



UvA-DARE (Digital Academic Repository)

Measure-Theoretic Musings Cannot Salvage the Full Bayesian Significance Test as a Measure of Evidence

Ly, A.; Wagenmakers, E.-J.

DOI

[10.1007/s42113-022-00154-1](https://doi.org/10.1007/s42113-022-00154-1)

Publication date

2022

Document Version

Final published version

Published in

Computational Brain and Behavior

License

CC BY

[Link to publication](#)

Citation for published version (APA):

Ly, A., & Wagenmakers, E.-J. (2022). Measure-Theoretic Musings Cannot Salvage the Full Bayesian Significance Test as a Measure of Evidence. *Computational Brain and Behavior*, 5(4), 583-289. <https://doi.org/10.1007/s42113-022-00154-1>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)



Measure-Theoretic Musings Cannot Salvage the Full Bayesian Significance Test as a Measure of Evidence

Rejoinder to Kelter

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

Accepted: 24 August 2022 / Published online: 28 September 2022
© The Author(s) 2022

Abstract

Ly and Wagenmakers (*Computational Brain & Behavior*:1–8, in press) critiqued the Full Bayesian Significance Test (FBST) and the associated statistic FBST ev: similar to the frequentist p -value, FBST ev cannot quantify evidence for the null hypothesis, allows sampling to a foregone conclusion, and suffers from the Jeffreys-Lindley paradox. In response, Kelter (*Computational Brain & Behavior*:1–11, 2022) suggested that the critique is based on a measure-theoretic premise that is often inappropriate in practice, namely the assignment of non-zero prior mass to a point-null hypothesis. Here we argue that the key aspects of our initial critique remain intact when the point-null hypothesis is replaced either by a peri-null hypothesis or by an interval-null hypothesis; hence, the discussion on the validity of a point-null hypothesis is a red herring. We suggest that it is tempting yet fallacious to test a hypothesis by estimating a parameter that is part of a different model. By rejecting any null hypothesis before it is tested, FBST is begging the question. Although FBST may be useful as a measure of surprise under a single model, we believe that the concept of *evidence* is inherently relative; consequently, evidence for competing hypotheses ought to be quantified by examining the relative adequacy of their predictions. This philosophy is fundamentally at odds with the FBST.

Keywords Full Bayesian significance test · Evidence · Optional stopping · Predictive irrelevance · Jeffreys-Lindley paradox

Introduction

In our original article we elaborated on how FBST ev violates several desiderata for scientific evidence (Ly & Wagenmakers, [in press](#)). Specifically, a proper measure of evidence should at least (i) be able to quantify support for both the null and the alternative hypothesis; (ii) increase in favor of the null as data accumulate that are (exactly) consistent with the null; (iii) allow practitioners to consult it at any moment in time without this practice predetermining the conclusions; and (iv) be exactly balanced when the data are evidentially irrelevant, that is, equally likely under both models.

In response, Kelter agrees with our assessment that FBST ev violates these desiderata, but counters that the FBST ev *does not need to adhere* to these natural rules:

“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.”

Kelter argues that the Bayes factor involves the assignment of non-zero mass to a point-null hypothesis, in many practical applications supposedly a dubious act that the FBST ev avoids. If the point-null hypothesis is never true, as FBST assumes, then it may not be relevant what ought to happen whenever the point-null hypothesis is true. We are grateful for Kelter’s comment on our paper: the central issue has been debated for decades, and Kelter’s main argument—the point-null hypothesis is never true, and this is a blow against Bayes factors—has been made many times before (see the

✉ 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

references in the Kelter article). We appreciate the opportunity to explain why we believe our conclusion stands, and the discussion on the validity on a point-null hypothesis is (and has always been) a red herring. Below we first examine Kelter's main claim, and then discuss several issues that we believe require clarification and sometimes rectification.

The Point-Null Hypotheses Is Not the Problem

Kelter agrees that FBST ev performs poorly whenever the null hypothesis holds true, but disputes the relevance of this fact. Indeed, Kelter surmises that point-null hypotheses are almost never realistic and can therefore be ignored.

There exist multiple counterarguments to this philosophical position. First, the very fact that the null hypothesis is being tested is an indication of it being possibly true. And if the null holds true, then the statistical method should be able to accumulate evidence in its favor. If a person can never be acquitted, then why the charade of taking them to court? We reiterate our position that even in FBST ev the null value is of special interest—one tests the hypothesis $\mathcal{H}_0 : \delta = 0$, and not, say, $\mathcal{H}_0 : \delta = 0.00177567$. In many statistical applications, the hypothesis $\mathcal{H}_0 : \delta = 0$ corresponds to the exclusion of a parameter, such as the effect of a covariate.

Second, the rejection of point-null hypotheses is an implicit rejection of *all* models. As detailed below, any model can be extended to a larger model, for instance, by adding a covariate or an across-trial parameter. Hence, every model can eventually be identified with a point-null hypothesis. Discarding all point-null hypotheses out of hand implies that researchers would always be forced to use the most complex model, and add a variable for every new data point. This would make impossible any ability for generalization and in fact prohibit any kind of scientific progress (e.g., Poincaré, 1913, pp. 119–120). Thus, tests against a point-null constitute an epistemic firewall against needless complexity. They protect against the whimsical adoption of vague models that overfit the data and generalize poorly (e.g., Myung et al., 2000; Roberts & Pashler, 2000). Point-null tests allow researchers to develop generalizable models in a controllable fashion (e.g., Haaf et al., 2019; Ioannidis, 2019).

Third, and crucially, the problems for FBST ev that we identified do not vanish when the point-null is abandoned. Suppose we accept, for the sake of argument, the common complaint that point-null hypotheses are never true exactly. Also suppose we also accept, again for the sake of the argument, that this realization would invalidate the usual Bayes factor of a point-null hypothesis \mathcal{H}_0 versus \mathcal{H}_1 . This knowledge should not cause any researcher to abandon the entire Bayes factor testing framework and embrace the FBST ev estimation framework instead; when a roofer examines their

work and finds a broken tile, they generally do not propose to remove the entire roof and declare the problem solved. Rather, in order to incorporate the advance knowledge that the point-null hypothesis is implausible we may replace it by a peri-null hypothesis.¹ For instance, the point-null hypothesis $\mathcal{H}_0 : \delta = 0$ may be replaced with the peri-null hypothesis $\mathcal{H}_0 : \delta \sim N(0, \epsilon)$, where ϵ accommodates the notion that \mathcal{H}_0 is used only as a mathematically convenient approximation. Importantly, such a change from point-null to peri-null leaves most of our critiques against FBST ev conceptually intact.

Specifically, a Bayes factor test with a peri-null still allows evidence to be collected in favor of the peri-null, and this evidence would increase as data accumulate that are more consistent with the peri-null than with the alternative. The evidence would not increase indefinitely, but be subject to an upper bound—such inconsistency is the price that is paid for adopting a peri-null rather than a point-null (Ly & Wagenmakers, 2022). Similarly, the Jeffreys-Lindley paradox also holds in the case of a peri-null hypothesis, albeit in a less extreme form—as sample size increases and the p -value remains fixed, the evidence will eventually support the peri-null hypothesis over the alternative hypothesis, but the extent of this support is bounded (see Eq. 3 and Fig. 2 in Wagenmakers & Ly, in press). This is a key observation, as it is a common misperception in the statistical literature that the Jeffreys-Lindley paradox is caused by the presence of a point-null hypothesis. It also remains the case that the Bayes factor with a peri-null hypothesis will indicate perfect indifference when confronted with data that are equally likely under the peri-null hypothesis and the alternative hypothesis.

Hence, we believe Kelter's main thesis is incorrect—our critique against FBST ev as a measure of evidence does not require the assignment of prior mass to a point. We wish to stress that our critique does not concern FBST ev wholesale; in fact, FBST ev may be used as a method to quantify *surprise* (e.g., Good, 1981, 1983). But when the quantification of *evidence* is in order, it is important not just to consider the surprise under a single model, but also the surprise under a competing model; specifically, the data provide evidence for model A over model B if and only if model A predicted those data better than model B.

We now turn to a more detailed analysis of Kelter's arguments, which demands that we also clarify some of our own assumptions.

The Nature of Testing

Kelter's arguments are motivated from an alternative-model only perspective. In contrast, we believe that a test, in its simplest form, involves at least *two* models: The null model

¹ See also <https://tinyurl.com/perinull>.

\mathcal{M}_0 corresponding to the null hypothesis \mathcal{H}_0 , and the alternative model \mathcal{M}_1 corresponding to the alternative hypothesis \mathcal{H}_1 . By hypotheses we mean assertions regarding (unobservable) parameters of interest, and by a (psychological) model we mean a (proposed) mathematical description of the underlying processes with which these parameters generate the (behavioral) data (e.g., Ly et al., 2017). Provided with data, a test can then be used to distinguish the competing models.

For instance, for speeded decision-making tasks both Ratcliff’s drift diffusion model (DDM, e.g., Boehm et al., 2018; Ratcliff, 1978), or Brown and Heathcote’s linear ballistic accumulator model (LBA, e.g., Brown & Heathcote, 2008) can be used to describe the distribution of response times. Similarly, to describe the relationship between physical dimensions, such as light intensity, and their psychological counterpart, such as brightness, Steven’s model or Fechner’s model can be used (Myung et al., 2000). Analogously, to describe normally distributed data Y , a Gaussian distribution with a mean fixed at zero, or, alternatively, a Gaussian distribution where the mean is free to vary can be used, that is,

$$\mathcal{M}_0 : Y \sim \mathcal{N}(0, \sigma^2), \text{ and } \mathcal{M}_1 : Y \sim \mathcal{N}(\mu, \sigma^2) \tag{1}$$

respectively. In the last example the models are easily seen to be nested. For the model comparison Eq. (1) the test addresses the question of *whether or not* the mean should be estimated, or—equivalently—whether the observed sample mean \bar{x} should either be regarded as random noise (i.e., when \mathcal{M}_0 holds true) or as a reliable indication of μ (i.e., when \mathcal{M}_1 holds true). In contrast, an estimation problem addresses the “*how much*” question, by providing a best guess of the magnitude of the population mean given that it is not zero, i.e., when \mathcal{M}_1 holds true.

Regardless of whether or not the models are nested, the Bayesian assessment of the uncertainty *between* models is the same. This procedure requires prior model probabilities $0 < P(\mathcal{M}_0), P(\mathcal{M}_1) < 1$ with $P(\mathcal{M}_0) + P(\mathcal{M}_1) = 1$ if only these two models are believed to be viable for the data. After observing outcomes y , these prior model probabilities are then updated using Bayes’ rule to yield posterior model probabilities, as follows:

$$P(\mathcal{M}_1|y) = \frac{BF_{10}(y)P(\mathcal{M}_1)}{BF_{10}(y)P(\mathcal{M}_1)+P(\mathcal{M}_0)} \text{ and } P(\mathcal{M}_0|y) = 1 - P(\mathcal{M}_1|y), \tag{2}$$

where $BF_{10}(y)$ is the Bayes factor indicating the evidence in the outcomes y for \mathcal{M}_1 over \mathcal{M}_0 .² The prior model probabilities $P(\mathcal{M}_0), P(\mathcal{M}_1)$ represent the researcher’s subjective beliefs, or uncertainty, regarding the *models* and associated hypotheses. For instance, before data observation, a

researcher might believe that the LBA and DDM provide equally viable descriptions of response time distributions, and accordingly set $P(\mathcal{M}_0) = P(\mathcal{M}_1) = 1/2$. Note that these prior model probabilities are *not* derived from the prior densities that describe the uncertainty *within* each model. In particular, $P(\mathcal{M}_0)$ is not derived from the prior density regarding the free parameters within \mathcal{M}_1 . The parameters of the LBA and DDM are interpreted within their own contexts (e.g., Osth et al., 2017).

Kelter’s Representations of the Bayes Factor

We believe that Kelter’s representation of the Bayes factor conflates the null hypothesis, the *statement* regarding the parameter of a, say, Gaussian model such as $\mathcal{H}_0 : \delta = \mu/\sigma = 0$, and the *null value* 0. Specifically, Kelter argues that $P(\mathcal{M}_0)$ has to be zero, once a continuous prior density $\pi(\delta|\mathcal{M}_1)$ is selected, because such a prior density gives zero prior probability to $\{0\}$, where $\{0\}$ is the set which only contains the null value. However, we dispute the idea that the prior density $\pi(\delta|\mathcal{M}_1)$ defined on the parameters *within the alternative model* should dictate the prior model probability of the *null model* $P(\mathcal{M}_0)$.

In his analysis, Kelter takes the alternative model as a point of departure, and from this perspective it should not come as a surprise that the null model is argued to be false. This alternative-only perspective is also apparent when Kelter discusses the Bayes factor as a ratio of marginal likelihoods. For concreteness, we consider again the model comparison Eq. (1), that is, the Bayesian *t*-test. The alternative model allows both the effect size $\delta \in (-\infty, \infty)$, and standard deviation $\sigma \in (0, \infty)$ to vary freely, whereas the null model has only one free parameter $\sigma \in (0, \infty)$. The specification of the Bayes factor now requires a *pair* of priors resulting in

$$BF_{10}(y) = \frac{\int_{-\infty}^{\infty} \int_0^{\infty} f(y|\delta, \sigma, \mathcal{M}_1) \pi(\delta, \sigma|\mathcal{M}_1) d\sigma d\delta}{\int_0^{\infty} f(y|\delta=0, \sigma, \mathcal{M}_0) \pi(\sigma|\mathcal{M}_0) d\sigma}. \tag{3}$$

Kelter’s presumption that only the alternative model matters translates to the claim that the marginal likelihood (i.e., the denominator in Eq. (3)) has to be zero.³

This claim neglects the fact that the priors densities $\pi(\delta, \sigma|\mathcal{M}_1)$ and $\pi(\sigma|\mathcal{M}_0)$ are conditional on their respective models that provide contexts to the parameters. In fact, direct computations with $\pi(\sigma|\mathcal{M}_0) \propto \sigma^{-1}$ show that the marginal likelihood is not zero (for details, see, e.g., Ly et al., 2016a, b).

² Equation (2) assumes that the marginal likelihood of the null model $p(y|\mathcal{M}_0)$ is non-zero, which is typical in theory and practice. No mixture prior is needed to define a Bayes factor.

³ Kelter (in press) writes “Now, from (A) in Eq. (5), it is immediate that whenever Θ_0 is a P_β -null-set, the Bayes factor BF_{01} will be zero because $\int_{\Theta_0} f_\beta dP_\beta = 0$ then.”

The marginal likelihood in fact is a predictive score, one that does not depend on either of the candidate models being “true” (e.g., O’Hagan and Forster, 2004, pp. 166-167; Kass & Raftery, 1995, p. 777). This can be appreciated by writing the marginal likelihood as a sum of one-step-ahead prediction errors (e.g., Wagenmakers et al., 2006), by interpreting the Bayes factor as a specific case of cross-validation (e.g., Gneiting & Raftery, 2007), or by noting the close relation with model comparison methods that obey the Minimum Description Length principle (e.g., Grünwald et al., 2005; Grünwald, 2007; Ly et al., 2017).

Suppose that researchers wish to compare Fechner’s law of psychophysics (which can account only for negatively accelerating curves) against Steven’s law of psychophysics (which can account for negatively and positively accelerating curves). The researchers view both accounts as sequential forecasting systems (e.g., Dawid, 1984) and use the Bayes factor to quantify the models’ relative predictive performance. They then consult the statistical literature and read, to their surprise, that Fechner’s law is nested under Stevens’ law (Kvålseth, 1992). Does the realization that the models are nested invalidate the conclusion that Fechner’s model has outperformed Steven’s model? Should we resign ourselves to the position that it is impossible to collect positive evidence for Fechner’s law over Stevens’ law? We don’t think so. In fact we believe that if the data had consistently showed negatively accelerating curves, Stevens’ law would not even have been proposed.

Null Sets

Kelter’s assertion that in practical applications the denominator in Eq. (3) is zero is argued based on the notion of null sets. A null set is a relative concept defined with respect to a certain (probability) measure and the space it works on. For instance, the natural measure on two-dimensional (Euclidean) space \mathbb{R}^2 is the so-called two-dimensional Lebesgue measure λ^2 , which gives the rectangle $[-6, -2] \times [0.5, 2.5]$, see Fig. 1, an area of $4 \times 2 = 8$ in, say, square meters m^2 . A lower dimensional object such as the bottom edge of the aforementioned rectangle (i.e., the line segment $[-6, -2] \times \{0.5\}$) has λ^2 -measure zero. This just means that this line segment has no area. It does *not* mean that the line itself has no intrinsic worth, nor that it is uninteresting. When we take into account the intrinsic dimension of the line we get the natural measure in one-dimensional Euclidean space \mathbb{R} .⁴ This natural measure is known as the one-dimensional

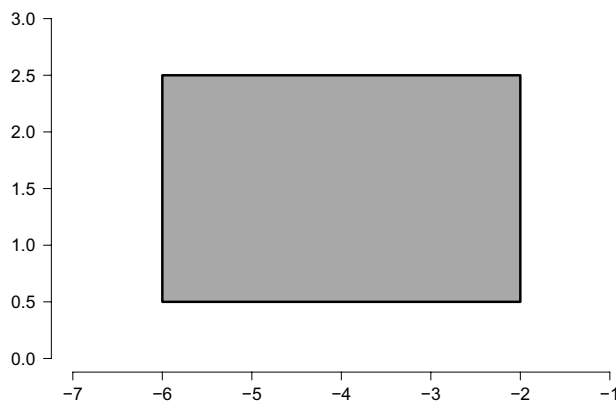


Fig. 1 The two-dimensional Lebesgue measure is the natural measure that gives rectangles with width w and height h an area of $w \times h$ in, say, square meters m^2

Lebesgue measure λ and assigns to the line segment a length of 4 in, say, meters m , as one might expect.

Kelter’s claim that the marginal likelihood of \mathcal{M}_0 is zero follows from him letting the two-dimensional prior density $\pi(\delta, \sigma | \mathcal{M}_1)$ abuse its null sets onto the one-dimensional prior density $\pi(\sigma | \mathcal{M}_0)$. Note again that the null sets of $\pi(\delta, \sigma | \mathcal{M}_1)$ defined for \mathcal{M}_1 are made to intrude on $\pi(\sigma | \mathcal{M}_0)$, which is defined on the different context given by \mathcal{M}_0 . Similarly, the within-model uncertainty of the LBA model should not dictate the priors on the parameters of the DDM model. The parameters in each of these models affect the data differently, even if they are given similar interpretations (e.g., Ost et al., 2017).

As mentioned above, Kelter’s argument of having the prior density $\pi(\sigma | \mathcal{M}_0)$ be violated by the higher-dimensional prior density $\pi(\delta, \sigma | \mathcal{M}_1)$ leads to an infinite regress. Concretely, suppose that in addition to the models in Eq. (1), we also entertain the model

$$\mathcal{M}_2 : Y|x_1 \sim \mathcal{N}(\mu + \beta_1 x_1, \sigma^2), \tag{4}$$

or equivalently

$$\mathcal{M}_2 : Y = \mu + \beta_1 x_1 + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{5}$$

where β_1 is the effect of x_1 on Y , where x_1 , for instance, could represent age, or a factor on which the data could be stratified, such as gender.

By taking $\beta_1 = 0$ in \mathcal{M}_2 , we retrieve \mathcal{M}_1 , and note that \mathcal{M}_2 is three-dimensional with free parameters μ, β_1, σ . Kelter’s argument leads to $\pi(\mu, \beta_1, \sigma | \mathcal{M}_2)$ forcing its null sets onto $\pi(\mu, \sigma | \mathcal{M}_1)$ making the marginal likelihood of \mathcal{M}_1 zero. Because the marginal likelihood of \mathcal{M}_1 is the normalizing constant of the posterior on μ, σ , Kelter would then have to conclude that the posterior density $\pi(\mu, \sigma | y, \mathcal{M}_1)$ does not exist. This implies that also no FBST ev can be computed. Observe that the whole argument is reasoned

⁴ This is in general known as disintegration (e.g., Chang & Pollard, 1997; Kallenberg, 2021) and forms the basis of conditional inference such as Bayesian statistics.

from the *perspective* of \mathcal{M}_2 in which both \mathcal{M}_1 and \mathcal{M}_0 are false to begin with. Every model can be further extended by including a possibly not relevant covariate, resulting in a higher-dimensional model. This argument can be repeated indefinitely, resulting in posteriors that are always ill-defined from a certain perspective. The way out is to give intrinsic value to each model, and have each their own prior densities in their own rights.

The FBST *ev* Behaves as a *p*-Value

In Ly and Wagenmakers (in press) we elaborated on a theorem derived by, amongst others, the advocates of FBST *ev* (Diniz et al., 2012). The clarification we added was the mode of convergence which is “in distribution” resulting in

$$ev \xrightarrow{d} 1 - F_{\dim(\theta_1)}[F_{\dim(\theta_1) - \dim(\theta_0)}^{-1}(1 - p)], \quad (6)$$

see the discussion surrounding Equation 3 in Ly and Wagenmakers (in press) for further details. Equation (6) implies that the sampling distribution of FBST *ev* can be accurately approximated by the distribution of a (transformed) *p*-value. It shows that for general priors, FBST *ev* *behaves* as a *p*-value. This qualitative statement provides sufficient insights to why FBST *ev* does not act as a proper measure of evidence. Furthermore, Kelter’s claim that Eq. (6) cannot hold for discrete models is unfortunately incorrect. Converge in distribution does not depend on the underlying probability spaces (e.g., van der Vaart, 1998, p. 5). This mode of convergence formalizes exactly what it means for the intuitive clear idea of a discrete uniform distribution on the interval (0,1) with *n* bins each occurring with chance 1/*n* to converge to the continuous uniform distribution on the interval (0,1). Another example is given by a Galton board, which demonstrates how a binomial distribution can be well approximated by a normal distribution Galton (p. 63, 1889). It is in this sense that the distribution of FBST *ev* can be well approximated by the distribution of a *p*-value based on the likelihood ratio statistic. In Ly and Wagenmakers (in press) we elaborated on how this relationship between FBST *ev* and the *p*-value causes FBST *ev* to not act as a proper measure of scientific evidence.

Concluding Comments

Kelter’s main counterargument to our critique on FBST *ev* as a measure of evidence is that in scientific practice, the point-null hypothesis is never true exactly. Hence, the epistemic defects of FBST *ev* are judged to be practically irrelevant. Presumably it does not matter that FBST *ev* cannot quantify evidence in favor of the point-null, because the

point-null is never true anyway. According to Kelter, the problem lies instead with methodologies such as the Bayes factor, that do assign credence to a point-null hypothesis. Kelter’s line of argumentation is relatively well-known in the field of statistics. For statisticians, it would be good news if the Kelter line of argumentation were correct, because hypotheses could then be evaluated using a methodology that is straightforward—much more straightforward than Bayes factors, which depend critically on the specification of prior distributions (e.g., Bayarri et al., 2012; Grünwald et al., 2020; Jeffreys, 1961; Ly et al., 2016b, 2016a, 2020).

In this rejoinder we have stressed the following points that are underappreciated in the statistical literature:

1. Bayes factors compare predictive performance between any two models, nested or non-nested. Researchers who abhor the point-null hypothesis may still prefer the Bayes factor as a method for model comparison.
2. Researchers allergic to the point-null may replace it with a peri-null. Crucially, this does not affect the key aspects of our critique on FBST *ev* (Ly & Wagenmakers, 2022). In particular, it does not change the fact that FBST *ev* suffers from the Jeffreys-Lindley paradox (Wagenmakers & Ly, in press).
3. The Bayes factor quantifies relative predictive performance and does not depend on any of the two models under consideration or the prior densities being “true”.⁵
4. The prior distribution for the test-relevant parameter in the alternative model does not dictate the prior model probability for the null hypothesis, nor the prior density within the null model. Doing so leads to an infinite regress.

Therefore we maintain that the epistemic defects of FBST *ev* are more than just a theoretical possibility. In practice, a skeptic who is confronted with compelling FBST *ev* support against the null has every right to remain doubtful; after all, the skeptic’s model—whether instantiated as a point-null or a peri-null hypothesis—may predict the data as well or better than the alternative model. This does not mean that the FBST *ev* is useless or inherently flawed; in fact, it may serve well as a measure of surprise (Good, 1981), attending the researcher to the fact that a single hypothesized model may be in need of reappraisal or adjustment. Thus, FBST *ev* may be particularly useful in an early stage of model-development, where clearly articulated alternative models are still lacking. However, in a later stage researchers will want to use the developed models to quantify evidence (e.g.,

⁵ The use of a Bayes factor also does not require any “specific measure-theoretic premise” and no restrictions are imposed on priors or any data generating model.

for a treatment effect in medical clinical trials). In this stage, a between-model perspective appears to be essential. In sum, we remain convinced that even though FBST ev may be useful as a within-model measure of surprise, it is not a genuine measure of evidence.

Acknowledgements The first author would like to thank Tyron Lardy, Daniel Gomon, and Udo Boehm for their discussion on and insights on the Skorokhod's representation theorem, which helped shape some of the ideas presented here. The authors would also like to thank Scott Brown, two anonymous reviewers, and Riko Kelter for their comments on an earlier version of this paper.

Author Contribution AL wrote the first draft, EJW revised the manuscript. Both authors discussed and approved the contents and AL finalized the manuscript.

Funding This research was supported by the Netherlands Organisation for Scientific Research (NWO; grant #016.Vici.170.083) and by the European Research Council (ERC Advanced; grant #743086 UNIFY).

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Ethics Approval Not applicable.

Consent for Publication Not applicable.

Consent to Participate Not applicable.

Conflict of Interest The authors declare no competing interests.

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.
- Boehm, U., Annis, J., Frank, M.J., Hawkins, G.E., Heathcote, A., Kellen, D., Krypotos, A.-M., Lerche, V., Logan, G.D., Palmeri, T.J., van Ravenzwaaij, D., Servant, M., Singmann, H., Starns, J.J., Voss, A., Wiecki, T.V., Matzke, D., & Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the diffusion decision model Expert advice and recommendations. *Journal of Mathematical Psychology*, 87, 46–75.
- Brown, S.D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, 57(3), 153–178.
- Chang, J.T., & Pollard, D. (1997). Conditioning as disintegration. *Statistica Neerlandica*, 51 (3), 287–317.
- Dawid, A.P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach (with discussion). *Journal of the Royal Statistical Society Series A*, 147, 278–292.
- Diniz, M., Pereira, C.A., Polpo, A., Stern, J.M., & Wechsler, S. (2012). Relationship between Bayesian and frequentist significance indices. *International Journal for Uncertainty Quantification*, 2 (2), 161–172.
- Galton, F. (1889). *Natural inheritance*. Macmillan and Company.
- Gneiting, T., & Raftery, E.A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378.
- Good, I. J. (1981). Some logic and history of hypothesis testing. In J.C. Pitt (Ed.) *Philosophical Foundations of Economics* (pp. 149–174). Dordrecht–Holland: D. Reidel Publishing Company.
- Good, I.J. (1983). *Good thinking: The foundations of probability and its applications*. Minneapolis: University of Minnesota Press.
- Grünwald, P. (2007). *The minimum description length principle*, MIT Press, Cambridge, MA.
- Grünwald, P., Myung, I. J. & Pitt M. A. (Eds.) (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.
- Grünwald, P., de Heide, R., & Koolen, W. (2020). *Safe testing*. arXiv:1906.07801.
- Haaf, J.M., Ly, A., & Wagenmakers, E.-J. (2019). Retire significance, but still test hypotheses. *Nature*, 567(7749), 461–462.
- Ioannidis, J.P. (2019). Retiring statistical significance would give bias a free pass. *Nature*, 567 (7749), 461–462.
- Jeffreys, H. (1961). *Theory of probability*, 3rd edn. Oxford, UK: Oxford University Press.
- Kallenberg, O. (2021). *Foundations of modern probability*, 3rd edn. Berlin: Springer.
- Kass, R.E., & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kelter, R. (in press). On the measure-theoretic premises of Bayes factor and full Bayesian significance tests: a critical reevaluation. *Computational Brain & Behavior*, pp 1–11.
- Kvålseth, T.O. (1992). Fechner's psychophysical law as a special case of Stevens' three-parameter power law. *Perceptual and Motor Skills*, 75, 1205–1206.
- Ly, A., & Wagenmakers, E.-J. (2022). Bayes factors for peri-null hypotheses. *TEST*. <https://doi.org/10.1007/s11749-022-00819-w>
- Ly, A., & Wagenmakers, E.-J. (in press). A critical evaluation of the FBST ev for Bayesian hypothesis testing. *Computational Brain & Behavior*, pp 1–8.
- Ly, A., Verhagen, A.J., & Wagenmakers, E.-J. (2016a). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, 72, 43–55.
- Ly, A., Verhagen, A.J., & Wagenmakers, E.-J. (2016b). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72, 19–32.
- 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., Komarlu Narendra Gupta, A.R., 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 & Behavior*, 3(2), 153–161.
- Myung, I.J., Balasubramanian, V., & Pitt, M. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, 97(21), 11170–11175.
- O'Hagan, A., & Forster, J. (2004). *Kendall's Advanced Theory of Statistics Vol. 2 B Bayesian Inference*, 2nd edn. London: Arnold.
- Osth, A.F., Bora, B., Dennis, S., & Heathcote, A. (2017). Diffusion vs. linear ballistic accumulation: Different models, different conclusions about the slope of the zROC in recognition memory. *Journal of Memory and Language*, 96, 36–61.
- Poincaré, H. (1913). *The Foundations of Science (G. B. Halsted, Trans.)* New York: The Science Press.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59–108.
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing in psychology. *Psychological Review*, 107, 358–367.
- van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Wagenmakers, E.-J., & Ly, A. (in press). History and nature of the Jeffreys–Lindley paradox. *Archive for History of Exact Sciences*, arXiv:2111.10191.
- Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, 50, 149–166.

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