



A Critical Evaluation of the FBST ev for Bayesian Hypothesis Testing

Critique of the FBST ev

Alexander Ly^{1,2} · Eric-Jan Wagenmakers¹

Accepted: 11 May 2021
© The Author(s) 2021

Abstract

The “Full Bayesian Significance Test e -value”, henceforth FBST ev , has received increasing attention across a range of disciplines including psychology. We show that the FBST ev leads to four problems: (1) the FBST ev cannot quantify evidence in favor of a null hypothesis and therefore also cannot discriminate “evidence of absence” from “absence of evidence”; (2) the FBST ev is susceptible to sampling to a foregone conclusion; (3) the FBST ev violates the principle of predictive irrelevance, such that it is affected by data that are equally likely to occur under the null hypothesis and the alternative hypothesis; (4) the FBST ev suffers from the Jeffreys-Lindley paradox in that it does not include a correction for selection. These problems also plague the frequentist p -value. We conclude that although the FBST ev may be an improvement over the p -value, it does not provide a reasonable measure of evidence against the null hypothesis.

Keywords Full Bayesian Significance Test · Evidence · Optional stopping · Predictive irrelevance · Jeffreys-Lindley paradox

Introduction

Over the past decades, the “FBST ev ” has been proposed as a Full Bayesian Significance Test (de Bragança Pereira & Stern, 1999; de Bragança Pereira et al., 2008) that offers distinct advantages over both frequentist p -values and Bayes factors (e.g., Madruga et al. 2003). The FBST ev has been applied in economics, genetics, signal processing, time series analysis, biology, astronomy, medicine, and other fields (see de Bragança Pereira & Stern J. M. *in press* and references therein). Recently, the FBST ev has also been introduced to the field of psychology (e.g., Kelter 2020b; Kelter *in press*). However, we believe that the method has specific limitations that warrant discussion. Below we first introduce the FBST ev and describe its similarities to conventional p -values. Next we outline four features that we perceive to be particularly problematic.

A Brief Definition of the FBST ev

The FBST ev (sometimes called “ e -value”¹) is a posterior tail probability, where the tail area is defined by a so-called posterior surprise function, which is defined below. For concreteness we focus on a point null hypothesis that fixes a parameter of interest to a specific value: $\mathcal{H}_0 : \theta = \theta_0$. The alternative hypothesis relaxes the restriction on θ , and it is assumed that a prior distribution $\pi(\theta | \mathcal{H}_1)$ is assigned to θ , which in light of the observed data can then be updated to a posterior $\pi(\theta | y, \mathcal{H}_1)$ using Bayes’ rule.

The posterior surprise function is the ratio of the posterior over a reference function. One of the most common reference functions, and the one we use throughout this manuscript, is the prior distribution, and leads to $\pi(\theta | \mathcal{H}_1, y) / \pi(\theta | \mathcal{H}_1)$ — a common Bayesian measure of evidence or support (e.g., Evans, 2015). The support associated with the null hypothesis is defined as the largest support for parameter values restricted by the null, that is, $\sup_{\theta \in \Theta_0} \pi(\theta | \mathcal{H}_1, y) / \pi(\theta | \mathcal{H}_1)$, where Θ_0 is the parameter space Θ_1 restricted according to \mathcal{H}_0 . For a point null hypothesis this simply yields $\pi(\theta_0 | \mathcal{H}_1, y) / \pi(\theta_0 | \mathcal{H}_1)$. The tail area T used to define the FBST ev consists of all

✉ Alexander Ly
a.ly@uva.nl

¹ Psychological Methods, University of Amsterdam, Amsterdam, the Netherlands

² Machine Learning Group, Centrum Wiskunde, Informatica, Amsterdam, the Netherlands

¹Not to be confused with the e -values produced by safe tests (Grünwald et al., 2020), where the e refers to expectation.

parameter values θ that have posterior support less than the support associated with the null hypothesis, that is,

$$T = \left\{ \theta \in \Theta_1 \mid \frac{\pi(\theta \mid y, \mathcal{H}_1)}{\pi(\theta \mid \mathcal{H}_1)} \leq \frac{\pi(\theta_0 \mid y, \mathcal{H}_1)}{\pi(\theta_0 \mid \mathcal{H}_1)} \right\}, \quad (1)$$

where Θ_1 is the unrestricted parameter space. The FBST ev for $\mathcal{H}_0 : \theta = \theta_0$ is defined as the posterior mass in area T , that is,

$$ev = \int_T \pi(\theta \mid y, \mathcal{H}_1) d\theta. \quad (2)$$

When ev is low, or $\bar{ev} = 1 - ev$ is high, the data are said to cast doubt on \mathcal{H}_0 .

As a concrete example, consider a binomial test of $\mathcal{H}_0 : \theta = \frac{1}{2}$ versus $\mathcal{H}_1 : \theta \in \Theta_1 = (0, 1)$ and a beta(1, 1) prior (i.e., a uniform prior from 0 to 1). The posterior surprise function then equals the posterior distribution. Consequently, the set T that defines the posterior tail area are those θ s that lead to a posterior value $\pi(\theta \mid y, \mathcal{H}_1) \leq \pi(\frac{1}{2} \mid y, \mathcal{H}_1)$. Equivalently, $\bar{ev} = 1 - ev$ defines the posterior probability of $\bar{T} = \Theta_1 \setminus T$, the complement of the tail area T .

For observed data consisting of $s = 15$ successes and $f = 5$ failures the posterior is a beta(16, 6) distribution. The support for the null is $\pi(\frac{1}{2} \mid y, \mathcal{H}_1) = 0.31$ and the area of parameter values that have more support is $\bar{T} = [\frac{1}{2}, 0.92]$; thus, the tail area is $T = (0, \frac{1}{2}) \cup (0.92, 1)$. The area under the curve shown in Fig. 1 equals $\bar{ev} = 1 - ev = .98$, and hence $ev = .02$, which combined with the standard threshold of .05 leads to a rejection of the null hypothesis (de Bragança Pereira et al., 2008, Definition 2.3).

The advantage of the FBST ev is that the procedure promises to work with non-subjective priors such as

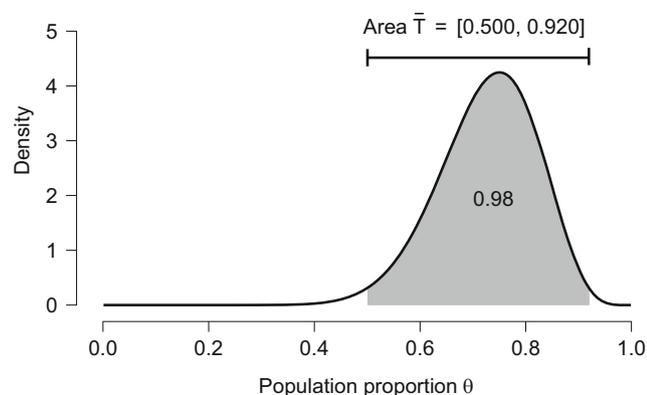


Fig. 1 Example of the FBST ev for a binomial test of $\mathcal{H}_0 : \theta = \frac{1}{2}$ versus $\mathcal{H}_1 : \theta \in \Theta_1 = (0, 1)$ with a beta(1,1) prior for $s = 15$ successes out of $n = 20$ observations. Shown is the beta(16, 6) posterior distribution. Area $\bar{T} = [\frac{1}{2}, 0.92]$ contains all values of θ with at least as much support from the data as $\theta = \frac{1}{2}$. The posterior mass on \bar{T} equals 0.98, and hence $ev = 1 - 0.98 = 0.02$. Figure based on the JASP module *Learn Bayes*

Jeffreys’s transformation-invariant priors (e.g., Jeffreys 1946; Ly et al., 2017), or reference priors (e.g., Berger et al. 2009; Bernardo 1979) as long as the posterior is proper. In contrast, Bayes factors prohibit the use of improper priors on the test-relevant parameter, and the construction of default Jeffreys’s Bayes factors requires the statistician to select a class of priors that fulfil certain desiderata (e.g., Bayarri et al. 2012; Jeffreys 1961; Li & Clyde 2018; Ly et al. 2016a; Ly et al. 2016b). Hence, the FBST procedure is more or less automatic whenever non-subjective priors are available (de Bragança Pereira & Stern J. M., in press), as it does not require additional difficulties of prior selection. Moreover, Diniz et al. (2012) mention that it is an advantage that the FBST procedure avoids the introduction of a prior probability on the null hypothesis. Below we show that the automatic use of the FBST can lead to problematic inferences. The root of the problem, we believe, is due to the fact that the behavior of FBST ev is similar to that of p -values.

The FBST ev as an Approximate p -value

Fig. 1 highlights the conceptual similarity between the FBST ev and the frequentist p -value from null hypothesis significance testing. One key difference is that the p -value violates the Likelihood Principle (e.g., Berger and Wolpert 1988; Wagenmakers, 2007), but the FBST ev does not, because it relies on the posterior distribution.²

Another key difference is that for p -values the tail event consists of the set of more extreme outcomes than what is observed across the sample space, whereas for the FBST ev the tail area consists of parameter values that receive less support than the null value. This difference between the spaces is noticeable in the binomial example, as the tail event of the p -value is comprised of discrete outcomes, e.g., the sets $\{0, 1, 2, 3, 4, 5\}$ and $\{15, 16, 17, 18, 19, 20\}$, whereas the tail area for FBST ev consists of two continuous intervals, e.g., $(0, \frac{1}{2})$ and $(0.92, 1)$.

There are specific cases where the FBST ev and the p -value (for a fixed- n sampling plan and without possible corrections for multiplicity) are exactly equal. In particular, this occurs when data are normally distributed with population mean μ and variance $\sigma^2 = 1$ and the null hypothesis is $\mathcal{H}_0 : \mu = 0$, whereas the alternative allows μ to vary freely, that is, $\mathcal{H}_1 : \mu \in \mathbb{R}$. Using the (improper) uniform prior (i.e., the “uninformative” Jeffreys’s transformation-invariant prior) then results in the FBST ev and the p -value being exactly equal (e.g., Diniz

²The Likelihood Principle holds that “all evidence, which is obtained from an experiment, about an unknown quantity θ , is contained in the likelihood function of θ for the given data.” (Berger & Wolpert, 1988, p. 1).

et al. 2012). More generally, consider Bayesian parameter estimation for the location parameter μ of a statistical model from the exponential family. Assume the prior on μ is uniform on the real line. Then the proportion of the posterior distribution with mass lower than zero equals the one-sided frequentist p -value (e.g., Lindley 1965; Pratt et al. 1995). This means that for a symmetric posterior distribution, the FBST ev equals the two-sided frequentist p -value. For non-uniform priors the relation between FBST ev and p -value is only approximate; however, as sample size grows the data quickly dominate the posterior distribution (e.g., Wrinch & Jeffreys 1921), such that the approximation will be increasingly close.

The large-sample connection between the FBST ev and the p -value holds even more generally, whenever the (asymptotic) p -value is derived from the (generalized) likelihood ratio statistic. This connection is made rigorous via Wilks' theorem (Wilks, 1938), which states that the likelihood ratio statistic asymptotically has a χ^2 -distribution, and the Bernstein von-Mises theorem (e.g., LeCam 1986; Ghosal & van der Vaart 2017; van der Vaart 1998), which states that, under general conditions, the posterior is asymptotically normal. Both theorems were used by Diniz et al. (2012) to derive the following (asymptotic) functional relationship

$$ev \xrightarrow{d} 1 - F_{\dim(\Theta_1)}[F_{\dim(\Theta_1) - \dim(\Theta_0)}^{-1}(1 - p)], \tag{3}$$

as $n \rightarrow \infty$, where p in the right-hand side is the asymptotic p -value of the generalized likelihood ratio statistic, F_k and F_k^{-1} are the cumulative distribution function and the quantile function of a χ^2 -distribution with k degrees of freedom, respectively. For instance, for the binomial example $\dim(\Theta_1) = 1$, since there is only one parameter that is free to vary in \mathcal{H}_1 and $\dim(\Theta_0) = 0$, since it is a singleton. As such, whenever $\dim(\Theta_1) = \dim(\Theta_1) - \dim(\Theta_0)$, Eq. (3) states that the difference between FBST ev and the asymptotic p -value becomes negligible as the sample sizes increase.

In the remainder of the paper we highlight how the FBST procedure can result in problematic inferences. Our conclusions are to some extent predicated on the notion that evidence is the degree to which the data change our conviction about competing accounts of the world (Evans, 2015; Morey et al., 2016).

Problem 1: The FBST ev Cannot Quantify Evidence in Favor of the Null Hypothesis

Our first demonstration concerns the case of *pure induction*: a universal generalization is stipulated and an unbroken string of n confirmatory instances is observed. The evidence in favor of the universal generalization ought to increase with n ; we consider this axiomatic. The mathematician

George Polya termed this regularity the “fundamental inductive pattern”:

“This inductive pattern says nothing surprising. On the contrary, it expresses a belief which no reasonable person seems to doubt: *The verification of a consequence renders a conjecture more credible. With a little attention, we can observe countless reasonings in everyday life, in the law courts, in science, etc., which appear to confirm to our pattern.*” (Polya, 1954, pp. 4–5)

The FBST ev violates the fundamental inductive pattern. Concretely, a universal generalization posits that all instances have property x , that is, $\mathcal{H}_0 : \theta = 1$ for a binomial likelihood kernel. Regardless of their number n , the confirmatory instances are best predicted by $\theta = 1$, and consequently this value receives more support than any other value of θ . This means that $\bar{ev} = 0$ and $ev = 1$ regardless of the number of confirmatory instances and regardless of the prior distribution under \mathcal{H}_1 . Note that the p -value suffers from the exact same problem (e.g., Jeffreys 1980; Ly et al. 2020; Wrinch & Jeffreys 1919).

Our second demonstration concerns a binomial test of $\mathcal{H}_0 : \theta = \frac{1}{2}$ versus $\mathcal{H}_1 : \theta \in \Theta_1 = (0, 1)$ and a beta(α, α) prior that is symmetric around $\frac{1}{2}$. Suppose that the data consist of just as many successes s as failures f (i.e., $s = f = \frac{n}{2}$). In this scenario, the evidence for \mathcal{H}_0 ought to increase with n . However, for any n the data are best predicted by $\theta = \frac{1}{2}$. As in the first demonstration, this means that $ev = 1$ regardless of n .

These demonstrations show that the FBST ev cannot quantify evidence in favor of a null hypothesis. Data maximally consistent with the null hypothesis ought to offer stronger support when sample size is high rather than low; however, this regularity is not reflected in FBST ev . This inability means that the FBST ev also cannot discriminate between evidence for absence and absence of evidence (e.g., Keyser et al. 2020).

Problem 2: The FBST ev is Susceptible to Sampling to a Foregone Conclusion

Not only is the FBST ev incapable of quantifying evidence for the null, it will lead to a sure rejection of the null if a sufficiently patient researcher monitors the FBST ev and stops whenever it dips below a fixed threshold such as $ev < .05$. When \mathcal{H}_0 is true, this is certain to happen (i.e., the decision to reject \mathcal{H}_0 based on an indefinite accumulation of observations is a foregone conclusion). This makes an observed rejection of the null based on a sequential application of the FBST procedure ambiguous and uninterpretable.

The reason why monitoring the FBST ev is problematic stems from its intimate (asymptotic) relationship with the p -value. As noted earlier, the asymptotic behavior of the p -value also describes the asymptotic behavior of the FBST ev ; and for p -values, it can be proven that there is no convergence to a single value — instead, p -values meander randomly in the $(0, 1)$ -interval (Feller, 1940; 1968). Hence, the FBST ev will eventually cross the threshold of, say, $ev < .05$. A concrete demonstration is given in Wagenmakers et al. (2018, intuition 5).

A Bayesian interpretation of this phenomena is that the FBST ev is approximately a test for the direction of an effect (Casella & Berger, 1987; Marsman & Wagenmakers, 2017). Recall that in many applications, the FBST ev consists of the sum of two posterior areas: the right tail and the left tail (cf. Fig. 1). Focusing on the left tail (i.e., the posterior mass on values of parameter values θ lower than θ_0), we note that it equals the probability that θ is lower than θ_0 as opposed to higher than θ_0 . Specifically, assume a prior distribution for θ under \mathcal{H}_1 that is centered on θ_0 , and assume that the posterior is symmetric such that the area in the left tail equals that in the right tail. The posterior probability that θ is lower than θ_0 then equals $ev/2$.

Consider now the scenario where data are generated under \mathcal{H}_0 , and accumulate indefinitely (cf. Kelter 2020a). In our interpretation, the FBST ev is an approximate test for direction; because the true value is exactly in between the positive and negative values, the FBST ev will therefore meander randomly — the probability (or odds) has nothing to converge to.

To conclude, the FBST ev can be interpreted as an approximate test for direction. This means that whenever data are generated under \mathcal{H}_0 , the truth is exactly in the middle and the FBST ev will drift randomly and be susceptible to sampling to a foregone conclusion (Rouder, 2014; Wagenmakers et al., 2018). Because the asymptotic behavior of the FBST ev equals that of the frequentist p -value, the proofs on sampling to a foregone conclusion in Feller (1968) also pertain to the FBST procedure.

Problem 3: The FBST ev Violates the Principle of Predictive Irrelevance

The principle of predictive irrelevance implies that if two models yield exactly the same predictions for to-be-observed data, then actually observing the data should not change our preference for one model over the other. As an example, consider the binomial test of $\mathcal{H}_0 : \theta = \frac{1}{2}$ versus $\mathcal{H}_1 : \theta \in \Theta_1 = (0, 1)$ with a $\text{beta}(\alpha, \alpha)$ prior, which is symmetric around $\frac{1}{2}$. Under \mathcal{H}_0 , the probability that the next observation is a success equals $\frac{1}{2}$. This is also the case under \mathcal{H}_1 , as the prior predictive of \mathcal{H}_1 is then a symmetric

beta-binomial distribution with mean $\alpha/(2\alpha) = \frac{1}{2}$ which equals the probability of the next observation being a success.

To see that the FBST ev violates the principle of predictive irrelevance, note that if $\alpha = 1$ and the first observation is a success then under \mathcal{H}_1 we have $\theta \sim \text{beta}(2, 1)$; ev is then the posterior mass on $\theta < \frac{1}{2}$, which equals $ev = .25$. If, instead, the first observation is a failure, then under \mathcal{H}_1 we have $\theta \sim \text{beta}(1, 2)$ and ev is also equal to .25. Hence, upon observing the new datum, the posterior changes and consequently the FBST ev changes — even though the datum is predictively irrelevant.

More generally, any symmetric posterior distribution $\theta \sim \text{beta}(\alpha + n/2, \alpha + n/2)$ under \mathcal{H}_1 has $ev = 1$ (see Problem 1 above). The occurrence of a new datum will decrease that FBST ev to indicate some evidence against \mathcal{H}_0 , even though the datum is predictively irrelevant and even though the decrease occurs regardless of the nature of the datum (i.e., success or failure). This analysis can be extended to data sequences of arbitrary length (Wagenmakers et al., 2020).

In contrast, Jeffreys's Bayes factor does behave in accordance with the principle of predictive irrelevance, because it is defined as the ratio of the prior predictives evaluated at the observed data, that is, the ratio of marginal likelihoods. Whenever \mathcal{H}_0 and \mathcal{H}_1 make identical predictions for the next datum, Jeffreys's Bayes factor is unaltered when that datum is observed, indicating perfect predictive equivalence: "The first member sampled is bound to be of one type or the other, whether the chance is $\frac{1}{2}$ or not, and therefore we should expect it to give no information about the existence of bias." (Jeffreys, 1961, p. 257). More generally, when the beta prior distribution is symmetric and $s = f$ observations have been made, the prior distribution for the analysis of the next observation is a $\text{beta}(\alpha + s, \alpha + f)$ distribution. Because this distribution is symmetric around $\theta = \frac{1}{2}$, the predictions for the upcoming observation are again identical under \mathcal{H}_0 and \mathcal{H}_1 , and therefore the Bayes factor remains unchanged irrespective of its outcome. "Thus if at a certain stage the sample is half and half, the next member, which is bound to be of one type or the other, gives no new information." (Jeffreys, 1961, p. 257). Another case consist of an asymmetric $\theta \sim \text{beta}(\alpha, \beta)$ distribution which is updated with unequal s and f such that $\alpha + s = \beta + f$; the next observation will not change the Bayes factor for \mathcal{H}_0 versus \mathcal{H}_1 .

In other words, whenever the distribution for θ under \mathcal{H}_1 is symmetric around $\frac{1}{2}$ (either before or during data collection), the next observation is predictively irrelevant, and will leave the Bayes factor unaffected (see Wagenmakers et al. 2020 and references therein). Jeffreys notes "In that case the posterior [model] probabilities are equal to the prior [model] probabilities; in other words the new data do nothing to help us decide between the

hypotheses. This is the case of *irrelevance*.” (Jeffreys 1973, p. 31; italics in original) and “...if the data were equally likely to occur on any of the hypotheses, they tell us nothing new with respect to their credibility, and we shall retain our previous opinion, whatever it was.” (Jeffreys, 1961, p. 29) In contrast, the FBST *ev* suggest that the observation of predictively irrelevant data ought to change our previous opinion.

Problem 4: The FBST *ev* Suffers from the Jeffreys-Lindley Paradox

The Jeffreys-Lindley paradox (e.g., Jeffreys 1935; Jeffreys 1937; Jeffreys 1939; Lindley 1957; see also Bernardo 1980; Cousins 2017; Good 1980a; Good 1980b) refers to the fact that, as sample size increases indefinitely and the threshold for the *p*-value remains constant at any non-zero value (e.g., $p < .005$), we inevitably arrive at a conflict between *p*-values and Bayes factors, in the sense that the *p*-value would suggest that \mathcal{H}_0 be rejected, whereas the Bayes factor would indicate that \mathcal{H}_0 decisively outpredicts \mathcal{H}_1 . This conflict will arise regardless of the prior distribution on the test-relevant parameter under \mathcal{H}_1 (under regularity conditions) and regardless of the *p*-value under consideration. As explained earlier, the FBST *ev* is an approximate *p*-value, and hence the conflict arises for the FBST *ev* as it does for the *p*-value.

For reasons that are unclear to us, the Jeffreys-Lindley paradox is sometimes been used to argue *against* the Bayes factor and motivate the search for a model comparison metric that does not ‘suffer’ from the paradox. For instance, de Bragança Pereira et al. (2008, p. 80) state: “However, whenever the posterior is absolutely continuous and the null hypothesis sharp, the use of Bayes Factors for significance testing is controversial” and then refer to Lindley (1957). In another article, de Bragança Pereira and Stern (1999, p. 109) state that the Jeffreys-Lindley paradox is due to the fact that the Bayes factor privileges the null hypothesis by giving it separate prior mass.

However, Lindley (1957) argued explicitly that the paradox revealed a shortcoming of the *p*-value, not of the Bayes factor:

Now in our example we have taken situations in which the significance level is fixed because, as explained above, we wish to see whether its interpretation as a measure of lack of conviction about the null hypothesis does mean the same in different circumstances. The Bayesian probability is all right, by the arguments above; and since we now see that it varies strikingly with *n* for fixed significance level, in an extreme case producing a result in direct conflict

with the significance level, the degree of conviction is not even approximately the same in two situations with equal significance levels. 5% in today’s small sample does not mean the same as 5% in tomorrow’s large one. (Lindley, 1957, p. 189)

Similarly, throughout his work on Bayes factor hypothesis testing in the 1930s Harold Jeffreys argued that a rational measure of evidence against the null hypothesis cannot be based on a constant multiple of the standard error, as is the case for the *p*-value. For instance,

“A constant significance limit, in relation to the standard error, would however be equivalent to saying that the prior probability of a zero value varies with the number of observations, which is absurd; or, alternatively, that the chance of a real difference exceeding the standard error is the same no matter how small the standard error is made by increasing the number of observations.” (Jeffreys, 1937, p. 259)

With respect to the question of prior model probabilities, note that the Bayes factor quantifies relative predictive performance of two rival models, and is thus independent of the prior beliefs about the hypotheses (e.g., Wrinch & Jeffreys 1921):

$$\frac{\underbrace{P(\mathcal{H}_1 | y)}_{\text{Posterior beliefs about the models}}}{\underbrace{P(\mathcal{H}_0 | y)}_{\text{Posterior beliefs about the models}}} = \frac{\underbrace{P(\mathcal{H}_1)}_{\text{Prior beliefs about the models}}}{\underbrace{P(\mathcal{H}_0)}_{\text{Prior beliefs about the models}}} \times \frac{\underbrace{p(y | \mathcal{H}_1)}_{\text{Bayes factor}}}{\underbrace{p(y | \mathcal{H}_0)}_{\text{Bayes factor}}} \tag{4}$$

One might argue that in order for the conditioning in the marginal likelihood $p(y | \mathcal{H}_i)$ to be well-defined, each model \mathcal{H}_i must at least have some prior probability $\epsilon > 0$.³ But when the aim is to test \mathcal{H}_0 it appears entirely reasonable for \mathcal{H}_0 to have separate prior mass. Instead, it seems remarkable to assign $\mathcal{H}_0 : \delta = 0$ a prior probability of zero, deeming the hypothesis impossible from the outset, and then proceeding to test it anyway. Why then test $\delta = 0$, and not, say, $\delta = 0.02144347729918\dots$, $\delta = 0.010101010101\dots$, or $\delta = e/30$? The reason is that $\mathcal{H}_0 : \delta = 0$ is a value of special interest: it represents a simple model in which an effect is entirely absent, and for that ‘sampling noise-only’ model to be abandoned the data must offer evidence against it. The Laplacean practice of not assigning the point-null value any separate prior mass, i.e., estimating only, “expresses a violent prejudice against any general law, a totally unacceptable description of the scientific attitude.” (Jeffreys, 1974, p. 1)

³We say *might*, because it could also be argued that the purpose of the Bayes factor is to compare the predictive performance of the sceptic’s \mathcal{H}_0 versus the proponent’s \mathcal{H}_1 . Such statements are about assigning predictive probability to data, and do not entail a commitment to any degree of relative plausibility of the rival positions from which the predictions originated.

Admittedly there exist many problems in which the point-null hypothesis is not believable, not even as an approximation. For instance, there is usually no reason to assume that interrater agreement equals zero exactly, or that two politicians are exactly equally popular. Jeffreys labelled such scenarios problems of estimation. But with problems of testing, a specific value of the parameter stands out; the desire to assess that value (out of an infinity of other values that could be assessed) already suggests that its prior probability is not exactly zero.

At its core, however, the Jeffreys-Lindley paradox is sufficiently general that it does not even require a Bayesian mindset. Good (1980a) (in modern notation) explains it in terms of simple versus simple likelihood ratio tests, as follows:

Dr. Deborah Mayo raised the following question. How could one convince a very naive student, Simplissimus, that a given tail-area probability (P-value), say $1/100$, is weaker evidence against the null hypothesis when the sample is larger? Although this fact is familiar in Bayesian statistics the question is how to argue it without (explicit) reference to Bayesian methods.

One can achieve this aim, without even referring to power functions, in the following manner.

Take a very concrete example, say the tossing of a coin, and count the number $[s]$ of heads (“successes”) in $[n]$ trials. Ask Simplissimus to specify any simple non-null hypothesis for the probability $[\theta]$ of a head. Suppose he gives you a value $[\theta = 0.5 + \epsilon]$. First compute a value of $[n]$ so that a ϵ value of $[s]$ approximately equal to $[n(0.5 + \epsilon/7)]$ would imply a tail-area probability close to $1/100$. Then point out that the fraction $0.5 + \epsilon/7$ of successes is much closer to 0.5 than it is to $0.5 + \epsilon$ and therefore must *support* the null hypothesis as against the specific rival hypothesis proposed by Simplissimus. Thus, for any specified simple non-null hypothesis, $[n]$ can always be made so large that a specified tail-area probability supports the null hypothesis more than the rival one. This should convince Simplissimus, if he had been listening, that the larger is $[n]$ the smaller the set S of simple non-null hypotheses that can receive support (as compared with $[\theta = 0.5]$) in virtue of a specified P-value. If the tail-area probability, for example $1/100$, is held constant, the set S converges upon the point $[\theta = 0.5]$ when $[n]$ is made larger and larger.” (Good 1980a, pp. 307–308; italics in original)

For instance, assume Simplissimus specifies their simple non-null hypothesis as $\theta = 0.57$ with $\epsilon = 0.07$. Then our target value for the number of successes s equals $n(0.5 + 0.07/7) = n \times 0.51$. So for a sample proportion of 0.51

we now seek n such that the two-sided tail area probability equals $.01$. We find that $n = 16700$ —consisting of 8517 heads, for a sample proportion of $s = 8517/16700 = 0.51$, as stipulated—yields a tail area just below $.01$. But the sample proportion of 0.51 is much closer to the null hypothesis (i.e., $\theta = 0.50$) than to the non-null hypothesis specified by Simplissimus (i.e., $\theta = 0.57$).

Returning to an explicit Bayesian perspective, suppose the estimate of the test-relevant parameter is within one standard error from zero. This will provide some evidence in favor of \mathcal{H}_0 . Now assume that sample size increases indefinitely, but the estimate remains within one standard error from zero. As implied by Good’s scenario concerning Simplissimus, the range of parameter values under \mathcal{H}_1 that are consistent with the observed data continually shrinks: with large sample size and an estimate within a standard error of the null value, parameter values far away from zero predict so poorly that they can be effectively excluded from consideration. For any sample size there will be a set of parameter values that are still in contention with the null value, but this set will shrink to zero. Of course, for any specific sample size it is possible to select with the maximum likelihood estimator and claim that the data support it over the null value; for instance, after seeing the data Simplissimus might argue that their non-null choice of 0.57 was overly optimistic, and that a choice of 0.51 would have been more apt. Of course, Good’s game could be played anew, and Simplissimus would be forced to select ever smaller values of θ . In other words, Simplissimus is cherry-picking. In the Bayes factor formalism, such cherry-picking is counteracted by averaging predictive performance across the prior distribution; when sample size increases but the estimate remains within a constant multiple of the standard error from the null value, a growing proportion of the prior distribution under \mathcal{H}_1 will start to predict poorly, driving down the average performance (cf. Jeffreys 1937, pp. 250–251).

Thus, the prior distribution can be viewed as an automatic method to prevent cherry-picking, that is, to prevent the selection of those parameter values that the data happen to support. In the Bayes factor methodology, any prior distribution fulfils this purpose, and will therefore produce the Jeffreys-Lindley paradox. Instead, the FBST *ev* is based on an assessment of the posterior distribution, and therefore lacks the Bayesian correction for cherry-picking. As a result, the evidence against the null hypothesis is overstated; moreover, the FBST *ev* is inconsistent, meaning that the support for \mathcal{H}_0 will not increase without bound if the data accumulate indefinitely and \mathcal{H}_0 is the true data-generating model. The FBST *ev* avoids the Jeffreys-Lindley conflict between a Bayesian measure of evidence and the p -value, but this is not something that needs avoiding — quite the opposite, it is something to embrace.

In sum, to paraphrase Lindley, we claim that the degree of conviction is not even approximately the same in two situations with equal FBST *evs*. An FBST *ev* of .05 in today's small sample does not mean the same as an FBST *ev* of .05 in tomorrow's large one.

Concluding Comments

The FBST *ev* provides a Bayesian hypothesis testing analogue of the frequentist *p*-value. The FBST *ev* is easy to use and it arguably offers several distinct Bayesian advantages (e.g., the result does not depend on the sampling plan). However, we believe that the FBST *ev* falls short on several counts. As detailed above, the FBST *ev* cannot quantify evidence in favor of \mathcal{H}_0 , it is susceptible to sampling to a foregone conclusion, it violates the principle of predictive irrelevance, and it suffers from the Jeffreys-Lindley paradox in the sense that its assessment of evidence is asymptotically equal to a constant multiple of the standard error.⁴ These limitations are fundamental.

In conclusion, we agree with the statement from de Bragança Pereira et al. (2008) that the FBST *ev* is a “genuine Bayesian measure of evidence” in the sense that the FBST *ev* is a genuinely Bayesian procedure; the FBST *ev* is not, however, a genuine measure of evidence.

Acknowledgements The authors would like to thank Scott Brown, John Dunn, Riko Kelter and an anonymous reviewer for their comments on an earlier version of this paper.

Funding This research was supported by the Netherlands Organisation for Scientific Research (NWO; grant #016.Vici.170.083).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3), 1550–1577.
- Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle*, 2nd edn. Hayward (CA: Institute of Mathematical Statistics.
- Berger, J. O., Bernardo, J. M., & Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, 37(2), 905–938.
- Bernardo, J. M. (1980). A Bayesian analysis of classical hypothesis testing (with discussion). *Trabajos de Estadística Y de Investigación Operativa*, 31, 605–647.
- Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2), 113–128.
- Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, 82(397), 106–111.
- Cousins, R. D. (2017). The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese*, 194, 395–432.
- de Bragança Pereira, C. A., & Stern, J. M. (1999). Evidence and credibility: Full Bayesian significance test for precise hypotheses. *Entropy*, 1, 99–110.
- de Bragança Pereira, C. A., & Stern, J. M. (in press). The *e*-value: A fully Bayesian significance measure for precise statistical hypotheses and its research program. *São Paulo Journal of Mathematical Sciences*.
- de Bragança Pereira, C. A., Stern, J. M., & Wechsler, S. (2008). Can a significance test be genuinely Bayesian. *Bayesian Analysis*, 3, 79–100.
- Diniz, M., Pereira, C. A. B., Polpo, A., Stern, J. M., & Wechsler, S. (2012). Relationship between Bayesian and frequentist significance indices. *International Journal for Uncertainty Quantification*, 2(2), 161–172.
- Evans, M. (2015). *Measuring statistical evidence using relative belief*. Boca Raton: CRC Press.
- Feller, W. (1940). Statistical aspects of ESP. *Journal of Parapsychology*, 4, 271–298.
- Feller, W. (1968). *An introduction to probability theory and its applications: Vol. I*, 3rd edn. Hoboken: Wiley.
- Ghosal, S., & van der Vaart, A. W. (2017). *Fundamentals of non-parametric Bayesian inference* Vol. 44. Cambridge: Cambridge University Press.
- Good, I. J. (1980a). The contributions of Jeffreys to Bayesian statistics. In A. Zellner (Ed.) *Bayesian analysis in econometrics and statistics: essays in honor of Harold Jeffreys* (pp. 21–34). Amsterdam: North-Holland Publishing Company.
- Good, I. J. (1980b). The diminishing significance of a *p*-value as the sample size increases. *Journal of Statistical Computation and Simulation*, 11, 307–313.
- Grünwald, P., de Heide, R., & Koolen, W. (2020). Safe testing. arXiv:1906.07801.
- Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, 31, 203–222.
- Jeffreys, H. (1937). *Scientific inference*, 1st edn. Cambridge: Cambridge University Press.
- Jeffreys, H. (1939). *Theory of probability*, 1st edn. Oxford: Oxford University Press.
- Jeffreys, H. (1961). *Theory of probability*, 3rd edn. Oxford: Oxford University Press.
- Jeffreys, H. (1973). *Scientific inference*, 3rd edn. Cambridge: Cambridge University Press.
- Jeffreys, H. (1974). Fisher and inverse probability. *International Statistical Review*, 42, 1–3.
- Jeffreys, H. (1980). Some general points in probability theory. In A. Zellner (Ed.) *Bayesian analysis in econometrics and statistics: essays in honor of Harold Jeffreys* (pp. 451–453). Amsterdam: North-Holland Publishing Company.

⁴This argument is based on the asymptotic equivalence between the *p*-value and the FBST *ev* detailed earlier (Diniz et al., 2012).

- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007), 453–461.
- Kelter, R. (2020a). Analysis of Bayesian posterior significance and effect size indices for the two-sample *t*-test to support reproducible medical research. *BMC Medical Research Methodology*, 20(88).
- Kelter, R. (2020b). How to choose between different Bayesian posterior indices for hypothesis testing in practice. Manuscript submitted for publication. arXiv:2005.13181.
- Kelter, R. (in press). fbst: An R package for the Full Bayesian Significance Test for testing a sharp null hypothesis against its alternative via the *e*-value. *Behavior Research Methods*. arXiv:2006.03332.
- Keyzers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, 23, 788–799.
- LeCam, L. M. (1986). *Asymptotic methods in statistical decision theory*. Berlin: Springer Science and Business Media.
- Li, Y., & Clyde, M. A. (2018). Mixtures of *g*-priors in generalized linear models. *Journal of the American Statistical Association*, 113(524), 1828–1845.
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44, 187–192.
- Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian viewpoint part: Vol. 2. Inference*. Cambridge: Cambridge University Press.
- Ly, A., Marsman, M., Verhagen, A. J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology*, 80, 40–55.
- Ly, A., Stefan, A., van Doorn, J., Dablander, F., van den Bergh, D., Sarafoglou, A., Kucharský, Š., Derks, K., Gronau, Q. F., Gupta, A.kash.R. K. N., Boehm, U., van Kesteren, E.-J., Hinne, M., Matzke, D., Marsman, M., & Wagenmakers, E.-J. (2020). The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the *p*-value hypothesis test. *Computational Brain and Behavior*, 3, 153–161.
- Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016a). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72, 19–32.
- Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016b). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, 72, 43–55.
- Madruca, M. R., Pereira, C. A. B., & Stern, J.M. (2003). Bayesian evidence test for precise hypotheses. *Journal of Statistical Planning and Inference*, 117, 185–198.
- Marsman, M., & Wagenmakers, E.-J. (2017). Three insights from a Bayesian interpretation of the one-sided *p* value. *Educational and Psychological Measurement*, 77, 529–539.
- 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.
- Polya, G. (1954). *Mathematics and plausible reasoning: Vol. II. Patterns of plausible inference*. Princeton: Princeton University Press.
- Pratt, J. W., Raiffa, H., & Schlaifer, R. (1995). *Introduction to statistical decision theory*, MIT Press, Cambridge.
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin and Review*, 21, 301–308.
- van der Vaart, A. W. (1998). *Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics*. Cambridge: Cambridge University Press.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin and Review*, 14, 779–804.
- Wagenmakers, E.-J., Gronau, Q. F., & Vandekerckhove, J. (2018). Five bayesian intuitions for the stopping rule principle. Manuscript submitted for publication.
- Wagenmakers, E.-J., Lee, M. D., Rouder, J. N., & Morey, R.D. (2020). The principle of predictive irrelevance or why intervals should not be used for model comparison featuring a point null hypothesis. In C. W. Gruber (Ed.) *The theory of statistics in psychology – applications, use and misunderstandings* (pp. 111–129): Springer.
- Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1), 60–62. 10.1214/aoms/1177732360.
- Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, 42, 369–390.
- Wrinch, D., & Jeffreys, H. (1919). On some aspects of the theory of probability. *Philosophical Magazine*, 38, 715–731.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.