and the discrete values are merely arbitrary choices of the analyst

(rather than the reader). Therefore, we judge it as suboptimal when researchers report or interpret results on the basis only of a set of discrete labels of evidence (as provided, e.g., in Jeffreys, 1961).

We discussed the results in a group meeting. The ratings among the five of us were largely in agreement. We focused on aspects for which some disagreement existed, as well as on things to adapt to make the assessment more streamlined. As a result, we decided on the following grading plan for all articles:

- Exclude the second (not specifying null and alternative hypotheses) and the eighth (mismatch between statistical and research hypotheses) criteria. The main argument in favor of the exclusion is that these criteria are not necessarily related to the Bayes factor per se (i.e., they could also be observed in articles resorting to NHST).
- The third criterion (incomplete reporting of prior distributions) was replaced by three more narrowed criteria:
  - #3a: The reason or justification for the chosen priors is not provided.

    Explanation: Ideally, the choice of the prior distributions taking part in a Bayesian model should be carefully justified. However, prior elicitation is a notoriously difficult endeavor (e.g., Falconer et al., 2022). Some authors seem to avoid this issue altogether and provide no explanation for the priors used in their analyses.
  - #3b: It is unclear which priors were used under either model.

    Explanation: Not providing a justification for the priors used is not the same as not declaring which priors were used; we think that, at a minimum, priors should be reported for the sake of reproducibility of the analyses.
  - #3c: The information on priors is incomplete (e.g., only the distribution family, but not the specific distribution used, is provided).

    Explanation: For example, stating to have used a Cauchy prior but omitting the corresponding scale parameter is not good practice. In such cases, the reader will need trial and error to disclose the missing information. We think this should be avoided.

- To attempt a thorough characterization of the practical use of the Bayes factor in applied research, we further included three extra criteria. These are descriptive in nature and do not necessarily reflect misuses of the Bayes factor. Instead, they are aimed at providing a more fine-grained

Table 2. Criteria Used

Criterion	Brief description
QRIP	
1—Describing the BF as posterior odds	Defining or elaborating on BFs as posterior odds ratios
3a—Missing explanation for the chosen priors	Reason or justification for the chosen priors not provided
3b—No mention of the priors used	Unclear which priors were used under either model
3c—Incomplete info regarding the priors used	Providing, e.g., only the distribution family ("Cauchy")
4—Not referring to the comparison of models	Presenting BFs as absolute evidence for one of the two models
5—Making absolute statements	Concluding that there is (not) an effect on the basis of the BF
6—Using the BF as posterior odds	Interpreting BFs as ratios of posterior model probabilities
7—Considering the BF as effect size	Associating the size of the BF to the size of the effect
9—Inconclusive evidence as evidence of absence	Stating that there is no effect when faced with inconclusive evidence (e.g., $1/3 < BF < 3$)
10—Interpreting ranges of BF values only	Interpreting the BF simply using cutoffs (e.g., 1–3, 3–10)
Usage	
A—Default prior	Justifying using a prior because it is "the" default
B—Null results	BFs as a follow-up to nonsignificant outcomes from NHST
C—Presence vs. absence	BFs to distinguish between the presence and absence of an effect

Note: QRIP = questionable reporting or interpreting practice; BF = Bayes factor; NHST = null hypothesis significance testing.

characterization about how and why the Bayes factor was used. Thus, we do not refer to them as QRIPs:

○ A: Justifying using a prior because it is "the" default

 Explanation: In practice, resorting to default priors (e.g., as suggested by available software) is commonplace. We wanted to learn how often this was done in practice.

○ B: Arguing to use the Bayes factor to be able to draw support for null findings from NHST

 Explanation: Some researchers seem to resort to the Bayes factor only after classical testing led to failure to reject the null hypothesis. Such researchers are then attracted to the ability of the Bayes factor to provide relative support in favor of the null hypothesis. We tallied the number of times this behavior was found in our sample.

○ C: Arguing that the Bayes factor allows distinguishing between the presence and the absence of an effect

 Explanation: It may be argued that a null model such as $\mathcal{M}_0 : \theta = 0$ completely captures the notion of a total absence of the effect operationalized by parameter θ. Perhaps surprisingly, however, operationalizing the complementary notion of the existence of an effect is not as straightforward. The problem is that, in Bayesian inference, simply stating the parameter support (such as $\mathcal{M}_1 : \theta \neq 0$) is insufficient; we must also supply a corresponding prior distribution for the parameter at hand. Because different choices of priors entail

different Bayes factor values, we must realize that one particular choice of a prior will lead to nothing more than one operationalization of what the researchers trust to represent the existence of an effect.

To make things further complex, it is also important to realize that the Bayes factor typically does not permit a strict separation between any two models under comparison. Bayesian model comparison proceeds by the accumulation of evidence either way; it does not logically function as proving a mathematical theorem does. Thus, authors claiming to use the Bayes factor to "establish," or "distinguish," between the existence or absence of an effect may be surprised to learn that their desideratum is quite difficult to achieve. In our study, we identified articles that explicitly claimed to have used the Bayes factor with this particular motivation in mind.

Table 2 lists the criteria used to classify the sampled articles. We kept, and extended, the original numeration from Wong et al. (2022) for consistency.

The above inspection was conducted by reading through all sections in the papers except for the abstract. The abstract is a rather condensed text for which we speculated that some types of reporting problems are more prone. After conducting the study, we decided to go through all the abstracts and flag all criteria separately from the rest of the articles. We report these results in a separate section.

All supporting files that complement this article can be found at https://osf.io/57ew4.

Table 3. Number of Articles Displaying the Corresponding Criterion

Criterion	Number of articles (%)
QRIP	
#1	22 (13.2)
#3a	18 (10.8)
#3b	50 (29.9)
#3c	10 (6.0)
#4	104 (62.3)
#5	59 (35.3)
#6	34 (20.4)
#7	7 (4.2)
#9	6 (3.6)
#10	9 (5.4)
Usage	
A	59 (35.3)
B	27 (16.2)
C	30 (18.0)

Note: QRIP = questionable reporting or interpreting practice.

Results

The frequencies and percentages associated to each evaluated criterion are given in Table 3. As can be seen, only four of the 10 QRIPs (3c, 7, 9, and 10) were relatively rare (less than 10% of the articles). Overall, 149 articles (89.2%) displayed at least one QRIP, and 104 articles (62.3%) displayed at least two QRIPs.

Table 4 shows the occurrence of pairs of criteria. Furthermore, the supplementary material available at the OSF further includes more tabulations for these data that help to better clarify the results. We refer to results from these tables in the discussion that follows to better characterize each identified problem.

Discussion of the Results

In what follows, we revisit each criterion that we included in our study. We list arguments that may help clarify why the observed issues are occurring as frequently as found in our study. This is the result of a joint discussion between the authors over these matters.

QRIPs 1 and 6

QRIP 1 concerns defining the Bayes factor as if it were a posterior odds. Equation 4 shows that the Bayes factor equates only to the posterior odds in the special case in which the prior odds is equal to 1. In other words, only when both models under comparison are a priori equally likely can the Bayes factor be interpreted as posterior model odds. However, in 13.2% of the articles we found that Bayes factors are simply introduced as if they were posterior odds, without having explicitly stated that prior odds equal to 1 were assumed. For example:

> These Bayes Factors can be readily interpreted as a ratio of evidence in favour of the experimental effect compared to the null effect. For example, a BF_{10} of 3 would represent that the experimental effect is three times more likely than the null, given the data. (P_9^5)

Or take this example: "For instance, a $BF_{10} = 10$ means that the H_1 is ten times more likely to be true than the H_0" (P_{130}). Relatedly, QRIP 6 concerns confusing the Bayes factor with the posterior odds when interpreting the results. This error was found relatively often—in 20.4% of the articles. Here is one example:

Table 4. Frequencies of the Occurrence of Pairs of Criteria

Criterion	#1	#3a	#3b	#3c	#4	#5	#6	#7	#9	#10	A	B	C
#1	—	1.2	3.0	1.2	7.8	4.2	5.4	0.6		0.6	5.4	3.6	3.6
#3a	2	—	0.6	1.8	6.6	1.2	1.2	0.6				3.0	1.8
#3b	5	1	—	0.6	21.0	14.4	6.0	3.6	1.8	4.2	0.6	1.8	4.2
#3c	2	3	1	—	3.6	0.6	1.2	0	0.6	0.6	0.6	3.6	
#4	13	11	35	6	—	21.6	12.6	2.4	1.2	5.4	22.2	12.0	14.4
#5	7	2	24	1	36	—	6.0	1.8	2.4	2.4	13.2	3.0	6.0
#6	9	2	10	2	21	10	—	1.2			7.8	4.8	3.0
#7	1	1	6		4	3	2	—					
#9			3	1	2	4			—		1.2	1.2	1.2
#10	1		7	1	9	4				—	0.6	0.6	0.6
A	9		1	1	37	22	13		2	1	—	6.6	7.2
B	6	5	3	6	20	5	8		2	1	11	—	2.4
C	6	3	7		24	10	5		2	1	12	4	—

Note: Dashes are used to indicate a diagonal line from the top left of the table to the bottom right. Counts are shown under this diagonal, percentages are shown above it, and missing entries are equal to 0.

Bayesian analyses . . . produced a JZS Bayes Factor of 3.74. According to Jeffreys (1961), this result indicates that there is some evidence for H0 over H1 (i.e., the hypothesis that gender is not associated with ODL scores is about three to four times more likely than the hypothesis that gender is associated with ODL scores, based on our sample's results). (P_{110})

Here is another example: "The alternative hypothesis is 2 times more likely than the null hypothesis (B_{+0} = 2.46; Bayesian 95 % CI [0.106, 0.896])" (P_{11}).

We discussed these findings and tried to explain them. We can summarize our main explanations in four points.

Lack of knowledge. It is entirely likely that practitioners still do not master the basics of the Bayes factor. This is a natural explanation that is also equally plausible to most of the coming QRIPs, and we do not repeat it further. The main argument is that Bayesian hypothesis testing is still relatively novel for most practitioners, and surely so compared with frequentist inference.

Principle of indifference. Some researchers may be implicitly assuming that prior odds equal 1, that is, that a priori both models under comparison are equally likely following the advice by Jeffreys.[6] If so, the problem may be perceived as one of lack of communication.

Bayesian versus classical approaches. Many introductory texts to Bayesian inference capitalize on the fact that the *p* value is based on the "wrong" conditional probability (of observed data—or more extreme—given a null hypothesis). Bayesian statistics, on the other hand, as the *theory of inverse probability* (Jeffreys, 1961), is touted as allowing the conditional and computing probabilities of hypotheses to be reversed given the observed data. This is at the essence of posterior probabilities and distributions and of the Bayesian credible interval. The above might create a false impression that *all* Bayesian statistical tools (including the Bayes factor) can be interpreted as the "inverse probability" of hypotheses given data. However, as shown in Equation 1, the Bayes factor is based on probabilities of the observed data conditional on the hypotheses. We suggest it is possible that this feature is not sufficiently well appreciated by practitioners.

Cognitive dissonance. It is possible that some researchers are aware of the issue. However, they also realize that they followed recommendations to use Bayes factors despite the fact that Bayes factors cannot be interpreted as posterior odds (as they actually wished). To alleviate this cognitive dissonance, they convince themselves that they are entitled to "somewhat extend" the realm of the Bayes factor to what Bayesian inference at large does.

QRIPs 3a, 3b, and 3c; Usage A

These four reporting styles concern how researchers deal with prior distributions when using Bayes factors. In almost one third of the articles nothing about priors was mentioned (QRIP3b; 29.9%). Incomplete available information regarding the priors used was not a frequently found issue (QRIP3c; 6%). It sometimes happened that the used priors were mentioned but no explanation was provided (QRIP3a; 10.8%), or the authors simply stated that they used the software's default priors (Usage A; 35.3%). In total, 130 articles (77.8%) displayed at least one of these reporting styles. Our arguments explaining this state of affairs are summarized as follows.

Too little space. Text space in most journals comes at a premium. Researchers are used to write succinctly whenever possible, saving space to highlight the main results from their studies. This fact may disadvantage a thorough presentation of the analytical details in the methods and results sections of articles. We found that, for articles reporting priors (i.e., not committing QRIP 3b), eight (6.8%) placed such information in supporting materials (supplements or appendices), although only one of these eight articles had a journal word limit. Furthermore, from articles reporting incomplete information regarding the priors used (QRIP 3c), three (30%) were published in journals with a strict word limit. Thus, at least to some extent, the pressure to write concisely may be conditioning the way explanations are provided. This argument may be a plausible explanation for QRIPs 3a, 3b, and 3c and to some extent to Usage A as well.

Habits inherited from NHST. Specifying alternative hypotheses and hypothesizing effect sizes of interest are essential to conducting power analyses in Neyman-Pearson-based NHST. Nevertheless, conducting power analyses is rare in practice. As a consequence, researchers pay relatively little attention to the alternative hypothesis already when conducting frequentist analyses. It is possible that this mindset is being carried over to NHBT, which would justify the neglect of the importance of priors in Bayesian testing as well.

QRIP 4

Bayesian evidence is relative. This means that the quantification of the merits of one model is strongly dependent on what other model is used for the comparison. As obvious as this may sound, it is very surprising that more than 60% of the articles seem to gloss over this fact. Here are two such examples: "With this 'stronger' VB05 prior, we found strong evidence for the null

hypothesis (BFs_{null} ranging from 12.7 to 22.7 for the 5 ROIs)" (P_{134}), and "These analyses revealed a Bayes factor of . . . $Bf_{1,0} = 0.19$ in the mindful attention condition, supporting the null hypothesis that sexual motivation does not affect partner judgments following mindful attention" (P_{85}). We also found that, among researchers who failed to mention what prior distributions were used (QRIP 3b), 70% also failed to explicitly refer to the relativeness of the evidence displayed by the Bayes factor outcome. This behavior is perhaps best explained by one or more of the reasons below.

Writing style. To some extent, we think that the economic way in which researchers write their articles can partly explain this result. Having to write repeatedly expressions such as "The Bayes factor indicates that the data are X times more likely under Model A than under Model B" is taxing after some time. It is very likely that some researchers objectively choose to omit parts of the text for the sake of convenience.

Implicitly assumed. This explanation is strongly tied with the previous one. We found examples of articles that in some instances explicitly referred to the relativeness of the evidence but in other cases did not. In addition to writing style, it is perhaps further assumed that the reader understands what is happening. As a consequence, dropping some words along the way may be perceived as "acceptable."

Increased impact. Ascribing evidence to one of the models only may also be a strategy to amplify the strength of the results found. The second example above is one good example of this. It feels stronger to only report "support for the null hypothesis of absence" than to report "support for the null hypothesis of absence over one possible operationalization of the alternative hypothesis of existence" instead. The shorter way of reporting the result is "fancier" and is easier to sell in an abstract or a talk, for example.

QRIP 5 and Usage C

As discussed before, there seems to be an irresistible appeal of researchers toward using the Bayes factor to establish the presence of an effect, or the lack thereof. Our account of Usage C indicates that 18% of the articles referred to this desideratum. In addition, 35% of the articles relied on the Bayes factor to make statements about the existence (or lack thereof) of effects (QRIP 5). Here are two examples: "For 6-year-olds, there was no difference between environments ($M_{smooth} = 2.11$ vs. $M_{rough} = 1.93$, $t(52) = 1.0$, $p = 0.31$, $d = 0.3$, $BF = .42$" (P_{76}), and "[a] Bayesian analysis found a reverse alignment effect with fewer

errors when the arrow pointed away from the object's handle (1.7% vs. 0.8%), $BF = 25.9$" (P_{20}). Several explanations seem plausible to us.

Increased impact. Like QRIP 4, one possible explanation is to enhance the results (i.e., to overclaim).

Avoiding uncertainty. Relatedly, the generalized lack of modesty that permeates published research (Hoekstra & Vazire, 2021) may also help explain this phenomenon. In fact, many researchers seem averse to acknowledging the uncertainty in their experiments and data analyses.

Writing style. We think that some authors may find that a misleading expression such as "there is a difference between the two groups ($BF = . . .$)" is interchangeable with the more adequate expression "the evidence supports the hypothesis that there is a difference between the two groups over the hypothesis that there is no difference ($BF = . . .$)." The former expression is unfortunate because it mixes the relative evidence found for an effect (the Bayes factor value) with the effect hypothesis itself.

Influence from NHST. This is directly related to the previous point. Old habits from reporting statistical results from NHST may also help clarify the situation. In rigor, a "statistically significant" outcome simply states that an effect of at least the magnitude that was observed would be too unlikely were the null hypothesis true. It is a statement about the data under a particular hypothesis and not about any of the hypotheses. Likewise, a similar situation occurs with the Bayes factor, and QRIP 5 is a way to express that.

Decision-making. Testing two hypotheses need not always end with a decision between the two. In many cases, reporting the relative plausibility between both hypotheses should suffice. But this strategy may be perceived as "too nuanced" or even "incomplete." Thus, instead of conducting a detailed cost-benefit analysis, and with the pressure to choose and discard between hypotheses, researchers may then fall into QRIP 5's trap and declare the existence or absence of the effect under study.

QRIP 7

Few articles (seven; 4.2%) considered the Bayes factor as an effect-size measure. Here is one example: "Pupil size was larger in a higher tracking load. . . . However, the Bayesian test showed only positive, but smaller, effect of Load on tracking pupil size ($BF_{incl.} = 7.506$)" (P_{104}).

p values and effect sizes. QRIP 7 may be the Bayesian counterpart to the wrongful association between statistical and practical significance. It is well known that even the

tiniest of effects may become "statistically significant" provided that we have access to enough data. Likewise, widely different effect sizes can be associated with similar levels of evidence as indicated by the Bayes factor, depending on the priors used (Wong et al., 2022). Some researchers may make the same mistake as they make with small p values and thus equate high values of BF_{10} with large effect sizes.

Bayes factor labels. It is possible that commonly used labels to qualify levels of evidence (e.g., $1 < BF_{10} < 3$ = anecdotal evidence for $\mathcal{H}_1$, $3 < BF_{10} < 10$ = moderate evidence for $\mathcal{H}_1$; Jeffreys, 1961) may create some confusion related to the magnitude of the associated effect size and perhaps foster the aforementioned wrongful association between test statistics and effect sizes.

QRIP 9

Bayes factors close to 1 imply that the evidence for either model under comparison is about the same. Erroneously, in a small set of articles (six; 3.6%), researchers instead concluded that they found evidence for the null model of no effect on reporting Bayes factor values close to 1. For example: "In contrast there was no difference in meaning between the thinking without examples and planning conditions; the Bayes factor provided anecdotal evidence in favor of the null ($BF_{10} = .86$)" (P_{105}). Here is another example:

> The difference was significant in the t-test ($t(55) = 2.14$, $p = .04$) but not when calculated on the basis of Cohen's d ($d = .29$, confidence interval between −.09 and .67) or according to a Bayesian test (Bayes factor $B_{10} = 1.2$). Since both the confidence interval and the Bayes' factor do not point towards a true difference and the t-test is borderline significant, this can be considered a very small or non-existent effect. (P_{12})

Influence from NHST. A nonsignificant outcome should imply a noncommital attitude toward the null hypothesis. However, too often researchers interpret nonsignificant findings as "evidence for the null" (e.g., Goodman, 2008). We think that it is possible that this unfortunate reasoning may be resurfacing within Bayesian testing in the form of QRIP 9.

Absence as default. This explanation is closely related to the previous explanation. From NHST tradition, the null model (typically, of absence) is the hypothesis that researchers try to nullify. Faced with an absence of evidence against the null model, researchers fail to reject the null model and retain it instead. The decision to retain the null model need not necessarily reflect belief in the null model, however. From a Neyman-Pearson point of view, retaining or accepting the null hypothesis reflects only a *behavioral* decision of action. This process of decision-making is unrelated to the notion of belief in the hypothesis retained (Neyman & Pearson, 1933). It may also be interpreted as a conservative decision. This "frequentist" attitude of retaining the null model in the absence of evidence is also what QRIP 9 could be based on.

Dichotomization. Hypothesis testing is inherently a dichotomic inferential exercise. Such dichotomization helps create a clear divide between a null model of absence and an alternative model of presence. It is then possible that, when faced with inconclusive evidence (i.e., Bayes factors close to 1), researchers are prone to choose the absence side of the dichotomy, also because of the two reasons below.

Increased impact. It sounds arguably stronger to say that there is "evidence of an absence of an effect" rather than to say "the evidence between absence and existence is ambiguous."

Preference for parsimony. The previous explanation not only sounds stronger but also simpler. We think that perhaps some form of Occam's razor is taking place here and researchers err for preferring the simpler way out (see, e.g., Gallistel, 2009). We note, however, that the Bayes factor already has a preference for simpler models (Jefferys & Berger, 1991), so an additional preference for parsimony should be justified explicitly.

QRIP 10

Basing the interpretation of Bayes factors on qualitative labels associated with ranges of values is the core of this QRIP. We observed this phenomenon in nine articles (5.4%). Here is one instance:

> Both disgust and fear were experienced more in the experimental group ($ps \leq .05$), but disgust showed clearly the largest difference. In terms of Bayes factor (BF), evidence for greater disgust in the experimental group was strong ($BF_{10} > 10$), but there was only weak evidence for a difference in other emotions (BF_{10}'s < 3). (P_{125})

Summary. In the article from which the example above was retrieved, there are six Bayes factors being interpreted (given in a table). The authors may have considered it to be too verbose to interpret each Bayes factor individually.

Seeking authority.
Resorting to interpretative labels has the major advantage of being able to quote others to back up one's own results. In this sense, researchers need less effort to determine the strength of the evidence that they found (i.e., they need not "think").

Avoid criticism.
Related to the previous explanation, using labels may be perceived as a means of protection against criticism aimed at the inherent subjectivity of interpreting Bayes factors. Thus, any questions concerning the perceived magnitude of the estimated effect can be deferred to the Bayes factor label system that was used.

Repeat literature.
Most introductions to Bayesian hypothesis testing refer to at least one label system for the Bayes factors. Some researchers may have found such systems compelling to the point of excessively relying on them.

NHST.
Using labels such as "significant" or "nonsignificant" is commonplace in frequentist inference. It is possible that some researchers are projecting the same kind of reporting behavior onto the Bayes factor.

Usage B

Twenty-seven articles (16.2%) mentioned that they used the Bayes factor as a follow-up to nonsignificant results from NHST. For example:

> In order to address the possibility that this study was underpowered (among other reasons), we also incorporated Bayesian analyses, which do not require a stopping rule (e.g., Rouder, 2014). If a *t*-test yielded a non-significant result, we conducted a Bayesian *t*-test ($r_{prior} = 0.707$). (P_{115})

Below are some considerations related to this particular motivation toward using the Bayes factor.

Support $\mathcal{H}_0$.
Very clearly, the desire to draw support for the null hypothesis is the most logical explanation. Supporting the null hypothesis is not allowed in NHST, and thus the Bayes factor is seen as advantageous (see, e.g., Dienes, 2014).

Trojan horse.
The Bayes factor's ability to draw relative support for the null hypothesis is one of its most touted advantages. We speculate whether, for some researchers, it was precisely this purported advantage that drew them to the Bayes factor.

Request from reviewers.
Given that the use of Bayesian hypothesis testing is growing, it is also possible that reviewers are explicitly requesting this type of analyses.

QRIPs in Abstracts

We also looked at the occurrence of each criterion in the abstracts. The most prominent QRIPs are those associated with short and catchy reporting: 24 (14.4%) QRIP 4 (evidence reported as absolute instead of relative) and 10 (6.0%) QRIP 5 (reporting the presence or absence of effects). Seven articles (4.2%) explicitly referred to a general goal of establishing the absence or presence of a particular effect, for which the Bayes factor would be of use (Usage C).

In general, the main questionable reporting practices that we identified in abstracts seem to be directly related to the fact that they are meant to be short. The pressure to write an appealing abstract may also help explaining our findings. Of course, authors should refrain from engaging in this habit in order to prevent distortions in the published literature.

Summary and Recommendations

In the previous section we outlined various possible causes for the problems we identified. In short, we think that the main causes for the problems include a basic lack of understanding, omission of important information, unfamiliarity on dealing with prior distributions, resorting to writing styles that overemphasize impact and deemphasize uncertainty and a desire to make a dichotomous decision as the final test's outcome.

In addition to the anticipated problems that we identified in our article reading (per Table 2), we also made note of a few other problems that we found (see the Supplementary Material available at the OSF). Here we mention three such occurrences. In one example, we identified a few instances of articles in which the authors seemed to conflate the concept of "evidence" (i.e., how the data allow us to update our belief) with that of "belief" (i.e., how likely we think each hypothesis is after observing the data). This is related to QRIP 1. In another example, there were authors who seemed to think that Bayesian statistics is less reliant on model assumptions. This is obviously misguided. In fact, Bayesian statistics has the potential of bringing models and their underlying assumptions to the analysis forefront. This is not always the case with frequentists statistics (e.g., the set of data "at least as extreme as" is not always clearly defined; Lindley, 1993). Finally, some authors were under the impression that Bayes factors could be used to test model fit. Perhaps surprisingly, Bayes factors do not fare well in what concerns model fit. The strength of the Bayes factor is to quantify the relative predictive ability between two models. One particular model may outpredict another competing model while at the same time fit the data quite poorly (but probably better than the model it outperformed).
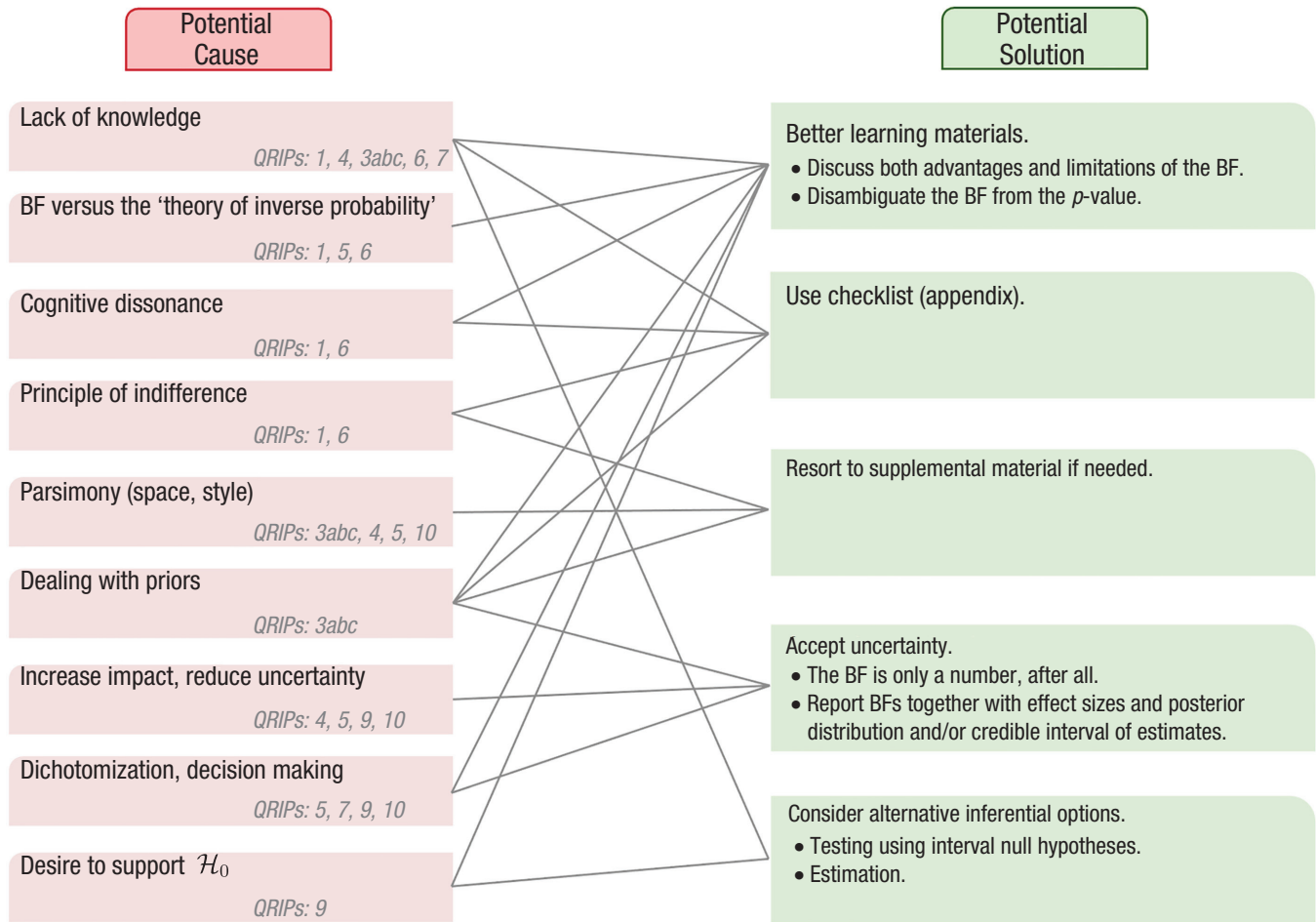
Fig. 2. Summary of the potential causes for the problems identified in the literature study and suggestions for potential solutions. For each potential cause (left), QRIPs that we anticipate that follow as a consequence are listed. Potential solutions (right) are linked back to the causes that we expect they most directly apply to. QRIP = questionable reporting or interpreting practice; BF = Bayes factor.

Our advice is to always consider model fit separately from testing through the Bayes factor.

All together, our findings provide a clearer image of the ongoing problems related to the use of the Bayes factor in practice. To address the current state of affairs, we also wish to offer some constructive suggestions aimed at improving things going forward. Figure 2 shows our suggestions and how they are meant to attend to each QRIP. Below we briefly summarize each of our proposals.

Learning materials

Introductions to the Bayes factor commonly start by highlighting problems with the *p* value. These limitations then motivate the use of the Bayes factor, which is then showcased. We think that this setup misses a crucial component, which is a *critical appraisal* of the Bayes factor. Some of us have written about this before (see Tendeiro & Kiers, 2019; Wong et al., 2022), but this is much more the exception than the rule. We suggest

that updated materials (e.g., articles, apps, training sessions) offering thoughtful discussions of the various QRIPs shown on Table 2 would go to great lengths to mitigate the problems we identified. In particular, we suggest that researchers learn the following:

- There is a difference between the concepts of the Bayes factor (the evidence) and posterior odds (the belief; QRIPs 1 and 6).[7]
- Prior odds must be specified whenever there is interest in the posterior odds. Reporting posterior odds without prior odds is, at best, not ideal because it requires that the reader must consider what the authors' priors odds were to start with.
- Reporting the priors used is crucial (QRIPs 3a, 3b, and 3c). Furthermore, and as much as possible, the motivation for choosing such priors should also be provided.
- It is important to conduct sensitivity analyses to assess the influence of the priors on the Bayes

factor. In our study, only 26 papers (15.6%) explicitly referred to sensitivity analysis.

- The Bayes factor is only a measure of the relative evidence between the two models under comparison (QRIP 4).
- It is most likely impossible that the Bayes factor of one isolated study can be used to *establish* the absence or presence of any effect (QRIP 5). Using the Bayes factor in this way should be deemed to be a severe error.
- It is important to always provide a full account of the interpretations in the article.[8] We do realize, however, that this is difficult without becoming overly repetitive. One suggestion is that authors add to the description of the statistical analysis in the methods section something like this: "Whenever we interpret a test result as providing support for one of the hypothesis, we mean to say that the evidence supports this hypothesis over the selected competing hypothesis." At the very least, we strongly suggest that authors follow our suggestion for the key outcomes of their studies.
- The Bayes factor is not an effect-size measure (QRIP 7).
- Understanding the difference between the absence of evidence and evidence of absence is essential (QRIP 9).
- The Bayes factor value should always be reported (QRIP 10). This is the Bayesian equivalent to requesting the exact p value instead of an inequality (e.g., $p < .05$). Providing the exact value of the Bayes factor has three immediate advantages: Readers may also make their own judgment concerning the strength of the evidence reported in articles, it facilitates future meta-analyses, and it allows the calculation of the posterior odds. Nevertheless, authors should still feel free to interpret the magnitude of the evidence as they see fit.

Checklist

We prepared a checklist that practitioners may use to guide them, at least throughout their first interactions with Bayesian hypothesis testing (see Appendix). This checklist highlights what aspects should be reported, either in the article itself or possibly in the supporting materials. We think that by using such a checklist researchers will feel reassured that they are taking all the important steps in their analysis. The checklist may also be of help to both journals and reviewers in developing standardized guidelines to which authors must abide. This may further contribute to increasing authors' awareness to these issues.

Supplementary material

Our checklist is thorough and possibly leads to more information than one is willing to incorporate in their articles. Relegating some information to the supporting materials is a valid solution in such cases. Authors may want to resort to free and publicly available repositories such as OSF for this purpose. Journals may also promote the practice of sharing supporting materials that include the information detailed on the checklist on their websites. One suggestion is to use supplementary material (if needed) to fully report the priors used and the motivation for choosing such priors. It is important to keep in mind that priors are part of the models; therefore, any inference is contingent on the chosen priors. In this sense, failing to report priors may be considered as much of an error as it is to fail to report that one assumes normally distributed data, for example. Another suggestion is to place the results of sensitivity analyses in the supplementary materials.[9]

Accept uncertainty

Statistical tools should be used within their own bounds. All the Bayes factor offers is a means of gathering evidence in favor of either hypothesis put up to a test. This does not equate to a formal proof as if it were a mathematical theorem. We suggest researchers adjust their expectations to what the Bayes factor permits. In particular, it is important to avoid the dichotomization trap that hypothesis testing typically entails. If a decision is really needed and in particular if the stakes are high, it is perhaps best to consider statistical-decision theory (Berger, 1993). It is also important to report effect sizes in order to complement test results.

Alternative inferential procedures

Testing, in particular null hypothesis testing, may not be what researchers need at all times. Some researchers have questioned the role of point null hypotheses (e.g., Vardeman, 1987). It is important to point out that alternatives do exist. One option is to use interval null hypothesis (Morey & Rouder, 2011). But often a research question may be well addressed by means of resorting to estimation instead. Arguably estimation may offer what testing does, and more (Tendeiro & Kiers, 2023).

Conclusion

In this article we charted the current state of affairs concerning the use of the Bayes factor in applied research. Our findings suggest that current practices are at best suboptimal. This happens in spite of Bayesian

inference in general, and the Bayes factor in particular, being often described as more intuitive than frequentist inference (Kruschke & Liddell, 2018). We think that the problem is real and needs to be addressed for the quality of research to increase.

Some of the numbers appear small; for example, we found that 3.6% of the articles committed the error (QRIP 9) of confusing Bayes factors of about 1 with evidence of absence. We note that the error rates we report are marginal error rates, but important error rates—such as the probability of committing an error given the situation is right—should be higher. For instance, one can commit QRIP 9 only if the Bayes factor is around 1. Following a suggestion from a reviewer, we computed the proportion of occurrences of QRIP 9 among all occurrences of Bayes factor values between $\frac{1}{3}$ and 3 in the text (values reported only in tables were not considered). The result was 14 occurrences in 429, or 3.3%. This finding does not fully align with our intuition. Based on this subanalysis only, it is yet unclear what the actual conditional error rates should be.

In addition to reporting the identified problems, we also attempted to explain what reasons may be behind each problem. Naturally, our arguments are not evidence-based. Future research aiming at a more fine-grained understanding of the current situation would be extremely helpful.

We have offered some suggestions for actions to be taken that may contribute toward improving the situation. We think that what is needed is a better understanding of the effect of prior distributions, the difference between posterior odds and the Bayes factor, the importance of providing thorough reports of the analyses conducted (Kruschke, 2021; van Doorn et al., 2021), the need to explain the choices made, the disconnect between Bayes factors and effect sizes, and what it takes to establish that a particular effect is absent or present. Also, carrying frequentist preconceptions over into the Bayesian world is not advisable.

The way forward is not to ban Bayesian inference from our toolbox. Instead, more and better education on Bayesian inference is needed. We think that future work should use findings from Wong et al. (2022) and this article to shape improved educational materials. Better showcasing how Bayesian inference can be correctly used will empower applied researchers and improve the quality of the published scientific findings.

Appendix

Below is an ordered list with the points that should be taken into account when conducting a Bayesian hypotheses test through the Bayes factor:

1. Check whether model assumptions hold well (e.g., independence of observations, normality).

2. Specify the two hypotheses that will be tested against each other ($\mathcal{H}_0$ and $\mathcal{H}_1$).
3. Completely specify the prior distributions for all parameters under either hypothesis.
4. Explain the choice of priors as much as possible.
5. (Optional) Specify prior odds in case you are interested in the final updated relative belief.
6. Specify the software used to compute the Bayes factor.
7. Report the Bayes factor using clear notation (e.g., use BF_{01} to denote the evidence in favor of $\mathcal{H}_0$ relative to $\mathcal{H}_1$).
8. Interpret the Bayes factor on the basis of either Equation 1 or Equation 4 (describe evidence).
9. Conduct sensitivity analyses to assess the effect of the priors on the Bayes factor. Consider varying both the width and the location of the priors.
10. (Optional) If prior odds were specified, compute the posterior probabilities of both hypotheses using Equation 5 (describe belief).
11. Report the estimated effect size together with a posterior distribution, or at least a credible interval.
12. Include a brief account of all steps above in your report. Some information (e.g., from Steps 3, 5, and 8) may be relegated to supplementary material.

ORCID IDs

Jorge N. Tendeiro ⓘD https://orcid.org/0000-0003-1660-3642
Henk A. L. Kiers ⓘD https://orcid.org/0000-0002-4995-9349
Rink Hoekstra ⓘD https://orcid.org/0000-0002-1588-7527
Tsz Keung Wong ⓘD https://orcid.org/0000-0003-0451-8328
Richard D. Morey ⓘD https://orcid.org/0000-0001-9220-3179

Notes

1. Here we treat θ_i as a single and continuous random variable for simplicity. The concept of marginal likelihood extends straightforwardly to the multiple random variables case by extending Equation 2 to multiple integration and to discrete random variables by replacing the integration by a summation.
2. Files are available at https://osf.io/57ew4.
3. Unlike the original article, here we did not perform corrections for multiple testing for simplicity.
4. Relatedly, we can also suggest the "Bayes factor workflow" from Schad et al. (2022), who provided guidance with respect to determining the computational stability of the Bayes factor.
5. We refer to specific articles in the sample using the codes $P_1, \ldots, P_{167}$.
6. "To take the prior probabilities different in the absence of observational reason for doing so would be an expression of sheer prejudice. The rule that we should then take them equal is not a statement of any belief about the actual composition of the world, nor is it an inference from previous experience; it is merely the formal way of expressing ignorance. It is sometimes referred to as the Principle of Insufficient Reason (Laplace) or the equal distribution of ignorance" (Jeffreys, 1961, pp. 33–34).
7. In fact, we also found some authors who seemed to clearly understand this distinction: "From Bayes' theorem, the odds of the two hypotheses given the data, $Pr(H_0 \mid D) / Pr(H_1 \mid D)$, are equal to the prior odds (that is, the odds before the current data were collected) multiplied by the Bayes factor" (P_{44}).
8. Here is one good example: "This analysis revealed a value of 6.08 to 1 in favor of the null hypothesis over the SLH for the present Experiments 2, 3 and 4. As such, the current results constitute 'some' evidence in favor of a null over the SLH" (P_{78}).
9. Or concisely in the article itself, as the following example illustrates: "Previously reported effect sizes for action language impairments in PD have been very large: approximately Cohen's $d = 2$. . . . However, we accepted the possibility that our effects would be smaller than this, given how well our control conditions were matched to the experimental conditions, and particularly in the metaphor conditions. Given this uncertainty, we report BFs under a range of Cauchy prior widths including 2 (based on previous effects), as well as the default (.707) to determine the robustness of the effects" (P_{123}).

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9
Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*(4), 389–396. https://doi.org/10.1037/1082-989X.10.4.389
Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., . . . Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z
Berger, J. O. (1993). *Statistical decision theory and Bayesian analysis* (2nd ed.). Springer-Verlag.
Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, Article 781. https://doi.org/10.3389/fpsyg.2014.00781
Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*(1), 5–34. https://doi.org/10.3758/s13423-017-1262-3
Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*(2), 313–329. https://doi.org/10.1214/16-STS599
Falconer, J. R., Frank, E., Polaschek, D. L. L., & Joshi, C. (2022). Methods for eliciting informative prior distributions: A critical review. *Decision Analysis*, *19*(3), 189–254. https://doi.org/10.1287/deca.2022.0451
Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*(1), 75–98. https://doi.org/10.1177/0959354395051004
Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.
Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003
Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3
Guo, D., & Ma, Y. (2022). The "p-hacking-is-terrific" ocean: A cartoon for teaching statistics. *Teaching Statistics*, *44*(2), 68–72. https://doi.org/10.1111/test.12305
Haeffel, G. J., Burke, H., Vander Missen, M., & Brouder, L. (2023). What diverse samples can teach us about cognitive vulnerability to depression. *Collabra: Psychology*, *9*(1), 71346. https://doi.org/10.1525/collabra.71346
Haeffel, G. J., Gibb, B. E., Metalsky, G. I., Alloy, L. B., Abramson, L. Y., Hankin, B. L., Joiner, T. E., & Swendsen, J. D. (2008). Measuring cognitive vulnerability to depression: Development and validation of the cognitive style questionnaire. *Clinical Psychology Review*, *28*(5), 824–836. https://doi.org/10.1016/j.cpr.2007.12.001
Haller, H., & Kraus, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, *7*(1), 1–20.
Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3
Hoekstra, R., & Vazire, S. (2021). Aspiring to greater intellectual humility in science. *Nature Human Behaviour*, *5*(12), 1602–1607. https://doi.org/10.1038/s41562-021-01203-8

JASP Team. (2023). *JASP* (Version 0.17) [Computer software]. University of Amsterdam. https://jasp-stats.org

Jefferys, W. H., & Berger, J. O. (1991). Sharpening Occam's razor on a Bayesian strop. *Bulletin of the American Astronomical Society*, *23*, 1259.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. https://doi.org/10.2307/2291091

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, *5*(10), 1282–1291. https://doi.org/10.1038/s41562-021-01177-7

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

Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*(1), 22–25. https://doi.org/10.1111/j.1467-9639.1993.tb00252.x

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. https://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*(4), 406–419. https://doi.org/10.1037/a0024377

Morey, R. D., & Rouder, J. N. (2021). *BayesFactor: Computation of Bayes factors for common designs*. Manual.

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

Oakes, M. W. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. John Wiley & Sons.

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

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., Speckman, P. L., Sun, D., & Morey, R. D. (2009). Bayesian t tests for accepting and rejecting the null hypothesis.

Psychonomic Bulletin & Review, *16*, 225–237. https://doi.org/10.3758/PBR.16.2.225

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2022). Workflow techniques for the robust use of Bayes factors. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000472

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Świątkowski, W., & Carrier, A. (2020). There is nothing magical about Bayesian statistics: An introduction to epistemic probabilities in data analysis for psychology starters. *Basic and Applied Social Psychology*, *42*(6), 387–412. https://doi.org/10.1080/01973533.2020.1792297

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795. https://doi.org/10.1037/met0000221

Tendeiro, J. N., & Kiers, H. A. L. (2023). With Bayesian estimation one can get all that Bayes factors offer, and more. *Psychonomic Bulletin & Review*, *30*, 534–552. https://doi.org/10.3758/s13423-022-02164-3

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1–2. https://doi.org/10.1080/01973533.2015.1012991

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development*, *85*(3), 842–860. https://doi.org/10.1111/cdev.12169

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š, Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., . . . Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*(3), 813–826. https://doi.org/10.3758/s13423-020-01798-5

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.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of P values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wellek, S. (2003). *Testing statistical hypotheses of equivalence*. CRC Press.

Wong, T. K., Kiers, H., & Tendeiro, J. (2022). On the potential mismatch between the function of the Bayes factor and researchers' expectations. *Collabra: Psychology*, *8*(1), Article 36357. https://doi.org/10.1525/collabra.36357