An Evaluation of Alternative Methods for Testing Hypotheses, from the Perspective of Harold Jeffreys

Alexander Ly, Josine Verhagen, Eric-Jan Wagenmakers[1]

University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, the Netherlands.

Abstract

Our original article provided a relatively detailed summary of Harold Jeffreys's philosophy on statistical hypothesis testing. In response, Robert (ress) maintains that Bayes factors have a number of serious shortcomings. These shortcomings, Robert argues, may be addressed by an alternative approach that conceptualizes model selection as parameter estimation in a mixture model. In a second comment, Chandramouli and Shiffrin (ress) seek to extend Jeffreys's framework by also taking into consideration data distributions that do not originate from either of the models under test. In this rejoinder we argue that Robert's (ress) alternative view on testing has more in common with Jeffreys's Bayes factor than he suggests, as they share the same "shortcomings". On the other hand, we show that the proposition of Chandramouli and Shiffrin (ress) to extend the Bayes factor is in fact further removed from Jeffreys's view on testing than the authors suggest. By elaborating on these points, we hope to clarify our case for Jeffreys's Bayes factors.

Keywords: Bayes factors, induction, model selection, replication, statistical evidence

In our original article (Ly et al., ress) we outlined how Harold Jeffreys constructed his hypothesis tests. Jeffreys's tests contrast a precise, point-null hypothesis $\mathcal{M}_0$ versus a more general alternative hypothesis $\mathcal{M}_1$. Here the point-null hypothesis represents a general law, an invariance, or a categorical causal claim (e.g., "apple trees always bear apples"; "people cannot look into the future"; "Alzheimer's disease is caused by a fungal infection of the central nervous system"), whereas the alternative hypothesis relaxes that law. Jeffreys's tests require a thoughtful specification of the prior distribution for the parameter of interest, and much of Jeffreys's work was concerned with providing good default specifications – "good" in the sense that they adhere to general common-sense desiderata (e.g., Bayarri et al., 2012). We are pleased that our summary attracted two comments by renowned researchers; below we respond to their ideas in a way that we hope is consistent with the overall philosophy of Harold Jeffreys himself.

[1]This work was supported by the starting grant "Bayes or Bust" awarded by the European Research Council. Correspondence concerning this article may be addressed to Alexander Ly. We thank Maarten Marsman for his comments on an earlier draft of the manuscript.

1. Rejoinder to Robert

In general, Robert's (ress) comments highlight the inevitable subtleties in constructing a Bayes factor. His alternative mixture model procedure is practical and may be immensely valuable for specific situations (i.e., hierarchical models) that are common in psychological research. Nevertheless, we believe Robert's suggestion about the demise of the Bayes factor to be an overstatement.

1.1. Robert's Critique on the Bayes Factor

Our understanding of Jeffreys's method is partly based on the work by Robert and colleagues (2009), and it should, therefore, not come as a surprise that Robert's view and ours overlap to a considerable degree. Robert's arguments for dismissing the Bayes factor can be grouped in terms of (1) its usage in making decisions and (2) the care that needs to be taken in choosing the priors.

1.1.1. First Critique: The Distinction Between Inference and Decision Making

We share Robert's discontent with the statistical practice that emphasizes all-or-none decisions at some arbitrary threshold, and we agree that scientific learning should instead be guided by a continuous measure of evidence. In the process of eviscerating p-value null hypothesis tests, Rozeboom (1960, pp. 422-423) already expressed a similar sentiment:

> "The null-hypothesis significance test treats 'acceptance' or 'rejection' of a hypothesis as though these were decisions one makes. But a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection of a hypothesis is a cognitive process, a degree of believing or disbelieving which, if rational, is not a matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true."

Our favorite continuous measure of evidence is of course a Bayes factor constructed from a pair of priors selected according to Jeffreys's desiderata, or a Jeffreys's Bayes factor in short. It is important to note that this measure provides only the first of three Bayesian ingredients needed for decision making. The other two ingredients are the prior model probabilities (which, combined with the Bayes factor, yield posterior model probabilities) and the specification of a loss function (or equivalently, a utility function; Berger, 1985, Lindley, 1977, and Robert, 2007).

For instance, consider a Bayes factor of $\mathrm{BF}_{10}(d) = 4.6$ for the observed data d. This Bayes factor can be converted to a posterior model probability of $P(\mathcal{M}_0 \mid d) = 0.17$ when we set $P(\mathcal{M}_0) = P(\mathcal{M}_1) = 1/2$ (Ly et al., ress). One possible subsequent decision rule is then to accept $P(\mathcal{M}_1 \mid d)$

because it has the highest posterior model probability. We did not intend to suggest such a procedure, as the decision is clearly sensitive to the prior model probabilities. Furthermore, we do not recommend uniform prior model probabilities regardless of scientific context. In fact, when decision making is desired, the assignment of prior model probabilities is left to the substantive researcher. Such flexibility in assignment introduces subjectivity, and this may be seen either as a disadvantage or as an advantage. At any rate, prior model probabilities can be used to formalize the adage that "extraordinary claims require extraordinary evidence" (e.g., Wagenmakers et al., 2011). Moreover, the prior model probabilities can be used to address the problem of multiplicity (e.g., Jeffreys, 1961; Scott and Berger, 2010; Stephens and Balding, 2009). A similar argument applies to utility functions: these may be subjective and hard to elicit, but such difficulties do not sanction the practice of ignoring utility functions altogether, at least not when the purpose is to make decisions.

Thus, Robert worries that computation of Bayes factors may tempt users to make all-or-none decisions while disregarding prior model probabilities or loss functions. We agree with Robert that there is a considerable difference between inference and decision making, and that scientific learning should be guided by a continuous measure of evidence that incorporates what we have learned from the observed data. The Bayes factor is such a measure.

1.1.2. Second Critique: The Jeffreys-Lindley-Bartlett Paradox

We suspect that the Jeffreys-Lindley-Bartlett (henceforth JLB) paradox is central to Robert's (1993; 2014) dismissal of the Bayes factor and it is the main motivation for the development of the mixture model alternative. We take a closer look at the JLB paradox and discuss two consequences foreseen by Jeffreys, who was keenly aware of the "paradox" from the very beginning (Etz and Wagenmakers, 2015).

First, the JLB paradox implies that we cannot use improper priors to construct a Bayes factor. For instance, to *estimate* μ within the normal model $\mathcal{M}_1 : X \sim \mathcal{N}(\mu, 1)$, we typically employ Jeffreys's (1946) prior $\mu \propto 1$. The reason to do so stems from the fact that Jeffreys's prior is translation-invariant, leading to a posterior that is independent on how researchers parameterize the problem (Ly et al., 2015b). The JLB paradox implies that we cannot use this same (estimation) prior on the test-relevant parameter for a Bayesian test. More specifically, when we pit the aforementioned model $\mathcal{M}_1$ against the null model $\mathcal{M}_0 : X \sim \mathcal{N}(0, 1)$ the improper prior $\pi_1(\mu) \propto 1$ then becomes useless. To see this we consider the Jeffreys's prior as the limit of proper priors $\mu \sim \mathcal{N}(0, \tau^2)$ with τ going to infinity. The Bayes factor for the observed data $d = (n, \bar{x})$ is then given by

$$\lim_{\tau \to \infty} \tilde{\text{BF}}_{10\,;\,\tau}(d) = \lim_{\tau \to \infty} \frac{\int \exp\left[-\frac{n}{2}(\bar{x}-\mu)^2\right] \exp\left[-\frac{1}{2\tau^2}\mu^2\right] \mathrm{d}\mu}{\sqrt{2\pi}\tau \exp[-\frac{n}{2}\bar{x}^2]} = 0, \tag{1}$$

$$= \lim_{\tau \to \infty} \frac{1}{\sqrt{1+n\tau^2}} \exp\left[\frac{(\tau n \bar{x})^2}{2(1+n\tau^2)}\right] = 0, \tag{2}$$

regardless of the fixed sample size n, the observed sample mean $\bar{x}$. As such, the Bayes factor constructed from the improper Jeffreys's prior will always favor the null model and this also holds for other improper priors. Moreover, Eq. (2) shows that for fixed data $d = (n, \bar{x})$ and a Bayes factor constructed from a normal prior with hyperparameter τ we can obtain a Bayes factor in favor of the null hypothesis of arbitrary size (i.e., $\tilde{\text{BF}}_{10\,;\,\tau}(d) < 1$) simply by taking τ large enough.

Hence, the JLB paradox effectively implies that a testing problem should be treated differently from one that is concerned with estimation. As such, when π_1 is interpreted as prior belief about the parameters θ_1, in the example above $\theta_1 = \mu$, one's belief about the parameter then changes depending on whether one is concerned with testing or estimating. More generally, this difference is due to the fact that estimation is typically a *within*-model affair. Recall that a model $\mathcal{M}_i$ specifies a relationship $f_i(d \,|\, \theta_i)$ that defines which parameters θ_i are relevant in the data generating process of the data d. Hence, the function f_i gives the (only) context in which the parameters θ_i can be perceived.

In essence, the f_i justifies that it is meaningful to calculate a posterior distribution for the parameter. To underline this point we added subscripts to the parameters to indicate model membership, that is, we take $\theta_0 = \sigma_0$ and $\theta_1 = (\mu_1, \sigma_1)$ for f_0 and f_1 and both normals. For example, when we assume that $\mathcal{M}_0 : X \sim \mathcal{N}(0, \sigma_0^2)$ only a posterior for the standard deviation σ_0 is worthwhile to be pursued, as the posterior for the population mean remains zero, regardless of the data. Within $\mathcal{M}_0$, the Jeffreys's prior for σ_0 is given by $\pi_0(\sigma_0) \propto \sigma_0^{-1}$, which can be updated to a posterior $\pi_0(\sigma_0 \,|\, d)$. On the other hand, under $\mathcal{M}_1 : X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ we are dealing with two parameters of interest. Within $\mathcal{M}_1$, the Jeffreys's prior for μ_1 is $\pi(\mu_1) \propto 1$, for σ_1 is $\pi_1(\sigma_1) \propto 1/\sigma_1$ and we take $\pi_1(\mu_1, \sigma_1) = \pi_1(\mu_1)\pi_1(\sigma_1)$. These priors can be updated to posteriors $\pi_1(\mu_1 \,|\, d)$ and $\pi_1(\sigma_1 \,|\, d)$. Even though the two priors $\pi_0(\sigma_0)$ and $\pi_1(\sigma_1)$ have the same form, they do not lead to the same posterior. In fact, due to the presence of μ_1 as a parameter, the posterior mean of $\pi_1(\sigma_1 \,|\, d)$ within $\mathcal{M}_1$ will be smaller or equal to the posterior mean of $\pi_0(\sigma_0 \,|\, d)$ within $\mathcal{M}_0$. Thus, when we are interested in the standard error σ_i, it matters whether we believe that $\mathcal{M}_0$ holds true or whether the population mean μ_1 plays a role in the data generating process in accordance to $\mathcal{M}_1$. The Bayes factor helps us distinguish which of the two models predict the data more adequately and which posterior for σ_i we should report. Hence, testing is a *between*-model matter. Jeffreys himself was very clear about the distinction between estimation and testing:

> "We are now concerned with the more difficult question: in what circumstances do observations support a change of the form of the law itself? This question is really logically prior to the estimation of the parameters, since the estimation problem presupposes that the parameters are relevant." (Jeffreys, 1961, p. 245)

Hence, testing implies that we are uncertain about which of the two functional relationships defined by the models $\mathcal{M}_0$ and $\mathcal{M}_1$ is adequate for the data under study. This uncertainty is expressed through the prior statement $P(\mathcal{M}_0), P(\mathcal{M}_1) > 0$ and when $\mathcal{M}_0$ and $\mathcal{M}_1$ are the only models under consideration we require that $P(\mathcal{M}_0) + P(\mathcal{M}_1) = 1$. The priors π_1, π_0 in a Bayes factor are, thus, chosen to guide scientific learning, that is, how one transitions from prior model odds to posterior model odds and are not designed to yield posteriors that are good for estimation. To simplify notation, we drop the subscripts indicating model membership when the context is clear.

Second, the separation of estimation and testing and the resulting separation of models led us to instantiate the hypotheses $\mathcal{H}_i$ with their respective models $\mathcal{M}_i$ as discussed in Ly et al. (ress). In effect, we have different contexts in which the respective parameters exist and, therefore, a philosophical conundrum in what is meant by common parameters. The difference between the posteriors $\pi_0(\sigma \,|\, d)$ and $\pi_1(\sigma \,|\, d)$ discussed above showed that one should not be fooled by the fact that the Greek letters are identical. We therefore agree with Robert's warning concerning the treatment of common parameters.

For the t-test the commonality between the two σs within $\mathcal{M}_0$ and $\mathcal{M}_1$ is given by their meaning as a scaling parameter within either models. Furthermore, the nesting of $\pi_0(\sigma)$ as $\pi_1(\mu, \sigma) = \pi_1(\delta)\pi_0(\sigma)$ can be considered a practical choice. In effect, the Bayes factor $\mathrm{BF}_{10}(d)$ is then given by the ratio of the following two marginal likelihoods

$$p(d \,|\, \mathcal{M}_1) = (2\pi)^{-\frac{n}{2}} \int_0^\infty \sigma^{-n} \int_{-\infty}^{\infty} \exp\left(-\tfrac{n}{2}\left[\left(\tfrac{\bar{x}}{\sigma} - \delta\right)^2 + \left(\tfrac{s}{\sigma}\right)^2\right]\right) \pi_1(\delta)\,\mathrm{d}\delta\,\pi_0(\sigma)\mathrm{d}\sigma, \tag{3}$$

$$p(d \,|\, \mathcal{M}_0) = (2\pi)^{-\frac{n}{2}} \int_0^\infty \sigma^{-n} \exp\left(-\tfrac{n}{2\sigma^2}\left[\bar{x}^2 + s^2\right]\right) \pi_0(\sigma)\mathrm{d}\sigma, \tag{4}$$

where $d = (n, \bar{x}, s^2)$. We would like to thank Robert for pointing out our notational inaccuracy, as Eq. (9) in Ly et al. (ress) should actually be Eq. (3) above, that is, the marginal likelihood of the alternative model, thus, the numerator of the Bayes factor $\mathrm{BF}_{10}(d)$, after $\pi_1(\delta)$ and $\pi_0(\sigma)$ are specified. In the original text we already filled in $\pi_0(\sigma) \propto \sigma^{-1}$, a choice which we elaborated on in Section 3.2.2 of Ly et al. (ress).

With the nesting of π_0 within π_1 we made the following recommendation explicit: "It is to be understood that in pairs of equations of this type [such as Eqns. (3, 4)] the sign of proportionality indicates the same constant factor, which can be adjusted to make the total probability 1." (Jeffreys, 1961, p. 247) More precisely, an improper prior $\pi_0(\sigma) \propto \sigma^{-1}$ has a suppressed normalization constant $\pi_0(\sigma) = c_0\sigma^{-1}$ and we not only take $\pi_1(\sigma) = c_1\sigma^{-1}$ of the same form, but also choose to set $c_1 = c_0$, which allows us to use improper priors on the nuisance parameters (see Berger et al., 1998 for a theoretical justification). More examples of this type of nesting can be found in Dawid and Lauritzen (2001), Consonni et al. (2008), and references therein.

1.2. Jeffreys's Common-Sense Desiderata

> "Rejection of a null hypothesis is best when it is interocular". Edwards et al. (1963, p. 240)

In conclusion, the JLB paradox prohibits the usage of improper priors for testing, separates the estimation practice from a testing concern, and challenges the idea of common parameters. As noted above, we first require a justification before we can use the same prior on the nuisance parameters. After doing so, we then create an exception on the ban of improper priors allowing us to assign improper priors to the nuisance parameters , say, $\theta_0 = \sigma$. Furthermore, let δ denote the test-relevant parameter with, say, $\theta_1 = (\theta_0, \delta)$. Hence, after specifying Jeffreys's translation-invariant priors on the nuisance parameters θ_0, which we would use for estimation within each model, we only require to set the prior $\pi_1(\delta)$ in order to define the Bayes factor $\text{BF}_{10}(d)$. We suspect that Jeffreys's underlying reasons for the choice of $\pi_1(\delta)$ was to have a test that passes "the interocular traumatic test; you know what the data mean when the conclusion hits you between the eyes." Edwards et al. (1963, p. 217).

We believe that the information consistency criterion makes explicit which data hit us right between the eyes. This criterion leads to a Bayes factor that is consistent for a finite sample, a requirement that is much harder to be fulfilled than the asymptotic consistency criterion, at least for parametric models (e.g., Bickel and Kleijn, 2012, Yang and Le Cam, 2000). We agree with Robert that information consistency is in some cases an approximate statement. In particular, when the data are either distributed according to $\mathcal{M}_0 : X \sim \mathcal{N}(0, \sigma^2)$ or $\mathcal{M}_1 : X \sim \mathcal{N}(\mu, \sigma^2)$ then the interocular data set with $n > 2$, $\bar{x} \neq 0$ and, in particular, $s^2 = 0$ occurs with zero probability under both models, due to having $\sigma > 0$. However, when $\mathcal{M}_0$ and $\mathcal{M}_1$ are the only two models under consideration, the observation $\bar{x} \neq 0$ with $n > 2$, in addition to $s^2 = 0$, then should lead to the logical exclusion of $\mathcal{M}_0$, thus, $\mathrm{BF}_{01}(d) = 0$.

To appreciate the information consistency criterion, we revisit the Bayesian t-test with Bayes factors $\tilde{\mathrm{BF}}_{10\,;\,\tau}(d)$ that lacks this property by constructing it from $\pi_0(\sigma) \propto \sigma^{-1}$ and $\pi_1(\delta, \sigma) = \pi_1(\delta)\pi_0(\sigma)$ where $\pi_1(\delta)$ is normal around zero with a standard deviation τ, i.e.,

$$\tilde{\mathrm{BF}}_{10\,;\,\tau}(d) = (1 + n\tau^2)^{\frac{n-1}{2}} \left(\frac{1 + \frac{n\bar{x}^2}{ns^2}}{(1 + n\tau^2) + \frac{n\bar{x}^2}{ns^2}} \right)^{\frac{n}{2}} \tag{5}$$

As before, letting τ tend to infinity, while keeping $n, \bar{x}$ and s^2 fixed, yields the JLB paradox, i.e., $\lim_{\tau \to \infty} \tilde{\mathrm{BF}}_{10\,;\,\tau}(d) = 0$.

To simplify the discussion we suppose that τ is set to one. The resulting Bayes factor $\tilde{\mathrm{BF}}_{10\,;\,\tau=1}(d)$ is then asymptotically consistent. This means that if we repeatedly sample from the null model, we let n tend to infinity and simultaneously let $n\bar{x}^2/(ns^2) = t^2/(n-1)$ tend to zero yielding a Bayes factor of zero, where t is the usual t-statistic $t = \sqrt{n}\bar{x}/s_{n-1}$. Similarly, if we repeatedly sample from the alternative model, we let n tend to infinity and simultaneously let $t^2/(n-1)$ tend to infinity yielding a Bayes factor of infinity. Thus, this Bayes factor $\tilde{\mathrm{BF}}_{10\,;\,\tau=1}(d)$ is able to detect the correct model when the number of data points tends to infinity.

The Bayes factor $\tilde{\mathrm{BF}}_{10\,;\,\tau=1}(d)$, however, is not information consistent. For the t-test information consistency is concerned with having a fixed number of data points $n > 2$, a observed sample mean, say, $\bar{x} \neq 0$ and s^2 tending to zero. With τ, n and $\bar{x}$ fixed, this Bayes factor $\tilde{\mathrm{BF}}_{10\,;\,\tau=1}(d)$ is an decreasing function of as a function of s^2 that attains its maximum when $s^2 = 0$. For instance, when $n = 4$, $\bar{x} = 7$ the maximum is given by $\lim_{s^2 \to 0} \tilde{\mathrm{BF}}_{10\,;\,\tau=1}(d) = 11.18$. Note that the data set with $n = 4$, $\bar{x} = 7$ and $s^2 \to 0$ is interocular as it leads to an observed sample effect size, an realization of the t-statistic, that tends to infinity, which should therefore lead to an infinite support for the alternative compared to the null model. The fact that the information inconsistent Bayes factor $\tilde{\mathrm{BF}}_{10\,;\,\tau=1}(d)$ is bounded makes it hard to be interpret. For instance, the observations $n = 4$, $\bar{x} = 7$ and $s^2 = 1$ yields a Bayes factor of $\tilde{\mathrm{BF}}_{10\,;\,\tau=1}(d) = 9.6$, which does not seem a lot of evidence against the null, but with respect to its maximum 11.81 might be considered substantial.

On the other hand, a Jeffreys's Bayes factor is by construction information consistent and has a supremum (i.e., maximum) at infinity, which makes it easier to be interpret. Jeffreys referred to this and other desiderata as common-sense as they came natural to him (Etz and Wagenmakers, 2015), but it took a long time before his intuition was formalized by Berger and Pericchi (2001) and extended by Bayarri et al. (2012).

Recall that information consistency in a t-test requires us to construct a Bayes factor from a heavy-tailed prior on δ and we agree with Robert that the Cauchy prior with scale $\gamma = 1$ is only one of many possible choices. This is why we included a robustness analysis in our open-source software package JASP (`https://jasp-stats.org/`). However, we believe that the merit of a Jeffreys's Bayes factor (with γ fixed) is due to the fact that it kickstarts scientific learning.

> "In any of these cases it would be perfectly possible to give a form of $[\pi_1(\delta)]$ that would express the previous information satisfactorily, and consideration of the general argument of [Chapter] 5.0 will show that it would lead to common-sense results, but they would differ in scale. As we are aiming chiefly at a theory that can be used in the early stages of a subject, we shall not at present consider the last type of case" (Jeffreys, 1961, p. 252).

Thus, Jeffreys was not opposed to incorporating previously acquired data in a Bayesian hypothesis test, but to do so he first designed a starting Bayes factor, for a first data set, say, d_{orig}. After observing d_{orig}, we can then straightforwardly update a Jeffreys's Bayes factor for a future, not yet observed, data set, say, d_{rep}. This informed Bayes factor $\text{BF}_{10}(d_{\text{rep}} \mid d_{\text{orig}})$ is then constructed from the priors $\pi_1(\theta_1 \mid d_{\text{orig}})$ and $\pi_0(\theta_0 \mid d_{\text{orig}})$. This idea forms the basis of the replication Bayes factors introduced in Verhagen and Wagenmakers (2014) and is further exploited in Ly et al. (2015a). Hence, the man who discovered the origin of the earth, thus, also provided us with the starting point for scientific learning.

1.3. Robert's Alternative Approach

> "Prior distributions must always be chosen with the utmost care when dealing with mixtures and their bearings on the resulting inference assessed by a sensitivity study. The fact that some noninformative priors are associated with undefined posteriors, no matter what the sample size, is a clear indicator of the complex nature of Bayesian inference for those models" (Marin and Robert, 2014, p. 199)

As an alternative to Bayes factors, Robert (ress) suggests to use a mixture model approach elaborated upon in Kamary et al. (2014). The data generating process of a mixture model can be envisioned as a stepwise procedure. First, a membership variable z_j is realized; in a two-component mixture, z_j assumes either the value zero or one. Next, given the outcome $z_j = 0$ (or $z_j = 1$) a data point x_j is generated according to $\mathcal{M}_0 : X_j \sim f_0(x_j \mid \theta_0)$ (or $\mathcal{M}_1 : X_j \sim f_1(x_j \mid \theta_1)$). This means that the complete data should therefore consists of n-pairs $(z_1, x_1), \dots, (z_n, x_n)$, but in reality we only have the observations $d = x_1, \dots, x_n$. As a result of not observing the membership variables z_j, the observations are perceived as if each of the data points were generated from the (arithmetic) mixture model $\mathcal{M}_a : X_j \sim (1-\alpha) f_0(x_j \mid \theta_0) + \alpha f_1(x_j \mid \theta_1)$, where α is the mixture proportion. The artificial encompassing model $\mathcal{M}_a$ therefore contains the two competing models, $\mathcal{M}_0$ and $\mathcal{M}_1$, as special cases; when $\alpha = 0$ and $\alpha = 1$ respectively. Hence, to uncover whether the observations are more consistent with $\mathcal{M}_0$ or $\mathcal{M}_1$, Kamary et al. (2014) suggest to focus on estimating α within the encompassing model $\mathcal{M}_a$.

Inferring α amounts to a missing data problem which is in principle computationally intensive as there are 2^n different combinations for the membership variables z_js. Luckily, one can resort to a completion method pioneered by Diebolt and Robert (1994). When this stochastic exploration method yields n_0 and n_1 numbers of observations allocated to $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively, the posterior for α is then given by $\mathcal{B}(a+n_0, a+n_1)$, when we use a beta prior on the mixture proportion, $\alpha \sim \mathcal{B}(a, a)$. When n_0 is large and n_1 small or zero, the posterior for α then concentrates most of its mass near zero indicating more evidence for $\mathcal{M}_0$ as one would expect.

Kamary et al. (2014) note that the data generative view of the mixture model is theoretically justified, but that the resulting natural Gibbs sampler has convergence problems when the hyperprior a is smaller than one. To circumvent this problem, Kamary et al. (2014) propose to use a Metropolis-Hastings algorithm instead and illustrate its use by examples followed by a proof that shows that the method is asymptotically consistent. Thus, the work by Kamary et al. (2014) impressively introduces an alternative view on testing, an algorithmic resolution, and a theoretical justification.

1.3.1. Testing versus Estimation

We believe that the Kamary et al. (2014) mixture approach will be especially useful in psychological research. In particular, consider a hierarchical model where each participant's performance x_j on a psychological task is captured by a particular model or strategy represented by f_i. The posterior for α then gives an indication of the prevalence of the model or strategy. When the posterior for α is near zero or near one, this suggests that one model or strategy is dominant; when the posterior for α is near 1/2, this suggests that some participants are better described by one strategy, and some are better described by another (for similar approaches see Friston and Penny, 2011; Lee et al., 2015).

The advantage of the mixture model approach is particularly acute when it is reasonable to assume that not all participants will follow one or the other strategy. In this special issue for *Journal of Mathematical Psychology* alone, the articles by Kary et al. (ress) and Turner et al. (ress) demonstrate considerable heterogeneity among participants: the behavior of some participants is predicted much better by one model, the behavior of other participants is predicted much better by the competing model, and the behavior of a third set of participants is predicted by the models about equally well (see also Steingroever et al., ress).

The standard Bayes factor tests determines whether *all* participants are better predicted by model $\mathcal{M}_0$ or whether *all* participants are better predicted by model $\mathcal{M}_1$. Therefore, one can construct situations in which the data support model $\mathcal{M}_0$ for 99 out of 100 participants, and nevertheless the Bayes factor strongly prefers model $\mathcal{M}_1$. We believe that in these hierarchical scenarios, the mixture model approach is a valuable technique that can offers additional insight.

The above considerations suggests that the mixture approach relaxes Jeffreys's conceptualization of a hypothesis test. More precisely, Jeffreys viewed the null hypothesis as a general law, which by definition implies that the membership variables z_j are either all zeroes or all ones. Note that by embedding the models into an artificial encompassing model, Kamary et al. (2014) transformed the testing problem into one of estimation. Jeffreys, however, did not feel that estimation is appropriate when the test of a general law is at hand:

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

Wrinch and Jeffreys (1919, 1921, 1923) argued that within an estimation framework, a general law such as $\mathcal{H}_0$:"All swans are white" cannot gain much evidence until almost all swans have been inspected.[2] Moreover, common sense prescribes that the plausibility of a general law increases

[2]We now know that this particular statement does not hold true, since Australia is home to many black swans. The statement itself however cannot be discarded until the first exception is actually observed.

with every observation in accordance with the law, that is, $s = n$ number of successes within n trials. Jeffreys (1961, p. 256) operationalized the general law as a binomial model $\mathcal{M}_0$ with θ_0 fixed and its negation as the binomial model $\mathcal{M}_1$ with a θ free to vary. With a uniform prior on θ this then leads to a Bayes factor of $\text{BF}_{01}(d) = \frac{(n+1)!}{s!f!}\theta_0^s(1-\theta_0)^f$, where n denotes the total number of trials, s the number of successes, and f the numbers of failures.

When only successes are observed (i.e., observations consistent with the general law $\mathcal{H}_0 : \theta_0 = 1$), the Bayes factor simplifies to $n+1$; a single failure, on the other hand, indicates infinite evidence against the general law: the observation of a single black swan is interocular, as it conclusively falsifies the general law "all swans are white". Hence, Jeffreys's Bayes factor formalizes inductive reasoning and the logic of proof by contradiction.

The discussion above indicates that the mixture model approach does not formalize inductive reasoning and the logic of proof by contradiction: after having observed 10,000 white swans, the observation of a single black swan will not greatly affect the mixture proportion – the mixture proportion still reflects the fact that there is a great preponderance of white swans. However, in Jeffreys conceptualization, the single exception utterly destroys the general law.

Another concern with the mixture model approach is that it is relatively insensitive to the shape of the prior distributions. Of course, this is also its strength, as this is needed to avoid the JLB paradox. However, models that make correct predictions should receive more reward when these predictions are risky, and the degree of risk is partly encoded in the shape of the prior distributions. For instance, suppose we model a binomial parameter θ and assume that $\mathcal{M}_1 : \theta \sim U[1/2, 1]$ and $\mathcal{M}_2 : \theta \sim U[0, 1]$; further, suppose the observed data are highly consistent with the simpler model $\mathcal{M}_1$. Because the predictions from $\mathcal{M}_1$ are twice as risky as those from $\mathcal{M}_2$ we would want to prefer $\mathcal{M}_1$ over $\mathcal{M}_2$, and in fact, the Bayes factor is asymptotically equal to 2 in favor of $\mathcal{M}_1$ (e.g., Heck et al., 2015; Shiffrin et al., ress).

1.4. Conclusion

Scientific learning involves more than just testing general laws and invariances. Estimation and exploration are important and the mixture approach has a lot to offer in this respect, particularly in hierarchical settings where the general law is unlikely to hold for all participants simultaneously. Other advantages of the mixture approach are apparent as well. For instance, Example 3.1 in Kamary et al. (2014) compares a Poisson distribution with parameter λ to a geometric distribution with parameter p (see Robert, 2015 for R code). The comparison begins by relating the parameterizations to each other by setting $p = (1 + \lambda)^{-1}$, which allows the use of the improper

Jeffreys's prior (with respect to the Poisson distribution) $\pi(\lambda) = \lambda^{-1}$ over the two models. Note how this procedure resembles Jeffreys's recommendation for common parameters even though the arguments differ. Moreover, its posterior $\pi(\lambda \mid d)$ will be mixture of the likelihoods of both models. The simulations show that the mixture approach performs well. We do not know how a Jeffreys's Bayes factor can be constructed to deal with a test between two models of different relational forms as Jeffreys was only concerned with nested model comparisons (e.g., Robert, ress).

The mixture approach is not fully automatic, however, and requires some thoughts on how the priors should be chosen. In particular, one cannot naively use improper priors on the test-relevant parameters, as this may yield posteriors that are also improper (Grazian and Robert, 2015). This was acknowledged by Robert (ress) who used an (arbitrary) standard normal prior on μ in a t-test. Our implementation of this recommendation lead to a posterior median ranging from 0.3 to 0.9, for the interocular data with $n = 4$, $\bar{x} = 7$ and $s^2 = 0$, while α should be 1.0 if it were information consistent. More recently, Kamary et al. (2016) proposed a noninformative reparametization for location-scale mixtures to resolve the aforementioned arbitrariness. Hence, as with a Jeffrey's Bayes factor, one should choose the priors carefully when one conceptualizes model selection as parameter estimation in a mixture model.

Lastly, Robert notes that the mixture approach is superior to the Bayes factor as it leads to a faster accumulation of α to the null. The parametric convergence rate of $\sqrt{n}$ follows immediate from casting the testing problem as one of estimation. Similarly, it should be noted that Johnson and Rossell (2010) also use the rate of convergence as a motivation for their Bayes factor approach. We are unsure whether this rate is relevant as we do not consider a testing problem as one of estimation. In the end the Bayes factor and the mixture approach of Kamary et al. (2014) simply answer different questions. The choice which method to use should not be based on the rate of convergence, but on the research question the user seeks to address.[3]

2. Rejoinder to Chandramouli and Shiffrin

Chandramouli and Shiffrin (ress) put forward a thought-provoking proposal which aims to explain and extend Bayesian induction using simple matrix algebra. We have given this novel idea considerable thought and outline some of our reservations below.[4]

[3] We thank Joris Mulder for attending us to this.

[4] The second and third authors are in a state of perpetual confusion regarding the details of the Chandramouli and Shiffrin proposal. All credit concerning this section goes to the first author, who, as such, takes full responsibility for any errors here. For a thorough understanding of our reply, we recommend to have the comment of Chandramouli and Shiffrin (ress) on hand.

We believe that Chandramouli and Shiffrin (henceforth C&S) put forward a belief propagation procedure that allows us to verify whether two given models, say, $\mathcal{M}_1$ and $\mathcal{M}_2$ align with a scientist's prior belief about the true data generating process $p^*(X)$. Instead of setting the priors onto the two given models $\mathcal{M}_1$ and $\mathcal{M}_2$ directly, C&S recommend to first elicit a scientist's prior belief about the true data generating $p^*(X)$ in the most general setting. This prior belief is then subsequently translated into priors on the models. Hence, the resulting prior model probabilities $P(\mathcal{M}_1)$ and $P(\mathcal{M}_2)$ are *derived*.

In contrast, a Jeffreys's Bayes factor follows from a top-to-bottom procedure, where the top level is concerned with the comparison *between* two models (i.e., model classes) for which one has to (subjectively) *choose* prior model probabilities $P(\mathcal{M}_i)$. Based on top level desiderata, i.e., a coherent comparison between the two models, we then *derive* the pair of priors π_1 and π_2 on the lower level that are concerned with the parameters (i.e., model instances) *within* the models $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. In effect, the sole purpose of the pair π_1, π_2 is to mediate scientific learning through the Bayes factor, that is, to update the prior model odds to posterior model odds.

On the other hand, the C&S induction scheme is a bottom-up approach based on the philosophy that the whole is the sum of its parts. At the lowest level, one has to *subjectively elicit* the scientist's prior belief about the true data generating process. The procedure then elaborates on how this lowest level belief can be used to *derive* the model instance priors π_1 and π_2 at the intermediate level. By aggregating the model instance priors of π_1 and π_2 we then get the prior model probabilities $P(\mathcal{M}_1)$ and $P(\mathcal{M}_2)$ at the top level. As such, this method is not free from subjective input on the lowest level.

Our major concern with the C&S method is the lack of invariance, which stems from their recommendation to operationalize their procedure with a seemingly innocent looking finite-dimensional matrix with M rows and W number of columns.[5] By using a finite-dimensional matrix, C&S basically made a choice in how they tackle the statistical modeling problem. The resulting model priors $P(\mathcal{M}_i)$ are sensitive to this choice. More specifically, by initializing their procedure with a finite-dimensional matrix, they use discretized approximations of quantities that are essentially continuous. The approximation error due to discretization is non-negligible, as it permeates through all subsequent steps due to the bottom-up nature leading to an ill-defined Bayes factor.

In brief, we believe that the C&S approach has to overcome some challenges before their

[5]We divert from the C&S notation, where the matrix is $M \times N$ dimensional. As the number of columns do *not* correspond with the number of samples in a data set. Instead, the number of columns refers to the number of possible outcomes a random variable can take on, we use w and W instead.

procedure can be perceived as an extension of a traditional Bayes factors, let alone Jeffreys's Bayes factors. We have three remarks: (1) The C&S procedure is not invariant to how one discretizes the statistical modeling problem; (2) the subjective assessment of the priors on the lowest level and the resulting prior model probabilities $P(\mathcal{M}_i)$ on the top level are, therefore, ill-defined, and (3) model selection based on posterior predictive p-statistics does not lead to a proper measure of evidence.

This paper continues as follows: We first apply the C&S induction scheme to a concrete example. Then we show that we get different results when we choose a different finite-dimensional matrix to operationalize the C&S induction scheme. Lastly, we argue that the implicit discretization necessary for the finite-dimensional matrix is the main culprit of the resulting lack of invariance.

2.1. Running Example

To illustrate why we believe that the C&S method is essentially a belief propagation procedure, we consider a random variable X with a finite number of outcomes W. This W is denoted as n in Chandramouli and Shiffrin (ress) and defines the number of columns in their matrix representation (i.e., their Figures 1 and 2). To simplify matters, we use an example (taken from Ly et al., 2015b) where X has $W = 3$ number of outcomes.

Example 1 (A Psychological Task with Three Outcomes)**.** *In the training phase of a source-memory task, the participant is presented with two lists of words on a computer screen. List $\mathcal{L}$ is projected on the left-hand side and list $\mathcal{R}$ is projected on the right-hand side. In the test phase, the participant is then presented with two words, side by side, that can stem from either list, thus, ll, lr, rl, rr. At each trial, the participant is asked to categorize these pairs as either:*

- x_1 *meaning both words come from the left list, i.e., "ll",*
- x_2 *meaning the words are mixed, i.e., "lr" or "rl",*
- x_3 *meaning both words come from the right list, i.e., "rr".*

Thus, the random variable X has $W = 3$ outcomes. To ease the discussion, we assume that the words presented to the participant are "rr". ◇

As model $\mathcal{M}_1$ we take the so-called individual-word strategy. A participant guided by this strategy will consider each word individually and compare it with list $\mathcal{R}$ only. Within this model $\mathcal{M}_1$, the parameter is given by $\theta_1 = \vartheta$, which we interpret as the participant's "right-list recognition ability". Hence, when the participant is presented with the pair "rr" she will respond x_1 with probability $(1-\vartheta)^2$, thus, two failed recollections; x_2 with probability $2\vartheta(1-\vartheta)$, thus, one failed and one successful recollection; x_3 with probability ϑ^2, thus, two successful recollections.

More compactly, a participant guided by this strategy generates the outcomes $[x_1, x_2, x_3]$ with the following three probabilities $p(X \mid \vartheta, \mathcal{M}_1) = [(1-\vartheta)^2, 2\vartheta(1-\vartheta), \vartheta^2]$, respectively. Note the data generative formulation. For instance, when the participant's true ability is $\vartheta^* = 0.9$, the three outcomes $[x_1, x_2, x_3]$ are then generated with the three probabilities $p(X \mid 0.9, \mathcal{M}_1) = [0.01, 0.18, 0.81]$ respectively. We call the function $p(X \mid \theta_i, \mathcal{M}_i)$ with θ_i fixed a probability mass function (pmf) or model instance of $\mathcal{M}_i$.[6] Hence, every ϑ in $(0, 1)$ yields a pmf that defines W number of probabilities. In effect, the model $\mathcal{M}_1$ consists of a collection of pmfs, which C&S refer to as a model class.

As a competing model $\mathcal{M}_2$, we take the so-called only-mixed strategy. Within this model $\mathcal{M}_2$, the parameter is given by $\theta_2 = a$, which we interpret as the participant's "mixed-list differentiability ability". With probability a the participant first checks whether the presented pair of words are mixed. If she perceives it as mixed, she then produces the outcome x_2 with probability a. If she does not perceive the pair of words as mixed, the participant then randomly chooses x_1 or x_3 each with probability $(1-a)/2$.

More compactly, a participant guided by this strategy generates the outcomes $[x_1, x_2, x_3]$ with the following three probabilities $p(X \mid a, \mathcal{M}_2) = [(1-a)/2, a, (1-a)/2]$, respectively. Again we formulated the model as a data generative process. For instance, when the participant's true ability is $a^* = 1/3$, the three outcomes $[x_1, x_2, x_3]$ are then generated with the same probability, i.e., $p(X \mid 1/3, \mathcal{M}_2) = [1/3, 1/3, 1/3]$. Note that this last pmf $p(X \mid 1/3, \mathcal{M}_2)$ is not in the collection of pmfs defined by $\mathcal{M}_1$. Similarly, the pmf $p(X \mid 0.9, \mathcal{M}_1)$ is not a member of the collection of pmfs defined by $\mathcal{M}_2$.

The two models $\mathcal{M}_1$ and $\mathcal{M}_2$ share only one pmf (model instance), that is, the pmf indexed by $\vartheta = 0.5$ within $\mathcal{M}_1$ and, coincidentally, when $a = 0.5$ within $\mathcal{M}_2$. We use these two models $\mathcal{M}_1$ and $\mathcal{M}_2$ to explain the C&S belief propagation procedure.

---

[6] C&S call the function $p(X \mid 0.9, \mathcal{M}_1)$ a data distribution predicted by the model instance $\vartheta = 0.9$. When we use a capital X we mean the three probabilities simultaneously. On the other hand, a small letter x refers to the probability with which it is generated, for instance, $p(x_w \mid 0.9, \mathcal{M}_2) = 0.18$ when $w = 2$.

2.2. Chandramouli and Shiffrin's Procedure for Induction

For a Bayesian analysis we need priors on the model instances, which we denote by $\pi_i(\theta_i)$ as we have done before,[7] and the priors on the models $P(\mathcal{M}_i)$. Instead of doing so directly, C&S recommend to first (Step 1) elicit the scientist's prior belief about the true data generating process $p^*(X)$ in the most general setting. Next (Step 2) this subjectively chosen prior belief about $p^*(X)$ is used to derive the model instance priors $\pi_i(\theta_i)$ and, subsequently, the model class priors $P(\mathcal{M}_i)$. Lastly, (Step 3) C&S recommend to use posterior p-statistics for inference.

2.2.1. Step 1: Eliciting the Prior on Candidate True Data Generating Pmfs

In our example, the true data generating pmf $p^*(X)$ defines three probabilities $p^*(X) = [p^*(x_1), p^*(x_2), p^*(x_3)]$ with which it generates the three outcomes $[x_1, x_2, x_3]$. For instance, a first candidate true data generating pmf could be $p(X \mid \psi_1) = [0.0, 0.0, 1.0]$, where ψ_1 is an indicator for later reference. A second candidate true data generating pmf could be $p(X \mid \psi_2) = [0.0, 0.1, 0.9]$ and so forth and so on. This method yields a candidate set of true pmfs that we depicted in Table 1. The "matrix" depicted in Table 1 is a simplification of the table in Figure 1 in Chandramouli and Shiffrin (ress) with $M = 66$ rows and $W = 3$ columns. Please ignore the quantities to the right of the double vertical line for the moment. Note that the number of rows $M = 66$ is a result of our arbitrary choice of using a step size of 0.1 on the probabilities. Furthermore, recall that the pmfs $p(X \mid \psi_m)$ are candidates for the true data generating pmf $p^*(X)$ and may not have any connection with the models $\mathcal{M}_1$ and $\mathcal{M}_2$ specified above. Of particular interest is the pmf $p(X \mid \psi_{62}) = [0.8, 0.1, 0.1]$, which is neither a member of $\mathcal{M}_1$ nor of $\mathcal{M}_2$,[8] but because it defines a valid pmf it is, nonetheless, a candidate true data generating pmf.

Given this finite-dimensional matrix of Table 1, C&S then recommend to elicit a scientist's prior belief by setting prior beliefs $\lambda(\psi_m)$ for $m = 1, \ldots, M$, thus, on each candidate true data generating pmf $p(X \mid \psi_m)$.[9] For example, $\lambda(\psi_{62}) = 0.7$ means that the scientist bestows a large portion of belief to the pmf indexed by ψ_{62} as being the true generating pmf $p^*(X)$. Furthermore, $\lambda(\psi_{61}) + \lambda(\psi_{62}) + \lambda(\psi_{63}) = 0.90$ means that the scientist is quite sure that the participant will generate the response x_1 with 80% chance. As λ represents the scientist's prior belief, we necessarily require

[7] C&S denote this by $p_0(\theta_i)$. Instead, we use the Greek letter π_i to distinguish this model instance prior from the prior model probability $P(\mathcal{M}_i)$ on the top level. The subscript i refers to the model membership.

[8] A pmf of $\mathcal{M}_1$ with $p(x_1 \mid \vartheta, \mathcal{M}_1) = 0.8$ requires $\vartheta \approx 0.11$. However, this automatically yields $p(x_2 \mid 0.11, \mathcal{M}_1) = 0.19$. Hence, there is no ϑ in $\mathcal{M}_1$ that leads to the pmf indexed by ψ_{62}. Similarly, a pmf of $\mathcal{M}_2$ necessarily has $p(x_1 \mid a, \mathcal{M}_2) = p(x_3 \mid a, \mathcal{M}_2)$, which is clearly not the case for the pmf indexed by ψ_{62}.

[9] C&S denote this prior pmf probability as $p_0(\psi_m)$. Instead, we use the Greek letter λ to distinguish this prior pmf probability on the lowest level from the model instance prior $\pi_i(\theta_i)$ on the intermediate level and the prior model probabilities $P(\mathcal{M}_i)$ on the top level.

Table 1: The matrix is a simplified version of the matrix found in Figure 1 of C&S with $M = 66$ and $W = 3$. The quantities under the columns with $G(\psi_m, \mathcal{M}_1)$ and $G(\psi_m, \mathcal{M}_2)$ at the top refer to the KL-divergences, see the main text. The parameter under θ_i refers to the model instance that the pmf $p(X \mid \psi_m)$ is allocated to within the model under $\mathcal{M}_i$. For example, the candidate true pmf $p(X \mid \psi_{18})$ is allocated to the model instance $p(X \mid \vartheta = 0.60, \mathcal{M}_1)$ of model class $\mathcal{M}_1$.

	x_1	x_2	x_3	$G(\psi_m, \mathcal{M}_1)$	$G(\psi_m, \mathcal{M}_2)$	θ_i	$\mathcal{M}_i$
ψ_1	0.0	0.0	1.0	0	0.693	$\vartheta = 1.00$	$\mathcal{M}_1$
ψ_2	0.0	0.1	0.9	0.002	0.624	$\vartheta = 0.95$	$\mathcal{M}_1$
ψ_3	0.0	0.2	0.8	0.011	0.555	$\vartheta = 0.90$	$\mathcal{M}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ψ_{11}	0.0	1.0	0.0	0.693	0	$a = 1.00$	$\mathcal{M}_2$
ψ_{12}	0.1	0.0	0.9	0.325	0.368	$\vartheta = 0.90$	$\mathcal{M}_1$
ψ_{13}	0.1	0.1	0.8	0.137	0.310	$\vartheta = 0.85$	$\mathcal{M}_1$
ψ_{14}	0.1	0.2	0.7	0.060	0.253	$\vartheta = 0.80$	$\mathcal{M}_1$
ψ_{15}	0.1	0.3	0.6	0.019	0.198	$\vartheta = 0.75$	$\mathcal{M}_1$
ψ_{16}	0.1	0.4	0.5	0.011	0.145	$\vartheta = 0.70$	$\mathcal{M}_1$
ψ_{17}	0.1	0.5	0.4	0.005	0.096	$\vartheta = 0.65$	$\mathcal{M}_1$
ψ_{18}	0.1	0.6	0.3	0.032	0.052	$\vartheta = 0.60$	$\mathcal{M}_1$
ψ_{19}	0.1	0.7	0.2	0.089	0.017	$a = 0.70$	$\mathcal{M}_2$
ψ_{20}	0.1	0.8	0.1	0.193	0	$a = 0.80$	$\mathcal{M}_2$
ψ_{21}	0.1	0.9	0.0	0.427	0.069	$a = 0.90$	$\mathcal{M}_2$
ψ_{22}	0.2	0.0	0.8	0.500	0.193	$a = 0.00$	$\mathcal{M}_2$
ψ_{23}	0.2	0.1	0.7	0.254	0.148	$a = 0.10$	$\mathcal{M}_2$
ψ_{24}	0.2	0.2	0.6	0.133	0.104	$a = 0.20$	$\mathcal{M}_2$
ψ_{25}	0.2	0.3	0.5	0.057	0.067	$\vartheta = 0.65$	$\mathcal{M}_1$
ψ_{26}	0.2	0.4	0.4	0.014	0.034	$\vartheta = 0.60$	$\mathcal{M}_1$
ψ_{27}	0.2	0.5	0.3	0.000	0.010	$\vartheta = 0.55$	$\mathcal{M}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ψ_{61}	0.8	0.0	0.2	0.500	0.193	$a = 0.00$	$\mathcal{M}_2$
ψ_{62}	0.8	0.1	0.1	0.137	0.310	$\vartheta = 0.15$	$\mathcal{M}_1$
ψ_{63}	0.8	0.2	0.0	0.011	0.555	$\vartheta = 0.10$	$\mathcal{M}_1$
ψ_{64}	0.9	0.0	0.1	0.325	0.368	$\vartheta = 0.10$	$\mathcal{M}_1$
ψ_{65}	0.9	0.1	0.0	0.003	0.624	$\vartheta = 0.05$	$\mathcal{M}_1$
ψ_{66}	1.0	0.0	0.0	0	0.693	$\vartheta = 0.00$	$\mathcal{M}_1$

that $\sum_{m=1}^{M} \lambda(\psi_m) = 1$.

2.2.2. *Step 2: Propagating the Prior Belief to yield the Prior Model Probabilities*

Once the prior beliefs $\lambda(\psi_m)$ about the true data generating $p^*(X)$ are chosen, C&S commence their belief propagation procedure by redistributing $\lambda(\psi_m)$ over the two models. Recall that a model (class) $\mathcal{M}_i$ defines a collection of pmfs (model instances) denoted as $p(X \mid \theta_i, \mathcal{M}_i)$. The allocation of the prior pmf belief of the first candidate true pmf in Table 1, that is, $\lambda(\psi_1)$, is easy, because the associated pmf $p(X \mid \psi_1) = [0.0, 0.0, 1.0]$ does not belong to $\mathcal{M}_2$, but it is a member of $\mathcal{M}_1$; the pmf indexed by ψ_1 is a model instance of $\mathcal{M}_1$ when $\theta_1 = \vartheta = 1$. C&S therefore allocate the prior pmf probability $\lambda(\psi_1)$ to the model instance $\pi_1(\vartheta = 1)$ of $\mathcal{M}_1$. On the other hand, the pmf $p(X \mid \psi_{62}) = [0.8, 0.1, 0.1]$ is neither a member of $\mathcal{M}_2$ nor does it belong to $\mathcal{M}_1$. To nonetheless allocate this prior pmf belief $\lambda(\psi_{62})$ to a model instance of either $\mathcal{M}_1$ or $\mathcal{M}_2$, C&S use a divergence measure denoted by G. For simplicity we take as G the Kullback-Leibler (KL) divergence, which is a measure of dissimilarity. The KL-divergence from a candidate true pmf indexed by ψ_m to a model instance of $\mathcal{M}_i$ is defined as

$$G(\psi_m, \theta_i \mid \mathcal{M}_i) = \sum_{w=1}^{W} p(x_w \mid \psi_m) \log \frac{p(x_w \mid \psi_m)}{p(x_w \mid \theta_i, \mathcal{M}_i)}, \tag{6}$$

and the larger this divergence, the more dissimilar the model instance $p(X \mid \theta_i, \mathcal{M}_i)$ is from the candidate true data generating pmf $p(X \mid \psi_m)$. For example, a direct calculation shows that the KL-divergence between the candidate true $p(X \mid \psi_1)$ in Table 1 to the model instance of $\mathcal{M}_1$ with $\theta_1 = \vartheta = 1.0$ is given by $G(\psi_1, \theta_1 = 1.0 \mid \mathcal{M}_i) = 0$. The KL-divergence is zero if and only if the pmfs indexed by ψ_m and the model instance indexed by θ_i are exactly the same, hence, their dissimilarity is zero.

The KL-divergence between the candidate true $p(X \mid \psi_m)$ and a collection of pmfs defined by the model $\mathcal{M}_i$ is given by $G(\psi_m, \mathcal{M}_i) = \min_{\theta_i} G(\psi_m, \theta_i \mid \mathcal{M}_i)$. That is, the dissimilarity between the candidate true data generating pmf ψ_m and the model $\mathcal{M}_i$ is the smallest dissimilarity between $p(X \mid \psi_m)$ and the model instances $p(X \mid \theta_i, \mathcal{M}_i)$ of model $\mathcal{M}_i$. For example, a direct calculation shows that the KL-divergence from the candidate true data generating pmf $p(X \mid \psi_{62})$ to $\mathcal{M}_1$ is given by $G(\psi_{62}, \mathcal{M}_1) = G(\psi_{62}, \theta_1 = 0.15, \mathcal{M}_1) = 0.137$. Similarly, the KL-divergence between the same candidate true data generating pmf to $\mathcal{M}_2$ is given by $G(\psi_{62}, \mathcal{M}_2) = G(\psi_{62}, \theta_2 = 0.1 \mid \mathcal{M}_2) = 0.310$. Because the divergence from the candidate true pmf $p(X \mid \psi_{62})$ to $\mathcal{M}_1$ is smaller than the divergence to $\mathcal{M}_2$, the C&S procedure implies that we should allocate the prior pmf probability $\lambda(\psi_{62})$ to the prior model instance probability $\pi_1(\vartheta = 0.15)$ belonging to $\mathcal{M}_1$.

We suspect that the underlying idea of this belief allocation procedure is based on the idea of chaining. Thus, if $\lambda(\psi_{62}) = 0.70$, the scientist has much fate in $p(X \mid \psi_{62})$ being the true data generating pmf. However, as $p(X \mid \psi_{62})$ is not in the model $\mathcal{M}_1$ nor in $\mathcal{M}_2$, the C&S procedure then recommends to go for the next best thing; assigning the pmf prior $\lambda(\psi_{62})$ to the model instance that is most similar, in this case, $\pi_1(\vartheta)$ with $\vartheta = 0.15$.

This redistribution of the pmf prior $\lambda(\psi_m)$ to model instance priors can be read from their table in Figure 1 in Chandramouli and Shiffrin (ress) from left to right.[10]

[10]We are unsure what φ in their table indicates.

In our Table 1 the numbers under $G(\psi_m, \mathcal{M}_1)$ and $G(\psi_m, \mathcal{M}_2)$ represents the KL-divergence from the candidate true pmf indexed by ψ_m to the models $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. The parameter value under θ_i indicates which parameter value ϑ within $\mathcal{M}_1$ or a within $\mathcal{M}_2$ corresponds to the model instance that is closest to the pmf of ψ_m. The last column indicates whether the ψ_m is eventually allocated to $\mathcal{M}_1$ or $\mathcal{M}_2$.

As in their table in Figure 1 of Chandramouli and Shiffrin (ress), note that there are multiple candidates ψ_ms allocated to certain parameter values in our Table 1. For example, the candidate pmfs indexed by ψ_3 and ψ_{12} are both allocated to the same model instance indexed by $\vartheta = 0.90$ within $\mathcal{M}_1$. As such, C&S derive the prior on the model instances as $\pi_1(\vartheta) = \sum \lambda(\psi_m)$, where the sum is over the candidates ψ_m which have the same ϑ in the column under θ_i. For example, $\pi_1(\vartheta = 0.90) = \lambda(\psi_3) + \lambda(\psi_{12})$.

After allocating all the M number of prior pmf probability $\lambda(\psi_m)$ to the model instances of either model classes, we have $\pi_1(\vartheta_k)$ and $\pi_2(a_{\tilde{k}})$ for $k = 1, \ldots, K$ and $\tilde{k} = 1, \ldots, \tilde{K}$. The K indicates the number of unique values of ϑs in the column under θ_i. As there are multiple candidates allocated to certain parameter values we typically have $K + \tilde{K} < M$. With the model instance priors at hand, the C&S scheme tells us to aggregate them to yield prior model probabilities, i.e., $P(\mathcal{M}_1) = \sum_{k=1}^{K} \pi_1(\vartheta_k)$ and $P(\mathcal{M}_2) = \sum_{\tilde{k}=1}^{\tilde{K}} \pi_2(a_{\tilde{k}})$. As a result of $\sum_{m=1}^{M} \lambda(\psi_m) = 1$ we have $P(\mathcal{M}_1) + P(\mathcal{M}_2) = 1$.

2.2.3. Step 3: Posterior Predictive p-Statistics

So far, we only discussed the C&S belief propagation procedure as a method to translate a scientist's prior belief $\lambda(\psi)$ about the true data generating $p^*(X)$ to prior beliefs on the model instances $\pi_i(\theta_i)$, which can then be used to define prior beliefs on the models $P(\mathcal{M}_i)$. These priors can be used for inference after we observe data d. As in C&S, we simplify the discussion by supposing that the data consist of one observation where the participant responded with x_1.

To invert the data generative view of pmfs, we fix the data part of each pmf at the observation $p(X \mid \psi_m) = p(d \mid \psi_m)$ and consider the pmfs as a function of ψ_m, i.e., as likelihood functions. Bayes' rule then allows us to update the subjectively chosen pmf prior to a pmf posterior using all specified candidate likelihood functions indexed by the ψ_ms, that is, $\lambda(\psi_m \mid d) = p(d \mid \psi_m)\lambda(\psi_m)/C$, for $m = 1, \ldots, M$, where the normalization constant C is given by $C = \sum_{m=1}^{M} p(d \mid \psi_m)\lambda(\psi_m)$. Recall that the rows $p(X \mid \psi_m)$, thus, the likelihood functions, themselves do not need to belong to the models $\mathcal{M}_1$ and $\mathcal{M}_2$. In fact, most of them do not, as most of the entries under $G(\psi_m, \mathcal{M}_1)$ and $G(\psi_m, \mathcal{M}_2)$ are non-zero.

For inference concerning replication studies, C&S recommend using posterior predictive p-statistics. For example, the observations d_{orig} of the original experiment might suggest that a participant's "right-list recognition ability" ϑ is a half. To test whether this postulate $\vartheta = 0.5$ can be reproduced, C&S recommend to update the subjectively chosen pmf prior about the true $p^*(X)$ to a posterior yielding $\lambda(\psi_m \,|\, d_{\text{orig}})$. Recall that this posterior is also based on likelihood functions $p(d \,|\, \psi_m)$ that do not belong to $\mathcal{M}_1$ as discussed above. For example, if $\lambda(\psi_{62}) > 0$ then $p(X \,|\, \psi_{62}) = [0.8, 0.1, 0.1]$ in Table 1 is used as a likelihood to relate the observations d_{orig} to ψ_{62}. Because there is no ϑ that leads to $p(X \,|\, \psi_{62})$, see the footnote at the end of Section 2.2.1, the likelihood function at ψ_{62} does not and *cannot* extract information about ϑ from d_{orig}.

Nonetheless, C&S use the posterior $\lambda(\psi_m \,|\, d_{\text{orig}})$ to weight all the candidate true pmfs in Table 1 resulting in a posterior predictive $p(x_w \,|\, d_{\text{orig}}) = \sum_{m=1}^{M} p(x_w \,|\, \psi_m)\lambda(\psi_m \,|\, d_{\text{orig}})$ for $w = 1, \ldots, W$. This posterior predictive is used as a sampling distribution, i.e., it defines the probabilities with which new data are generated. If the actually observation d_{rep} is very improbable under this predictive, then the C&S procedure prescribes this as a failure of reproducibility. The problem with this prediction is that it is also calculated from the predictions of $p(X \,|\, \psi_{62})$, even though this pmf ψ_{62} has no connection to ϑ whatsoever.

In sum, it seems that the C&S recommendation for replication boils down to comparing the observed data d_{rep} in a replication attempt using the posterior predictive as a sampling distribution, which is based on irrelevant likelihood functions and subjective belief $\lambda(\psi_m)$. Moreover, by using the posterior predictive as a sampling distribution to assess replication, this method shares many pitfalls with common p-value tests and therefore does not quantify evidence (e.g., Bayarri and Berger, 2000; Wagenmakers, 2007).

2.2.4. C&S Bayes factors

Although C&S do not recommend to use Bayes factors for inference, they note that Bayes factors can be constructed from their belief propagation procedure. The main idea is to reuse the belief propagation procedure, but this time to redistribute the posterior beliefs $\lambda(\psi_m \,|\, d)$ about the true data generating $p^*(X)$ to posterior beliefs for the "model instances" $\pi_i(\hat{\theta}_i \,|\, d)$, which can then be used to define posterior beliefs on the "models" $P(\hat{\mathcal{M}}_i \,|\, d)$. We are reluctant to call $P(\hat{\mathcal{M}}_i \,|\, d)$ the posterior model probabilities, because they are calculated using likelihood functions that do not belong to $\mathcal{M}_i$ (hence, the hats in our notation). There are now two ways to derive a Bayes factor based on the quantities resulting from the C&S belief propagation procedure.

The first method involves the ratio of the posterior and prior model odds, that is,

$$\hat{\mathrm{BF}}_{12}(d) = \frac{P(\hat{\mathcal{M}}_1 \,|\, d)/P(\hat{\mathcal{M}}_2 \,|\, d)}{P(\mathcal{M}_1)/P(\mathcal{M}_2)}. \tag{7}$$

This Bayes factor depends on the subjectively chosen prior beliefs $\lambda(\psi_m)$ about $p^*(X)$, the chosen divergence measure G, and –most troublesome– on the collection of candidate likelihood functions $p(d \,|\, \psi_m)$ rather than on the likelihood that belong to the respective models.

The second method involves the ratio of marginal likelihoods, that is,

$$\tilde{\mathrm{BF}}_{12}(d) = \frac{\sum_{k=1}^{K} p(d \,|\, \vartheta_k, \mathcal{M}_1)\pi_1(\vartheta_k)}{\sum_{\tilde{k}=1}^{\tilde{K}} p(d \,|\, a_{\tilde{k}}, \mathcal{M}_2)\pi_2(a_{\tilde{k}})}. \tag{8}$$

In contrast to $\hat{\mathrm{BF}}_{12}(d)$, this Bayes factor is calculated from the likelihoods $p(d \,|\, \theta_i, \mathcal{M}_i)$ that actually do belong to the respective models. Hence, $\hat{BF}_{12}(d)$ and $\tilde{BF}_{12}(d)$ will differ from each other.

We have some reservations about the Bayes factor as defined in Eq. (7) or Eq. (8) as a generalization of traditional Bayes factors. First, a traditional Bayes factor leads to the same quantity whether it is computed as the ratio of the posterior and prior model odds or as the ratio of marginal likelihoods. Second, a traditional Bayes factor would involve continuous integrals, whenever the parameters ϑ and a are free to vary in continuous intervals. The replacement of the integrals by finite sums is an artefact of only considering a finite number M of candidate true pmfs $p(X \,|\, \psi_m)$.

2.3. Lack of Invariance

Our major concern with Bayes factors calculated from the C&S approach, however, is rooted in its operationalization using a finite-dimensional matrix (e.g., Table 1), as it causes a lack of invariance affecting every step of their belief propagation procedure. As such, two scientist with the same subjective belief $\lambda(\psi)$ about the true $p^*(X)$ using the same divergence measure G, but with a different finite-dimensional matrix will calculate different Bayes factors.

We appreciate the attempt by C&S to assess how well models represent the true data generating process. Their procedure considers all possible data generating pmfs and as such can account for model misspecification. Although attractive, such an unrestrictive view leads to complications when one is concerned with testing models for which one has to set priors. The C&S recommendation is to do so subjectively, which we consider nigh impossible. More specifically, the collection of all possible data generative pmfs $\mathcal{P}$ is typically hard to describe and without a proper description even harder to subjective assign prior beliefs to. Our paper continuous as follows: (1) We first characterize $\mathcal{P}$ and simplify it with a parameterization; (2) a different parameterization of $\mathcal{P}$ is then given leading to a different finite-dimensional matrix. (3) In effect, this leads to different prior beliefs and (4) different allocations, thus, different Bayes factors. (5) Lastly, we remark how this problem is related to the invariance problem already solved by Jeffreys (1946) and what his solution implies for the C&S procedure.

2.3.1. Characterizing the Collection of All Possible Pmfs

When X has $W = 3$ number of outcomes, its true distribution $p^*(X)$ can then be characterized by $W - 1 = 2$ parameters. Recall that a pmf for X then defines the three chances $p(X) = [p(x_1), p(x_2), p(x_3)]$ with which it generates the outcomes $[x_1, x_2, x_3]$. The pmf must therefore satisfy two conditions: (i) it has to be non-negative and bounded by one, i.e., $0 \leq p(x_w) \leq 1$ for each outcome x_w of X with $w = 1, \ldots, W$, and (ii) the probabilities must sum to one, i.e., $\sum_{w=1}^{W} p(x_w) = 1$. Note that this holds true for any candidate true pmf $p(X \mid \psi_m)$ in Table 1. We call the collection of functions for which the conditions (i) and (ii) hold the collection of all possible pmfs or the full model and denote it by $\mathcal{P}$. The collection $\mathcal{P}$ has an uncountably infinite number of members, each capable of being the true data generating process $p^*(X)$. By using a finite-dimensional matrix such as the one in Table 1, C&S, thus, only restrict their prior belief elicitation to only $M = 66$ candidate true pmfs.

To show that even for $W = 3$ the full model $\mathcal{P}$ is uncountable, we first parameterize $\mathcal{P}$, that is, we identify each possible true pmf of $\mathcal{P}$ with a two dimensional parameter $\psi = (b, c)$. Given any pmf $p(x) = [p(x_1), p(x_2), p(x_3)]$, we define $b = p(x_1)$, $c = p(x_2)$ and set $\psi = (b, c)$. This construction is essentially a function ξ that maps a member of the full model $\mathcal{P}$ into a parameter space Ψ of dimension $W - 1 = 2$. Using the inverse parameterization ξ^{-1} we can identify every parameter $\psi = (b, c)$, where (i') $0 \leq b, c \leq 1$ and (ii') $b + c \leq 1$, with a pmf such that the three outcomes $[x_1, x_2, x_3]$ are generated with the probabilities $p(X \mid \psi) = [b, c, 1 - b - c]$. As there are an uncountable number of $\psi = (b, c)$s for which the conditions (i') and (ii') holds, we conclude that there are also an uncountable number of pmfs $p(X \mid \psi)$ in the full model $\mathcal{P}$ for which (i) and (ii) holds.

2.3.2. Different Parameterizations, Different Representation of $\mathcal{P}$: A Different Set of Candidate True Pmfs

The aforementioned parameterization $\xi : \mathcal{P} \to \Psi$ relates to the candidate true pmfs of Table 1 as we have actually chosen $\psi_1 = (0.0, 0.0)$, $\psi_2 = (0.0, 0.1), \ldots, \psi_{62} = (0.8, 0.1), \psi_{63} = (0.8, 0.2), \psi_{64} = (0.9, 0.0), \psi_{65} = (0.9, 0.1), \psi_{66} = (1.0, 0.0)$. The resulting $M = 66$ number of columns is due to the dependence between b and c.

A different parameterization $\tilde{\xi}$ from the full model $\mathcal{P}$ to a parameter space $\tilde{\Psi}$ is based on a "stick-breaking" approach. Given a $p(x)$ we then choose $\tilde{b} = p(x_1)$, $\tilde{c} = p(x_2)/[1 - p(x_1)]$ and define $\tilde{\psi} = (\tilde{b}, \tilde{c})$.[11] Using the inverse parameterization $\tilde{\xi}^{-1}$ we can also identify every parameter $\tilde{\psi} = (\tilde{b}, \tilde{c})$, where (i'+ii') $0 \leq \tilde{b}, \tilde{c} \leq 1$, with a pmf such that the three outcomes $[x_1, x_2, x_3]$ are generated with the probabilities $p(X \mid \psi) = [\tilde{b}, (1 - \tilde{b})c, (1 - \tilde{b})(1 - \tilde{c})]$. Note that every parameter $\tilde{\psi}$ lies within the unit square $\tilde{\Psi} = [0, 1] \times [0, 1]$ and that $\tilde{b}$ and $\tilde{c}$ can be chosen independently from each other. Again, as there are an uncountable numbers in the unit square, we have an uncountable collection of candidate true pmfs $\mathcal{P}$. With this stick-breaking representation of $\mathcal{P}$ and a step size of 0.1 we get the matrix depicted in Table 2.

This new matrix differs substantially from the previous one. First, it has more rows, thus, a larger number of candidate true pmfs; $M = 111$ compared to $M = 66$ in Table 1. Second, there are more candidate pmfs that imply that the first response x_1 is generated with 80% chance; eleven in Table 2 compared to three in Table 1.

2.3.3. Different Representation, Different Prior Beliefs

Expanding on these observations, we suspect that a scientist would subjectively set different prior beliefs depending on whether she is confronted with the matrix of Table 1 or with the matrix of Table 2. In particular, when confronted with the matrix of Table 1 the scientist might subjectively set $\lambda(\psi_{61}) = \lambda(\psi_{63}) = 0.1$ and $\lambda(\psi_{62}) = 0.7$ meaning that she is quite sure, that the participant will generate the response x_1 with 80% chance, i.e., $P(p^*(x_1) = 0.80) = 0.9$. To cohere to this belief the scientist would simply set $\lambda(\tilde{\psi}_{89}) = \lambda(\tilde{\psi}_{99}) = 0.1$ and $\lambda(\tilde{\psi}_{94}) = 0.7$ and, subsequently, set the prior belief of all the "in-between" pmfs that generate x_1 with 80% to zero in the Table 2. We highly doubt that any scientist would be so specific in formulating her prior beliefs and, thus, doubt that a subjective assessment of the prior beliefs will work here.

As an alternative, we might think that we are noninformative if we give each candidate true

[11] This only works if $p(x_1) \neq 1$. When $p(x_1) = 1$, we simply set $\tilde{c} = 0$ and define $\tilde{\psi} = (1, 0)$.

Table 2: The matrix is a simplified version of the matrix found in Figure 1 of C&S based on the different parameterization $\tilde{\xi}$ defined in text. Note how the pmf $p(X \mid \tilde{\psi}_{19})$ is allocated to $\mathcal{M}_2$.

	x_1	x_2	x_3	$G(\psi_m, \mathcal{M}_1)$	$G(\psi_m, \mathcal{M}_2)$	θ_i	$\mathcal{M}_i$
$\tilde{\psi}_1 = (0.0, 0.0)$	0.00	0.00	1.00	0	0.693	$\vartheta = 1.00$	$\mathcal{M}_1$
$\tilde{\psi}_2 = (0.0, 0.1)$	0.00	0.10	0.90	0.003	0.624	$\vartheta = 0.95$	$\mathcal{M}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\tilde{\psi}_{17} = (0.1, 0.5)$	0.10	0.45	0.45	0.000	0.120	$\vartheta = 0.68$	$\mathcal{M}_1$
$\tilde{\psi}_{18} = (0.1, 0.6)$	0.10	0.54	0.36	0.013	0.078	$\vartheta = 0.63$	$\mathcal{M}_1$
$\tilde{\psi}_{19} = (0.1, 0.7)$	0.10	0.63	0.27	0.046	0.041	$a = 0.63$	$\mathcal{M}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\tilde{\psi}_{88} = (0.7, 1.0)$	0.70	0.30	0.00	0.027	0.485	$\vartheta = 0.15$	$\mathcal{M}_1$
$\tilde{\psi}_{89} = (0.8, 0.0)$	0.80	0.00	0.20	0.500	0.193	$a = 0.00$	$\mathcal{M}_2$
$\tilde{\psi}_{90} = (0.8, 0.1)$	0.80	0.02	0.18	0.393	0.212	$a = 0.02$	$\mathcal{M}_2$
$\tilde{\psi}_{91} = (0.8, 0.2)$	0.80	0.04	0.16	0.315	0.233	$a = 0.04$	$\mathcal{M}_2$
$\tilde{\psi}_{92} = (0.8, 0.3)$	0.80	0.06	0.14	0.248	0.256	$\vartheta = 0.17$	$\mathcal{M}_1$
$\tilde{\psi}_{93} = (0.8, 0.4)$	0.80	0.08	0.12	0.189	0.281	$\vartheta = 0.16$	$\mathcal{M}_1$
$\tilde{\psi}_{94} = (0.8, 0.5)$	0.80	0.10	0.10	0.137	0.310	$\vartheta = 0.15$	$\mathcal{M}_1$
$\tilde{\psi}_{95} = (0.8, 0.6)$	0.80	0.12	0.08	0.091	0.342	$\vartheta = 0.14$	$\mathcal{M}_1$
$\tilde{\psi}_{96} = (0.8, 0.7)$	0.80	0.14	0.06	0.053	0.378	$\vartheta = 0.13$	$\mathcal{M}_1$
$\tilde{\psi}_{97} = (0.8, 0.8)$	0.80	0.16	0.04	0.022	0.421	$\vartheta = 0.12$	$\mathcal{M}_1$
$\tilde{\psi}_{98} = (0.8, 0.9)$	0.80	0.18	0.02	0.002	0.474	$\vartheta = 0.11$	$\mathcal{M}_1$
$\tilde{\psi}_{99} = (0.8, 1.0)$	0.80	0.20	0.00	0.011	0.555	$\vartheta = 0.10$	$\mathcal{M}_1$
$\tilde{\psi}_{100} = (0.9, 0.0)$	0.90	0.00	0.10	0.325	0.368	$\vartheta = 0.10$	$\mathcal{M}_1$
$\tilde{\psi}_{100} = (0.9, 0.1)$	0.90	0.01	0.09	0.263	0.384	$\vartheta = 0.10$	$\mathcal{M}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\tilde{\psi}_{110} = (0.9, 1.0)$	0.90	0.00	0.10	0.003	0.623	$\vartheta = 0.05$	$\mathcal{M}_1$
$\tilde{\psi}_{111} = (1.0, 0.0)$	1.00	0.00	0.00	0	0.693	$\vartheta = 0.00$	$\mathcal{M}_1$

pmf the same prior probability. This means that we then give each candidate true pmf of Table 1 a prior probability of $\lambda(\psi_m) = 1/66 \approx 0.0152$. The pmfs that the participant will generate the response x_1 with 80% chance then get a total prior probability of $3/66 \approx 0.0455$. On the other hand, in Table 2 a uniform prior on $\lambda(\tilde{\psi}_m) = 1/111 \approx 0.009$ and the pmfs that the participant will generate the response x_1 with 80% chance then gets prior probability of $11/111 \approx 0.099$. Hence, a different set of candidate true pmfs will lead to a different assessment of prior beliefs. This lack of invariance depends on how many and which true candidate pmfs are chosen from $\mathcal{P}$ in constructing the finite-dimensional matrices of Table 1 and Table 2.

2.3.4. Different Representation, Different Prior Model Probabilities, thus, Bayes Factors

Applying the C&S belief propagation procedure to the matrix of Table 1 yields different allocations, thus, different Bayes factors then when we use the matrix of Table 2. For example, a scientist might believe that the true data generating pmf is close to $p(X \mid \psi_{18}) = [0.1, 0.6, 0.3]$ of

Table 1, thus, chooses $\lambda(\psi_{18}) = 0.50$. This prior belief then gets allocated to the model instance $p(X \mid \vartheta = 0.6, \mathcal{M}_1)$ of $\mathcal{M}_1$. Similarly, we would expect that the scientist would also set $\lambda(\tilde{\psi}_{19}) \approx 0.50$ when confronted with Table 2, because the candidate pmf $p(X \mid \tilde{\psi}_{19}) = [0.10, 0.63, 0.27]$ in the second matrix does not differ that much from the pmf $p(X \mid \psi_{18})$ of the first matrix. However, according to the second matrix the prior pmf probability $\lambda(\tilde{\psi}_{19})$ is then allocated to the model instance $p(X \mid a = 0.63, \mathcal{M}_2)$ of $\mathcal{M}_2$. In effect, a different representation leads to a different belief allocation, thus, different priors $\pi_i(\theta_i)$, $P(\mathcal{M}_i)$ and different posteriors $P(\hat{\mathcal{M}}_i \mid d)$ and, consequently, different Bayes factors. As such, our understanding of the C&S belief propagation procedure leads to an inadequate definition Bayes factors, which depends on how we choose to represent $\mathcal{P}$.

2.3.5. Jeffreys's Prior and the C&S Procedure

The reason for this lack of invariance is due to an error incurred from (1) the parameterization ξ itself, and (2) the discretization of the parameter space. For example, the matrices depicted in Table 1 and Table 2 were derived from the parameterizations ξ and $\tilde{\xi}$, respectively, followed by a discretization of the parameter space with a step size of 0.1 in each coordinate. The first point can be repaired, as Jeffreys (1946) showed that the Fisher information can be used to neutralize the parameterization error. This solution is more commonly known as the Jeffreys's prior. In Ly et al. (2015b) we showed that the Jeffreys's prior on the parameters, say, $\psi = (b, c)$ in Ψ leads to a uniform prior on pmfs in $\mathcal{P}$. The second point however cannot be fixed.

To elaborate on this latter point, recall that the collection of all data generating pmfs $\mathcal{P}$ is uncountably large, which means that the scientist's actual prior belief $\lambda(\psi)$ is a continuous quantity. By using a finite number M of candidate true data generating pmfs, the target continuous random variable $\lambda(\psi)$ is then approximated by a discretized version $\lambda(\psi_m)$. The corresponding discretization errors are comparable to the errors incurred when histograms are used to approximate a smooth density function. Moreover, because the actual belief about ψ is continuous, we have zero probability of having the true data generating process $p^*(X)$ being exactly equal to one of the finite number of candidate pmfs $p(X \mid \psi_m)$. As such, we cannot construct the actual belief $\lambda(\psi)$ from point masses. Note that this continuity issue was already alluded to in Section 2.3.3 as one would expect that if the pmfs indexed by $\tilde{\psi}_{89}, \tilde{\psi}_{94}, \tilde{\psi}_{99}$ in Table 2 are assigned some prior mass, the pmfs in between would also receive some prior mass. The implication is that the C&S procedure might only work if we use a "matrix" with an uncountable number of rows.

Furthermore, the discretization leads to another type of approximation error that we refer to as geometric approximation error due to the chosen divergence measure G. This error was alluded to in Section 2.3.4, where a small change in the candidate true data generating pmf $p(X \mid \psi_{18}) = [0.1, 0.6, 0.3]$ to $p(X \mid \tilde{\psi}_{19}) = [0.10, 0.63, 0.27]$ leads to a completely different allocation of the prior belief; from a model instance of $\mathcal{M}_1$ to $\mathcal{M}_2$ instead. The geometrical interpretation stems from the fact that KL-divergence can be thought of as a generalization of the Fisher information metric.[12] Moreover, it follows directly from the geometric interpretation that the C&S belief propagation procedure favors the more complex model, as it will attracts a larger number of candidate data generating pmfs indexed by ψ_m, see Ly et al. (2015b). This a priori boosting of the more complex model is at odds with the simplicity postulate that seems to be central in the foundations of the C&S procedure, see Shiffrin et al. (ress) in this special issue.

The fact that we cannot construct the actual belief $\lambda(\psi)$ from point masses is at odds with the C&S idea that $P(\mathcal{M}_i)$ is the sum of its parts. This bottom-up view is what caused Shiffrin et al. (ress) to avoid overlapping models; when $\mathcal{M}_1$ and $\mathcal{M}_2$ share a pmf and the shared instance receives some prior mass, this prior mass will be accounted for twice. As a result, the prior model probabilities will then exceed one, i.e., $P(\mathcal{M}_1) + P(\mathcal{M}_2) > 1$. To deal with overlapping models Shiffrin et al. (ress) suggested to remove the common pmfs from the larger model. This idea is elaborated on with a toy example where $\mathcal{M}_3$ is a binomial model with the chance of success θ fixed at $\theta = 0.5$ and where $\mathcal{M}_4$ represents the binomial model in which θ is free to vary between zero and one. They then reformulate $\mathcal{M}_4$ as the binomial model $\tilde{\mathcal{M}}_4$ in which θ is free to vary between $(0, 0.49)$ and $(0.51, 1)$. This replacement of $\mathcal{M}_4$ by $\tilde{\mathcal{M}}_4$ leads to another approximation error. One solution would be to allow $\tilde{\mathcal{M}}_4$ to converge to $\mathcal{M}_4$ by allowing θ to be in $(0, 0.5 - \epsilon) \cup (0.5 + \epsilon, 1)$. This construction however depends on how ϵ goes to zero and induces the Borel-Kolmogorov paradox (e.g., Lindley, 1997; Wetzels et al., 2010). This paradox is another indication of how the C&S belief propagation scheme depends on how we as scientists represent the problem in terms of the chosen parameterization and, subsequently, discretize the parameter space.

In other words, we believe that the lack of invariance is inescapable when the C&S approach is operationalized with a finite-dimensional matrix leading to an oversimplication of the problem resulting in a representation that is not on par with the sophisticated ideas behind the C&S approach.

[12]The KL-divergence is not a metric in the formal sense, only its infinitesimal version can be related to the Fisher information as a metric, i.e., Jeffreys's prior.

2.4. Conclusion

Based on the different strategies used to set priors $\pi_i(\theta_i)$ within the models $\mathcal{M}_i$, we conclude that the C&S belief propagation procedure answers a different question than a traditional Bayes factor. We believe that C&S are mostly concerned with how a scientist's subjective knowledge of the true data generating $p^*(X)$ is permeated in the models $\mathcal{M}_1$ and $\mathcal{M}_2$. Hence, C&S focus on checking whether the models $\mathcal{M}_1$ and $\mathcal{M}_2$ give a good representation of expert knowledge.

As such, we think that the C&S approach can be valuable at the preliminary stage of model building. In particular, by considering all possible data generating pmfs for the random variable X, the C&S procedure forces the statistician to focus on building a model that is relevant for the problem at hand, rather than being restricted by the standard models. We would like to emphasize that our remarks are not aimed at the aspiration of C&S to construct good models that mimic nature well.

Our major concern deals with the finite-dimensional representation that C&S use to operationalize their procedure and the recommendations to set $\lambda(\psi)$ subjectively. The idea to consider the full model $\mathcal{P}$ is to account for misspecification; as a result, however, the subjective assessment of prior beliefs is nigh impossible. Note that the subjective belief $\lambda(\psi)$ is necessary a continuous random variable, because the full model $\mathcal{P}$ contains an uncountable number of candidate true pmfs $p(X \mid \psi)$. To make their procedure viable C&S oversimplify the problem with a finite-dimensional matrix yielding approximation errors that cannot be ignored.

The problem worsens when X is also continuous. In that case, the full model should then be represented by a "matrix" with an uncountable number of rows and columns. Moreover, this full model is far too complex as it does not even allow for consistent inference (Dvoretzky et al., 1956). This is why regularization methods were invented and alternative models were proposed that grow with the number of samples (e.g., Bickel et al., 2006). The goal set by C&S to compare models in a totally unrestrictive setting is ambitious and an active area of research that is progressing slowly, see Borgwardt and Ghahramani (2009), Ghosal et al. (2008), Holmes et al. (2015), Labadi et al. (2014), Salomond (2013) and Salomond (2014) for some recent results.

For estimation problems, one solution would be to forgo the finite matrix representation and consider the prior on $\mathcal{P}$ as a continuous random variable instead. As a replacement of the subjective assessment, we then recommend Jeffreys's prior as it is uniform on $\mathcal{P}$ when X has a finite number of outcomes W. A Jeffreys prior for the full model $\mathcal{P}$ is viable when $W < \infty$, as the distribution of X is then at most a multinomial distribution with W categories. When X is continuous the

Jeffreys prior can then be extended by a method described in Ghosal et al. (1997), which has been used successfully to justify Bayesian nonparametric estimation methods, see also Ghosal et al. (2000) and Kleijn (2013). However, this replacement of the discretizated $\lambda(\psi_m)$ by a continuous version $\lambda(\psi)$ is at odds with the philosophy that the prior on the whole, $P(\mathcal{M}_i)$, is a sum of its parts $\pi_i(\theta_i)$ as the individual model instances then necessarily receive zero mass. Furthermore, we do not know how to translate a continuous $\lambda(\psi)$ on all pmfs $\mathcal{P}$ to the model instances π_i of $\mathcal{M}_i$ without an explicitly defined relationship between the true data generating $p^*(X)$ and the model instances of $\mathcal{M}_i$. In effect, we doubt that the C&S procedure extends traditional Bayes factors and that it is capable of yielding a Jeffreys's Bayes factors that formalizes inductive reasoning and the logic of proof by contradiction. The reason for this doubt is due to the fact that C&S do not focus on the two models under test, instead, they embed these two models within a larger encompassing model as Robert did, see Section 1.

In conclusion, we believe that a Jeffreys's Bayes factor remains the preferred method of inference, because a Jeffreys's Bayes factor does not depend on how the full model $\mathcal{P}$ is represented and discretized. Thus, it does not suffer from the lack of invariance as discussed above. Furthermore, a Jeffreys's Bayes factor does not require a subjectively elicitation of prior beliefs. Note that the Bayes factor focuses on comparing the models $\mathcal{M}_1$ and $\mathcal{M}_2$, no reference is made to any true data generating process $p^*(X)$. Jeffreys was mostly concerned with quantifying the (relative) evidence provided by the observations for either model. The Bayes factor is not concerned with the true data generating process $p^*(X)$ and it does not aspire to do so. Both $\mathcal{M}_1$ and $\mathcal{M}_2$ could be poor descriptions of the true data generating pmf $p^*(X)$, but fortunately it has been shown that the model selected with a Bayes factor is the model closest to the true $p^*(X)$ in terms of KL-divergence (e.g., Dass and Lee, 2004). Hence, the model that is preferred by the Bayes factor will be able to generalize better to yet unseen data –a guarantee that aligns with the spirit of the C&S approach.

3. Conclusion

We would like to thank the authors of both comments for their stimulating remarks and for their creative alternatives and extensions to Jeffreys's Bayes factors. We hope this discussion has resulted in a renewed appreciation for Harold Jeffreys's foundational contributions to model selection and hypothesis testing, and we look forward to future developments in this exciting and important area of research.

References

Bayarri, M., Berger, J., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of statistics*, 40(3):1550–1577.

Bayarri, M. and Berger, J. O. (2000). P values for composite null models. *Journal of the American Statistical Association*, 95(452):1127–1142.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer Verlag.

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. *Lecture Notes-Monograph Series*, pages 135–207.

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 307–321.

Bickel, P., Li, B., Tsybakov, A., van de Geer, S., Yu, B., Valdés, T., Rivero, C., Fan, J., and van der Vaart, A. (2006). Regularization in statistics (with discussion). *Test*, 15(2):271–344.

Bickel, P. J. and Kleijn, B. J. K. (2012). The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, 40(1):206–237.

Borgwardt, K. M. and Ghahramani, Z. (2009). Bayesian two-sample tests. *arXiv preprint arXiv:0906.4032*.

Chandramouli, S. and Shiffrin, R. (in press). Extending Bayesian induction. *Journal of Mathematical Psychology*.

Consonni, G., Veronese, P., et al. (2008). Compatibility of prior specifications across linear models. *Statistical Science*, 23(3):332–353.

Dass, S. C. and Lee, J. (2004). A note on the consistency of Bayes factors for testing point null versus non-parametric alternatives. *Journal of statistical planning and inference*, 119(1):143–152.

Dawid, A. P. and Lauritzen, S. L. (2001). Compatible prior distributions. *Bayesian Methods with Applications to Science, Policy and Official Statistics*, pages 109–118.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–375.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70:193–242.

Etz, A. and Wagenmakers, E.-J. (2015). Origin of the Bayes factor. *arXiv preprint arXiv:1511.08180.*

Friston, K. J. and Penny, W. (2011). *Post hoc Bayesian model selection.* NeuroImage*, 56:2089–2099.*

Ghosal, S., Ghosh, J., and Ramamoorthi, R. (1997). Non-informative priors via sieves and packing numbers. In Advances in statistical decision theory and applications*, pages 119–132. Springer.*

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. Annals of Statistics*, 28(2):500–531.*

Ghosal, S., Lember, J., Van Der Vaart, A., et al. (2008). Nonparametric bayesian model selection and averaging. Electronic Journal of Statistics*, 2:63–89.*

Grazian, C. and Robert, C. P. (2015). Jeffreys' priors for mixture estimation. In Bayesian Statistics from Methods to Models and Applications*, pages 37–48. Springer.*

Heck, D., Wagenmakers, E.-J., and Morey, R. D. (2015). Testing order constraints: Qualitative differences between Bayes factors and Normalized Maximum Likelihood. Statistics and Probability Letters*, 105:157–162.*

Holmes, C. C., Caron, F., Griffin, J. E., Stephens, D. A., et al. (2015). Two-sample bayesian nonparametric hypothesis testing. Bayesian Analysis*, 10(2):297–320.*

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.*

Jeffreys, H. (1961). Theory of Probability*. Oxford University Press, Oxford, UK, 3 edition.*

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

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

Kamary, K., Eun, J., and Robert, Christian, P. (2016). Non-informative reparameterisations for location-scale mixtures. arXiv preprint arXiv:1601.01178.

Kamary, K., Mengersen, K., Robert, C. P., and Rousseau, J. (2014). Testing hypotheses via a mixture estimation model. arXiv preprint arXiv:1412.2044.

Kary, A., Taylor, R., and Donkin, C. (in press). Using Bayes factors to test the predictions of models: A case study in visual working memory. Journal of Mathematical Psychology.

Kleijn, B. (2013). Criteria for Bayesian consistency. arXiv preprint arXiv:1308.1263.

Labadi, L. A., Masuadi, E., and Zarepour, M. (2014). Two-sample bayesian nonparametric goodness-of-fit test. arXiv preprint arXiv:1411.3427.

Lee, M. D., Lodewyckx, T., and Wagenmakers, E.-J. (2015). Three Bayesian analyses of memory deficits in patients with dissociative identity disorder. In Raaijmakers, J. R., Criss, A., Goldstone, R., Nosofsky, R., and Steyvers, M., editors, Cognitive Modeling in Perception and Memory: A Festschrift for Richard M. Shiffrin*, pages 189–200. Psychology Press.*

Lindley, D. V. (1977). The distinction between inference and decision. Synthese*, 36:51–58.*

Lindley, D. V. (1997). Some comments on Bayes factors. Journal of Statistical Planning and Inference*, 61(1):181–189.*

Ly, A., Etz, A., Marsman, M., Epskamp, S., Gronau, Q., Matzke, D., Selker, R., Steingroever, H., Love, J., Jamil, T., and Wagenmakers, E.-J. (2015a). Replication Bayes factors. Manuscript in preparation.

Ly, A., Marsman, M., Verhagen, A., Grasman, R., and Wagenmakers, E.-J. (2015b). A tutorial on Fisher information. Manuscript submitted for publication at the Journal of Mathematical Psychology.

Ly, A., Verhagen, A., and Wagenmakers, E.-J. (in press). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. Journal of Mathematical Psychology.

Marin, J.-M. and Robert, C. P. (2014). Bayesian essentials with R. *Springer.*

Robert, C. (2007). The Bayesian choice: from decision-theoretic foundations to computational implementation. *Springer Science & Business Media.*

Robert, C. (in press). The expected demise of the Bayes factor. Journal of Mathematical Psychology.

Robert, C. P. (1993). A note on Jeffreys-Lindley paradox. Statistica Sinica, *3(2):601–608.*

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

Robert, C. P. (2015). The Metropolis–Hastings algorithm. arXiv preprint arXiv:1504.01896.

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

Rozeboom, W. W. (1960). The fallacy of the null–hypothesis significance test. Psychological Bulletin, *57:416–428.*

Salomond, J.-B. (2013). Bayesian testing for embedded hypotheses with application to shape constrains. arXiv preprint arXiv:1303.6466.

Salomond, J.-B. (2014). Adaptive bayes test for monotonicity. In The Contribution of Young Researchers to Bayesian Statistics, *pages 29–33. Springer.*

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. The Annals of Statistics, *38(5):2587–2619.*

Shiffrin, R., Chandramouli, S., and Grünwald, P. (in press). Bayes factors, relations to minimum description length, and overlapping model classes. Journal of Mathematical Psychology.

Steingroever, H., Wetzels, R., and Wagenmakers, E.-J. (in press). Bayes factors for reinforcement–learning models of the Iowa Gambling Task. Decision.

Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. Nature Reviews Genetics, *10:681–690.*

Turner, B. M., Sederberg, P. B., and McClelland, J. L. (in press). Bayesian analysis of simulation-based models. Journal of Mathematical Psychology.

Verhagen, J. and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. Journal of Experimental Psychology: General*, 143(4):1457.*

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. Psychonomic Bulletin & Review*, 14:779–804.*

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., and van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. Journal of Personality and Social Psychology*, 100:426–432.*

Wetzels, R., Grasman, R. P. P. P., and Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage–Dickey density ratio test. Computational Statistics & Data Analysis*, 54:2094–2102.*

Wrinch, D. and Jeffreys, H. (1919). On some aspects of the theory of probability. Philosophical Magazine*, 38:715–731.*

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

Wrinch, D. and Jeffreys, H. (1923). On certain fundamental principles of scientific inquiry. Philosophical Magazine*, 45:368–375.*

Yang, G. L. and Le Cam, L. (2000). Asymptotics in statistics: Some basic concepts. Berlin, German: Springer.