Advantages Masquerading as 'Issues' in Bayesian Hypothesis Testing: A Commentary on Tendeiro and Kiers (2019)

Don van Ravenzwaaij[1] and Eric–Jan Wagenmakers[2]

[1]University of Groningen [2]University of Amsterdam

Correspondence concerning this article should be addressed to:

Don van Ravenzwaaij

University of Groningen, Department of Psychology

Grote Kruisstraat 2/1, Heymans Building, room 169

9712 TS Groningen, The Netherlands

Ph: (+31) 50 363 7021

E–mail should be sent to d.van.ravenzwaaij@rug.nl.

Abstract

Tendeiro and Kiers (2019) provide a detailed and scholarly critique of Null Hypothesis Bayesian Testing (NHBT) and its central component –the Bayes factor– that allows researchers to update knowledge and quantify statistical evidence. Tendeiro and Kiers conclude that NHBT constitutes an improvement over frequentist p-values, but primarily elaborate on a list of eleven 'issues' of NHBT. We believe that several issues identified by Tendeiro and Kiers are of central importance for elucidating the complementary roles of hypothesis testing versus parameter estimation and for appreciating the virtue of statistical thinking over conducting statistical rituals. But although we agree with many of their thoughtful recommendations, we believe that Tendeiro and Kiers are overly pessimistic, and that several of their 'issues' with NHBT may in fact be conceived as pronounced advantages. We illustrate our arguments with simple, concrete examples and end with a critical discussion of one of the recommendations by Tendeiro and Kiers, which is that "estimation of the full posterior distribution offers a more complete picture" than a Bayes factor hypothesis test.

Keywords: Bayes factors, Bayesian hypothesis testing, parameter estimation.

but who attempts to eat an orange without first disposing of the peel, or what manner of a dwelling could be erected unless an adequate foundation be first provided?

Ernest Bramah Smith, *Kai Lung's Golden Hours*

In recent years, Bayesian hypothesis tests have become increasingly visible as a way to supplement or even supplant the standard 'frequentist' p-value hypothesis tests (e.g., Vandekerckhove, Rouder, & Kruschke, 2018; Hoijtink & Chow, 2017). One of the most prominent Bayesian tests, explicitly developed to offer an alternative to p-values, quantifies evidence for two rival hypotheses by comparing their predictive success for the observed data (i.e., the Bayes factor; Jeffreys, 1939; Kass & Raftery, 1995). In the context of significance testing, the rival hypotheses usually consist of a point null hypothesis $\mathcal{H}_0$, instantiating the skeptic's position that the effect is absent, and an alternative hypothesis $\mathcal{H}_1$, which postulates that the effect is present but of unknown size; in order to be able to evaluate the predictive adequacy of $\mathcal{H}_1$, the unknown effect size is assigned a prior distribution. In what follows, the Bayes factor hypothesis test for $\mathcal{H}_0$ versus $\mathcal{H}_1$ is denoted as Null Hypothesis Bayesian Testing (i.e., NHBT).

In their article "A review of issues about Null Hypothesis Bayesian Testing", Tendeiro and Kiers (2019; henceforth TK) provide an in-depth discussion of NHBT. We agree with TK that "NHBT is an improvement over NHST [Null Hypothesis Significance Testing]". We also agree that NHBT is important to discuss, and that it is important for practitioners to know more about what NHBT is and what it is not. However, we disagree with TK when they identify eleven issues associated with NHBT and interpret all of them as potentially problematic (i.e., in their abstract, TK equate the term 'issues' with "limitations or sources of misinterpretation" and "shortcomings"). In our opinion, many of the 'issues' listed by TK are a blessing

Table 1
The eleven issues of NHBT according to TK. The left column contains TK's assessment of each of these issues, and the right column contains our own assessment. ✗ *indicates a disadvantage;* = *indicates a neutral property;* ✔ *indicates an advantage.*

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

rather than a curse. The issues identified by TK are listed in Table 1. All eleven issues are interpreted as problematic by TK; in contrast, we believe that five of the eleven issues are actually advantages, and six are neutral. Below we first discuss each issue in turn, with the express purpose of providing the reader with a perspective that balances and complements that of TK. We then provide a more detailed critique of TKs alternative inference procedure that bases inference solely on the continuous posterior distribution under $\mathcal{H}_1$. This alternative procedure is simple and easy to apply, but Bayesian literature (e.g., Berger & Delampady, 1987; Jeffreys, 1961) suggest that it is beguiling and can easily fool practitioners into believing that they answered a question that they in fact never asked (e.g., Wagenmakers, Lee, Rouder, & Morey, 2020).

1. "Bayes factors can be hard to compute"

TK point out that Bayes factors are hard to compute, as they involve a multidimensional integral across the parameter space. This statement is true for some Bayes factors (e.g., those that involve composite hypotheses and lack an analytic expression for the marginal likelihood), but we teach Bayes factors to our first-year undergraduate students with examples that can be calculated by hand. The first author of this paper habitually uses an example of a Bayes factor as a simple likelihood ratio based on the umbrella scenario presented in chapter 17 of Navarro (2015). In the example, the student sees me walking into the lecture hall (that has no windows) with an umbrella. The student wishes to know whether or not it rains outside. It is given that the prior that it rains is 45% (based, for instance, on data from the Dutch weather institute, the way the sky looked six hours ago when they entered the building, etc.), that the probability of me carrying an umbrella if it rains is 80% (sometimes I forget), and that the probability of me carrying an umbrella if it does not rain is 10% (better safe than sorry). The Bayes factor is the simple ratio of the likelihoods, i.e. $0.8/0.1 = 8$ (and the resulting posterior probability that it rains, not relevant here, is $(8 \times .45) \,/\, (8 \times .45 + (1 - .45)) \approx .87$).

Bayes factors for more realistic scenarios can be hard to compute, but this is no different for NHST (we know of no-one that can calculate an exact p-value without the help of software). More to the point, the difficulty in computation is relevant mostly for mathematical statisticians who wish to develop novel Bayes factor tests or for those who wish to use idiosyncratic prior distributions (e.g., a trimodal prior distribution on a binomial chance parameter θ). And even for this relatively small group of statistical experts, there exist convenient ways to obtain the Bayes factor (e.g., the Savage-Dickey density ratio test, Dickey & Lientz, 1970; Wagenmakers, Lodewyckx,

Kuriyal, & Grasman, 2010; bridge sampling, Gronau, Singmann, & Wagenmakers, 2020; Gronau et al., 2017; for an overview see Gamerman & Lopes, 2006; Martin, Frazier, & Robert, 2020; see also Chib, 1995; Chib & Jeliazkov, 2001). Furthermore, introductory papers on some of the more technical aspects of Bayesian inference are on the rise (e.g., van Ravenzwaaij, Cassey, & Brown, 2018; Etz & Vandekerckhove, 2018; van Ravenzwaaij & Etz, in press), gradually expanding the group of statistical experts who are comfortable with computing their own Bayes factor.

For the practitioner interested in applying existing statistical models, however, the fact that Bayes factors involve a multidimensional integral is irrelevant, as the computation is usually done by software programs in a matter of seconds. As acknowledged by TK, specialized software such as JASP (JASP Team, 2020, `jasp-stats.org`) or the BayesFactor R-package (Morey et al., 2018) allow users to obtain Bayes factors with relatively little effort. Other recently developed Bayes factor software includes BaIn (e.g., Gu, Mulder, & Hoijtink, 2018; Gu, Hoijtink, Mulder, & Rosseel, 2019; Hoijtink, Mulder, van Lissa, & Gu, 2019), BFPack (e.g., Mulder, Gu, et al., 2019; Mulder, Hoijtink, & Gu, 2019), BAS (e.g., Clyde, Ghosh, & Littman, 2011; Clyde, 2016; van den Bergh, Clyde, et al., in press), and the Bayes factor calculator developed by Zoltan Dienes available at `http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm`. A comprehensive overview at `https://cran.r-project.org/web/views/Bayesian.html` lists many other packages that produce Bayes factors with ease. For linear models, the theoretical work by Zellner and Siow (1980), Liang, Paulo, Molina, Clyde, and Berger (2008), and Rouder, Speckman, Sun, Morey, and Iverson (2009) has further reduced the need for time-consuming numerical integration techniques.

In general, almost all statistical analyses are hard to compute *by hand*. Structural equation models, network models, machine-learning methods, analysis of vari-

ance (ANOVA) – all of these would constitute formidable or even insurmountable challenges for practitioners if the analyses had to be conducted by hand. However, in the modern era of statistical computing, virtually all analyses are conducted *in silico*, such that the practitioner need not worry about the details of the computation but is free to focus on the substantive interpretation of the results. In sum, we agree with TK when they state "Given the availability of Bayes factor-friendly software (...), we may conclude that the difficulty [of computing the marginal likelihood] is a feature of Bayes factors that does not offer major problems for practitioners, at least for the most common types of tests used in the social sciences."

One could argue that perhaps the issue is not that NHBT is harder to *compute* than NHST, but that it is harder to *understand*.[1] We find this argument unconvincing: there is ample evidence of the difficulty that scientific practitioners experience when trying to understand NHST (Gigerenzer, 2004, 2018; Oakes, 1986; Haller & Krauss, 2002; Lyu, Xu, Zhao, Zuo, & Hu, 2020). Our admittedly limited experience suggests that if taught from day one, NHBT may be more intuitive and easier to understand than NHST (see also Haucke, Miosga, Hoekstra, & van Ravenzwaaij, in press).

2. "Bayes factors are sensitive to within-model priors"

In NHBT, the exact value of the Bayes factor depends on the prior of the effect size parameter under the alternative model $\mathcal{H}_1$. Considered an issue by TK, we believe this is an advantage (e.g., Etz & Vandekerckhove, 2016; Vanpaemel, 2010). The reason that Bayes factors are sensitive to within-model priors is that Bayes factors evaluate models by the predictions they make, and predictions are determined partly by the prior.[2] So the negative statement that "the Bayes factor is sensitive to the

[1] We thank a reviewer for bringing this to our attention.

[2] The connection is implicit in the linguistic similarity between the words "prediction" and "prior".

within-model prior" is the same as the positive statement "the Bayes factor evaluates models based on their predictive adequacy". The argument may therefore be turned on its head: *any method that fails to take into account the predictions that a hypothesis makes is seriously deficient.*

Within-model priors determine the nature of a model (Box, 1980). In particular, narrow priors instantiate parsimonious models that make risky predictions. When such risky predictions come true, the Bayes factor will indicate that the model outperformed more flexible models who were hedging their bets instead. As a concrete example, consider the study reported by Tawakol et al. (2017) on the relation between amygdalar activity and perceived stress in 13 patients with post-traumatic stress disorder (see also van Dongen et al., 2019). The top panel of Figure 1 shows the result of a default two-sided analysis, where the prior on effect size under $\mathcal{H}_1$ is symmetric around zero; specifically, $\mathcal{H}_1 : \rho \sim \text{Uniform}[-1, 1]$ (Jeffreys, 1961). In other words, the Bayes factor test compares the predictive performance of $\mathcal{H}_0$ (i.e., the correlation is absent) against that of an alternative hypothesis which states that all values of the population correlation coefficient are equally likely a priori. This comparison yields a Bayes factor in favor of $\mathcal{H}_1$ of about 2.

However, it may be argued that the alternative hypothesis ought to be amended in order to incorporate the theoretical knowledge that the relation between amygdalar activity and perceived stress is assumed to be positive rather than negative. In other words, we can define the alternative hypothesis so that it allows only positive population correlation coefficients: $\mathcal{H}_+ : \rho \sim \text{Uniform}[0, 1]$ (see also Jeffreys, 1961, p. 283; Wagenmakers et al., 2010). By allocating all prior mass to positive coefficients, the sharpened alternative hypothesis makes predictions that are more risky and better reflect the theory under test. The middle panel of Figure 1 shows that sharpening the alternative hypothesis has increased its predictive advantage over $\mathcal{H}_0$ and yields a

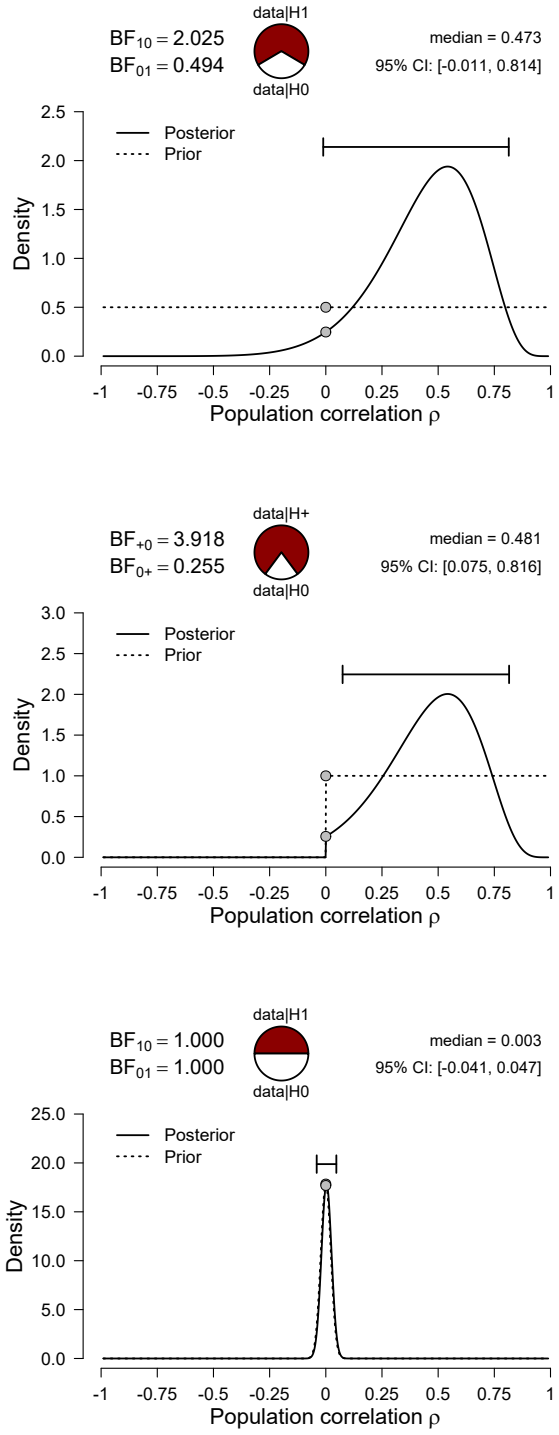

Figure 1. A model's predictive performance depends on the prior distribution: a demonstration for the correlation between Amygdalar Activity and Perceived Stress reported in Tawakol et al. (2017). Top panel: Results from a two-sided prior, $\mathcal{H}_1 : \rho \sim \text{Uniform}[-1, 1]$. Middle panel: Results from a one-sided prior, $\mathcal{H}_1 : \rho \sim \text{Uniform}[0, 1]$. Bottom panel: Results from a highly peaked prior, $\mathcal{H}_1 : \rho \sim \text{Stretched-Beta}[1000, 1000]$. Figures from JASP.

Bayes factor of about 3.9, almost a doubling of the evidence.[3] Note that, in this particular case, sharpening the alternative hypothesis does not result in a large change in the posterior distribution for the correlation coefficient (see van Ravenzwaaij, Monden, Tendeiro, & Ioannidis, 2019 for many other applied examples of the effect of one-sided versus two-sided hypotheses on the Bayes factor). One could state that the posterior distribution is 'robust' to the specification of the model as two-sided (i.e., undirectional) or one-sided (i.e., directional), but the epithet 'insensitive' appears to be more apt.

To underscore the importance of the prior distribution for assessing relative predictive success, the bottom panel of Figure 1 shows the result when the alternative hypothesis is specified to be extremely similar to $\mathcal{H}_0$; here, we used $\mathcal{H}_1 : \rho \sim \text{Stretched-Beta}[1000, 1000]$. The predictions from this alternative hypothesis are virtually identical to those from the point-null hypothesis $\mathcal{H}_0$, and with a Bayes factor of approximately 1, the data are almost perfectly nondiagnostic.[4] Consequently, it hardly matters whether we use the point $\mathcal{H}_0$ or the peaked $\mathcal{H}_1$ for inference and prediction; this is an important point that we will revisit later when we discuss a pragmatic solution to the critique that 'the point null hypothesis is never true'.

The upshot is that the prior distribution does affect the model predictions (and, consequently, the Bayes factor), but that this is hardly an 'issue' (Rouder, Morey, & Wagenmakers, 2016; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016; Rouder & Morey, 2019). Moreover, reasonable choices of priors will have a relatively moderate effect on the Bayes factor (see e.g., Stefan, Katsimpokis, Gronau, & Wagenmakers, 2021). The demonstration of TK was taken from Liu and Aitkin (2008)

[3]Nevertheless, this degree of evidence remains insufficient to convince a skeptic, as the data lower a 50% prior probability for $\mathcal{H}_0$ to about 20%, a non-negligible number.

[4]These Bayesian analyses can be obtained from JASP with a handful of mouse clicks and key strokes.

who used two extreme prior distributions – the Jeffreys's prior and the Haldane prior – that were proposed for estimation, and emphatically not for testing (e.g., Jeffreys, 1961, p. 251). No software package should offer these two prior distributions as a default for testing, and researchers with the statistical and technical skills to specify custom priors are unlikely to use these priors for testing. The fact that Liu and Aitkin (2008) employed these priors for testing (and the fact that TK followed their lead) bolsters our argument, outlined in the next section, that it is useful to offer practitioners a default prior distribution from which they can deviate only in the presence of strong prior information.

When strong prior information exists, it may come from theoretical constraints (e.g., M. D. Lee & Vanpaemel, 2018) or from expert knowledge (e.g., O'Hagan et al., 2006). The development of reliable procedures to elicit strong prior knowledge is challenging but holds considerable promise, as it makes the test more diagnostic (e.g., Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019; Stefan, Evans, & Wagenmakers, in press).

3. "Use of 'default' Bayes factors"

TK worry that the existence of 'default' Bayes factors might give practitioners the false impression that there exists only a single reasonable Bayes factor. First, although TK are right to worry about the blind application of Bayes factors, this worry applies to *any* statistical technique. Most data sets can be analyzed using a large variety of different statistical models (e.g., Silberzahn et al., 2018) and present a multitude of data selection options (e.g., Gelman & Loken, 2014). One may argue that any statistical method whatsoever requires experience, quantitative understanding, and sound judgment.

Second, in our experience the adoption of reasonable non-default prior distribu-

tions has only a modest impact on the Bayes factor (e.g., Stefan et al., 2021; Gronau, Ly, & Wagenmakers, 2020). This impact is typically much smaller than that caused by a change in the statistical model, by variable transformations, by different treatment of outliers, and so forth. To explain why the impact of prior distributions is often surprisingly modest, consider TK's critique that the default prior for the t-test –a Cauchy distribution centered at zero with scale parameter .707– is too wide. Specifically, this distribution assigns 50% of its mass to values larger than |.707|: if this is unrealistically wide, maybe the default prior distribution is of limited use, and the resulting Bayes factor misleading? Indeed, we ourselves have been worried in the past that the default Cauchy distribution is too wide, despite literature reviews showing that large effect sizes occur more often than one may expect (e.g., Aczel, 2018, slide 20; Wagenmakers, Wetzels, Borsboom, Kievit, & van der Maas, 2013). However, we recently realized that the impact of the 'wideness' is much more modest than one may intuit.

Consider two researchers, A and B, who analyze the same data set. Researcher A uses the default zero-centered Cauchy prior distribution with inter-quartile range of .707; researcher B uses the same prior distribution, but truncated to have mass only within the interval from $-.707$ to $+.707$. Assume that, in a very large sample, the observed effect is relatively close to zero. Researcher A reports a Bayes factor of 3.5 against the null hypothesis. It is now clear that the truncated default prior used by researcher B will provide better predictive performance, because no prior mass is 'wasted' on large values of effect size that are inconsistent with the data. As it turns out, truncating the default Cauchy to its interquartile range increases the predictive performance of the alternative hypothesis by a factor of at most 2. This means that the Bayes factor for B's truncated alternative hypothesis versus A's default 'overly wide' alternative hypothesis is at most 2; consequently, B will report a Bayes

factor against the null hypothesis that cannot be any larger than $2 \times 3.5 = 7$. This means that the potential predictive benefit of truncating the default distribution to its interquartile range is just as large as the potential predictive benefit of conducting a one-sided test instead of a two-sided test.[5] In other words, suppose a very large data set has an effect size of 0.3 with almost all posterior mass ranging from 0.2 to 0.4; the predictive benefit of knowing in advance the direction of the effect is just as large as the predictive benefit of knowing in advance that it falls within the prior interquartile range; consequently, the Bayes factor from a one-sided default Cauchy distribution is virtually identical to the Bayes factor from a two-sided default Cauchy distribution that is truncated to the $[-.707, +.707]$ interval.

What the foregoing shows is that although the width of the prior distribution is important for assessing relative predictive success, its impact should not be overstated. In particular, prior distributions that are overly wide (but not grotesquely wide) will nevertheless yield informative outcomes. We recommend that future critiques of the Bayes factor being dependent on the prior distribution include a careful robustness analysis across a range of plausible options.

Third, as acknowledged by TK, the default prior distributions are not chosen haphazardly, but are constructed to fulfill general desiderata that apply across a wide range of scientific contexts (e.g., Bayarri, Berger, Forte, & García-Donato, 2012; Consonni, Fouskakis, Liseo, & Ntzoufras, 2018). Consequently, the default prior allows a reference analysis that provides an objective point of departure in the possible formulation of alternative hypotheses that make riskier predictions. Default priors also act as an anchor in the sense that, if the non-default prior distribution yields a Bayes factor that is qualitatively different from that associated with the default prior,

[5]Note that this happens for exactly the same statistical reason: in both cases, 50% prior mass of poorly predicting values is eliminated.

this encourages the researcher to acknowledge explicitly that the result is substantially affected by the prior knowledge that was used to define the alternative hypothesis.

Fourth, default Bayes factors allow efficient communication of evidence that many researchers may consider less sensitive to human bias than more subjective or 'informed' alternatives. Fifth, without default distributions every researcher would be required to define their own prior distributions for every application, an unfamiliar activity that may deter many researchers from using Bayes factors at all. In our opinion, a default Bayes factor, however blindly applied, is usually preferable over no Bayes factor at all.

Sixth, for relatively complex models such as repeated-measures ANOVA and linear regression with many predictors, subjective specification of the prior distribution may be near impossible, and there is no realistic alternative to a default specification.

Seventh, software programs that offer Bayesian analyses, such as JASP (JASP Team, 2020), usually offer a sensitivity analysis, that is, a comprehensive investigation of the extent to which the Bayes factor differs across alternative specification of the prior distribution. The presence of a default prior allows this sensitivity analysis to proceed in a more systematic fashion.

4. "Bayes factors are not posterior model probabilities"

TK warn that Bayes factors quantify the relative probability of the data given two rival models, and not the relative probability of two rival models given the data, which TK believe is the more relevant number. In other words, posterior model probabilities quantify the models' relative plausibility after the data have been observed, whereas the Bayes factor quantifies the *change* in this plausibility brought about by the observed data. Thus, posterior probabilities quantify belief, whereas Bayes factors quantify evidence (e.g., Evans, 2015).

Although it is clearly important to distinguish between *belief* and *evidence*, both concepts are important. The question at hand is what concept is more relevant when researchers wish to communicate the results of their experiments to their peers. In our opinion, what researchers should refrain from doing is to present posterior model probabilities without clarifying the extent to which those probabilities are a result of the evidence. Thus, whenever posterior model probabilities are reported, so should its constituent elements: the prior model probabilities and the evidence. And in such a case, one may further debate how much added value is provided by reporting the prior model probabilities, as these mostly measure the researchers' enthusiasm for the hypotheses under consideration. Barring exceptional cases such as extra-sensory perception, it therefore seems prudent for empirical reports to emphasize the evidence, and leave it to individual readers to construct their own posterior probabilities by combining that evidence with one's prior conviction, information, or enthusiasm. An alternative reporting format is to present the Bayes factor and accompany it by a range of prior model probabilities, yielding an associated range of posterior model probabilities (Etz, Haaf, Rouder, & Vandekerckhove, 2018). A recent survey suggests that among 31 authors of articles published in *Nature Human Behaviour*, none believe their key claims were either outlandish or trivial a priori, with the lowest prior probability at 0.20 and the highest at 0.75 (van Doorn et al., in press). Regardless, we personally believe that for drawing conclusions from data in most scientific applications, evidence is the key concept. When the goal is to convince a skeptical scientific audience that a particular claim has empirical support, it is evidence –and not belief– which is required. The extent to which practitioners agree with this statement is an open empirical question.

In sum, we agree with TK that "It is essential to understand this distinction [between evidence and belief – DvR & EJW] in order to avoid erroneous interpretations

of Bayes factors. As stated in one of the first statistics articles on the Bayes factor, "To raise the probability of a proposition from 0.01 to 0.1 does not make it the most likely alternative." (Jeffreys, 1935, p. 221). We disagree with TK that researchers are primarily interested in degree of belief and not evidence when communicating and consuming scientific results.

5. "Bayes factors do not imply a model is probably correct"

TK warn that Bayes factors are a model comparison tool that ignores absolute model performance. Specifically, the Bayes factor concerns the *relative* predictive adequacy of rival models. Model A may be favored by the Bayes factor and decisively outpredict model B, but this does not mean that model A is likely to be the "true model", as both models may be dramatically misspecified (for a concrete example see Wagenmakers et al., 2018, p. 50).

In general, Bayes' rule shows that *all* Bayesian inference is inherently relative – probability is distributed across a range of alternatives, and knowledge updates are relative to the performance of those alternatives. As explained by Lindley (1993, p. 25):

> "The Bayesian method is comparative. It compares the probabilities of the observed event on the null hypothesis and on the alternatives to it. In this respect it is quite different from Fisher's approach which is absolute in the sense that it involves only a single consideration, the null hypothesis. All our uncertainty judgements should be comparative: there are no absolutes here. A striking illustration of this arises in legal trials. When a piece of evidence E is produced in a court investigating the guilt G or innocence I of the defendant, it is not enough merely to consider the probability of E assuming G; one must also contemplate the probability

> of E supposing I. In fact, the relevant quantity is the ratio of the two probabilities. Generally if evidence is produced to support some thesis, one must also consider the reasonableness of the evidence were the thesis false. Whenever courses of action are contemplated, it is not the merits or demerits of any course that matter, but only the comparison of these qualities with those of other courses."

So the comparative nature of the Bayes factor pertains to all Bayesian inference, and is also a feature of the posterior distribution under $\mathcal{H}_1$, whose inspection is the method TK recommend in the model misspecification context when they state, "Instead, analyzing the posterior distribution for the parameter being tested offers a much richer insight". In fact, we believe the opposite is true: the TK-posterior-inspection method is arguably more susceptible to model misspecification than the Bayes factor, because the Bayes factor does not commit fully to a single model –in this case, $\mathcal{H}_1$– entirely on a priori grounds. The Bayes factor involves the specification of two models, not just one, and the Bayes factor methodology may be seamlessly expanded to include very many models –hundreds, thousands, or even millions, depending on the application– thereby increasing robustness and decreasing the probability of problematic model misspecification (e.g., Hinne, Gronau, van den Bergh, & Wagenmakers, 2020). In contrast, the TK-posterior-inspection method takes one of the models from the Bayes factor model space and ignores all others. Ignoring alternative models is not a good method to detect model misspecification and make statistical inference more robust. This disadvantage of the TK-posterior-inspection method becomes clear with informed priors. Suppose we estimate a binomial chance θ and assign it a highly informed beta$(500, 500)$ prior distribution. The data consist of 20 successes, which yields a beta$(520, 500)$ posterior distribution, hardly different

from the prior distribution. Inspecting only the beta$(520, 500)$ posterior distribution, one may conclude "a posteriori, the binomial chance θ is likely to be very close to 0.5". The considerable conflict between data and prior (i.e., the model misspecification; Box, 1980) is not evident from the posterior distribution.

One may generalize the TK-posterior-inspection method and include a visual inspection of the *prior* distribution. This is better, but in the above scenario it can still easily lead to the erroneous conclusion that "the data are not very informative; after all, the posterior distribution is very close to the prior distribution". Again, this conclusion ignores the model misspecification, that is, the data-prior conflict. So the TK-posterior-inspection method is not a panacea, and the spectre of model misspecification haunts almost all of statistical inference, and will continue to do so in the foreseeable future.

In sum, for *all* statistical work it is essential to check for model misspecification and confirm that the inference is valid (e.g., Anscombe, 1973). Putting all one's inferential eggs into one model basket, as the TK-posterior-inspection method does, is not a solution to this long-standing and pernicious problem.

6. "Qualitative interpretation of Bayes factors"

Bayes factors provide a continuous measure of evidence. TK are concerned that –despite several proposals in the literature (Jeffreys, 1961; Kass & Raftery, 1995)– no objective amount of evidence counts as 'strong', or 'much'. In general, it is rarely the case that a quantitative measure can be categorized in a qualitative fashion that is context-independent, objective, or universally accepted. A notable exception is Bayesian decision analysis (Berger, 1985; Lindley, 1985), where a specific criterion level of evidence is required in order to trigger a particular all-or-none decision (e.g.,

allowing a particular drug on the market).[6]

We believe that the evidence categories proposed by Jeffreys (1961) are useful as a rough heuristic, a rule-of-thumb guideline that helps prevent enthusiastic researchers from overinterpreting Bayes factors that represent weak evidence. However, any discretization of a statistical measure that is inherently continuous (e.g., p-values, bounds on confidence intervals, Bayes factors, or indeed posterior probabilities) will be somewhat arbitrary and lead to a loss of information. This was mentioned by Jeffreys himself when he discussed thresholds on BF_{01} (which he termed K): "we are at liberty to surround $K = 1$ by two other values and say that within this range the data are not sufficiently decisive, and even this device would be purely one of convenience and sacrifice some information given by the actual values of K." (Jeffreys, 1938, p. 378). To facilitate the interpretation of the continuous strength of evidence, JASP (JASP Team, 2020) presents Bayes factors by probability wheels.

TK conclude that "the best solution is to not report Bayes factors only, but to also report posterior model probabilities". The main advantage of this proposal is that the Bayes factor evidence is placed in context. For instance, a Bayes factor of 200 may be 'extreme' (M. D. Lee & Wagenmakers, 2013) when it concerns the test of list-strength effects in recall, but it may be inconsequential when the hypothesis is that neutrinos travel faster than the speed of light. As explained above under issue 4, our preference is to emphasize the evidence (cf. Aczel et al., 2020; Jeffreys, 1935), but we acknowledge that prior model probabilities are currently undervalued and that they can be useful when it comes to interpreting the evidence in qualitative terms.

[6] A Bayesian decision analysis requires not only prior belief and evidence, but also *utilities*, that is, the costs associated with acting on incorrect beliefs (e.g., approving an ineffective drug and rejecting an effective drug).

7. "Bayes factors test model classes"

NHBT compares a single null hypothesis (typically $\mathcal{H}_0 : \delta = 0$) to a composite alternative hypothesis, or what TK call a model class (typically $\mathcal{H}_1 : \delta \neq 0$, $\mathcal{H}_- : \delta < 0$, or $\mathcal{H}_+ : \delta > 0$ with associated prior distributions for δ). TK provide an example of how, in the case of a sample size typically found in psychology, it can happen that when the true parameter value is close to the point null value, the Bayes factor favors the point null model over the composite alternative due to the fact that the composite model is a weighted average over all parameter values, many of them much larger than the true parameter value. This finding may appear undesirable, but as it follows directly from Bayes' rule, it warrants a closer look.

Consider the researcher interested in computing the Bayes factor, that is, the relative predictive performance of $\mathcal{H}_0$ and $\mathcal{H}_1$. The researcher is unable to specify the alternative hypothesis as a single point, because the true population value is usually highly uncertain a priori. This uncertainty translates into a prior distribution on the test-relevant parameter δ. An immediate benefit of this prior distribution is that it acts as an implicit correction for multiple testing: the price the researcher pays for not committing to a single δ is that the prior resources have to be distributed across the different options. This is similar to when a risk-averse gambler distributes his stake across many potential winners; doing so will inevitably reduce his payoff compared to putting his entire stake on the winner.

The prior distribution therefore reflects one's uncertainty about the size of the effect; it can be interpreted as a bet on effect size, given that $\mathcal{H}_1$ holds. It follows that when the observed effect is surprisingly small (compared to the expectations encoded in the prior distribution), the bet was poor and, with a typical sample size, the evidence may therefore favor $\mathcal{H}_0$. We regard this as desirable rather than

problematic. What is problematic is that the expectations were overly optimistic and the sample is small. In a comment online, Richard Morey concludes that "the 'bias' toward the null for small effect sizes is exactly what must happen for any reasonable method...Although it is counter intuitive, we would be worried if it *didn't* happen for some measure of evidence."[7] Similarly, Jeffreys (1961, p. 248) addresses the TK concern directly when he explains the rationale of the Bayes factor hypothesis test:

> "The difficulty pointed out before (...) about the uniform assessment of the prior probability was that even if α [DvR & EJW: the true value of the test-relevant parameter] was 0, a [DvR & EJW: the sample estimate] would usually be different from 0, on account of random error, and to adopt a as the estimate would be to reject the hypothesis $\alpha = 0$ even if it was true. We now see how to escape from this dilemma. Small values of $|a|$ up to some multiple of s [DvR & EJW: the standard error] will be taken to support the hypothesis $\alpha = 0$, since they would be quite likely to arise on that hypothesis, but larger values support the need to introduce α. In suitable cases high probabilities may be obtained for either hypothesis. The possibility of getting actual support for the null hypothesis from the observations really comes from the fact that the value of α indicated by it is unique. q' [DvR & EJW: $\mathcal{H}_1$] indicates only a range of possible values, and *if we select the one that happens to fit the observations best we must allow for the fact that it is a selected value.* [italics ours] If $|a|$ is less than s, this is what we would expect on the hypothesis that α is 0, but if α was equally likely to be anywhere in a range of length m it requires that an event with a probability $2s/m$ shall have come off. If $|a|$ is much larger

[7] See `https://richarddmorey.org/category/bayes-factor/`, post "All about that 'bias, bias, bias' (it's no trouble)", April 10, 2015.

> than s, however, a would be a very unlikely value to occur if α was 0, but no more likely than any other if α was not 0. In each case we adopt the less remarkable coincidence."

In a later chapter, Jeffreys mentions that as sample size grows, our judgement can be revised:

> "It is worth while to devote some attention to considering *how* a law [DvR & EJW: $\mathcal{H}_0$], once well supported, can be wrong. A new parameter rejected by a significance test [DvR & EJW: Jeffreys's Bayes factor significance test] need not in fact be zero. All that we say is that on the data there is a high probability that it is. But it is perfectly possible that it is not zero but too small to have been detected with the accuracy yet attained. We have seen how such small deviations from a law may be detected by a large sample when they would appear to have been denied by any sub-sample less than a certain size [DvR & EJW: this refers to the concrete example on p. 333], and that this is not a contradiction of our general rules." (Jeffreys, 1961, p. 367)

8. "Mismatch between Bayes factors and parameter estimation"

TK observe that there can be situations where 95% credible intervals include the null value, but Bayes factors indicate support for the alternative hypothesis, or vice versa. This is again a direct consequence of Bayes' rule. The reason for the discrepancy is that the 95% credible interval ('estimation') is usually conditional on $\mathcal{H}_1$, whereas Bayes factors ('testing') concern the tenability of $\mathcal{H}_0$ vs. $\mathcal{H}_1$. As the two analyses depart from different assumptions and address different questions, there is no reason to expect them to produce the same answer. Indeed, the discrepancy be-

tween these two approaches was the main motivation for Jeffreys's pioneering work in Bayesian inference (e.g., Etz & Wagenmakers, 2017; Howie, 2002; Jeffreys, 1935, 1939; Ly, Verhagen, & Wagenmakers, 2016; Ly et al., 2020; Robert, Chopin, & Rousseau, 2009; Wrinch & Jeffreys, 1921). As remarked by Berger (2006, p. 383): "Bayesians cannot test precise hypotheses using confidence intervals. In classical statistics one frequently sees testing done by forming a confidence region for the parameter, and then rejecting a null value of the parameter if it does not lie in the confidence region. This is simply wrong if done in a Bayesian formulation (and if the null value of the parameter is believable as a hypothesis)." (see also Bayarri & Berger, 2013). Because classical confidence intervals are often numerically close (or even identical) to Bayesian credible intervals, Berger's admonition still holds, that is, Bayesians cannot test precise hypotheses using credible intervals (for an extended discussion see Wagenmakers et al., 2020).

Practically speaking, a 95% confidence interval that does not overlap with zero can be interpreted as a two-sided significance test rejecting the null hypothesis with significance level $\alpha = .05$. What sort of parallel interpretation could novice practitioners of NHBT make? If the Bayesian 95% credible interval does not encompass zero, it is not immediately obvious what 'significant Bayes factor' level would be falsely implied. One common misinterpretation is as follows. Suppose 96% of the continuous posterior mass is on positive effect sizes, and 4% is on negative effect sizes. This could be misinterpreted to mean that the Bayes factor against the null hypothesis is $.96/.04 = 24$. After all, the posterior probability that the effect is positive is .96. This interpretation is incorrect, because the continuous prior distribution did not assign any mass to the null hypothesis a priori; the null hypothesis was deemed false from the outset, and hence no amount of data can either support or undercut it. Instead, with a continuous prior distribution that does not express a preference for positive or

negative effect sizes, the Bayes factor of 24 quantifies the evidence that the effect is positive rather than negative, under the assumption that the effect is not zero. This misinterpretation is not, however, a risk of using a Bayes factor; instead, it is a risk of abusing the posterior distribution to address a question (i.e., "what evidence do the data provide for the null hypothesis versus an alternative hypothesis?) that only a Bayes factor can answer. In other words, a question of hypothesis testing ought not be answered by estimating a parameter.

The estimation and testing frameworks can be brought in line, however, by considering the *unconditional* prior and posterior distributions for effect size δ, that is, a "spike and bell" distribution, where the mass on the spike at $\delta = 0$ represents the plausibility of the null hypothesis. Traditionally, in parameter estimation, the focus is firmly on the location and width of the bell (i.e., the posterior knowledge about δ under $\mathcal{H}_1$) and no attention is paid to the height of the spike. But, as recently pointed out by Rouder, Haaf, and Vandekerckhove (2018, p. 108) "If one admits the possibility of the spike, then assuredly it should affect posterior estimation as well." (for a more detailed treatment see van den Bergh, Haaf, Ly, Rouder, & Wagenmakers, in press). Another way to bring the estimation and testing frameworks in line is by constructing an interval that encompasses only values that receive a certain minimum Bayes factor support from the data (e.g., Evans, 2015; Wagenmakers, Gronau, Dablander, & Etz, in press). The final way is to recognize that Bayesian parameter estimation, just as Bayes factor hypothesis testing, is fundamentally about changes in plausibility brought about by relative predictive success (Rouder & Morey, 2019; Wagenmakers, Morey, & Lee, 2016): regardless of whether they are labelled parameters or models, accounts of the world that predict observed data relatively well experience a gain in plausibility, whereas accounts that predict relatively poorly suffer a decline.

9. "Bayes factors favor the point null model"

TK observe that default Bayes factors are more conservative than p-values (see also van Ravenzwaaij & Ioannidis, 2017). That is, whereas p-values just below .05 may prompt the conclusion to "reject the null hypothesis", the corresponding Bayes factors usually indicate the evidence against the null hypothesis is only weak; for large sample sizes, the Bayes factor may even reveal that the point null hypothesis is supported rather than undercut (e.g., Jeffreys, 1935; Lindley, 1957). Several remarks are in order. First, as far as p-values are concerned, their all-or-none nature in Neyman-Pearson α-level NHST implies that strong decisions will sometimes be made on weak evidence. We agree with Robinson (2019, p. 246): "In my opinion, inference (as opposed to merely choosing) must always allow the additional option of stating that the evidence is not sufficiently strong for a reliable choice to be made between the hypotheses." (see also Rouder, Morey, & Wagenmakers, 2016; Rouder, Morey, Verhagen, et al., 2016). Second, it is important that the Bayes factor is a unique and deterministic consequence of the prior distribution and the data; complaints about the behavior of the Bayes factor are therefore indirectly complaints about the prior distribution or the data. When the prior distribution is specified to one's satisfaction, the data then give rise to a single Bayes factor that reflects the evidence.

This 'backward propagation of discontent' is a general property of Bayesian inference that is worth emphasizing. Bayesian inference is *coherent*, in the sense that it disallows conclusions that are internally inconsistent. As stated by Lindley (2006, p. 37), "Coherence is the most important tool that we have today for the measurement of uncertainty, in that it enables you to pass from simple, measurable events to more complicated ones. Coherence plays a role in probability similar to the

role Euclidean geometry plays in the measurement of distance."[8] Imagine a perfect chef who creates the best possible dish (tailored to your tastes) given the available ingredients. If you nevertheless strongly dislike the dish, this can only mean that the ingredients were poor, and it would be inappropriate to critique the chef. Similarly, if you find posterior conclusions entirely unacceptable, this signals a problem with prior knowledge or with the data, but not with the learning mechanism itself.

Consequently, we maintain that is is impossible to devise a statistical scenario where a rational actor would be pleased, a posteriori, with the statistical specification of the rival models and the data, but at the same time displeased with the resulting Bayes factor—coherence forbids this scenario from arising.

Quite correctly, in our opinion, TK do not blame the Bayes factor for being conservative; instead, they conclude that p-values are trigger-happy: "p-values overstate the evidence against the point null model, both in terms of the Bayes factor as in terms of posterior model probabilities. In this regard, NHBT does offer an advantage over NHST." TK also point out that, for tests of direction (i.e., is the effect larger or smaller than zero), Bayes factors and p-values are relatively similar (Casella & Berger, 1987). Indeed, from a Bayesian perspective, the p-value is an approximate Bayes factor test for direction (Marsman & Wagenmakers, 2017). A test for the direction of an effect, however, is qualitatively different from a test for the presence of an effect.

Where we disagree is that TK argue that the point null hypothesis is almost never true, and consequently should not be assigned separate prior mass. This popular argument has met with several rejoinders (e.g., Kass & Raftery, 1995, pp. 788-789; Iverson, Wagenmakers, & Lee, 2010, pp. 175-176) and it would have been interesting

[8] cf. "This theorem [DvR & EJW: Bayes's theorem] is to the theory of probability what Pythagoras's theorem is to geometry." (Jeffreys, 1931, p. 19).

to learn why TK believe these do not carry much weight. Briefly, we believe that there is scientific merit to Jeffreys's razor, which states that "variation must be taken as random until there is positive evidence to the contrary" (Jeffreys, 1939, p. 345). The notion that variation is entirely random constitutes the idealized position of a skeptic; one may dismiss this position on a priori grounds (i.e., in nature, 'everything affects everything else'), but we believe that it is more compelling to dismiss a skeptic's position based on data. The legitimacy of the point null hypothesis is bolstered by the fact that several effects from psychology failed to replicate, even in very large samples (e.g., Camerer et al., 2018; Klein et al., 2018; Wagenmakers, Beek, et al., 2016); empirically, it appears that the point null hypothesis is not so easily discarded after all. In addition, it should be stressed that the Bayes factor is based on a comparison of *predictive* performance that is independent of the notion of absolute or relative model truth. (Wagenmakers, Grünwald, & Steyvers, 2006). The idea that we should not use models that we know to be false can easily result in an inferential impasse, because all statistical models are ultimately false.

Another rejoinder, one that was acknowledged explicitly by TK, is that the point null is merely a mathematically convenient approximation, "a hazily defined small region rather than a point" (Edwards, Lindman, & Savage, 1963, p. 235). As explained by Cornfield (1966, p. 582):

> "There is a psychological difficulty felt by some to the concentration of a lump of probability at a single point. Thus, even though entirely convinced of the ineffectiveness of whiskey in the treatment of snake bite they would hesitate to offer prior odds of p to $1-p$ that the true mortality difference between treated and untreated is zero to an arbitrarily large number of decimal places.[DvR & EJW: Note that 'p' here refers to the

> posterior probability for the null hypothesis, not to the classical p-value] But if the concentration is regarded as the result of a limiting process it appears unexceptional. To say that the treatment is ineffective means that the hypothesis $\mathcal{H}_\delta : |\theta| \neq |\delta|$ is true, where δ is quite small, perhaps of the order of 1 death among all persons bitten by venomous snakes in a decade, but not specifiable more precisely. For finite sized samples the probability of rejecting either $\mathcal{H}_0$ or $\mathcal{H}_\delta$ will be nearly equal, and concern about the high probability of rejecting one is equivalent to concern about rejecting the other."

One reviewer pointed out that it would nevertheless be worthwhile to specify the null hypothesis as a sharply peaked prior distribution (i.e., a 'peri-null hypothesis') rather than a spike (i.e., a 'point-null hypothesis'). The predictive advantage of specifying a peri-null hypothesis over a point-null hypothesis is given by the Bayes factor that directly compares both null hypotheses (Morey & Rouder, 2011).[9] When the peri-null hypothesis is sharply peaked around the null value, and the number of observations is not astronomical, the predictive performance of both null models will be highly similar (e.g., van Ravenzwaaij & Etz, in press), and consequently (1) the Bayes factor for the peri-null hypothesis versus the point-null hypothesis will be near 1; and (2) the Bayes factor against the alternative hypothesis will be virtually the same regardless of whether one uses the peri-null or the point-null hypothesis.

As a concrete example, consider again the analysis of the correlation between amygdalar activity and preceived stress reported by Tawakol et al. (2017). The top panel of Figure 1 showed that the Bayes factor in favor of $\mathcal{H}_1$ over the point-null hypothesis was 2.025. But instead of the point-null hypothesis we could have specified a peri-null hypothesis, for instance as $\mathcal{H}_0^{'} : \rho \sim \text{Stretched-Beta}[1000, 1000]$. The

[9] See also `https://jasp-stats.org/2017/10/25/test-interval-null-hypotheses-jasp/`.

bottom panel of Figure 1 shows that for the Tawakol et al. data set of 13 patients, the predictive performance of the point-null $\mathcal{H}_0$ is almost equal to that of the peri-null $\mathcal{H}_0'$; in fact, the rounded value of the Bayes factor between both null hypotheses is 1.000. Consequently, the Bayes factor for $\mathcal{H}_1$ over the peri-null hypothesis is also 2.025, the same value that was obtained for a comparison to the point-null hypothesis.

Although the predictive advantage of the peri-null over the point-null will generally be modest at best, we acknowledge that the peri-null specification may prove useful from a rhetorical point of view, as it nips in the bud any discussion about 'the null hypothesis is never true'. However, use of the peri-null will naturally spark a new discussion, one about the extent to which the peri-null resembles the point-null (i.e., the width of the peri-null). If the peri-null is adopted merely to respect the theoretical argument that, with an astronomical number of observations, it will be discovered that the point-null is never exactly true, then we believe that the peri-null should be extremely narrow.

Alternatively, there may be practical reasons to replace the point-null hypothesis with an interval-null hypothesis that restricts effect size to lie within a certain range from zero. This interval-null represents effect sizes that are too small to matter for the specific practical application under consideration (e.g., Morey & Rouder, 2011). The parameter space covered by the interval-null hypothesis usually does not overlap with that covered by the alternative hypothesis. Bayes factors for the resulting *equivalence tests* can be obtained using various software programs (e.g., baymedr, Linde & van Ravenzwaaij, 2019; BaIn, Hoijtink et al., 2019; BayesFactor, Morey et al., 2018, and JASP).

10. "Bayes factors favor the alternative"

TK observe that when sample size is very large, Bayes factors will favor the alternative hypothesis even when the non-zero true effect size is minuscule. Note that this is how Bayes factors should behave – as sample size goes to infinity, default Bayes factors will increasingly support the true model (e.g., Consonni et al., 2018).

Consider an example even more extreme than that offered by TK: The predictions from a model that stipulates the true population effect size to be $\delta = .01$ are virtually indistinguishable from the predictions of the null model. Such highly similar accounts can only be discriminated with an overwhelming amount of informative data. For instance, if the sample effect size equals the population effect size of 0.01, for a one-sample t-test one needs $114,035$ participants before a default JZS Bayes factor starts supporting the alternative hypothesis (see also Figure 5 in Rouder et al., 2009).[10]

Nevertheless, we agree with TK that it is good practice to report effect sizes. This is also consistent with Jeffreys's strategy, which was to start the statistical investigation with a test in order to determine whether the null hypothesis could be discarded, meaning some effect exists that is worthy of estimation (Haaf, Ly, & Wagenmakers, 2019). Only if the data provided clear evidence against the null hypothesis would Jeffreys turn to estimation, which he then based on the alternative hypothesis. Thus, for the case of $\delta = .01$ with an overwhelmingly large data set, Jeffreys would have concluded "Yes, the effect exists (this follows from testing), but the size of the effect is tiny (this follows from estimation)." In our opinion, this conclusion is intuitive and eminently reasonable – we are unsure what qualitatively different conclusion could be drawn.

[10]This can be verified in the R-package BayesFactor (Morey & Rouder, 2018): "exp (ttest.tstat(t=.01*sqrt(114035), n1=114035, rscale = 1/sqrt(2))[['bf']])".

Mistakes in interpretation can arise when either testing or estimation is applied to the exclusion of the other. With a sole focus on testing, one would report only a Bayes factor and state "the effect exists". Without any other information, this statement may tempt the reader to conclude that the effect is large or practically relevant. In other words, the "test-only" approach stops the inference process prematurely. With a sole focus on estimation, on the other hand, one would report only a credible interval and state "95% of the posterior mass falls between .006 and .014 (say)". This leaves unaddressed the question of whether the effect exists in the first place; the statement may tempt the reader to conclude that the data support the hypothesis that the effect is present. In the estimation framework, this conclusion (i.e., "the effect exists") is already assumed from the outset. In other words, the "estimate-only" approach jumps the gun and skips the first stage of the scientific process – to establish that a phenomenon is worth estimating in the first place. (e.g., Fisher, 1928, p. 274; Jeffreys, 1939, p. 345).

Related to this, a reviewer reminded us that in their paper, TK state that "Johnson and Rossell (2010) show the following: Bayes factors accumulate evidence in favor of true $\mathcal{M}_1$ much faster than they do in favor of true $\mathcal{M}_0$ as the sample size increases, for fixed sample-based estimates of θ. That is, although Bayes factors allow drawing support for either $\mathcal{M}_0$ or $\mathcal{M}_1$, they do so asymmetrically. This property is in contrast with the commonly praised feature of Bayes factors being symmetric (in the sense that they allow accumulating evidence for either model), unlike p-values."

The contrast that TK allude to does not exist. Bayes factors do quantify evidence, either for $\mathcal{H}_0$ or for $\mathcal{H}_1$, but they do not need to do this at an equal rate, nor is it clear why this would be at all desirable. In general, the claim that something is absent is more difficult to support than the claim that something is present, at least when one is uncertain about the size of the phenomenon that is present. Consider,

for instance, the null hypothesis "There is no animal in this room", tested against the alternative hypothesis: "There is an animal in this room, but it could be as small as an ant or as big as a cow". Now if the "effect" is of medium size (say a cat), it can be quickly discovered and $\mathcal{H}_1$ then receives decisive support. But if a cursory inspection does not reveal any animal, then support for $\mathcal{H}_0$ will only be weak (after all, it is easy to miss an ant). Now there is a way to collect strong evidence for $\mathcal{H}_0$, but it requires more effort – a systematic search with a magnifying glass, for instance. So instead of being problematic, the asymmetry in the rate of increase in evidence is desirable, in line with common sense, and indeed a direct mathematical consequence of how the competing models were constructed.

11. "Bayes factors often agree with p-values"

TK argue that in certain cases, Bayes factors and p-values lead to similar conclusions. This is both trivially true and trivially false. It is trivially true because of what is known as Berkson's interocular traumatic test (Edwards et al., 1963, p. 217): "you know what the data mean when the conclusion hits you between the eyes. The interocular traumatic test is simple, commands general agreement, and is often applicable; well-conducted experiments often come out that way." The interocular traumatic test also brings to mind Lord Rutherford's statement that "if your experiment needs statistics, you ought to have done a better experiment". Simply put, careful experimental design will reduce measurement error to such a degree that any reasonable method of inference (and even some unreasonable ones) will arrive at the same conclusion.

At the same time, the fact that p-values often lead to conclusions that are similar to those from Bayes factors is also trivially false. After all, conclusions that follow from p-values are usually *dichotomous*: the result is either statistically significant

or non-significant. In contrast, conclusions that follow from Bayes factors are of a graded nature, as Bayes factors quantify relative predictive performance of two rival hypotheses. The coarsest categorization of Bayes factors applied in practice is *trichotomous*: the result yields a satisfactory level of evidence for $\mathcal{H}_0$, a satisfactory level of evidence for $\mathcal{H}_1$, or an unsatisfactory level of evidence, meaning that both hypotheses are supported by the data to a degree that is about equal, and more observations are needed to reach a definite conclusion.

As a demonstration of the difference between dichotomous p-values and trichotomous Bayes factors, consider the following two scenarios that may arise from a two-sided one-sample test with $\mathcal{H}_1 : \delta \sim \text{Cauchy}(0, \sqrt{2}/2)$:

- For $t = 2$, $n = 10$: p-value $= 0.077$ ("not statistically significant"), but $\text{BF}_{10} = 1.28$ (a smidgen of evidence for $\mathcal{H}_1$).
- For $t = 2$, $n = 100$: p-value $= 0.048$ ("statistically significant"), but $\text{BF}_{10} = 0.75$ (a smidgen of evidence for $\mathcal{H}_0$).

The Core of the Disagreement

The core of our disagreement with TK is twofold. First, TK argue that "We find it ill-advised to suggest that researchers move from NHST towards NHBT without enduring the growing pains of learning the basics of this new statistical tool." Although this remark appears eminently reasonable, we worry that in practice, researchers will refer to the TK paper mainly as an excuse for avoiding NHBT altogether and not learning anything new at all. Suggestions by reviewers to report a Bayes factor may be countered by referring to the TK paper, listing the eleven 'issues' and suggesting that advanced statistical knowledge is required to safely traverse a veritable Bayes factor mine field. But the relevant question is not "will NHBT be used responsibly,

after careful deliberation, and without resorting to statistical rituals?" Instead, the relevant questions are "does NHBT constitute a step forward compared to the status quo?" and "can NHBT usefully *supplement* the inference from NHST?" or even "when it comes to answering the questions that researchers care about, is there any alternative to NHBT at all?" We note that if a thorough understanding were a prerequisite for statistical reporting, the use of p-values and confidence intervals would be all but forbidden (e.g., Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016, and references therein). As summarized by Sellke, Bayarri, and Berger (2001, p. 71), "The most important conclusion is that, for testing 'precise' hypotheses, p values should not be used directly, because they are too easily misinterpreted. The standard approach in teaching—of stressing the formal definition of a p value while warning against its misinterpretation—has simply been an abysmal failure."

We believe that NHBT, even if executed as a thoughtless ritual, still markedly improves on the status quo. We believe that in most situations, the default prior distributions are reasonable, and informed prior distributions will generally not greatly alter the qualitative pattern of results (see also Dawid, 2011). The ability of NHBT to obtain evidence in favor of the null hypothesis, to discriminate between absence of evidence and evidence of absence, and to monitor the evidence as data accumulate are profound steps forward compared to the standard report of '$p < .05$' or '$p > .05$' (e.g., Wagenmakers, Morey, & Lee, 2016). In fact, by being able to deviate from the default settings and specify an informed prior distribution, NHBT invites statistical thinking more than does NHST, whose procedures are etched in stone.

In sum, one may certainly argue that NHBT is like handing a toddler the keys to a Tesla. However, the Tesla comes equipped with an autopilot (i.e., the default prior settings and the intuitive interpretation of the results); moreover, up to this

point the toddler has been driving a 1970s Trabant (i.e., NHST). The choice is not between Tesla or binky – the status quo is the Trabant. We do support the call for more statistical thinking, but we are pessimistic about the prospects of substantially reducing the use of statistical rituals. It is true that TK advocate NHBT over NHST (e.g., "we want to state very clearly that NHBT is an improvement over NHST") but we are afraid that, for many readers, the take-away message is exactly the opposite.

Our second disagreement centers on the TK claim that Bayesian estimation (e.g., inspecting the posterior for effect size under $\mathcal{H}_1$) constitutes an adequate alternative to NHBT. For instance, TH state "Having the full posterior distribution can suffice for model comparisons too. For instance, comparing the model $M_0 : |\theta| < \epsilon$ that the parameter is close to zero, with its complement, $M_1 : |\theta| > \epsilon$ can be done by simply directly computing the posterior odds ratio for these two models on the basis of the full posterior distribution." and "estimation of the full posterior distribution offers a more complete picture". The procedure that TK propose is simple, intuitive, and has a history that dates back to Bayes and Laplace. Unfortunately, the method can also be seriously misleading, as it begs the question – it assumes the falsity of the null hypothesis, which is often the very target of inference. The deficiencies of the posterior estimation methodology were first outlined by Wrinch and Jeffreys (1921), and it was to address these deficiencies that Jeffreys later developed NHBT (see also Etz & Wagenmakers, 2017). To regress from Jeffreys's approach (i.e., first do testing, then follow it up by estimation in case the data undercut $\mathcal{H}_0$) back to Laplace's approach (i.e., only use estimation) requires that Jeffreys's arguments are discussed and refuted – which is something that neither TK nor anybody else has done.

We grant TK that in many scenarios, the presence of an effect is not of interest; the value of 0 is not special, and, consequently, what is needed is Laplacian estimation by means of a continuous posterior distribution. For instance, when the purpose is

to learn about someone's IQ, interrater reliability, or the proportion of Starbucks customers who order a latte, hypothesis tests are simply not appropriate.

Nevertheless, many psychological experiments are designed to answer a question formulated as a hypothesis test. In such scenarios, the value of 0 (i.e., the absence of an effect) stands out as special and warrants separate attention. For instance, D. S. Lee, Kim, and Schwarz (2015, Study 2) examined the hypothesis that people perform better on the Wason rule discovery task when they are exposed to the smell of fish (supposedly arousing suspicion; "something smells fishy"). The authors reported that "As predicted, participants were more likely to generate at least one negative test in the fishy condition (21 out of 44, 47.7%) than in the control condition (13 out of 47, 27.7%), $\chi^2(1, N = 91) = 3.91$, $p = 0.048$".[11] Suppose that these results were reported by means of a posterior distribution on the log odds ratio, assuming that the effect is present. We believe that researchers who interpret this posterior distribution would find it almost impossible to resist the temptation to include some sort of comparison to the value of 0, for instance by stating "most of the posterior distribution is located away from 0" or "most posterior mass is assigned to values higher than 0". After all, the primary claim under scrutiny is that there exists an effect. In sharp contrast, there would be no such temptation to draw comparisons to any other value of effect size; for instance, the statement that "most of the posterior distribution is located away from 0.0257251" would be preposterous. This state of affairs suggests that there is something special about 0, and that skeptical scientists will demand that the data undercut this special value before entertaining the proposition that fishy smells indeed affect performance on the Wason rule discovery task.

[11] NB. A Bayesian reanalysis with a one-sided default alternative hypothesis (i.e., a folded standard normal prior on the log odds ratio) yields a Bayes factor of 4.14 in favor of $\mathcal{H}_+$ over $\mathcal{H}_0$; with equal prior model probability this leaves 19.5% for $\mathcal{H}_0$. For details see Gronau, Raj, and Wagenmakers (in press).

More generally, Harold Jeffreys developed NHBT because he felt it represented the process of scientific learning, where new effects are accepted by the scientific community only after the data have undercut the null hypothesis: "The principle of testing possible complications by significance tests and not introducing them unless they pass them has important consequences for scientific procedure. In the first place, the onus of proof is always on the advocate of the more complicated hypothesis." (Jeffreys, 1977, p. 90). Similarly, Jeffreys (1937, p. 252) states that, until the Bayes factor favors the alternative hypothesis, "the simpler hypothesis holds the field".

We are in agreement with Jeffreys that hypothesis testing is an essential component of scientific work (see also Morey, Rouder, Verhagen, & Wagenmakers, 2014). Specifically, we believe that questions of estimation ought to be addressed by a methodology of estimation, but that questions of hypothesis testing ought to be addressed by a methodology of testing, and that testing usually precedes estimation. TK find this procedure "overly complex and unnecessary". We disagree and refer the reader to the epigraph of this manuscript, taken from Jeffreys (1950).

A pragmatic researcher may feel that in practice, little harm could come from omitting the testing stage and directly proceeding to the estimation stage. This is not true. First, consider the care of polynomial regression: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ... + \beta_m x^m + \epsilon$. How many terms should be included in the model? An estimation perspective does not address this question, and offers no principled way to whittle down the number. From an estimation point of view, it is asserted, without any empirical evidence whatsoever, that it is necessarily better to include one predictor than no predictor at all; but also that it is better to include two predictors rather than one, three rather than two, and so on. This sequence of preferences, each of them rash and unsupported by data, ultimately results in the unshakable conviction that the model that should be used for inference and prediction should contain (at a

minimum) as many predictors as there are observations. In other words, by violating the statistical razor that "the onus of proof is always on the advocate of the more complicated hypothesis", the estimation perspective results in the recommendation to use models that dramatically overfit the data, unless special precautions are taken.[12]

A second practical problem with the "estimate-only" perspective is that it tempts researchers into believing that they conclusively answered a key question that was in fact never asked. For instance, at the time of writing, medical trials are conducted to examine the effect of remdesivir on COVID-19. The goal of these trials is not to assess the efficacy of remdesivir assuming that it works. Neither is the goal to determine whether remdesivir helps or harms the progression of COVID-19. Instead, the goal is to establish that remdesivir works, and the only way to do so is for the data to undercut the null hypothesis that remdesivir is ineffective. If remdesivir proves effective, then one may proceed, either by developing similar drugs or changing the drug dosage. However, the estimation-only framework invites reports such as "the posterior probability that remdesivir is effective against COVID-19 is 98%". This report is formally correct, but it is easily forgotten that the complementary probability of 2% refers to remdesivir actually being harmful, and that the proposition that remdesivir is ineffective was ruled out and deemed utterly irrelevant from the very beginning (Berger, 2006; Berger & Delampady, 1987; Jeffreys, 1961; Wagenmakers et al., 2020). In sum, we believe that the estimation-only framework is beguiling and usually not the correct methodology to answer the primary question that most experimental psychologists wish to address, namely, "am I just looking at random noise or are my sample estimates indicative of an effect that generalizes to

[12]Specifically, within the estimation framework one may apply "shrinkage priors" that drive the estimate of the regression coefficients to 0 (van Erp, Oberski, & Mulder, 2019). Among these shrinkage priors, the "spike-and-slab" prior is consistent with the approach to take seriously the proposition that a predictor may not be needed.

the population?"

Concluding Comments

Tendeiro and Kiers (in press) have written a thought-provoking, scholarly article on Null Hypothesis Bayesian Testing (NHBT). We agree on the statistical facts presented by TK, and we support several of their recommendations – some cautiously (e.g., to be explicit about the prior model probabilities), and some enthusiastically (e.g., to prefer Bayes factors over p-values). Superficially, our disagreement mostly concerns the interpretation of Bayes factor features. TK identified eleven 'issues' that they deemed potentially problematic, whereas our perspective is much more positive (cf. Table 1). More deeply, we feel that only Bayes factors can address the key question common to most empirical research in psychology: "to what extent do the data support the hypothesis that there is an effect?". Alternative approaches such as visualizing and summarizing the continuous posterior distribution under $\mathcal{H}_1$, although easy, are fundamentally unable to address the key question of interest. In other words, researchers who wish to make an evidential omelette need to break the Bayes factor eggs (see also Evans, 2015).

We worry that the TK article can be misused as an excuse not to present Bayes factors. We agree that Bayes factors bring their own conceptual challenges and opportunities, but more important is the insight that to avoid Bayes factors is to skirt the key question of interest. Researchers will nevertheless want to answer this question, however, and consequently they may be tempted to misapply an estimation-only approach to a situation that demands a hypothesis test.

We expect and hope that Bayes factors will earn themselves a permanent position in the statistical toolkit of the experimental psychologist. Future practice will ultimately have to reveal whether the Bayes factor features identified by TK are more

of a help or a hindrance. In the mean time, as with any relatively new method, it is important that early adopters are aware of the features – or, as we believe, the advantages – of the Bayes factor.

Acknowledgements

This research was supported by a Dutch scientific organization VIDI fellowship grant (016.Vidi.188.001) to DvR and a Dutch scientific organization VICI fellowship grant (016.Vici.170.083) to EJW.

References

Aczel, B. (2018). *Assessing the properties of psychological research from collections of results.* Retrieved from `https://osf.io/8ahfz/` (Slide 20 of paper presented at IMPS conference)

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., ... Wagenmakers, E.-J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, *4*, 4–6.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*, 17–21.

Bayarri, M. J., & Berger, J. O. (2013). Hypothesis testing and model uncertainty. In P. Damien, P. Dellaportas, N. G. Polson, & D. A. Stephens (Eds.), *Bayesian theory and applications* (pp. 361–400). Oxford: Oxford University Press.

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*, 1550–1577.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer.

Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, vol. 1 (2nd ed.)* (pp. 378–386). Hoboken, NJ: Wiley.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, *143*, 383–430.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M.,

... Wu, H. (2018). Evaluating replicability of social science experiments in *Nature* and *Science*. *Nature Human Behaviour*, *2*, 637–644.

Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one–sided testing problem. *Journal of the American Statistical Association*, *82*, 106–111.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*, 1313–1321.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, *96*, 270–281.

Clyde, M. A. (2016). BAS: Bayesian adaptive sampling for Bayesian model averaging [Computer software manual]. (R package version 1.4.1)

Clyde, M. A., Ghosh, J., & Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, *20*, 80–101.

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, *13*, 627–679.

Cornfield, J. (1966). A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *Journal of the American Statistical Association*, *61*, 577–594.

Dawid, A. P. (2011). Posterior model probabilities. In D. M. Gabbay & J. Woods (Eds.), *Handbook of the philosophy of science: Philosophy of statistics* (pp. 607–630). San Diego: Elsevier.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for

psychological research. *Psychological Review*, *70*, 193–242.

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*, *1*, 281–295.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. *PLoS ONE*, *11*, e0149794.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34.

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*, 313–329.

Evans, M. (2015). *Measuring statistical evidence using relative belief*. Boca Raton, FL: CRC Press.

Fisher, R. A. (1928). *Statistical methods for research workers* (2nd ed.). Edinburgh: Oliver and Boyd.

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

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio–Economics*, *33*, 587–606.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, *1*, 198–218.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian t-tests. *The American Statistician*, *74*, 137–143.

Gronau, Q. F., Raj, A., & Wagenmakers, E.-J. (in press). Informed bayesian inference for the a/b test. *Journal of Statistical Software*. Retrieved from `https://`

arxiv.org/abs/1905.02068

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ... Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, *92*.

Gu, X., Hoijtink, H., Mulder, J., & Rosseel, Y. (2019). Bain: A program for Bayesian testing of order constrained hypotheses in structural equation models. *Journal of Statistical Computation and Simulation*, *89*, 1526–1553.

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximate adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*, 229–261.

Haaf, J. M., Ly, A., & Wagenmakers, E.-J. (2019). Retire significance, but still test hypotheses. *Nature*, *567*, 461.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, *7*, 1–20.

Haucke, M., Miosga, J., Hoekstra, R., & van Ravenzwaaij, D. (in press). Bayesian frequentists: Examining the paradox between what researchers can conclude versus what they want to conclude from statistical results. *Collabra: Psychology*. Retrieved from https://psyarxiv.com/escvy/

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*, 200–215.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*,

1157–1164.

Hoijtink, H., & Chow, S.-M. (2017). Bayesian hypothesis testing: Editorial to the Special Issue on Bayesian data analysis. *Psychological Methods*, *22*, 211–216.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*, 539–556.

Howie, D. (2002). *Interpreting probability: Controversies and developments in the early twentieth century*. Cambridge: Cambridge University Press.

Iverson, G. J., Wagenmakers, E.-J., & Lee, M. D. (2010). A model averaging approach to replication: The case of p_{rep}. *Psychological Methods*, *15*, 172–181.

JASP Team. (2020). *JASP (Version 0.12)[Computer software].* Retrieved from `https://jasp-stats.org/`

Jeffreys, H. (1931). *Scientific inference* (1st ed.). Cambridge, UK: Cambridge University Press.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, *31*, 203–222.

Jeffreys, H. (1937). *Scientific inference* (1st ed.). Cambridge, UK: Cambridge University Press.

Jeffreys, H. (1938). The comparison of series of measures on different hypotheses concerning the standard errors. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *167*, 367–384.

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

Jeffreys, H. (1950). *Earthquakes and mountains* (2nd ed.). London: Methuen.

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

Jeffreys, H. (1977). Probability theory in geophysics. *Journal of the Institute of*

Mathematics and its Applications, *19*, 87–96.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Klein, R., Vianello, M., Hasselman, F., Adams, B., Adams, R., Alper, S., ... Nosek, B. (2018). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490.

Lee, D. S., Kim, E., & Schwarz, N. (2015). Something smells fishy: Olfactory suspicion cues improve performance on the Moses illusion and Wason rule discovery task. *Journal of Experimental Social Psychology*, *59*, 47–50.

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*, 114–127.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.

Linde, M., & van Ravenzwaaij, D. (2019). baymedr: An R package for the calculation of Bayes factors for equivalence, non–inferiority, and superiority designs. *Manuscript submitted for publication*. Retrieved from `https://arxiv.org/abs/1910.11616`

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192.

Lindley, D. V. (1985). *Making decisions* (2nd ed.). London: Wiley.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*, 22–25.

Lindley, D. V. (2006). *Understanding uncertainty*. Hoboken: Wiley.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.

Ly, A., Stefan, A., van Doorn, J., Dablander, F., van den Bergh, D., Sarafoglou, A., ... 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*, 153–161.

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

Lyu, X.-K., Xu, Y., Zhao, X.-F., Zuo, X.-N., & Hu, C.-P. (2020). Beyond psychology: Prevalence of p value and confidence interval misinterpretation across different fields. *Journal of Pacific Rim Psychology*, *14*.

Marsman, M., & Wagenmakers, E.-J. (2017). Three insights from a Bayesian interpretation of the one–sided p value. *Educational and Psychological Measurement*, *77*, 529–539.

Martin, G. M., Frazier, D. T., & Robert, C. P. (2020). Computing Bayes: Bayesian Computation from 1763 to the 21st Century. Manuscript submitted for publication.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.

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

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor 0.9.12-4.2.* Comprehensive R Archive Network. Retrieved from `http://cran.r-project.org/web/packages/BayesFactor/index.html`

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). *Bayes Factor (Version 0.9.12-4.1) [computer software].* Retrieved from `https://CRAN.R-project.org/package=BayesFactor`

Morey, R. D., Rouder, J. N., Verhagen, A. J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming. *Psychological Science*, *25*, 1289–1290.

Mulder, J., Gu, X., Olsson–Collentine, A., Tomarken, A., Böing–Messing, F., Hoijtink, H., ... van Lissa, C. (2019). BFpack: Flexible Bayes factor testing of scientific theories in R. *Manuscript submitted for publication.* Retrieved from `https://arxiv.org/abs/1911.07728`

Mulder, J., Hoijtink, H., & Gu, X. (2019). Default Bayesian model selection of constrained multivariate normal linear models. *Manuscript submitted for publication.* Retrieved from `https://arxiv.org/abs/1904.00679`

Navarro, D. J. (2015). *Learning statistics with R.* University of Adelaide: Lulu.com.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* New York: Wiley.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities.* Chichester, UK: John Wiley & Sons.

Robert, C. P., Chopin, N., & Rousseau, J. (2009). Harold Jeffreys's Theory of Probability revisited. *Statistical Science*, *24*, 141–172.

Robinson, G. K. (2019). What properties might statistical inferences reasonably be expected to have?—crisis and resolution in statistical inference. *The American Statistician*, *73*, 243–252.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic*

Bulletin & Review, *25*, 102–113.

Rouder, J. N., & Morey, R. D. (2019). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, *73*, 186–190.

Rouder, J. N., Morey, R. D., Verhagen, A. J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*, 520–547.

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, *2*, 1–12.

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

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

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337–356.

Stefan, A. M., Evans, N. J., & Wagenmakers, E.-J. (in press). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*, 1042–1058.

Stefan, A. M., Katsimpokis, D., Gronau, Q. F., & Wagenmakers, E.-J. (2021). Expert agreement in prior elicitation and its effects on Bayesian inference. *Manuscript submitted for publication*. Retrieved from `https://psyarxiv.com/8xkqd`

Tawakol, A., Ishai, A., Takx, R. A. P., Figueroa, A. L., Ali, A., Kaiser, Y., ... Pitman,

R. K. (2017). Relation between resting amygdalar activity and cardiovascular events: a longitudinal and cohort study. *The Lancet*, *389*, 834–845.

Tendeiro, J. N., & Kiers, H. A. L. (in press). A review of issues about Null Hypothesis Bayesian Testing. *Psychological Methods*.

van den Bergh, D., Clyde, M. A., Raj, A., de Jong, T., Gronau, Q. F., Marsman, M., ... Wagenmakers, E.-J. (in press). A tutorial on Bayesian multi–model linear regression with BAS and JASP. *Behavior Research Methods*.

van den Bergh, D., Haaf, J. M., Ly, A., Rouder, J. N., & Wagenmakers, E.-J. (in press). A cautionary note on estimating effect size. *Advances in Methods and Practices in Psychological Science*.

van Dongen, N. N. N., van Doorn, J. B., Gronau, Q. F., van Ravenzwaaij, D., Hoekstra, R., Haucke, M. N., ... Wagenmakers, E.-J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician*, *73*, 328–339.

van Doorn, J., van den Bergh, D., Dablander, F., van Dongen, N., Derks, K., Evans, N. J., ... Wagenmakers, E.-J. (in press). Strong public claims may not reflect researchers' private convictions. *Significance*. Retrieved from `https://psyarxiv.com/pc4ad`

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*, *25*, 143–154.

van Ravenzwaaij, D., & Etz, A. (in press). Simulation studies as a tool to understand Bayes factors. *Advances in Methods and Practices in Psychological Science*. Retrieved from `https://psyarxiv.com/27ndb/`

van Ravenzwaaij, D., & Ioannidis, J. P. A. (2017). A simulation study of the strength of evidence in the recommendation of medications based on two trials with

statistically significant results. *PLoS ONE*, *12*, e0173184.

van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC: Medical Research Methodology*, *19*, 71.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4.

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R., ... Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*, 917–928.

Wagenmakers, E.-J., Gronau, Q. F., Dablander, F., & Etz, A. (in press). The support interval. *Erkenntnis*. Retrieved from `https://psyarxiv.com/zwnxb/`

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.

Wagenmakers, E.-J., Lee, M. D., Rouder, J. N., & Morey, R. D. (2020). The principle of predictive irrelevance or why intervals should not be used for model comparison featuring a point null hypothesis. In C. W. Gruber (Ed.), *The theory of statistics in psychology – applications, use and misunderstandings* (pp. 111–129). Cham: Springer.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method.

Cognitive Psychology, *60*, 158–189.

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

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Kievit, R. A., & van der Maas, H. L. J. (2013). Yes, Psychologists Must Change the Way They Analyze Their Data: Clarifications for Bem, Utts, and Johnson (2011). Retrieved from `https://psyarxiv.com/tvarg/`

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390.

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