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

Author Note

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.

Abstract

Hoijtink, van Kooten, and Hulsker (in press) present a method for choosing the prior distribution for an analysis with Bayes factor that is based on controlling error rates, which they advocate as an alternative to our more subjective methods (Morey & Rouder, 2014; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). We show that the method they advocate amounts to a simple significance test, and that the resulting Bayes factors are not interpretable. Additionally, their method fails in common circumstances, and has the potential to yield arbitrarily high Type II error rates. After critiquing their method, we outline the position on subjectivity that underlies our advocacy of Bayes factors.

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

In a recent paper for *Multivariate Behavioral Research*, Hoijtink, van Kooten, and Hulsker(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 contend that the priors we advocate are not subjective, and advocate “calibrating” the prior distribution and Bayes factor (henceforth, we call this HKH calibration).

In this note, we will show that HKH calibration corresponds to significance testing and produces Bayes factors that are uninterpretable. We will then discuss why HKH are incorrect in their belief that the priors we advocate are non-subjective.

The scaled-information Bayes factor

HKH present their method in the context of a one-sample design. Let y_i, $i = 1, \ldots, n$ be independent samples from a normal population:

$$y_i \overset{indep.}{\sim} \text{Normal}(\mu, \sigma^2)$$

where μ and σ^2 are the mean and variance of the normal population, respectively. A so-called non-informative prior is placed on the parameter σ^2:

$$p(\sigma^2) \propto (\sigma^2)^{-1}.$$

For the standardized effect size $\delta = \mu/\sigma$, we can define the scaled-information Bayes factor as a comparison of two hypotheses: $H_0 : \delta = 0$ and $H_1 : \delta \neq 0$. A Bayesian analysis requires that H_1 be represented by a prior distribution that assigns various plausibilities to ranges of true δ; in the case of the scaled-information Bayes factor, the prior distribution on δ is Normal with a mean of 0 and a standard deviation[1] of τ:

$$\delta \sim \text{Normal}(0, \tau^2).$$

[1] We use τ to be consistent with HKH; this is not the same as the prior precision, which is often also denoted τ.

Changing the parameter τ introduces subjective information into the test, as shown in Figure 1A: a small τ corresponds to the belief that the true standardized effect δ is small, and a large τ corresponds to the belief that the true effect is large. For a given τ, the scaled-information Bayes factor indicates how much more likely the observed t statistic would be under H_0 compared to H_1. This ratio is a measure of the relative strength of evidence provided by the data for H_0 over H_1.

The logarithm of the scaled-information Bayes factor (BF), in favor of the null hypothesis, is

$$\begin{aligned}\log(BF) &= -\frac{N}{2}\log\left(1+\frac{t^2}{N-1}\right)+\frac{N}{2}\log\left(1+\frac{t^2}{(1+N\tau^2)(N-1)}\right)\\ &+\frac{1}{2}\log\left(1+N\tau^2\right). \qquad (1)\end{aligned}$$

In this equation there are four unknown quantities: the Bayes factor, t^2, N, and τ. Specifying any three of these quanties determines the fourth, assuming a solution exists. In a typical data analysis, t and N are provided by the data and τ is chosen *a priori*. When the resulting $\log(BF)$ is large and positive, the null hypothesis is preferred; when $\log(BF)$ is large and negative, the alternative hypothesis is preferred. Note that the data affects the Bayes factor only through t and N, and that for a given N the Bayes factor is a monotone function of t^2: as the observed effect size increases, the Bayes factor will decrease, becoming more favorable to the alternative hypothesis.

HKH calibration

HKH suggest that the scaled-information Bayes factor can be "calibrated" by choosing τ according to some rule related to error rates. HKH are not very clear on how this would be done in practice, but they offer two definitions of "calibration" that will help find an HKH-calibrated τ value. Consider their first definition:

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 a effect size that is strictly unequal to zero.*

As Figure 1B shows, however, 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 (which we denote t_0) or equivalently, the critical standardized effect size (denoted d_0) used for the decision, HKH's Definition 1 is equivalent to "the Bayes factor is well-calibrated if the critical t_0 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_0 statistic from an balanced-error significance test against $H_1 : \delta = \delta_1$.*

HKH calibration therefore amounts to the following procedure:

i. Assert that a specific $|t| = t_0$ statistic for a given N should yield a Bayes factor of 1. This leaves only τ in Eq. 1 unknown.

ii. Solve Eq. 1 for τ, yielding an HKH-calibrated τ value.

iii. Compute the Bayes factor BF_{HKH} for the observed t, N, and the HKH-calibrated τ value.

iv. Decide in favor of H_0 if $|t| > t_0$ and in favor of H_1 if $|t| < t_0$.

Notice that step iv is disconnected from the other steps and involves only checking for statistical significance. To the extent that HKH-calibration has interesting properties, it must be in the interpretation of the resulting Bayes factors. To shed light on the properties of a Bayes factor calibrated using HKH's Definition 1, it will help to define in detail the equivalent significance test. Under the null hypothesis H_0 that $\mu = 0$, we know that

$$d = \bar{y}/s \sim \text{Student's } t_{N-1}/\sqrt{N},$$

where $\bar{y}$, s, and d are the sample mean, sample standard deviation, and an estimate of the standardized effect size $\delta = \mu/\sigma$, respectively. For our purposes it will be useful to think in terms of d instead of t, but this is exactly parallel to the standard one-sample null hypothesis, because $d = t/\sqrt{N}$.

Suppose we would like to test the null hypothesis $H_0 : \delta = \mu = 0$ against a two-sided alternative hypothesis $H_1 : |\delta| = \delta_1$ for some positive δ_1. It is important to note that this is not a typical significance test with alternative hypothesis $\delta \neq 0$; rather, it is a test of two simple point hypotheses, $\delta = 0$ vs. $|\delta| = \delta_1$. Under the specific alternative hypothesis δ_1, d has a scaled noncentral t distribution with noncentrality parameter $\delta_1\sqrt{N}$:

$$d \sim \text{Noncentral } t_{N-1}(\delta_1\sqrt{N})/\sqrt{N}.$$

An example of this setup is shown in Figure 2A for $\delta_1 = .4$ and $N = 36$. The critical bounds that determine when the test rejects are shown as vertical lines; in order to build a test with specified Type I or Type II error rates, it is only necessary adjust these bounds.

Under either HKH Definition 1 or Definition 2, one needs only to specify a single number to determine the critical bounds. Under Definition 1, one can specify either δ_1 or the Type I error rate α; from either of these numbers the other follows. Under Definition 2, one need only specify α.[2] HKH calibration then asserts that an observation at one of the critical bounds yields a Bayes factor of 1.

The central issue with HKH calibration is now obvious: the calibration *imposes* an arbitrary quantification of the evidence on the data — *Bayesian evidence should be equivocal at an arbitrary frequentist decision criterion, determined by a significance test* — rather than justifying this through an appeal to any principle related to evidence itself. In our opinion, this fact alone should be enough to dissuade anyone from using HKH-calibrated Bayes factors; however, we can also explore the inferential properties of HKH-calibrated Bayes factors and show they violate the requirements one would have for a reasonable quantification of the strength of evidence.

[2] HKH present their Definition 2 with a fixed $\alpha = .05$. Because .05 is arbitrary, we use a generalization of Definition 2 that allows any α.

HKH Calibration Yields Bad Inferences

HKH are not clear on how they expect researchers to implement calibrated analyses; however, understanding that the equal-error calibrated analysis is equivalent to a significance test of $\delta = 0$ vs. $|\delta| = \delta_1$ reveals that an HKH-calibrated analysis for a given data set is completely determined by one of two settings: the stipulated Type I error rate or the chosen δ_1. Choosing either will determine τ. In practice, therefore, HKH calibration must amount to choosing either α or δ_1. Regardless of which the analyst chooses to set, the resulting Bayes factor has bad properties. We discuss both options in turn.

Calibrating by choosing error rates. In describing what they call the "rational" option for choosing τ (p. 13 of the manuscript), HKH suggest that desired error rates can be used as a guide to calibration. They suggest choosing δ such that the error rates are neither too high nor too low.

Under Definition 1, choosing desirable error rates completely determines the alternative standardized effect size δ_1. The idea of choosing one's error rates and using them to choose an alternative hypothesis is thinking backward: in both frequentist and Bayesian analysis, the analyst chooses an alternative hypothesis first, based on substantive knowledge. The error rates are deductions based on the alternative hypothesis and the sample size (e.g. Neyman, 1957). Although we believe choosing the error rates first is not defensible, we can examine the properties of the resulting Bayes factor under such a calibration scheme. The results here hold for both of HKH's definitions of calibration.

Because the precision with which δ is estimated increases as the sample size N increases, in order to obtain a fixed-error significance test as $N \to \infty$, the alternative standardized effect size δ_1 and the critical effect size $\pm d_0$ must decrease and approach 0. As $\delta_1 \to 0$, the alternative hypothesis is more and more similar to the null hypothesis; to accommodate this, the calibrated τ must also approach 0. Figure 3A shows how the calibrated τ approaches 0 as $N \to \infty$ assuming that the desired Type I error rate is $\alpha = .2$.

The ever-decreasing calibrated τ, in turn, induces strange behaviors in the Bayes factor. Consider that a basic desiderata of a Bayes factor is that an observation *most consistent* with the null hypothesis — in this case, $d = 0$ — should yield an increasing amount of evidence for the null hypothesis as $N \to \infty$: that is a null observed effect size should be more convincing with 1000 participants than with 100. The scaled information Bayes factor has this property, as shown in Figure 3B (solid line). The HKH-calibrated Bayes factor, however, does not. The null effect size $d = 0$ yields *decreasing* amounts of evidence for the null hypothesis as $N \to \infty$, due to the fact that $\tau \to 0$. Figure 3B (dashed line) shows that the Bayes factor for $d = 0$ continually decreases, approaching a constant $BF < 2$ as N grows.

This strange behavior implies that the HKH-calibrated Bayes factor cannot be interpreted as a measure of the strength of evidence provided by the data, if calibrated against a fixed error rate. In addition, calibrating against fixed error rates would not yield error rates that decrease to 0 as the sample size increases: under Definition 1, α and β will both be fixed and hence cannot approach 0; under Definition 2, β will decrease to 0, but α will be fixed.

Calibrating by choosing δ_1. HKH also suggest that it is possible to translate prior knowledge about standardized effect sizes into an “optimal” τ. For instance, they note that if one is looking for an effect size between .2 and .3, one could choose τ such that a test against $\delta_1 = .25$ yields equal error rates (p. 12 of the manuscript). In line with HKH’s suggestion, we can examine the properties of the calibrated Bayes factor when we select a reasonable *a priori* δ_1 against which to calibrate and then choose the τ according to their Definition 1.

In order to see the problem with calibration against a fixed δ_1, the following fact is helpful: as $N \to \infty$, if we keep the error rates balanced as suggested under Definition 1, then the critical bounds $\pm d_0$ will approach $\pm\delta_1/2$ (proof provided in the supplemental materials). Figure 2B shows the significance test for $\delta_1 = .4$ and $N = 180$, and indeed, the critical bounds $\pm d_0$ are very close to half-way between the two hypotheses. This makes sense: for a test of $\delta = 0$ against $|\delta| = \delta_1$, the null hypothesis is preferred when d

is closer to 0; otherwise, $|\delta_1| = \delta_1$ is preferred.

Because the critical effect sizes $\pm d_0$ are forced to correspond to $BF = 1$, and these will be about half-way between 0 and $\pm\delta_1$, 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 when calibrated against a fixed δ_1, HKH's procedure can never find evidence against the null hypothesis when the observed d is closer to the null than the absolute value of the HKH-calibrated effect size.

To see this, suppose we perform a HKH test calibrated against $\delta_1 = 0.4$, and consider the Bayes factor for $d = 0.19$; that is, we observe a standardized effect size just under half the calibrated effect size. We choose $\delta_1 = 0.4$ because HKH used it in their figures; any other value of δ_1 would suffice.

For any reasonable analysis, as $N \to \infty$ an observation of $d = 0.19$ should lead to increasing evidence against the null hypothesis as our uncertainty about δ diminishes. The scaled-information Bayes factor exhibits the desired behavior, as shown in Figure 4A. 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.

This strange behavior is caused by the fact that the prior scale τ must be increased to bias the analysis increasingly toward the null hypothesis in order to accommodate the unreasonable requirement that an observed effect size of $d = 0.2$ yield equivocal evidence as $N \to \infty$. Figure 4B shows the increase in the prior scale as a function of the sample size. 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 standardized 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 extreme prior is needed to yield such an extreme result. The prior is so extreme, in fact, that we do not see how it can be called a "prior" at all. The resulting Bayes factor is uninterpretable.

Another way to view this is from a frequentist perspective. Ignoring the Bayes factor calibration and simply focusing on the equal-error significance test, if the true δ

is less than one-half the calibrated δ_1, then the probability of a Type II error — as defined as making a decision for H_0 when $\delta \neq 0$ — approaches 1 as $N \to \infty$. Consequently, the power of the test goes to 0. Only if the true δ is closer to $-\delta_1$ or δ_1 will the power of the test increase as sample size increases. This is why no one uses significance tests with point-alternative hypotheses in practice.

To make this more concrete, imagine a group of researchers attempting to replicate a result in the psychological literature that has an estimated standardized effect size of 0.2. In order to accurately assess the effect, they obtain $N = 1000$ participants using an online experimental methodology. Following HKH, they decide to use Definition 1 compute a calibrated Bayes factor against $\delta_1 = 0.2$, obtaining an HKH $\tau = 6$. The critical bounds for rejection are $|d| > 0.103$; as expected, very close to $\delta_1/2$. The conditional error rates are

$$Pr(\text{Deciding } H_1 \mid \delta = 0) = Pr(\text{Deciding } H_0 \mid \delta = \delta_1) = 0.0011.$$

Given the large sample size, the researchers are happy with this high power and low type I error rate.

Suppose our researchers observe a standardized effect size of $d = 0.099$ ($SE_d = 0.032, t_{999} = 3.131, p < .001$). Using the scaled information Bayes factor with a $\tau = 0.297$ chosen using the prior studies (yielding a 50% probability that $|\delta| < 0.2$)[3], we obtain a Bayes factor of $BF_{10} = 13$. This Bayes factor favors the alternative hypothesis, as would be expected from the moderately large t value. The HKH-calibrated Bayes factor in favor of the alternative hypothesis, on the other hand, is $BF_{HKH} = 0.65$. Because this is less than 1, the researchers would decide in favor of the null hypothesis, in spite of having evidence *against* it, simply because the observed effect size was $d < \delta_1/2$.

[3] In this situation, we would argue that a one-sided test with this τ is appropriate. However, the question of one-sided vs. two-sided test is irrelevant to the point we are making; to remain consistent with the reported t test and the HKH-calibrated test, we report a two-sided Bayes factor.

Problems with small τ values

More basic problems with HKH calibration exist when the calibrated τ is small. Consider calibrating a Bayes factor when $N = 36$, against the alternative standardized effect sizes $\delta_1 = .2$ and $\delta_1 = .25$. This sample size and these standardized effect sizes were used by HKH to demonstrate the calibration method.

For $\delta_1 = .2$ and $N = 36$, the only τ that meets HKH's calibration Definition 1 is $\tau = 0$, as shown in Figure 5A. 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 hypothesis H_1 is exactly the same as the null H_0. If the null and alternative hypotheses are identical, *all data* will yield a Bayes factor of 1. In fact, whenever the critical t_0 for the HKH-calibrated Bayes factor is $|t| \leq 1$ the HKH-calibrated Bayes factor always exactly 1. Figure 5B 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.

It is important to emphasize that this lack of a solution is caused solely by the calibration of the scaled-information Bayes factor. If we wanted to use a significance test without using HKH calibration, we could simply find the critical t_0 statistic for the test with equal errors, and use that to make decisions. This would not be a good procedure, due to the fact that when $\delta < \delta_1/2$ the Type II error rate will approach 1, but it is still possible. HKH, however, use the scaled-information Bayes factor to make decisions, so when no non-trivial τ exists the entire procedure fails.

Even if a HKH calibration exists such that $\tau > 0$, the calibrated τ 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 5A). As Figure 5C shows, this τ is so close to zero that the scaled-information Bayes factor for the alternative hypothesis can never exceed about 7, regardless of the size of the t statistic.

Whether one uses a fixed error rate or fixed alternative effect size δ_1 against which to calibrate, the resulting HKH-calibrated Bayes factor has properties that render it inappropriate for drawing inferences. When the Type I error rate is fixed, the amount of evidence yielded by a completely null effect in favor of the null will decrease as sample size increases; when the alternative effect size δ_1 is fixed, the procedure will yield evidence for the null for all effect sizes $d < \delta_1/2$. These behaviors contradict the requirements we would have for any reasonable quantification of evidence, and hence the HKH-calibrated Bayes factor is uninterpretable.

Even if we ignore the Bayes factor and consider the procedure a way to generate significance tests, the significance test implied by Definition 1 is not a significance test anyone would actually use, and the procedure implied by Definition 2 is merely a Student's t test. Furthermore, the technique often fails to provide solutions for the decision solely because of the unnecessary calibration, even in cases considered by HKH. The HKH-calibrated Bayes factor is neither a defensible significance test nor a defensible Bayesian procedure.

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 "default family of priors" 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 "default family of prior distributions" approach differs markedly from the "single default" approach that HKH suggests we advocate.[4] We agree that subjective information is important. We also agree that researchers need assistance in understanding the subjective implications of using a default family of 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, the prior scale has a direct interpretation in terms of the prior probabilities of ranges of standardized effect sizes: in the case of the Bayes factors they describe, $|\delta| > \tau$ has a prior probability of 1/2.[5] 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 standardized 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. One may contrast this to the τ values in Figure 4B.

For instance, 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 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

[4] 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." (p. 9 of the manuscript)

[5] Typically, a t prior with 1 degree of freedom is used instead of the normal prior described here, and the prior scale is called r (de Vries & Morey, 2013; Morey & Rouder, 2011; Rouder et al., 2009). The prior probabilities described here are for the t prior; the prior probabilities will be slightly different for the scaled-information Bayes factor's normal prior.

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 (see also Goldstein, 2006).

As Rouder et al. (2009) pointed out, the default family of priors are useful precisely because they have built-in subjective information: standardized 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 researcher. 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 no alternative at all due to its undesirable properties. 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 peculiar hybrid, which in practice amounts to a significance test and which does not have any good Bayesian properties. We believe that researchers exploring new methods should avoid HKH-calibrated Bayes factors.

References

de Vries, R. M. & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods*, *18*(2), 165–185. Retrieved from http://dx.doi.org/10.1037/a0031037

Edwards, A. (1972). *Likelihood: an account of the statistical concept of likelihood and its application to scientific inference.* London: Cambridge University Press. Retrieved from http://www.ams.org/mathscinet-getitem?mr=348869

Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. *Bayesian Analysis*, *1*, 403–420. Retrieved from http://dx.doi.org/10.1214/06-ba116

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. Retrieved from http://dx.doi.org/10.1111/j.2044-8317.2012.02067.x

Morey, R. D. & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419. Retrieved from http://dx.doi.org/10.1037/a0024377

Morey, R. D. & Rouder, J. N. (2014). Bayesfactor: computation of Bayes factors for common designs. R package version 0.9.9. Retrieved from http://CRAN.R-project.org/package=BayesFactor

Neyman, J. (1957). “Inductive behavior” as a basic concept of philosophy of science. *Review of the International Statistical Institute*, *25*, 7–22. Retrieved from http://dx.doi.org/10.2307/1401671

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. Retrieved from http://dx.doi.org/10.1007/978-1-4612-0919-5_6

Rouder, J. N. & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877–903. Retrieved from http://dx.doi.org/10.1080/00273171.2012.734737

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. Retrieved from http://dx.doi.org/10.1016/j.jmp.2012.08.001

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. Retrieved from http://dx.doi.org/10.3758/PBR.16.2.225

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. Retrieved from http://dx.doi.org/10.1037/a0022790

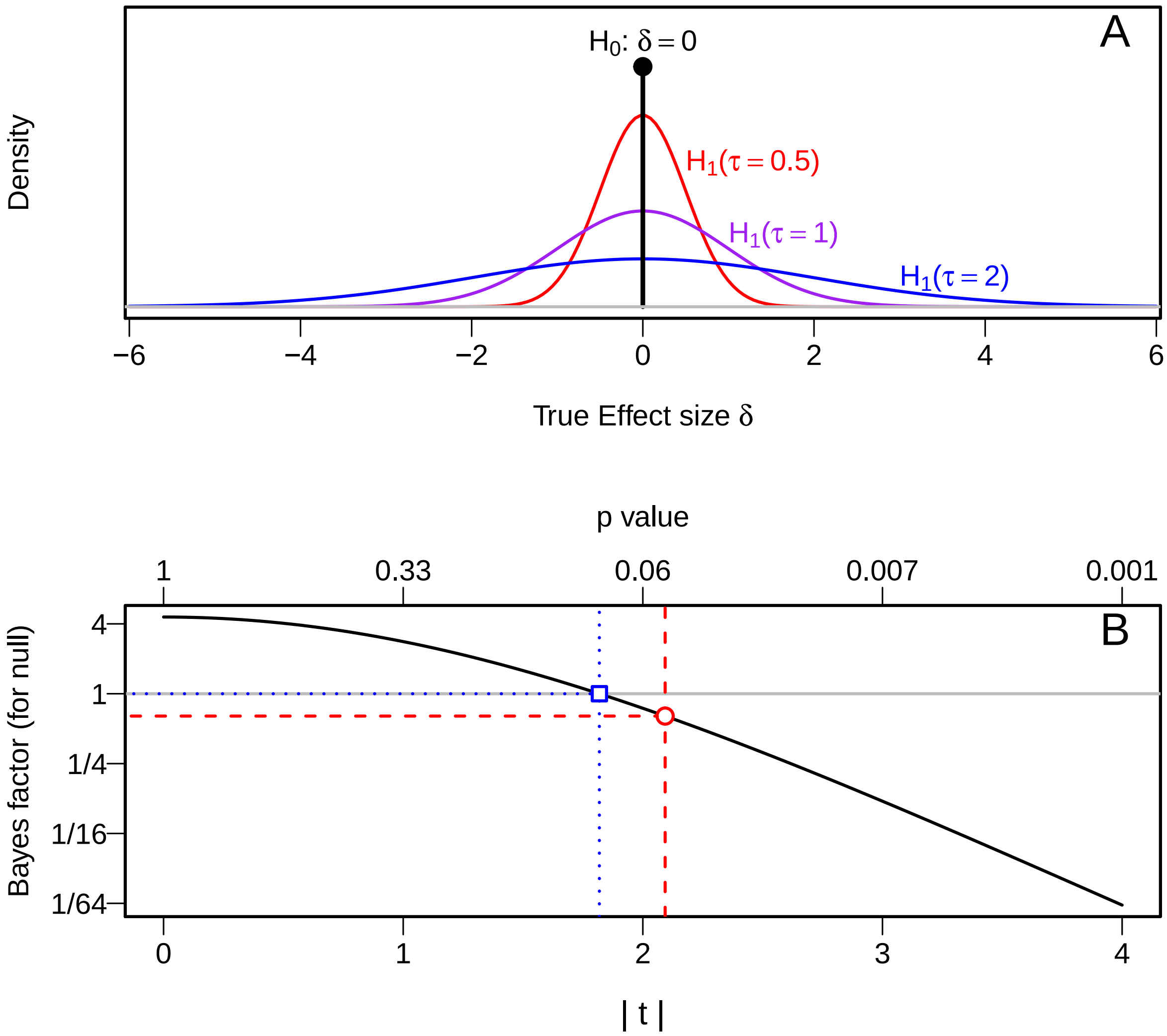

Figure 1. A: Null hypothesis and three different alternative hypotheses for the scaled-information Bayes factor. B: Scaled-information Bayes factor ($\tau = 1$) as a function of $|t|$, for $N = 20$. The dashed lines and open circle show the critical $|t_0|$ statistic and Bayes factor for $p = 0.05$; the dotted lines and open square show the critical $|t_0|$ and Bayes factor for $BF = 1$.

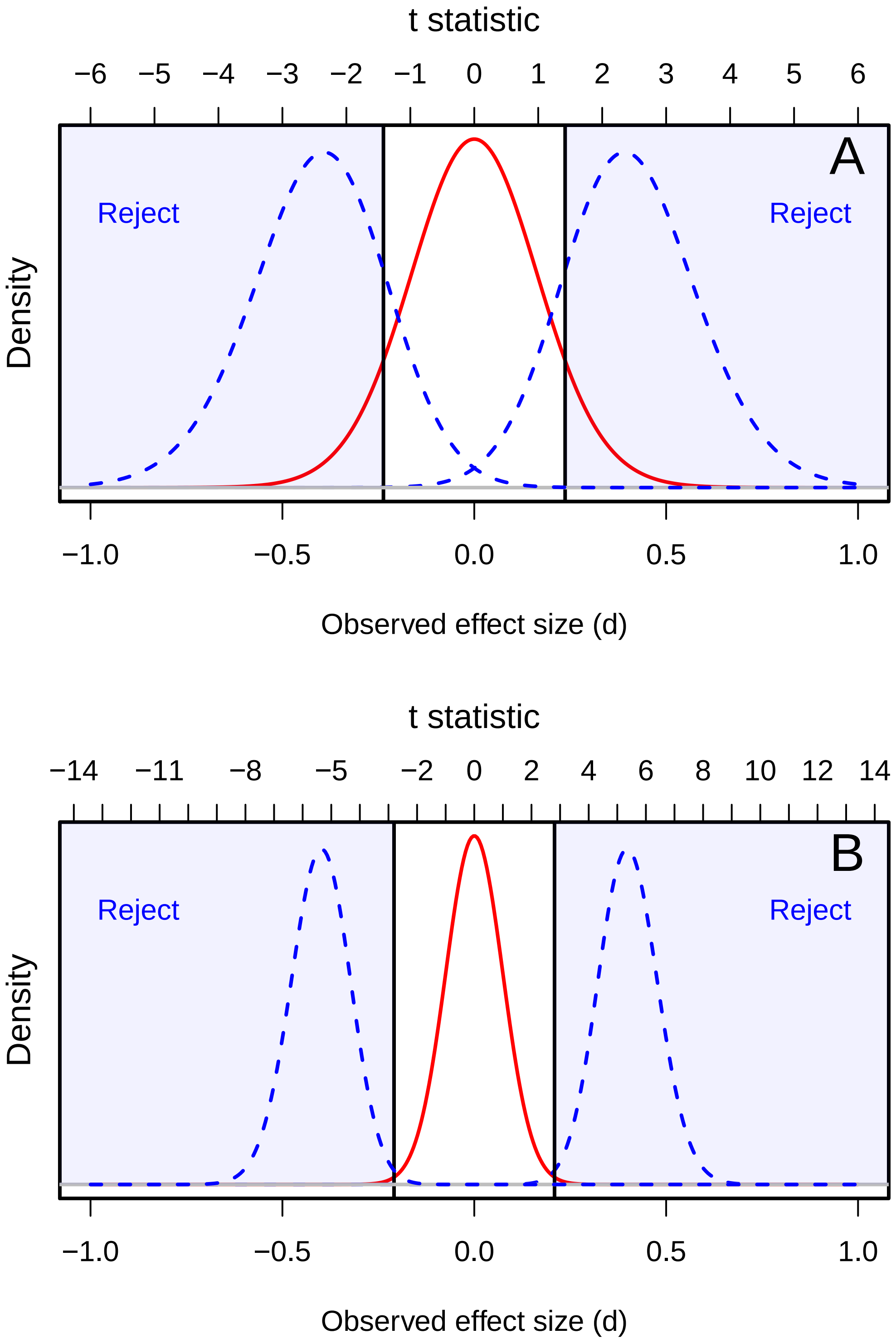

Figure 2. A frequentist two-sided test of $\delta = 0$ vs. $|\delta| = .4$. The solid density is the distribution of d under the null hypothesis; the left and right dashed densities are the distribution of the d under the hypothesis $\delta = -.4$ and $\delta = .4$, respectively. Vertical lines show the critical bounds, beyond which (shaded region) the test will reject the null hypothesis. This particular choice for critical bounds yields balanced Type I and Type II errors. A: $N = 36$; B: $N = 180$.

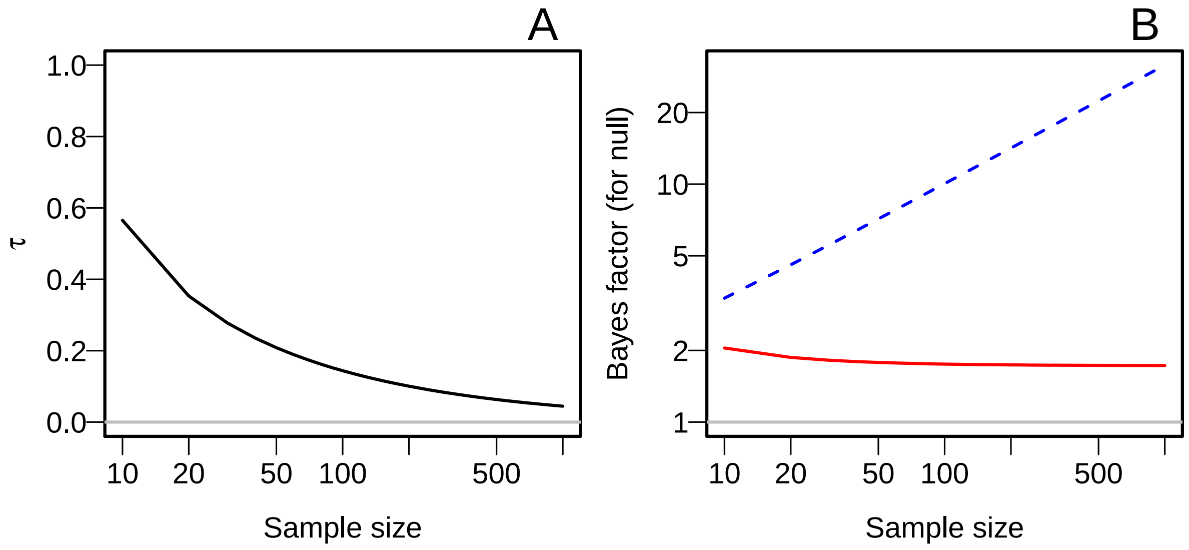

Figure 3. A: The HKH-calibrated prior scale τ as a function of sample size. B: HKH-calibrated (solid line) and scaled-information ($\tau = 1$; dashed line) Bayes factors for an observed standardized effect size of $d = 0$ as a function of sample size. In both plots, HKH calibration is performed against Type I error rate $\alpha = .2$, and holds for both definitions of calibration.

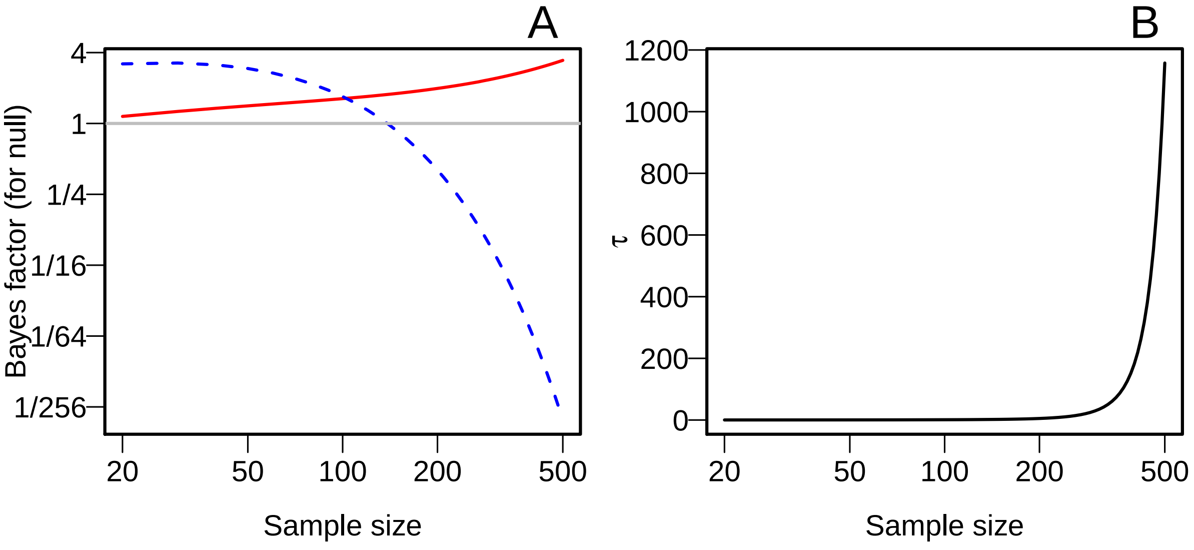

Figure 4. A: HKH-calibrated (solid line) and scaled-information ($\tau = 1$; dashed line) Bayes factors for an observed standardized effect size of $d = .19$ as a function of sample size. B: The HKH-calibrated prior scale τ as a function of sample size. In both plots, HKH-calibration is performed against $\delta_1 = 0.4$.

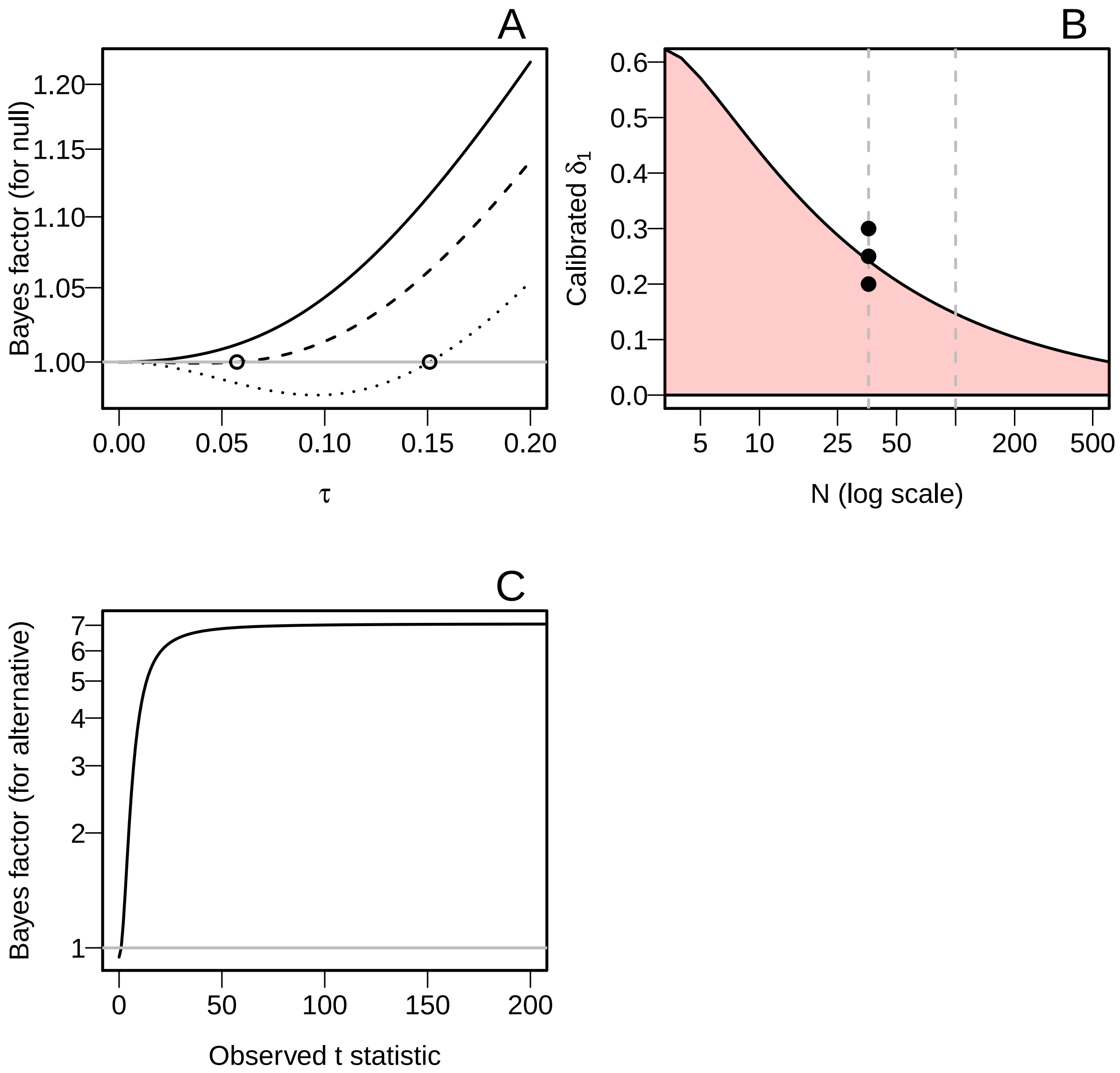

Figure 5. A: Bayes factor as a function of τ for three t statistics (top to bottom): 0.91, 1.03, and 1.15, corresponding to the critical t_0 values for HKH calibration to δ_1 values of .2, .25, and .3. Circles show $\tau > 0$ values where $BF = 1$. For $t = 0.91$, no such value exists. B: Combinations 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 $\delta_1 = .25$.