“Calibrated” Bayes factors should not be used: a reply to Hoijtink, van Kooten, and Hulsker

Richard D. Morey
Cardiff University

Eric-Jan Wagenmakers
University of Amsterdam

Jeffrey N. Rouder
University of Missouri

Abstract

Hoijtink, van Kooten, and Hulsker (in press) present a method for choosing the prior distribution for a Bayes factor analysis that is based on controlling error rates, which they advocate as an alternative to more subjective methods (Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011; Morey & Rouder, 2014). We show that the method they advocate is trivial from a frequentist perspective and yields inferences that do not make sense. We then outline a position on subjectivity that underlies our advocacy of Bayes factors.

In a recent paper for *Multivariate Behavioral Research*, Hoijtink et al. (in press, henceforth HKH) argued that “Bayesian psychologists should change the way they use the Bayes factor”. HKH’s disagreement with us revolves around the choice of priors for use in calculating Bayes factors. HKH advocate the use of so-called “calibrated” priors (henceforth HKH-calibrated), as opposed to the priors we advocate.

In this note, we will show that HKH-calibration is trivial and leads to tests with indefensible properties. We will then discuss the types of priors we advocate, as opposed to the types of priors that HKH characterizes us as advocating.

A Redundant, Flawed Method

We begin with a review of the one-sample scaled-information Bayes factor t test. The scaled-information Bayes factor is a comparison of two hypotheses: $H_0 : \delta = 0$

Address correspondence to Richard D. Morey (richarddmorey@gmail.com). The source code for this paper, all code needed to compile it, and an online supplement containing proofs may be downloaded at https://github.com/richarddmorey/responseToHoijtink.

and $H_1 : \delta \neq 0$. Because $\delta \neq 0$ does not make any predictions with respect to the data, it cannot accumulate evidence from the data. A Bayesian analysis therefore requires that H_1 be represented by a prior distribution. In the case of the scaled-information Bayes factor, the prior distribution is Normal with a mean of 0 and a standard deviation of τ. This prior constraint will impose predictions on the observed effect sizes. Tweaking τ introduces subjective information into the test, as shown in Figure 1A: a small τ corresponds to expecting that the effect is small, and a large τ corresponds to expecting that the effect is large. The scaled-information Bayes factor is how much more likely the observed t statistic would be under H_0 compared to H_1. One-sided and other extensions are straightforward (Morey & Wagenmakers, 2014).

As an alternative to choosing priors based on subjective information, HKH suggest that Bayes factors can be "calibrated" by adjusting τ to obtain equal Type I and Type II error rates. These error rates are defined against a specific alternative δ_1, using a $BF = 1$ as a critical value to decide between the null and the alternative hypothesis.

The procedure advocated by HKH has a number of peculiarities. Among these are the incoherent specification of the alternative (both a normal distribution and point are used, at different times); the dependence of the prior on sample size; the dependence of the prior, and hence the Bayes factor, on the decision criterion used on the Bayes factor; and two definitions of calibration that lead to radically different procedures. HKH-calibration *imposes* an arbitrary quantification of the evidence using error rates — "Bayesian evidence should be equivocal at an (arbitrary) frequentist decision criterion" — rather than justifying this through an appeal to any principle.

HKH-Calibration Yields Trivial Frequentist Tests

Underneath all of these problems lies a more fundamental issue: the method is nothing more than an overly-complicated way of constructing significance tests. By this we do not mean that the resulting method is *similar* to significance tests; we mean the method is *equivalent* to certain kinds of significance tests. To intuit why, consider HKH's first definition of "calibration":

Definition 1 *The Bayes factor for the comparison of H_0 and H_1 is well-calibrated if τ is chosen such that $P(BF_{01} > 1 \mid \delta = 0) = P(BF_{01} < 1 \mid \delta = \delta_1)$, where δ_1 denotes an effect size that is strictly unequal to zero.*

Figure 1B, however, shows that the scaled information Bayes factor is simply a monotonic function of the $|t|$ statistic. Because the error rates in the test are determined solely by the critical $|t|$ statistic used for the decision, HKH's Definition 1 is equivalent to "the Bayes factor is well-calibrated if the critical t statistic for the balanced-error significance test yields $BF_{01} = 1$." This, in turn, is equivalent to:

Equivalent Definition 1 *The Bayes factor is well-calibrated if using $BF = 1$ as a critical statistic would lead to the same decision as using the critical $|t|$ statistic from an balanced-error significance test against $H_1 : \delta = \delta_1$.*

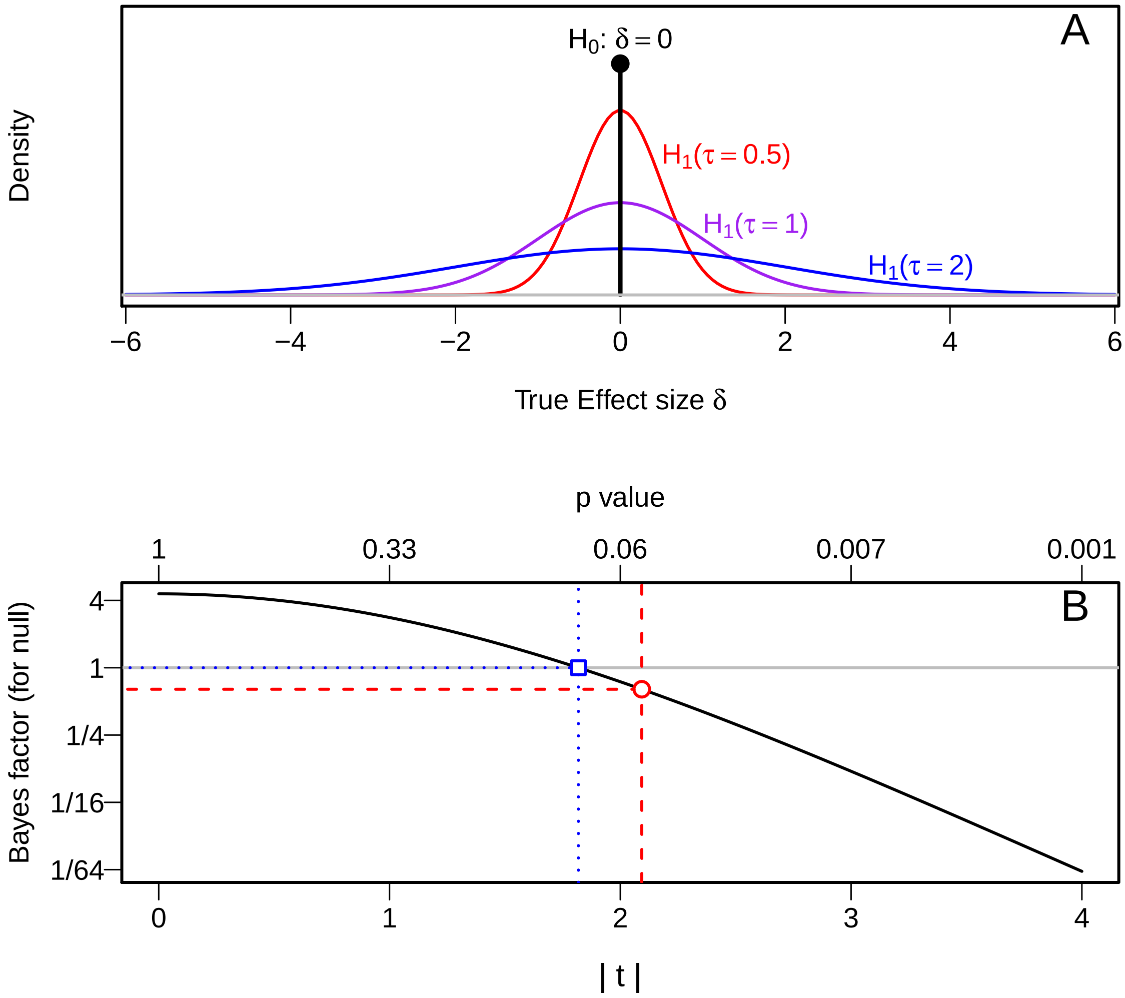

Figure 1. A: Null hypothesis and three different alternative hypotheses for the scaled-information Bayes factor. B: Scaled-information Bayes factor ($\tau = 1$) as a function of $|t|$, for $N = 20$. The dashed lines and open circle show the critical $|t|$ statistic and Bayes factor for $p = 0.05$; the dotted lines and open square show the critical $|t|$ and Bayes factor for $BF = 1$.

For the same reason, HKH's Definition 2 — which requires that the Type I error rate of a decision using a Bayes factor be held constant — yields nothing more than a significance test. There is nothing special about HKH's use of the Bayes factor; because HKH advocate making decisions with particular error rates, all that matters is the t statistic.

What of the complicated sampling and rhetoric about "calibrating" the prior? The only purpose that HKH-calibration serves is to tweak the prior in such a way that the critical t statistic needed for the significance test is the same as the critical t statistic needed for the Bayes factor to be $BF = 1$. This has some odd consequences, besides the wasted time actually spent finding these HKH-calibrated priors. For instance, it is obvious that using HKH-calibration Definition 1 (balanced-error rates), the critical observed effect size $\hat{\delta} = t/\sqrt{N}$ must be about half-way between the null value and the δ_1 value used for HKH-calibration of the priors. As $N \to \infty$, the critical $\hat{\delta}$ approaches $\delta_1/2$ to arbitrary precision. The HKH procedure is, in effect, just a balanced-error simple vs. simple test of $\delta = 0$ vs. $\delta = \delta_1$ (with the strange feature that it is two-tailed).

HKH-Calibration Yields Nonsensical Inferences

In contrast to the simple vs. simple logic of HKH-calibration, the scaled-information Bayes factor is a test of a simple null hypothesis $\delta = 0$ versus the *composite* alternative $\delta \sim \text{Normal}(0, \tau)$. This mismatch between the simple vs. composite logic of the scaled-information Bayes factor and the hidden simple vs. simple logic of the HKH method has serious consequences. Because the critical $\hat{\delta}$ corresponding to $BF = 1$ is forced to be about half-way between 0 and δ_1, no matter how large the sample size, the Bayes factor for an observed effect size of one-half the HKH-calibrated effect size will always be equivocal. This, in turn, means that HKH's procedure can never find evidence against the null hypothesis when the observed $\hat{\delta}$ is closer to the null than the absolute value of the HKH-calibrated effect size.

To see how absurd the conclusions from an HKH-calibrated Bayes factor can be, suppose we perform a HKH test calibrated against $\delta_1 = 0.4$, and consider the Bayes factor for $\hat{\delta} = 0.19$; that is, we observe an effect size just under half the calibrated effect size. For any reasonable analysis, as $N \to \infty$ this observation should lead to increasing evidence against the null hypothesis. The scaled-information Bayes factor exhibits the desired behavior, as shown in Figure 2A. The adjustment of the criterion in the HKH test to $\delta_1/2$, however, means that as more data is obtained, an ever-increasing amount of evidence for the *null* is obtained from this decidedly non-null observation, as shown by the solid line in Panel A. The HKH-calibrated Bayes factor is not a reasonable analysis method.

The cause of this strange behavior is HKH's "calibration" of the prior scale parameter τ. Figure 2B shows the increase in the prior scale as a function of the sample size. To accommodate the unreasonable requirement that an observed effect size of $\hat{\delta} = 0.2$ yield equivocal evidence as $N \to \infty$, the prior scale must be increased to bias the analysis increasingly toward the null hypothesis. At a sample size of 500, the HKH-calibrated scale is well over 1000. To put this in perspective, a prior scale of 1000 means that a priori, true effect sizes of $\delta > 100$ — several orders of magnitude larger than those actually seen in experiments — are expected with 12 to 1 odds. Such an absurd prior is needed to yield

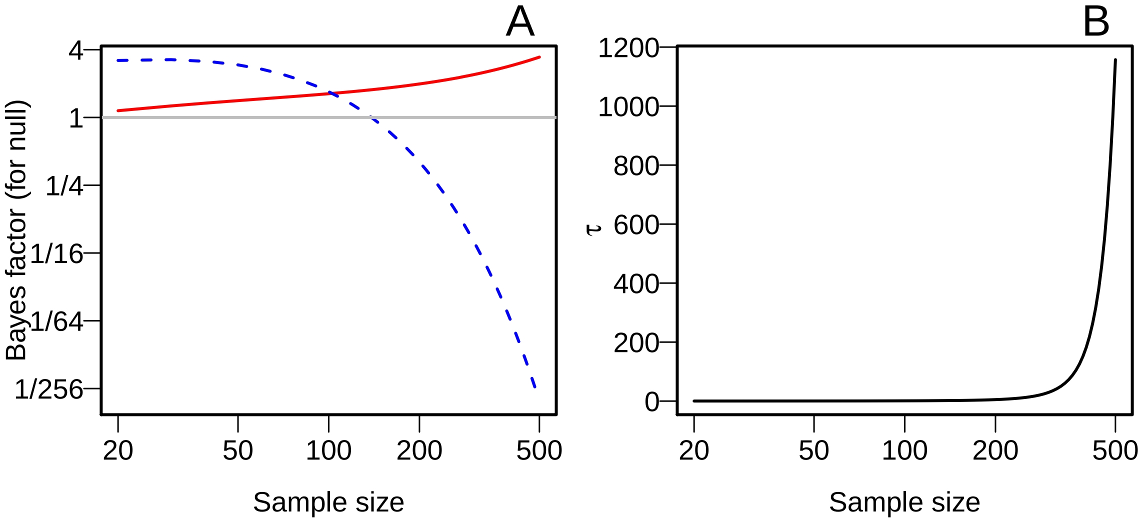

Figure 2. A: HKH-calibrated (solid) and scaled-information (dashed) Bayes factors for an observed effect size of $\hat{\delta} = .19$ as a function of sample size. B: The HKH-calibrated prior scale τ as a function of sample size. In both plots, HKH-calibration is performed against $\delta_1 = 0.4$.

such an absurd result. The prior is so extreme, in fact, that it is charitable to even call it a "prior"; it is simply a deterministic function of N and δ_1 that serves no discernible purpose other than to complicate an otherwise simple significance test.

More basic problems with HKH-calibration exist with small τ values. In their Figure 2, HKH show curves for a number of possible calibration effect sizes δ_1, assuming a sample size of $N = 36$. The smallest two HKH-calibration effect sizes shown are $\delta_1 = .2$ and $\delta_1 = .25$. What HKH do *not* show is the actual calibrated τ for these two effect sizes; these calibrated τ values fall to the left of the graph, beyond the plotted values. The reader is left to guess what those calibrated τ values might be.

For $\delta_1 = .2$ and $N = 36$, the only τ that meets HKH's calibration Definition 1 is $\tau = 0$, as shown in Figure 3A. Recall that τ is the standard deviation of the Normal prior under H_1. If $\tau = 0$, then H_1 is represented by a Normal with mean 0 and standard deviation 0; that is, the alternative H_1 is exactly the same as the null H_0. If the null and alternative are identical, *all data* will yield a Bayes factor of 1. In fact, whenever the critical $|t|$ for the HKH-calibrated Bayes factor is $|t| \leq 1$ the HKH-calibrated Bayes factor always exactly 1. Figure 3B shows for which combinations of δ_1 and N this occurs; for all values in the shaded region, the HKH-calibrated Bayes factor will equal 1, regardless of the data.

Even if a HKH-calibration exists for $\tau > 0$, it may be so close to 0 that the resulting Bayes factor has strange properties. For $\delta_1 = .25$ and $N = 36$ under HKH's Definition 1, the calibrated $\tau = 0.057$ (see Figure 3A). As Figure 3C shows, this τ is so close to zero that the scaled-information Bayes factor for the alternative can never exceed 7.06, regardless of the size of the t statistic.

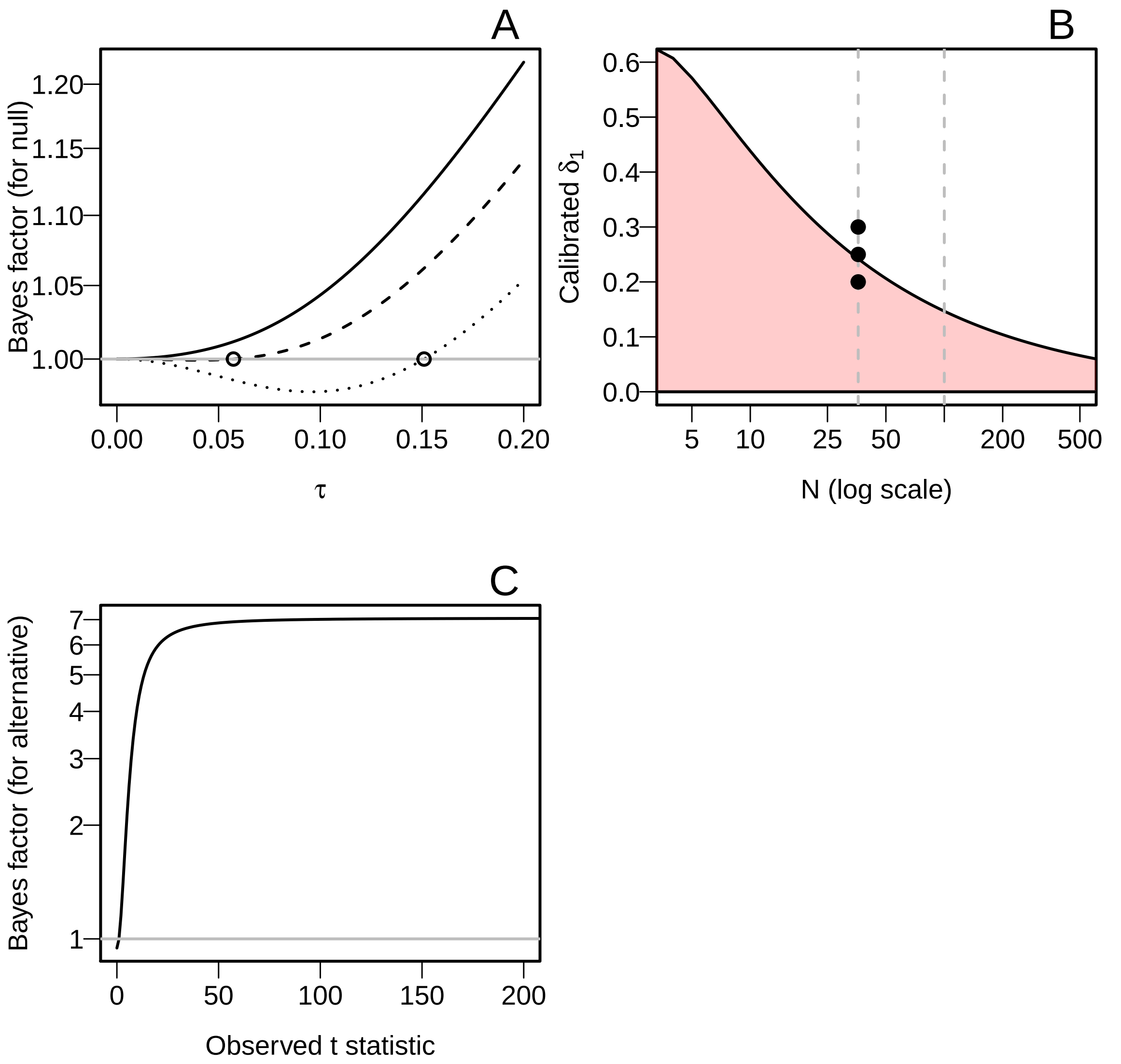

Figure 3. A: Bayes factor as a function of τ for three t statistics (top to bottom): 0.91, 1.03, and 1.15, corresponding to the critical t values for HKH-calibration to δ_1 values of .2, .25, and .3. Circles show $\tau > 0$ values where $BF = 1$. For $t = 0.91$, no such value exists. B: Combinations of δ_1 and N that have only trivial ($\tau = 0$) calibrations are shown as the shaded region. Vertical lines denote sample sizes used on HKH's figures; points represent δ_1 values of .2, .25, and .3. C: Bayes factor as a function of t for $\tau = 0.057$, the HKH-calibrated τ for $N = 36$ and $\delta_1 = .25$.

We believe it goes without saying that no one should use a method that accumulates infinite evidence for the null hypothesis when $\hat{\delta}$ approaches a non-zero, possibly even large, effect size as sample size grows. We also do not believe anyone should use a method that forces all Bayes factors to be small, or identically 1, even when t is arbitrarily large. HKH-calibration is indefensible.

Our Subjective Approach

The authors of this rejoinder do not agree on all philosophical points. However, we do agree on a few major points related to the subjective nature of Bayesian priors. Perhaps it would be profitable to state our consensus as it currently stands.

We advocate a "family default" approach

The default prior is better thought of as a default *family* of priors. The parameter τ indexes a single, intuitive parameter that can be changed to inject subjective information into the analysis. Bayesian prior elicitation is always a trade-off of flexibility with plausibility (Goldstein, 2006); too much flexibility, and analyses become impossible because there are too many options to specify. Too little flexibility, however, yields difficult-to-interpret analyses because the "Bayesian egg" remains unbroken. If a researcher finds that the families of priors we advocate are too constraining, they should not use them. As we have stated in our previous work, however, we believe that the subjective properties of the priors we advocate are appealing.

Our "family default" approach differs markedly from the "single default" approach that HKH suggests we advocate.[1] We agree that subjective information is important. We also agree that researchers need assistance in understanding the subjective implications of the "family default" priors, and in our previous work we have tried to provide guidance (Rouder et al., 2009, 2012). As Morey and Rouder (2011) and de Vries and Morey (2013) note, τ in the JZS Bayes factor (typically denoted r) has a direct interpretation in terms of the prior probabilities of ranges of effect sizes: $|\delta| > \tau$ has a prior probability of 1/2. This is tremendously useful in interpreting the prior subjectively, and makes clear that the value of τ is much less arbitrary than HKH would have us believe. If τ is 1, then there is a 50% prior probability that $|\delta| > 1$. If τ is 2, then there is a 50% prior probability that $|\delta| > 2$. Given that $\delta = 2$ is a very large effect size in most settings, it is immediately obvious that τ cannot be too much larger than 1, unless very large effect sizes are expected.

In discussing the subjective nature of the Bayesian prior, de Vries and Morey (2013) state that "[t]he scaling factor [] allows the adjustment of the weighting distribution for different areas of study, across which plausible effects may vary... [P]lausible effect sizes

[1] Confusingly, HKH appear to admit that their reading is unfair. They write that "[i]t has to be noted that Rouder et al. (2009) and Wagenmakers et al. (2011) either in the publications referred to in the current paper or in other publications, also note that the choice $\tau = 1$ is to some degree arbitrary and other choices could/should be considered."

may vary from study to study, and the [] scale can be adjusted accordingly," and they discuss their choice of prior scale in the context of single-subject research.

Rouder and Morey (2012) give interpretations of τ in terms of the true correlation coefficient in regression models, allowing researchers to use this familiar metric in setting their priors. Wagenmakers et al. (2011) explicitly show how the Bayes factor changes across many values of τ. Finally, the authors have written in less formal forums about the subjective interpretation of τ, even providing an applet to allow anyone to visually see how the prior, posterior, and Bayes factor change in response to changes in τ (Morey, 2014). We continue to work hard to make these methods usable and transparent to everyone, and this has included helping researchers understand the subjective interpretation of the prior.

We advocate "consensus" priors

The second point on which the present authors agree is that a particular researcher's subjective prior is of limited use in the context of a public scientific discussion. Statistical analysis is often used as part of an argument. Wielding a fully personal, subjective prior and concluding "If you were me, you would believe this" might be useful in some contexts, but in others it is less useful. In the context of a scientific argument, it is much more useful to have priors that approximate what a reasonable, but somewhat-removed researcher would have in the situation. One could call this a "consensus prior" approach. The need for broadly applicable arguments is not a unique property of statistics; it applies to all scientific arguments. We do not argue to convince ourselves; we should therefore make use of statistical arguments that are not pegged to our own beliefs. Subjective Bayesianism is *always* a model of an idealized person (Morey, Romeijn, & Rouder, 2013); in some situations, we might model our own beliefs, but in others we might choose to model someone else's belief. Both approaches are subjective Bayesian approaches.

As Rouder et al. (2009) pointed out, the default family priors are useful precisely because they have built-in subjective information: effect sizes tend to be small, and increasingly-large effect sizes are increasingly unlikely. This is the sort of subjective information a reasonable colleague would have available to them, and our priors reflect this.

It should now be obvious how we make our "Bayesian omelet"; we break the eggs and cook the omelet for others in the hopes that it is something like what they would choose for themselves. With the right choice of ingredients, we think our Bayesian omelet can satisfy most people; others are free to make their own, and we would be happy to help them if we can.

Conclusion

Reasonable prior distributions are a critical aspect of Bayesian analysis. In our work, we advocate prior distributions that would garner broad agreement as being reasonable, without being highly tailored to any individual. We provide ways of changing the subjective content to suit researchers in different fields, without adding so much complexity that the analyses become unwieldy. This approach is grounded in the subjective Bayesian viewpoint.

HKH's alternative to this approach, however, is indefensible. If one is interested in controlling error rates, Neyman and Pearson (1933) outlined a comprehensive theory of frequentist testing that can be used to do so. If one is interested in statistical evidence, likelihoodism (Edwards, 1972; Royall, 1997) and Bayesian theories provide adequate account of such ideas. HKH present an incoherent hybrid, which in practice amounts to a significance test no one would use (a two-tailed simple vs. simple hypothesis test) and which does not have any good Bayesian properties. Anyone seeking to "change the way they use Bayes factor" would be ill-advised to seriously consider the method that HKH advocate.

References

de Vries, R. M. & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods*, *18*(2), 165–185.

Edwards, A. (1972). *Likelihood: an account of the statistical concept of likelihood and its application to scientific inference*. London: Cambridge University Press.

Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. *Bayesian Analysis*, *1*, 403–420.

Hoijtink, H., van Kooten, P., & Hulsker, K. (in press). Why Bayesian psychologists should change the way they use the Bayes factor. *Multivariate Behavioral Research*.

Morey, R. D. (2014). Bayes factor *t* tests, part 2: Two-sample tests. Retrieved from http://bayesfactor.blogspot.co.uk/2014/02/bayes-factor-t-tests-part-2-two-sample.html.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2013). The humble Bayesian: model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, *66*, 68–75.

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

Morey, R. D. & Rouder, J. N. (2014). Bayesfactor: computation of Bayes factors for common designs. R package version 0.9.9.

Morey, R. D. & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics & Probability Letters*, *92*, 121–124.

Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, *231*, 289–337.

Rouder, J. N. & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877–903.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374.

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

Royall, R. (1997). *Statistical evidence: a likelihood paradigm*. New York: CRC Press.

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