Computational Brain & Behavior (2022) 5:572–582
https://doi.org/10.1007/s42113-021-00110-5

ORIGINAL PAPER

On the Measure-Theoretic Premises of Bayes Factor and Full Bayesian Significance Tests: a Critical Reevaluation

Commentary to Ly and Wagenmakers

Riko Kelter[1]

Accepted: 11 May 2021 / Published online: 7 July 2021
© The Author(s) 2021

Abstract

The Full Bayesian Significance Test (FBST) and the Bayesian evidence value recently have received increasing attention across a variety of sciences including psychology. Ly and Wagenmakers (2021) have provided a critical evaluation of the method and concluded that it suffers from four problems which are mostly attributed to the asymptotic relationship of the Bayesian evidence value to the frequentist p-value. While Ly and Wagenmakers (2021) tackle an important question about the best way of statistical hypothesis testing in the cognitive sciences, it is shown in this paper that their arguments are based on a specific measure-theoretic premise. The identified problems hold only under a specific class of prior distributions which are required only when adopting a Bayes factor test. However, the FBST explicitly avoids this premise, which resolves the problems in practical data analysis. In summary, the analysis leads to the more important question whether precise point null hypotheses are realistic for scientific research, and a shift towards the Hodges-Lehmann paradigm may be an appealing solution when there is doubt on the appropriateness of a precise hypothesis.

Keywords Full Bayesian significance test · Statistical evidence · Bayes factor · Mixture prior · Point null testing · Hodges-Lehmann-paradigm

Introduction

Over the last two decades, the Full Bayesian Significance Test (FBST) has been developed as a Bayesian counterpart to a frequentist point null hypothesis test (Pereira & Stern, 1999). The FBST and its associated Bayesian evidence value, the e-value, were designed to replace a frequentist hypothesis test while being coherent with the likelihood principle and a fully Bayesian philosophy (Pereira et al., 2008). In their paper, Ly and Wagenmakers (2021) identify four problems of the FBST and conclude that the Bayes factor (BF) avoids these. In this paper, it is shown that the problems are observed only under a measure-theoretic premise which is required for a Bayes factor test, but explicitly avoided under the FBST. This renders the criticism inadequate in practical data analysis and leads to the question whether testing precise hypotheses is realistic in scientific research.

Measure-Theoretic Background—the Bayes Factor

Let $(\varTheta, \mathcal{G})$ be the parameter space (which is a measurable space) parameterizing a statistical model $\varTheta \rightarrow \mathcal{P} : \theta \mapsto P_\theta$, where $\mathcal{P} := \{P_\theta : \theta \in \varTheta\}$ is the family of distributions the statistician deems plausible for describing the observed data. The observed data Y themselves are located in the sample space $(\mathcal{Y}, \mathcal{B})$ (which also is a measurable space), that is, $Y \in \mathcal{Y}$, and in the Bayesian approach a prior distribution $P_\vartheta : \mathcal{G} \rightarrow [0, 1]$ is chosen which maps from the parameter space $\varTheta$ (actually, the associated σ-algebra $\mathcal{G}$) to the values $[0, 1]$. In hypothesis testing, we consider the measurable partition $\{\varTheta_0, \varTheta_1\}$ of $\varTheta$ such that $P_\vartheta(\varTheta_0) > 0$ and $P_\vartheta(\varTheta_1) > 0$. The prior and posterior odds ratios in favour of $\varTheta_0$ are defined as $P_\vartheta(\varTheta_0)/P_\vartheta(\varTheta_1)$ and $P_{\vartheta|Y}(\varTheta_0)/P_{\vartheta|Y}(\varTheta_1)$ respectively (Robert, 2007). The

✉ Riko Kelter
riko.kelter@uni-siegen.de

[1] Department of Mathematics, University of Siegen, Walter-Flex-Street 3, 57072 Siegen, Germany

Bayes factor in favour of Θ_0 is then defined as the ratio of posterior and prior odds

$$\mathrm{BF}_{01} = \frac{P_{\vartheta|Y}(\Theta_0)}{P_{\vartheta|Y}(\Theta_1)} \Big/ \frac{P_\vartheta(\Theta_0)}{P_\vartheta(\Theta_1)} \tag{1}$$

where $P_{\vartheta|Y}$ is the posterior distribution of ϑ given Y.[1] An important condition in the definition of the Bayes factor is that both Θ_0 and Θ_1 receive strictly positive prior probability mass, that is, $P_\vartheta(\Theta_0) > 0$ and $P_\vartheta(\Theta_1) > 0$ both need to hold. Otherwise, it is clear that the prior odds $P_\vartheta(\Theta_0)/P_\vartheta(\Theta_1)$ are either zero (whenever $P_\vartheta(\Theta_0) = 0$ and $P_\vartheta(\Theta_1) > 0$) or not even well-defined (whenever $P_\vartheta(\Theta_0) > 0$ and $P_\vartheta(\Theta_1) = 0$). In the former case, the posterior odds are zero, too, because these can be rewritten as

$$\frac{P_\vartheta(\Theta_0)}{P_\vartheta(\Theta_1)} \cdot \mathrm{BF}_{01} = \frac{P_{\vartheta|Y}(\Theta_0)}{P_{\vartheta|Y}(\Theta_1)} \tag{2}$$

so that no matter what value the Bayes factor $\mathrm{BF}_{01}(y)$ takes for observed data $Y = y$, the posterior odds $P_{\vartheta|Y}(\Theta_0)/P_{\vartheta|Y}(\Theta_1)$ are then zero, too. In the latter case, the posterior odds are not well-defined because one would have to divide by zero. Now, consider a dominated statistical model $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ for sample data $Y \in \mathcal{Y}$ where the parameter space is a subset of the real numbers, $\Theta \subset \mathbb{R}$ (we could also use $\mathbb{R}^n$ for $n \in \mathbb{N}$ to generalize the argument), and more specifically, an interval (a, b) for $a < b \in \mathbb{R}$. Suppose we test the precise point null hypothesis[2]

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0 \tag{3}$$

Now, if the prior P_ϑ is absolutely continuous with respect to Lebesgue measure λ on Θ, the prior odds are zero, and the posterior odds, too. This follows because of the absolute continuity of P_ϑ with respect to λ, as for every Lebesgue-null-set A with $\lambda(A) = 0$ it follows that $P_\vartheta(A) = 0$. However, it is well-known from standard measure theory that lower-dimensional submanifolds and, in particular, countable sets have Lebesgue measure zero (Bauer, 2001). As a consequence, as $H_0 : \theta = \theta_0$ is a countable set with a single element $\theta_0 \in \Theta$, it follows that $\lambda(\Theta_0) = 0$ and thereby $P_\vartheta(\Theta_0) = 0$, where $\Theta_0 := \{\theta_0\}$. The prior odds of H_1 are strictly positive, because $\lambda((a, b)) > 0$ for any interval with $a < b$.[3] From Eq. (2), it follows that the posterior odds $P_{\vartheta|Y}(\Theta_0)/P_{\vartheta|Y}(\Theta_1)$ are then zero, too.[4] Importantly, the Bayes factor BF_{01} is not even well-defined in this case: as the prior odds are zero, and BF_{01} is the ratio of posterior to prior odds which is seen from Eq. (2), we would have to divide by zero in this case to obtain BF_{01}. In this context it is important to note that

> "When the parameter space is uncountable, prior distributions are typically continuous. This means that the prior (and posterior) probability of $\Theta = \theta_0$ is 0." (Schervish (1995), p. 221)

That is, in the majority of continuously parameterized models $\mathcal{P}$ (or equivalently, whenever we have uncountable parameter spaces), the prior distributions P_ϑ are typically absolutely continuous with respect to the Lebesgue measure λ and have a continuous Radon-Nikodým density with respect to λ. Examples include a normal prior, exponential prior, Cauchy prior, or Student-t prior with corresponding λ-densities. Under any of these priors, we cannot test a simple null hypothesis versus its alternative. To circumvent this problem, one is forced to introduce an arbitrary prior probability $\varrho > 0$ that $\theta = \theta_0$ and additionally select a prior distribution $P_\vartheta^{\Theta_1}$ under the alternative hypothesis Θ_1. The introduction of prior probability ϱ is primarily justified to be able to calculate a Bayes factor.[5] Proceeding with the subjective assignment of positive probability ϱ to a Lebesgue-null-set $\{\theta_0\}$, the resulting prior distribution P_ϑ is then a convex combination of a Dirac-measure[6] $\mathcal{E}_{\Theta_0}$ under Θ_0 and the prior distribution $P_\vartheta^{\Theta_1}$ under Θ_1:

$$P_\vartheta(G) := \varrho \cdot \mathcal{E}_{\Theta_0}(G) + (1 - \varrho) \cdot P_\vartheta^{\Theta_1}(G) \tag{4}$$

for all parameter sets $G \subseteq \mathcal{G}$ in the parameter space $(\Theta, \mathcal{G})$.

Even interpreting the Bayes factor as the ratio of marginal likelihoods under H_0 and H_1 does not help, although it seems to avoid the assignment of an explicit prior probability to Θ_0 and Θ_1: equivalent to (1), the BF can be expressed as (compare Robert (2007, p. 227))

$$BF_{01} = \underbrace{\frac{\int_{\Theta_0} f_\theta dP_\vartheta}{\int_{\Theta_1} f_\theta dP_\vartheta}}_{=:(A)} \overset{(1)}{=} \frac{\int_{\Theta_0} f_\theta(y) p_0(\theta) d\lambda}{\int_{\Theta_1} f_\theta(y) p_1(\theta) d\lambda} \overset{(2)}{=} \frac{\int_{\Theta_0} f_\theta(y) p_0(\theta) d\theta}{\int_{\Theta_1} f_\theta(y) p_1(\theta) d\theta} \tag{5}$$

[1] Recall that the parameter is a random variable $\vartheta : (\Omega, \mathcal{A}, P^*) \to (\Theta, \mathcal{G})$ from the product space $(\Omega, \mathcal{A}, P^*)$ where $\Omega := \mathcal{Y} \times \Theta$ is the product space of the data and parameter, $\mathcal{A}$ the associated product-σ-algebra and the joint probability measure P^* is implicitly defined via the selection of prior distribution P_ϑ and the statistical model $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ for the observed data $Y \in \mathcal{Y}$, that is, $Y \sim P_\theta$, compare Schervish (1995).

[2] The below notation implies $\Theta_0 := \{\theta_0\}$ and $\Theta_1 := \Theta \setminus \{\theta_0\}$ and is mainly used because it is widely established.

[3] When $\Theta = \mathbb{R}$, $\lambda(H_1) = \infty$ and the prior odds are zero, too, when adopting the convention $0/\infty := 0$.

[4] This also follows from the fact that the posterior distribution $P_{\vartheta|Y}$ is absolutely continuous with respect to the prior distribution P_ϑ, that is, $P_\vartheta(A) = 0$ implies $P_{\vartheta|Y}(A) = 0$ for $A \subset \Theta$, see Schervish (1995, Theorem 1.31).

[5] However, Schervish (1995) emphasizes that a (possibly) more realistic alternative would be to "replace the hypothesis with (what might be more reasonable) an interval hypothesis of the form $H' : \Theta \in [\theta_0 - \varepsilon, \theta_0 + \varepsilon]$." (Schervish, 1995, p. 221), see also Berger (1985, p. 148).

[6] The Dirac-measure $\mathcal{E}_{\Theta_0}$ is defined as follows: $\mathcal{E}_{\Theta_0}(A) := 1$ for $A \in \Theta_0$ and $\mathcal{E}_{\Theta_0}(A) := 0$ for $A \notin \Theta_0$, for any $A \subseteq \Theta$, where for a precise hypothesis $H_0 := \Theta_0 = \{\theta_0\}$ the former simplifies to $\mathcal{E}_{\Theta_0}(A) := 1$ for $A = \theta_0$ and $\mathcal{E}_{\Theta_0}(A) := 0$ for $A \neq \theta_0$.

where equality (1) follows from denoting the Lebesgue-density of the prior distribution P_ϑ on Θ_i as $p_i := dP_\vartheta/d\lambda$ for $i = 0, 1$, and equality (2) by writing integration of θ with respect to λ as usual as $d\theta$. Now, from (A) in Eq. (5), it is immediate that whenever Θ_0 is a P_ϑ-null-set, the Bayes factor BF_{01} will be zero because $\int_{\Theta_0} f_\theta dP_\vartheta = 0$ then. Thus, the Bayes factor cannot provide any evidence anymore for Θ_0.[7] While the definition of the BF could easily be adapted to incorporate this scenario, the more alerting situation occurs when instead $BF_{10} = 1/BF_{01}$ is calculated for a P_ϑ-null-set $\Theta_0 \in \Theta$: Again, from (A) in Eq. (5), it follows that BF_{10} has a singularity in Θ_0 then, and one would arrive at $BF_{10} = 1/0$, causing the BF not to be well-defined anymore. This is the reason why by definition, the Bayes factor is only defined for measurable partitions $\{\Theta_0, \Theta_1\}$ with both $P_\vartheta(\Theta_0) > 0$ and $P_\vartheta(\Theta_1) > 0$ (Schervish, 1995, p. 220–221). Thus, it is crucial that the prior distribution P_ϑ assigns positive mass to Θ_0 for the BF to be well-defined, and only then does Eq. (5) provide the usual interpretation of the Bayes factor as the ratio of marginal likelihoods under $H_0 := \Theta_0$ and $H_1 := \Theta_1$.

Also, the more familiar-looking representation $BF_{01} = \frac{f_{\theta_0}(y)}{\int_{\Theta_1} f_\theta(y) p_1(\theta) d\theta}$ of the Bayes factor for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ holds if and only if a mixture prior with Dirac-measure component $\mathcal{E}_{\theta_0}$ for the value θ_0 is assigned to the parameter with positive mixing weight $\varrho > 0$ (Robert, 2007, p. 231). Robert phrases this as follows: “Testing of point-null hypotheses and the like thus impose a drastic modification of the prior distribution ... This modification of the prior is puzzling from a measure-theoretic point of view, since it puts some weight on a set previously of measure 0.” (Robert, 2007, p. 229), see also Berger (1985, p. 148).

The above line of thought clarifies why the Bayes factor is only defined for hypotheses Θ_0 and Θ_1 which receive strictly positive prior mass. Importantly, it plays no role for this requirement whether the BF is interpreted as the ratio of posterior to prior odds or as the ratio of marginal likelihoods of both hypotheses under consideration.

Measure-Theoretic Background—the FBST

The idea of the FBST is to use the e-value $\overline{\text{ev}}(H_0)$, which quantifies the Bayesian evidence against H_0 as a Bayesian replacement of the frequentist p-value. Importantly, the FBST was designed as a Bayesian replacement of frequentist point null tests and thus should be capable only to reject a point null hypothesis. In the FBST, the Bayesian surprise function $s(\theta) := p(\theta|y)/r(\theta)$ normalizes the posterior density $p(\theta|y)$ by a reference function $r(\theta)$. Possible choices include a flat reference function $r(\theta) := 1$ or any prior probability density $p(\theta)$ for the parameter θ. In most settings, $r(\theta)$ will be the probability density of an absolutely continuous probability distribution with respect to the Lebesgue measure λ, as the introduction of a mixture prior as detailed above for the BF is not required for the FBST. As a consequence, the FBST does not separate between hypothesis testing and estimation with regard to the choice of prior distribution, and is not forced to assign positive probability ϱ to a Lebesgue-null-set $\{\theta_0\}$. The calculations of the e-value are explicitly possible under absolutely continuous priors: s^* is defined as the maximum of the surprise function $s(\theta)$ over the null set Θ_0 which belongs to the hypothesis H_0, that is, $s^* := s(\theta^*) = \max_{\theta \in \Theta_0} s(\theta)$. Thus, for a precise hypothesis $H_0 := \Theta_0 := \{\theta_0\}$ the maximum s^* is given as $s(\theta_0)$.

The tangential set $\overline{T}(\nu)$ to the hypothesis H_0 is defined as $\overline{T}(\nu) := \Theta \setminus T(\nu)$ where $\Theta \setminus T(\nu)$ is the set complement of Θ and $T(\nu)$. In the above, $T(\nu) := \{\theta \in \Theta | s(\theta) \leq \nu\}$. $T(s^*)$ includes all parameter values θ with surprise smaller or equal to the maximum s^* of the surprise function under the null set, and $\overline{T}(s^*)$ includes all parameter values θ which attain a surprise larger than the maximum s^* of the surprise function under the null set. The cumulative surprise function $W(\nu)$ is defined as

$$W(\nu) := \int_{T(\nu)} p(\theta|y) d\theta \tag{6}$$

and setting $\nu := s^*$, $W(s^*)$ is the integral of the posterior density $p(\theta|y)$ over $T(s^*)$. The Bayesian e-value, which measures the evidence *against* the null hypothesis H_0, is then defined as $\overline{\text{ev}}(H_0) := \overline{W}(s^*)$ where $\overline{W}(\nu) := 1 - W(\nu)$. The null hypothesis H_0 is rejected when the evidence against it is large, as then H_0 traverses low posterior density (respectively surprise) regions.

The crucial difference to the BF is that the FBST does not force the introduction of an arbitrary prior probability mass ϱ on the point null value $\{\theta_0\}$. Thus, the FBST is coherent with a Bayesian parameter estimation perspective (which would not accept the assignment of arbitrary probability mass to a point null value θ_0) and does not separate Bayesian hypothesis testing and parameter estimation. The use of priors which are absolutely continuous with respect to the Lebesgue measure λ is explicitly possible and a standard choice when adopting the FBST.[8]

[7] Note that because $\Theta_1 := \Theta \setminus \Theta_0$ (where $\setminus$ denotes the set complement, that is $A \setminus B$ includes all elements which are located in A but not in B), the set Θ_1 has P_ϑ-measure one, because $P_\vartheta(\Theta_1) = P_\vartheta(\Theta \setminus \Theta_0) = P_\vartheta(\Theta) - P_\vartheta(\Theta_0) = 1 - 0 = 1$. Thus, $BF_{01} = 0/1 = 0$ in Eq. (5) when $P_\vartheta(\Theta_0) = 0$.

[8] The FBST could also be used with a mixture prior which is necessary to calculate the BF, but there is no need to assume such a prior. In general, the FBST is compatible both with absolutely continuous priors and non-absolutely continuous priors with regard to λ.

The e-value as an Approximate p-value

The first two problems identified by Ly and Wagenmakers (2021) are connected to the alleged relationship of the e-value to the frequentist p-value. As an example where the e-value and p-value coincide, they provide the example of Diniz et al. (2012) who assumed normally distributed data $\mathcal{N}(\mu, 1)$ and concluded that under a uniform prior the two quantities become exactly equal. However, as Diniz et al. (2012) stressed:

> "This result is a consequence of the fact that the normal density depends on t and μ only on $(t-\mu)^2$." (Diniz et al. (2012), p. 162)

Diniz et al. (2012) considered a Gauß test for normal data with unknown mean $\mu \in \mathbb{R}$ and variance $\sigma^2 = 1$, and assumed that the sample mean $t = 3.9$ is observed for $n = 3$ observations. The null hypothesis $H_0 := \{(\mu, \sigma^2) : \mu = 5, \sigma^2 \in \mathbb{R}_+\}$ is considered. The standard error of μ under H_0 is $\sigma/\sqrt{n}$ — compare Held and Sabanés-Bové (2014, Example 3.4) — and thus the sampling density of the statistic t (not the original data, which we can replace with the minimal sufficient statistic t instead) under H_0 is a $\mathcal{N}(5, 1/3)$ density. The density of the posterior distribution based on t under an improper flat prior $p(\mu) = 1$ is also a normal density $\mathcal{N}(3.9, 1/3)$, because the likelihood $L(\mu) \propto \exp[-\frac{n}{2\sigma^2}(\mu - t)^2]$, and the latter is a kernel of a $\mathcal{N}(3.9, 1/3)$ density for observed $t = 3.9$ and $\sigma^2 = 1$. Now, the two-sided p-value results in $P_{\mu=5,\sigma^2=1/3}(t < 3.9) \cdot 2 = 0.0567$, which is twice the tail-area of $t < 3.9$ of the sampling density under H_0 (we could equivalently compute $P_{\mu=5,\sigma^2=1/3}(t < 3.9) + P_{\mu=5,\sigma^2=1/3}(t > 6.1)$). The e-value $\text{ev}(H_0)$ in favour of H_0 results in $P_{\vartheta=(3.9,1/3)|Y}(\mu > 5) \cdot 2 = 0.0567$ which is twice the tail area of $\mu > 5$, where 5 is the value under H_0 (we could equivalently compute $P_{\vartheta=(3.9,1/3)|Y}(\mu > 5) + P_{\vartheta=(3.9,1/3)|Y}(\mu < 2.8)$). Thus, we arrive at $p = \text{ev}(H_0)$, and the only reason for this coincidence is that the tail probabilities we compute for p and $\text{ev}(H_0)$ depend only on the distance $(t-\mu)^2$ because the variances are identical in the likelihood $\mathcal{N}(5, 1/3)$ and posterior $\mathcal{N}(3.9, 1/3)$ (equivalently, omit the quadratic term and replace it by $|t - \mu|$, because we are interested in the absolute differences between the true parameter and t). In both cases, we compute tail probabilities which only depend on the difference $|t - \mu| = 1.1$ (or equivalently $(t-\mu)^2 = 1.1^2$) and thus $p = \text{ev}(H_0)$. Very informally speaking, the "bell-shape" of the normal distribution is equal for identical variances in the likelihood and posterior, and the tail probabilities of both coincide when the only difference are the means, which are shifted. However, as soon as we use a proper prior, the relationship disappears.

Also, it is well-known that testing based on improper priors quickly leads to problems, which is why some authors have argued that "improper priors should not be used *at all* in tests." (Robert, 2007, p. 232), compare Degroot (1973). Widening their criticism from a single model to the general case, Ly and Wagenmakers (2021) argue that the asymptotic results of Diniz et al. (2012) pose a more serious problem, while admitting that for "non-uniform priors the relation between FBST ev and p-value is only approximate" (Ly & Wagenmakers, 2021, p. 5). Diniz et al. (2012) showed based on the Bernstein-von-Mises-Theorem (van der Vaart, 1998, Section 10) that for large samples sizes n

$$\text{ev}(H_0) \approx 1 - F_m[F^{-1}_{m-h}(1-p)] \qquad (7)$$

where p is the frequentist p-value, $m := \dim(\Theta)$, $h := \dim(\Theta_0)$ and F_m is the cumulative distribution function (c.d.f.) of the χ^2_m distribution and F^{-1} denotes the generalized inverse c.d.f. Ly and Wagenmakers (2021) argue that whenever $\dim(\Theta_0) = 0$ and $\dim(\Theta) = k$ for $k \in \mathbb{N}_0$, Eq. (7) implies that $\text{ev}(H_0) \approx 1 - F_m[F^{-1}_{m-h}(1-p)] = 1 - F_k[F^{-1}_k(1-p)] = 1 - (1-p) = p$, so that for large sample sizes n, the p-value and e-value become equal. Admittedly, the binomial example chosen by Ly and Wagenmakers (2021) exploits this condition as $\dim(\Theta_0) = 0$ ($H_0 : \theta = \theta_0$ is a single point which has dimension zero) and $\dim(\Theta) = 1$ which shows that at least in univariate models, the asymptotic relationship seems to hold.[9] The simulations in Diniz et al. (2012) confirm this relationship, but they also show that even under perfectly controllable simulation settings, there remain differences: they try to model the functional relationship between p-value and e-value via a Beta c.d.f., and results show that the asymptotic relationship (not equality!) starts to hold only for $n \geq 1000$ observations in each group in some models (Diniz et al., 2012, Figure 8).

However, the more important aspect is that even when $\dim(\Theta_0) = 0$ the relationship is, in general, a purely theoretical result. The Bernstein-von-Mises theorem assumes that the sample $X_1, X_2, ...$ is generated i.i.d. from P_{θ_0} from some $\theta_0 \in \Theta$ for the dominated and parameterized statistical model $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$. However, under an absolutely continuous prior P_ϑ with regard to λ on θ, the probability of the parameter taking the value θ_0 is zero: $P_\vartheta(\theta_0) = 0$, because $\lambda(\theta_0) = 0$. By Definition of a sharp hypothesis as a submanifold[10], this holds for *any* such

[9] Note that this still excludes the majority of widely used models, for example, standard tests like Student's t-test or the test for coefficients in a regression model, or in general, any test where there is at least a second parameter, which ensures that $\dim(\Theta_0) > 0$.

[10] See Definition 1 in Diniz et al. (2012) and Definition 2.1 in Pereira et al. (2008).

hypothesis $H_0 := \Theta_0 := \{\theta_0\}$. Thus, the critical condition that H_0 is true, under which the result of Diniz et al. (2012) is established, has probability zero under any prior which is absolutely continuous with respect to Lebesgue measure λ. As a consequence, data is generated with probability zero from P_{θ_0}, where θ_0 is the value specified in $H_0 := \Theta_0 := \{\theta_0\}$.[11] As this is the usual prior choice in continuous models $\mathcal{P}$ in the FBST, even in univariate models, the asymptotic relationship (not equality) identified by Diniz et al. (2012) will be observed with probability zero in practice. That is, Eq. (7) holds with probability zero under this prior choice. As Cohen stressed, the null hypothesis

> "taken literally (and that's the only way you can take it in formal hypothesis testing) is always false. It can only be true in the bowels of a computer processor running a Monte Carlo study."
> Cohen (1990, p. 1308)

Although the relationship identified by Diniz et al. (2012) is interesting from a theoretical perspective, it will only be observed in the laboratory conditions of a computer processor. Furthermore, even *then* the relationship does not hold whenever the assumptions of the Bernstein-von-Mises theorem are violated, in particular, in *all* models $\mathcal{P}$ with discrete parameters. In summary, in continuous models the relationship in Eq. (7) thus (1) holds only in univariate models and (2) occurs with probability zero outside of a Monte Carlo simulation with perfectly controllable conditions, whenever an absolutely continuous prior is assumed. Point (2) thus resolves the problem also for univariate models. In discrete models, Eq. (7) does not hold at all, even without the above arguments.

Note, however, that from a measure-theoretic point of view when testing with Bayes factors, the last criticism stays valid at least in univariate models $\mathcal{P}$: the point null value θ_0 then has positive measure as specified in the Dirac component of the mixture prior, compare Eq. (4), and the Bernstein-von-Mises theorem can be applied to recover the asymptotic relationship. However, unless the mixing weight is chosen as $\varrho = 1$ in the mixture prior, the probability of obtaining an i.i.d. sample $X_i \sim P_{\theta_0}$ goes to zero for large enough sample size $n \to \infty$. Thus, even under a mixture prior the assumption of an i.i.d. sample cannot be reconciled with the Bayesian approach. However, when taking a frequentist measure-theoretic stance, θ is fixed but unknown so it becomes meaningful again when saying the sample $X_1, X_2, ...$ is i.i.d. $\sim P_{\theta_0}$.[12] This is also the assumption in the law of the iterated logarithm proven in Feller (1968, p. 204–205), which shows that the assertion that "the proofs on sampling to a foregone conclusion in Feller (1970) also pertain to the FBST procedure" Ly and Wagenmakers (2021, p. 6) does not hold, as the i.i.d. assumption $X_1, X_2, ... \sim P_{\theta_0}$ under an absolutely continuous prior P_ϑ – which is the standard choice in the FBST – is not valid.

Revisiting Problem 1: Quantifying Evidence in Favour of the Null Hypothesis

Now, the first problem in Ly and Wagenmakers (2021) is revisited. It is argued that a defect of the FBST and e-value is that it cannot quantify evidence in favour of the null hypothesis $H_0 : \theta = \theta_0$. However, from the above measure-theoretic analysis it is clear that this is a mere consequence of the FBST not assigning arbitrary probability mass ϱ to Θ_0 as is assigned by the mixture prior in Eq. (4). Under absolutely continuous priors, Θ_0 has zero prior probability $P_\vartheta(\Theta_0)$, so it cannot have positive posterior probability $P_{\vartheta|Y}(\Theta_0)$. Thus, acceptance of $H_0 := \Theta_0$ is not possible. However, the FBST was designed as a Bayesian replacement of frequentist point null tests and thus should be capable only to reject a point null hypothesis. Also, the mixture prior in Eq. (4) was historically chosen to become able to confirm general laws (Wrinch & Jeffreys, 1921; Etz & Wagenmakers, 2015). Thus, the introduction of positive probability ϱ for the value θ_0 (which in a general law referred to a boundary of the parameter space like $\theta = 1$ in a binomial experiment) was due to the fact that

> "... Broad used Laplace's theory of sampling, which supposes that if we have a population of n members, r of which may have a property φ, and we do not know r, the prior probability of any particular value of r (0 to n) is $1/(n+1)$. Broad showed that (...) if we take a sample of number m and find all of them with φ, the posterior probability that all n are φ's is $(m+1)/(n+1)$. A general rule would never acquire a high probability until nearly the whole of the class had been sampled. We could never be reasonably sure that apple trees would always bear apples (...). The result is preposterous, and started the work of Wrinch and myself in 1919-1923."
> Jeffreys (1980, p. 452)

[11] Note that from the Bayesian perspective, the sample is not i.i.d., but only i.i.d. conditional upon having observed a parameter value θ. Unconditionally, data is distributed as the marginal distribution (the prior predictive distribution) of the data, $P^\vartheta(B) := \int_\Theta P_\theta(B) dP_\vartheta$ for all $B \in \mathcal{B}$, where $(\mathcal{Y}, \mathcal{B})$ is the sample space. Denote $B := y \in \mathcal{Y}$ as the observed data, then $P_\theta(B) := P_\theta(y) = f(y|\theta)$ is the value of the λ-density of $P_\theta \in \mathcal{P}$ and denote $p(\theta) := dP_\vartheta / d\lambda$ as the prior density with respect to the Lebesgue measure λ. Then, $P^\vartheta(B)$ is the more familiar looking marginal likelihood $P^\vartheta(B) = f(B) = f(y) = \int_\Theta f(y|\theta) p(\theta) d\theta$ of the observed data $B = y \in \mathcal{Y}$.

[12] This latter point shows that the Bernstein-von-Mises theorem primarily provides an asymptotic justification of Bayesian methods for frequentists, see Edwards (1992).

While the mixture prior structure renders the statistician able to confirm general laws, it is today commonly used for what could be called "arbitrary laws": For example, when testing for a difference between two groups in a clinical trial, "it will virtually never be the case that one seriously entertains the possibility that $\theta = \theta_0$ exactly (c.f. Hodges and Lehmann (1954)." (Berger, 1985, p. 148), where θ could model the effect size between the treatment and control group. Thus, whenever no general law like all apple trees bear apples or all swans are white are considered, it remains questionable to impose the mixture prior.[13]

Importantly, next to the inadequate use of the mixture prior in situations where no general law is tested, the assignment of positive mass $\varrho > 0$ to θ_0 separates the prior beliefs for hypothesis testing and parameter estimation (where for the latter task an absolutely continuous prior like a normal prior often is much more reasonable than a mixture prior), while the FBST does not separate testing and estimation.[14] Importantly, the FBST does not even *need* to be able to quantify evidence *in favour* of H_0 by measure-theoretic design. From the above, we know that H_0 has probability zero under an absolutely continuous prior P_ϑ with respect to λ. As the posterior $P_{\vartheta|Y}$ is absolutely continuous with respect to the prior P_ϑ (Schervish, 1995, Theorem 1.31), the posterior probability of H_0 will be zero, too. Thus, quantifying evidence in favour of H_0 becomes obsolete under the measure-theoretic assumptions of the FBST.[15] Admittedly, one might argue what the use is to test H_0 if it is a priori is known to have probability zero. However, the e-value quantifies the discrepancy between observed data and H_0, and the FBST rejects a null hypothesis only when the data display sufficient evidence against it. The BF bypasses this issue via the assignment of probability ϱ to H_0, but this does not render H_0 more realistic in practice. The relevance of the ability of the BF to quantify evidence in favour for H_0 may thus be overstated, because in practice, the parameter value is most probably *not* exactly equal to θ_0. Thus, for large enough sample size n, the BF will reject H_0, too, even if the true parameter is $\theta_0 + \varepsilon$ for a tiny ε and thus the choice $H_0 : \theta = \theta_0$ was very close to the true parameter $\theta = \theta_0 + \varepsilon$. Thus, from this perspective, both the FBST and BF experience a similar form of sampling to a foregone conclusion which could be called sampling to trivial effect sizes. Admittedly, the FBST is somewhat begging the question by not assigning prior mass to $H_0 : \theta = \theta_0$ and being only able to reject H_0. The BF is, however, similarly begging the question by assuming from the outset that H_0 is true with probability $\varrho > 0$ a priori, and as a consequence being able to confirm $H_0 : \theta = \theta_0$ then.

Revisiting Problem 2: Susceptibility to Sampling to a Foregone Conclusion

Revisit problem 2: Based on the asymptotic relationship to the frequentist p-value it is argued that the FBST will sample to a foregone conclusion. Thus, by only collecting enough samples one will eventually produce an e-value $\text{ev}(H_0)$ which rejects $H_0 : \theta = \theta_0$. However, the measure-theoretic premises circumvent this caveat again. While under a frequentist perspective, $H_0 : \theta = \theta_0$ may be true and from a Bayes factor perspective, θ in $H_0 : \theta = \theta_0$ has positive prior probability (compare Eq. (4)), under an absolutely continuous prior P_ϑ with respect to λ used in the FBST $P_\vartheta(H_0) = 0$ holds. Therefore, sampling to a foregone conclusion will not occur as data is generated as i.i.d. from P_{θ_0} with probability zero, except inside the bowels of a computer processor running a Monte Carlo simulation, to recite Cohen (1990). Under the mixture prior which is used for the Bayes factor, the problem can occur as H_0 has positive prior probability. However, assuming such a prior is simply not necessary when using the FBST, and thus the problem vanishes when choosing an absolutely continuous prior with respect to λ.[16] Even when adopting a mixture prior, unless $\varrho = 1$ (which implies the prior probability of θ_0 is one, and all uncertainty vanishes) for large enough $n \in \mathbb{N}$ the probability of sampling $X_1, ..., X_n$ i.i.d. as $X_i \sim P_{\theta_0}$ goes to zero, compare Eq. (4). Thus, even under the measure-theoretic assumptions of the mixture prior which is used in the Bayes factor test the problem vanishes.[17]

[13] Note also that a small interval hypothesis is clearly not more realistic than a point null hypothesis in the case a *general* law is tested. For example, testing $H_0 : \theta = 1$ is in sharp contrast to $\tilde{H}_0 : \theta \in [0.95, 1]$ and the interpretation is radically different (e.g. now only 95–100% of swans are white when confirming $\tilde{H}$ compared to all swans are white when confirming H_0).

[14] Which is again close to the frequentist paradigm, where for example confidence intervals are obtained by inverting hypothesis tests (Robert, 2007).

[15] As a sidenote, the convergence rate under H_0 and H_1 for the Bayes factor differ, which is why solutions like non-local priors have been introduced to establish exponential convergence rates under both hypotheses, see Johnson and Rossell (2010). The faster convergence of the e-value under H_1 compared to the BF, see Kelter (2020), is a consequence of the FBST not assigning prior probability to H_0.

[16] Importantly, this argument shows that the criticism of sampling to a foregone conclusion pertains to p-values, as the assumption $X_1, X_2, ... \sim P_{\theta_0}$ remains possible: No probability statements about θ_0 being the true parameter can be made by frequentists and one can assume such an i.i.d. sampling mechanism.

[17] When n is only moderate, convergence in the Bernstein-von-Mises theorem is questionable because it is an asymptotic result (van der Vaart, 1998). Thus, no sampling to a foregone conclusion occurs under a mixture prior even in moderately sized samples.

Revisiting Problem 3: the Principle of Predictive Irrelevance

The third problem concerns the principle of predictive irrelevance which goes back to Jeffreys (1961). The binomial example used by Ly and Wagenmakers (2021) shows that the FBST is, in general, not predictively matched. However, although predictive matching is an appealing property, it is at best loosely related to more profound principles of statistical inference (Berger and Wolpert, 1988). Furthermore, predictive matching depends on the definition of a sample. Suppose that in the experiment, tupels (n_1, n_2) are observed (e.g. simultaneous sensor measurements) instead of single measurements n for each step in the sequence. Observing an uninformative tupel which consists of a success and a failure based on the symmetric beta posterior will result in another symmetric beta posterior, leaving $\text{ev}(H_0)$ unchanged at $\text{ev}(H_0) = 1$. The fallacy is to interpret $\text{ev}(H_0) = 1$ as evidence in favour for H_0, while it actually only implies the weakest possible evidence *against* H_0. When the entire posterior has posterior surprise values which are smaller or equal to the surprise value under H_0, there is essentially no evidence against H_0. Thus, $\text{ev}(H_0) = 1$. As the FBST aims at rejection of H_0 (and not confirmation of H_0), the concept of predictive matching is not helpful. Any modification of the posterior which causes the surprise function of the null value to shift away from the mode indicates (and should indicate) some magnitude of evidence against H_0. More generally, Robert (2016) argued that

> "predictive matching is not a well-defined concept. That both predictives take the same values for such "completely uninformative" data thus sounds more like a post-hoc justification than a way of truly calibrating the Bayes factor."
> Robert (2016, p. 3)

Thus, as the FBST cannot accept H_0, it cannot be calibrated in this way. In fact, predictive matching is important whenever one commits the "all too common sin of assuming that θ can be assigned the same prior distribution, $\pi(\theta)$, under H_0 and H_1." Berger and Guglielmi (2001, p. 177). A partial justification of such an assumption is thus given by predictive matching, see Berger & Guglielmi (2001, p. 177) as then a completely uninformative "sample of size m should always yield a Bayes factor of 1 (implying that the two models are equally supported by the data)" (Berger & Guglielmi, 2001, p. 177), which justifies the assignment of identical priors under H_0 and H_1.[18] Thus, predictive matching post-hoc justifies the selection of identical priors (for nuisance parameters) under H_0 and H_1, but in the FBST, we do not need separate priors for nuisance parameters because in the FBST there is only single prior and no separate priors under H_0 and H_1 exist. Thus, there is no need for predictive matching after all. Ultimately, predictive matching is an important concept for the BF, but a controversial one as the BF itself is not always predictively matched, see Gronau et al. (2019), Wang and Liu (2016), and Berger and Guglielmi (2001).

Revisiting Problem 4: the Jeffreys-Lindley Paradox

Problem four is formulated as the FBST avoiding the Jeffreys-Lindley paradox. It is argued that "the FBST ev is based on an assessment of the posterior distribution, and therefore, lacks the Bayesian correction for cherry-picking" (Ly & Wagenmakers, 2021, p. 11), where the correction is the prior distribution which prevents the selection of parameter values that the data happen to support. However, the implicit premise is that a flat reference function $r(\theta) = 1$ is used (which is equal to an improper prior), so that the surprise function results in the posterior density. Whenever a proper absolutely continuous prior (e.g. normal prior, Cauchy prior) is used for $r(\theta)$, the correction of the prior applies as it prevents the inclusion of parameter values inside the tangential set that the data happen to support. The assumption is similar to the flat prior assumption made in problem 1 for the binomial test, and it shows that improper priors can be problematic for the FBST. However, the ultimate question is what can be learned from the Jeffreys-Lindley paradox, and as argued by Robert (2014) "divergences between different statistical theories of inference and their numerical conclusions are to be expected", and even more importantly, the Jeffreys-Lindley-paradox "points at the poor (and even unacceptable) behaviour of improper prior distributions when testing point-null hypotheses" (Robert, 2014, p. 2).

In fact, the Jeffreys-Lindley paradox can be interpreted as the consequence of the failing approximation of a small interval hypothesis through a precise null hypothesis. When replacing the precise hypothesis $H_0 : \theta = \theta_0$ with a small interval hypothesis $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$ for $b > 0$, the Jeffreys-Lindley paradox does not occur because the approximation is rendered unrealistic before the paradox blends in. For illustration purposes, an example of Berger (1985) is used: Suppose a sample $X_1, ..., X_n$ is observed from a $\mathcal{N}(\theta, \sigma^2)$ distribution with known σ^2. The observed likelihood function is then proportional to a $\mathcal{N}(\bar{x}, \sigma^2/n)$

[18] Compare also the intrinsic Bayes factor approach of Berger and Pericchi (1996).

density for θ, and given that we really should be testing $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, we need to know when it is suitable to approximate H_0 by $H_0 : \theta = \theta_0$. The "only sensible answer to this question is – the approximation is reasonable if the posterior probabilities of H_0 are nearly equal in the two situations." (Berger, 1985, p. 149) and this happens when the observed likelihood function is approximately constant on $(\theta_0 - b, \theta_0 + b)$, because then the posterior probabilities will be equal when b is small enough. Berger (1985) showed that the likelihood function varies by no more than 5% on $(\theta_0 - b, \theta_0 + b)$ if

$$b \leq (0.024) z^{-1} \sigma / \sqrt{n} \tag{8}$$

where $z = \sqrt{n}|\bar{x} - \theta_0|/\sigma$ is the classical test statistic for a Gauß test. Turning to the Jeffreys-Lindley paradox now, when we assign a $\mathcal{N}(\mu, \tau^2)$ density under $H_1 : \theta \neq \theta_0$ to θ, and set $\mu := \theta_0$, the posterior probability of $H_0 : \theta = \theta_0$ can be computed as

$$P_{\vartheta|Y}(H_0) = \left[1 + \frac{1-\pi_0}{\pi_0} \cdot \frac{\exp[(\bar{x}-\theta_0)^2 n\tau^2/(2\sigma^2[\tau^2 + \sigma^2/n])]}{(2\pi\sigma^2/n)^{-1/2} \exp[-(\bar{x}-\theta_0)^2/(2\sigma^2/n)]}\right]^{-1} \tag{9}$$

for details see Berger (1985, p. 150). Now, suppose a fixed z is observed, which for example for $z = 1.960$ corresponds to a two-sided p-value of $p = 0.05$. Berger (1985) provides the posterior probabilities (9) for varying sample sizes n and these start at $P_{\vartheta|Y}(H_0) = 0.35$ for $n = 1$ and then grow steadily to $P_{\vartheta|Y}(H_0) = 0.80$ for $n = 1000$ (Berger, 1985, p. 151, Table 4.2). In fact, "this phenomenon that $\alpha_0 \rightarrow 1$ as $n \rightarrow \infty$ and the P-value is held fixed, actually holds for virtually any fixed prior and point null testing problem." Berger (1985, p. 156), where α_0 equals the posterior probability $P_{\vartheta|Y}(H_0)$ of $H_0 : \theta = \theta_0$. As a consequence of $P_{\vartheta|Y}(H_0) \rightarrow 1$ for fixed z (or p-value) under $n \rightarrow \infty$, it follows that $BF_{01} \rightarrow \infty$. Now, the reason of the Jeffreys-Lindley paradox occurring (that is, $p = 0.05$ seems to reject H_0 while $BF_{01} \rightarrow \infty$ for $n \rightarrow \infty$) is that the assumption of a precise null hypothesis simply is not feasible for even moderately large n:

> "This large n phenomenon provides an extreme illustration of the conflict between classical and Bayesian testing of a point null. One could classically reject H_0 with a P-value of 10^{-10}, yet, if n were large enough, the posterior probability of H_0 would be very close to 1. This surprising result has been called "Jeffreys' paradox" and "Lindley's paradox," ... We will not discuss this "paradox" here because the point null approximation is rarely justifiable for very large n."
>
> Berger (1985, p. 156)

Thus, when n is small, the Jeffreys-Lindley paradox does not occur for a fixed z (or p-value). When we let $n \rightarrow \infty$, the Jeffreys-Lindley paradox blends in as $P_{\vartheta|Y}(H_0) \rightarrow 1$ and the p-value remains constant, but then the validity of our approximation of the more realistic interval hypothesis $(\theta_0 - b, \theta_0 + b)$ by $H_0 : \theta = \theta_0$ breaks. In fact, it breaks even for moderate sample sizes. For the situation $n = 1$ and $z = 1.960$ under $\sigma = 1$, we arrive at the posterior probability $P_{\vartheta|Y}(H_0) = 0.35$, which is not in conflict with $p = 0.05$, but only somewhat weaker evidence against H_0 than assured by the p-value. Increasing sample size to $n = 50$, the paradox blends in as then $P_{\vartheta|Y}(H_0) = 0.52$, so that the BF concludes to favour H_0 (assuming prior weights of 0.5 for both H_0 and H_1) while $p = 0.05$ states evidence against $H_0 : \theta = \theta_0$. However, for $n = 50$, we have $b \leq 0.0017$ based on Eq. (8) and we would have to accept the tiny interval hypothesis $(\theta_0 - 0.0017, \theta_0 + 0.0017)$ as the hypothesis we actually want to test, for $H_0 : \theta = \theta_0$ to be a reasonable approximation. Thus, even for moderate samples sizes like $n = 50$, the innocuous looking approach to approximate a realistic interval hypothesis via a precise hypothesis can become untrustable unless one has extremely good reasons to assume a tiny interval hypothesis. In the majority of research, we will not be willing to accept such a precise interval hypothesis. Thus, as the point null approximation is rarely justifiable for large n, it matters little that the Jeffreys-Lindley paradox is observed. In these situations, we will already hesitate to trust the test of a precise hypothesis because the approximation has become unreliable.

The reason that the FBST does not suffer from the paradox is simply due to the fact that it does *not* make use of an approximation. Whereas the prior probability $\varrho > 0$ in the mixture prior used for a Bayes factor test is conceptualized as the prior probability assigned to the more realistic interval hypothesis $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, the FBST does not assign mass to $H_0 : \theta = \theta_0$. However, the prior probability of the more realistic interval hypothesis $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, of course, has positive prior probability under the absolutely continuous prior P_ϑ (with respect to the Lebesgue measure λ) which is employed in the FBST. Thus, no paradox occurs.[19]

[19] Importantly, the paradox also does not occur when the approximation is checked in a Bayes factor test. As outlined above, when the approximation imposes an unrealistically high precision for $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, the paradox blends in. However, then one would not be willing to trust a Bayes factor test because the approximation becomes unreliable.

Admittedly, the FBST cannot make it easier for an a priori highly unprobable hypothesis to be rejected. In contrast, the Bayes factor approach allows to balance the prior probability $\varrho > 0$ accordingly to incorporate such a priori scepticism.[20] However, while the FBST does not allow for such a straightforward incorporation of lower prior probability, modifying the prior distribution P_ϑ which is used for the FBST allows for such a Bayesian calibration. In fact, in a variety of situations the modification of the prior probability (e.g. assigning the bulk of the mass to values inside $[0.4, 0.6]$ for the parameter θ of a newborn being a boy) is straightforward and under the FBST this prior perspective holds both when opting for Bayesian parameter estimation and hypothesis testing. Also, the modification of the reference function $r(\theta)$ allows to incorporate such a priori beliefs in a second step.

Conclusion—the Validity of Precise Hypotheses

In this paper, it was shown that the problems identified by Ly and Wagenmakers (2021) are mostly consequences of making the measure-theoretic premise of assigning positive probability $\varrho > 0$ to a point null value of a precise hypothesis $H_0 : \theta = \theta_0$. Thus, they hold only under a perspective which is required to apply a Bayes factor test, but not under the absolutely continuous priors available for use with the FBST. An important lesson from the analysis of Ly and Wagenmakers (2021) is that absolutely continuous priors could be preferred in the FBST to avoid these problems. Whenever a prior is chosen which assigns positive probability mass ϱ to the null value, sampling to a foregone conclusion as well as the asymptotic relationship to the p-value could hold. However, as discussed in this paper, due to the requirements of the Bernstein-von-Mises theorem and unless $\varrho = 1$, they will not hold in practice. Also, improper priors are questionable when the Jeffreys-Lindley paradox and predictive matching are considered.

Importantly, the mathematical arguments brought forward by Ly and Wagenmakers (2021) and in this paper hold depending on which prior beliefs are held, and not depending on whether one wants to be able to only reject or to both reject and confirm a statistical hypothesis. For example, we could hold the prior beliefs that data are normally distributed as $\mathcal{N}(\mu, \sigma^2)$, but we are not willing to assign positive mass $\varrho > 0$ to a point null value θ_0. Thus, we cannot employ a Bayes factor, and we need to use the FBST. Alternatively, we could have prior beliefs which are reflected by a mixture prior that assigns mass $\varrho > 0$ to the theoretically interesting value θ_0 as proposed by Ly and Wagenmakers (2021).[21] Then, the more natural choice is the Bayes factor. Thus, our prior beliefs determine whether we can only reject or both reject and confirm a hypothesis. In contrast, our desire to be able to either only reject, or both reject and confirm a hypothesis should *not* determine our prior beliefs.[22] Importantly, the deficits which are brought forward by Ly and Wagenmakers hold under the latter prior beliefs where we use a mixture prior, and as shown in this paper no mixture prior is needed when employing the FBST.

However, the discussion puts the magnifying glass onto a more important problem. Are point null hypotheses realistic for scientific research? Both the BF and the FBST test a precise hypothesis, and criticisms on the appropriateness of such hypotheses range back at least to Good (1950). The validity of point null hypotheses was challenged seriously the first time by Hodges and Lehmann (1954), and the more realistic test of small interval hypotheses $H_0 : \theta \in [\theta_0 - b, \theta_0 + b]$ versus $H_1 : \theta \notin [\theta_0 - b, \theta_0 + b]$ was termed the Hodges-Lehmann paradigm consequently. Good (1992,1993,1994), Anderson and Hauck (1983), Berger and Delampady (1987), and Rao and Lovric (2016) challenged the appropriateness of precise hypotheses, and Rousseau (2007) showed that the approximation of interval BFs — see Morey and Rouder (2011) — through precise BFs does hold only under very small intervals and is unreliable for large sample size n, see also Berger (1985), Bernado (1999), and Sellke et al. (2001, p. 64). A shift towards the Hodges-Lehmann paradigm may improve the reliability of research while simultaneously removing most of the measure-theoretic bypasses which become necessary for precise hypothesis testing. For the BF, this bypass is the assignment of arbitrary probability mass to a Lebesgue-null-set which can be seen as the price which is paid to be able to confirm H_0. For the FBST, the bypass is the resulting inability to confirm H_0 because no such assumption is made. An important open problem is to clarify the "vexing issue of the relevance of point null hypotheses" (Robert, 2016, p. 5), which is neither achieved by the BF nor the FBST. I agree with Ly and Wagenmakers (2021) that

[20] Three examples illustrate this scepticism, see Savage (1961). A musician states he can tell whether a sheet of music is from Mozart or Haydn and succeeds in ten out of ten cases. A lady drinking tea states she can tell whether the tea or the milk was filled in the cup first and she succeeds in ten out of ten cases. A drunken friend states at 3 p.m. that he can tell the result of a flipped coin and succeeds in ten out of ten cases. In the last case, the prior probability of $H_0 : \theta = 1$ will be a priori smaller than in cases one and two for most people.

[21] In any case, we should then interpret the assignment of mass $\varrho > 0$ to $H_0 : \theta = \theta_0$ as a proxy for the assignment of $\varrho > 0$ to $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$ and check the quality of this approximation.

[22] Note that for any small interval hypothesis $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, the prior probability is *implied* by the choice of prior distribution P_ϑ and the width b when P_ϑ is absolutely continuous with respect to λ, and we do not need to explicitly assign a prior probability to H_0.

whenever there are strong a priori grounds to believe in the point null value (e.g. when testing a general law), testing a point null hypothesis is reasonable. Also, testing a point null hypothesis is reasonable when interpreted as an approximation to a small interval hypothesis. However, whenever there is suspicion about the validity of such an approximation, the FBST may be an attractive alternative because it does not assign positive prior mass $\varrho > 0$ to the null value θ_0. Thus, checking the approximation which conceptualizes ϱ as the mass which should actually be assigned to the more realistic small interval hypothesis $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, and which fails for $n \to \infty$ becomes obsolete when using the FBST (which was one reason for the Jeffreys-Lindley paradox not to occur under the FBST as shown above).

In fact, a shift to the Hodges-Lehmann paradigm is mostly a sociological contribution to statistical science. When conducting a precise hypothesis test, the statistician (whether frequentist or Bayesian) in the majority of cases must check the quality of the approximation of the more realistic interval hypothesis by the point null hypothesis. However, "it is usually easier for a Bayesian to deal directly with the interval hypothesis than to check the adequacy of the approximation." (Berger, 1985, p. 149). More importantly, "there are many (...) problems that would lead to a hypothesis of the above interval form with *large b*, but such problems will rarely be well approximated by testing a point null." (Berger, 1985, p. 149), where b is the width of the interval hypothesis around the point null value $\theta_0 \in \Theta$. Thus, shifting to a Hodges-Lehmann test of an interval hypothesis requires first to elicit the interval hypothesis boundaries, and in a second step hypothesis testing is performed. Importantly, performing a test without checking the quality of the approximation becomes impossible.

As Alan Birnbaum stressed in Savage et al. (1962, p. 322), "each scientist and interpreter of experimental results bears ultimate responsibility for his own concepts of evidence and his own interpretation of results.", and which method to choose needs to be decided by practitioners themselves. However, I suspect both the BF and FBST to provide similar conclusions in the majority of cases, and a shift towards the Hodges-Lehmann paradigm would be beneficial in a variety of situations.

Acknowledgements I am grateful to Eric-Jan Wagenmakers and Alexander Ly for their comments and review of a first draft of this manuscript. I also thank two anonymous reviewers for helpful comments, and I am grateful to Scott Brown for his openness to bring this topic to a wider audience.

Funding Open Access funding enabled and organized by Projekt DEAL.

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

Anderson, S., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics - Theory and Methods*, 12(23), 2663–2692.

Bauer, H. (2001). *Measure and integration theory*. Berlin: De Gruyter.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Berger, J., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2(3), 317–335.

Berger, J., & Wolpert, R. L. (1988). *The likelihood principle*. California, Hayward: Institute of Mathematical Statistics.

Berger, J. O., & Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, 96(453), 174–184.

Berger, J. O., & Pericchi, L. R. (1996). On the justification of default and intrinsic Bayes factors. In *Modelling and Prediction Honoring Seymour Geisser, pages 276–293. Springer New York.*

Bernado, J. (1999). Nested hypothesis testing: the Bayesian reference criterion. In J. Bernado, J. Berger, A. Dawid, & A. Smith (Eds.) *Bayesian Statistics (Vol. 6), pages 101–130 (with discussion). Oxford University Press, Valencia.*

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304–1312.

Degroot, M. H. (1973). Doing what comes naturally: interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association*, 68(344), 966–969.

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.

Edwards, A. (1992). *Likelihood*. Baltimore: The Johns Hopkins University Press Maryland. expanded edition.

Etz, A., & Wagenmakers, E.-J. (2015). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, 32(2), 313–329.

Feller, W. (1968). *An introduction to probability theory and its applications*, 3rd edn. Vol. i. New York: John Wiley & Sons.

Good, I. (1950). *Probability and the weighing of evidence*. London: Charles Griffin.

Good, I. J. (1992). The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, 87(419), 597–606.

Good, I. J. (1993). C397. Refutation and rejection versus inexactification, and other comments concerning terminology. *Journal of Statistical Computation and Simulation*, 47(1-2), 91–92.

Good, I. J. (1994). C420. The existence of sharp null hypotheses. *Journal of Statistical Computation and Simulation*, 49(3-4), 241–242.

Gronau, Q. F., Ly, A., & Wagenmakers, E. -J. (2019). Informed Bayesian t -tests. *The American Statistician*, 74(2), 137–143.

Held, L., & Sabanés-Bové, D. (2014). *Applied Statistical Inference*. Berlin, Heidelberg: Springer.

Hodges, J. L., & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society:, Series B (Methodological)*, 16(2), 261–268.

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

Jeffreys, H. (1980). Some general points in probability theory. In A. Zellner, & J. B. Kadane (Eds.) *Bayesian Analysis in Econometrics and Statistics : Essays in Honor of Harold Jeffreys, pages 451–453. North-Holland Publishing Company, Amsterdam, The Netherlands*.

Johnson, V. E., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society:, Series B (Statistical Methodology)*, 72(2), 143–170.

Kelter, R. (2020). 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).

Ly, A., & Wagenmakers, E.-J. (2021). A critical evaluation of the FBST ev for Bayesian hypothesis testing. PsyArxiv Preprint: https://psyarxiv.com/x9t6n.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4), 406–419.

Pereira, C. A. d. B., & Stern, J. M. (1999). Evidence and credibility: Full Bayesian significance test for precise hypotheses. *Entropy*, 1(4), 99–110.

Pereira, C. A. d. B., Stern, J. M., & Wechsler, S. (2008). Can a significance test be genuinely Bayesian? *Bayesian Analysis*, 3(1), 79–100.

Rao, C. R., & Lovric, M. M. (2016). Testing point null hypothesis of a normal mean and the truth: 21st Century perspective. *Journal of Modern Applied Statistical Methods*, 15(2), 2–21.

Robert, C. P. (2007). *The Bayesian Choice*, 2nd edn. Paris: Springer New York.

Robert, C. P. (2014). On the Jeffreys-Lindley paradox. *Philosophy of Science*, 81(2), 216–232.

Robert, C. P. (2016). The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, 72(2009), 33–37.

Rousseau, J. (2007). Approximating interval hypothesis : p-values and Bayes factors. In J. Bernado, J. Berger, A. Dawid, & A. Smith (Eds.) *Bayesian Statistics (Vol. 8), pages 417–452. Oxford University Press, Valencia*.

Savage, L. (1961). The subjective basis of statistical practice. Technical report, Dept of Statistics, University of Michigan, Michigan.

Savage, L. J., Barnard, G., Cornfield, J., Bross, I., Box, G. E. P., Good, I. J., Lindley, D. V., Clunies-Ross, C. W., Pratt, J. W., Levene, H., Goldman, T., Dempster, A. P., Kempthorne, O., & Birnbaum, A. (1962). On the foundations of statistical inference: Discussion. *Journal of the American Statistical Association*, 57(298), 307–326.

Schervish, M. J. (1995). *Theory of statistics*. New York: Springer.

Sellke, T., Bayarri, M. J., & Berger, J.O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician*, 55(1), 62–71.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wang, M., & Liu, G. (2016). A simple two-sample Bayesian t-test for hypothesis testing. *American Statistician*, 70(2), 195–201.

Wrinch, D., & Jeffreys, H. (1921). XLII. On certain fundamental principles of scientific inquiry. *The London Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 42(249), 369–390.

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