Dear Dr. Hoijtink,  
  
Thank you very much for submitting your manuscript "A Tutorial on Testing Hypotheses Using the Bayes Factor" for review and consideration for publication in *Psychological Methods*. I sincerely appreciate the opportunity to review the manuscript. I have now received the reviews of your manuscript and am able to make an editorial decision at this time.  
  
We were fortunate to receive reviews from three incredible scholars. They all provided very thoughtful feedback which is included below. Based on that information and my own independent reading and assessment, I am inviting you to revise and resubmit your paper to the journal. While the three reviewers differed in their opinion about how much work remained to be done to make this the most impactful paper it could be, all three reviewers were positive about the manuscript. I concur with Reviewer 1 that this paper could serve a very nice role as an "add-on" to introductory statistics courses taken by many doctoral students. I won't reiterate all the many excellent points raised by the reviewers, but wanted to point out that many are suggestions about improving clarity. I think that is your primary objective in the revision. I also note that Reviewer 3 felt you could do a little more to motivate the Bayesian approach and say a little more about the challenges/limitations. I agree with both sentiments and think they can be reasonably addressed without adding massive amounts of text.  
  
Reviewer comments are insightful and constructive; please address all of them them in a letter of reply that includes their comments, interleaved with your responses and revisions, quoting from the revised text as needed, for easy reference. All three reviewers were positive about this paper and think it can do some real good in the world. Please keep that in mind as you respond to their thoughts and suggestions.  
  
If you have not already done so as part of your Author Note, please provide the details (2-4 sentences) of prior dissemination of the ideas and data appearing in the manuscript (e.g., if some or all of the data and ideas in the manuscript were presented at a conference or meeting posted on a listserv, shared on a website, etc.).  
  
We request that your revision arrive by July 28, 2018. If for some reason you find that you cannot meet this deadline, please contact us as soon as possible in order to make other arrangements.  
  
To submit a revision, go to <https://met.editorialmanager.com/> and log in as an Author. You will see a menu item called Submission Needing Revision. You will find your submission record there.  
  
Sincerely,  
Michael C Edwards, PhD  
Special Action Editor  
*Psychological Methods*  
  
  
Reviewers' comments:  
  
Reviewer #1:  
  
Overall, I found this paper to be effective for its intended goal and intended audience. It is very easy to make discussion of Bayesian topics highly technical without really trying, but this paper was able to avoid this potential pitfall and was able to consistently present the material in a very comprehensible fashion. The writing style should make the content easy to follow even for a reader with very little knowledge of Bayesian statistics. The coverage of topics had a good logical flow and I thought the paper did a good job of starting basic and sequentially building up to more realistic situations. The frequent use of code and R output was also very useful to directly see how a researcher would obtain and interpret the results. One criticism is that the type of models to which the tutorial could be applied is rather limited - mostly ANOVA designs. Nonetheless, for the intended audience, I do not think that this is a huge problem because many psychological researchers heavily rely on ANOVA designs and the paper remains very informative for this group of researchers. I could also see the paper being quite useful as supplementary reading at the end of an introductory course on ANOVA/the general linear model as an introduction to the advantages that Bayesian methods possess over their frequentist counterparts.  
  
In summary, I found that the paper was written at a very appropriate level given its intended audience and I think that it could be useful for researchers who are new to Bayes or who have limited experience with running and interpreting Bayesian models.  
  
Detailed Comments  
  
Page 7, Line 9: Should f\_0 and c\_0 be changed to f\_0Alpha and c\_0Alpha for consistency? Otherwise, it seems like f and c only correspond to the null hypothesis, at least based on my reading. Part of this may have been due to the R output labeling the columns only as "f" and "c", so the "0" subscript made me think that these belonged to the H0 hypothesis.  
  
Page 7, second paragraph: I got a little lost here, I wasn't immediately clear to me what Figure 1 was showing or which results were being discussed.  
  
Page 7, Line 17: For a tutorial on Bayes Factors, it seems that a cursory definition of prior and posterior distribution could be useful to include in the main text instead of outsourcing readers to a rather dense textbook.  
  
Page 7, Line 22: It was not clear to me at this point whether large values of f\_0 were indicative of good fit or bad fit. It also was not clear if "large fit" is good or bad ; this can vary across different types of fit measures in statistical models (e.g., large CFI is good, large RMSEA is bad in an SEM context). It might be easier to use relative terms when talking about fit like "better" or "worse" than "large" or "small".  
  
Page 8, Line 3: Can more explanation be provided about why the prior variance changes the complexity of the hypotheses? It was not immediately clear to me that something the researcher controls like the prior could affect the complexity of otherwise static hypotheses  
  
Page 9, Line 1: The betting analogies were a little confusing to me since the amount of money being bet was not stated. My guess is that it is $1 but stating this in the paper would be helpful.  
  
Page 12, last line: At this point, an unaddressed question is 'how much support is enough support?' Some discussion on this topic comes up later in the paper, but I think it would be useful if it were placed earlier since I think this is a central idea when using Bayes Factors. When I was reading through the paper pretending that this was the first time I had heard of Bayes Factors, this was the one lingering question that I wish were answered earlier.  
  
Page 16, Results 4 box: Can the font of the plot's axes be increased? I had trouble reading the axes even when zoomed in 200%. The same issue also exists in the Results 5 box on Page 18.  
  
Page 21, first paragraph: This section seemed to go a little off point. It started out about the size of Bayes Factors and ended up talking about issues in reproducible science. I agree that reproducibility is important, but everything from the last paragraph on Page 19 to the end of the section seemed unnecessary to the topic of the section.  
  
Typographical/Grammatical Errors  
  
Page 7, Line 6: simples = simplest?  
  
Page 10, last paragraph: morel rebel = moral rebel?  
  
Page 14, last paragraph, second line: prize = price?  
  
Page 17, last paragraph, fourth line: advice = advise?  
  
Page 19, Line 7: "an Bayes" = "a Bayes"?  
  
Page 22, first paragraph, second to last line: extend = extent?  
  
Page 28, Point 3 and 4: the text uses the British spelling of "labelled" at times and the American version "labeled" at other times. Either is fine, but the spelling should be consistent.  
  
Page 31, last paragraph, Line 5: ont he = on the?  
  
Signed,  
Dan McNeish  
  
  
Reviewer #2:  
  
I found the paper to be on a topic worthy of attention, interesting, nicely written, and well positioned to be a useful references for those familiar and unfamiliar with BFs. Below I list some areas where improvements would further strengthen the paper.  
  
1. Some framing would help. The authors refer to complexity (p. 8) as being driven by prior variance. First, most readers will associate complexity with an attempt to count the number of parameters in the model, and find it odd that the prior variance of H\_a could somehow affect the complexity of H\_0. But that's not what is going on here. So some additional framing might avoid confusion. This could take the flavor of saying the figure (and perhaps procedure) looks only at fitting the model expressed by H\_a, and considers H\_0 as being at a particular point in the distribution, so that we are considering how complex H\_0 is with respect to H\_a, not in and of itself.  
  
2. I was momentarily confused by analysis surrounding the "one mean" hypothesis on p. 8, because I had the Monin data described on pp. 5-6 in mind, and wasn't sure if the "one mean" null also referred to the null for the Monin data on p. 6. I would move the description of the Monin data until later, when it is actually used, to avoid such possible confusion.  
  
3. Having the R code in 'BFTutorial.R' is of course important, and quite helpful. As this will also be a vehicle for readers to learn about Bayes factors, readers will likely take this exact code and modify it to their situation. Accordingly, I would encourage the authors to annotate their code even further than they have so that readers know what is specific to their contexts, what can be changed, etc. For example, the code in Step 4 appears to fit the model, extract estimates of some kinds, make things into a table, and so on. Even small comments that declare what's going here (e.g., "Analyze a model corresponding to H\_a") would be helpful. Some of this is contained in the paper in Results 14; it would be useful to have in the BFTutorial.R file as well. What I would hate to have happen is someone just take this code, and because they are not aware of what each line is doing, do something (or change something) that is inappropriate for their situation.  
  
4. On p. 11 the authors state that "BF.c denotes the Bayes factor of a hypothesis against its complement." Then in Step 5 (p. 13) they indicate that, the BF.c gives the value for the hypothesis against H\_a. Does this imply that H\_a is the complement of each of H\_0, H\_a1, H\_a2, and H\_a3? How could that be? There is a discussion later about the complement of hypotheses with inequality as opposed to equality constraints. Some clarity or foreshadowing early would be helpful.  
  
5. The discussion of how the Bayes factors provide evidence for a hypothesis relative to the other hypotheses under consideration (p. 13-14) is quite good, and (still) needed in the field.  
  
6. Tutorial Step 7. The comments in the code says this is for the Holubar data. That should be the Monin data, right? Also, when I ran the code the plot was created, but the legend was shifted a bit to the right and cut off a bit.  
  
7. Some of the discussion regarding not using thresholds for the BF are a little bit tangential to the main focus of the article, but are useful as the intended audience includes researchers learning about BF. I recommend connecting the recommendation for preregistration back to the main focus by saying explicitly that, when preregistering a study with the planned use of BF, the final paper would include the BF value(s), whatever they may be.  
  
8. Quite aside from concerns one might have about a prior based on fraction of the data, it would be quite useful to have a mechanism where users could define their own priors directly. This is more a recommendation for future work, than a revision to the paper.  
  
Minor points  
  
9. P. 13 'prize' should be 'price'  
10. P. 39 'evaluation' should be 'evaluate'  
  
  
Reviewer #3:  
  
Interest in Bayesian methods continues to increase in psychological research. As such, this tutorial paper on Bayes factors is timely and definitely has clear potential to make a worthwhile contribution to the methodological literature.  
  
A major strength of this paper is that it is very clearly written. The authors have intentionally avoided technical details, and as a result the presentation will be quite accessible to the broad readership of Psychological Methods. The manuscript does not claim to present any new methodological developments, but this absence is not a problem for a paper intended to be a tutorial.  
  
Despite being timely and well written, I do not believe the paper in its present form achieves its potential. I will present several thoughts about ways in which I believe the paper could be strengthened.  
  
First, I believe the paper could reach a broader audience if it provided more motivation for prospective readers to learn about Bayes factors. Although the abstract and the introduction are clearly written, they provide almost no preview of why psychological researchers might benefit from learning more about Bayes factors. Although some readers will have heard of Bayes factors and perhaps have some intrinsic interest in learning more about this approach, many other readers will have no intrinsic reason to read a methodological tutorial on Bayes factors. Unfortunately, nothing in the abstract or the first few pages of the body of the paper are likely to attract the attention of such readers. As a consequence, the paper may miss a large proportion of its intended readership. I would encourage the authors to consider how they can revise the abstract and introductory section of the paper to alert prospective readers to reasons (or examples) of why learning about Bayes factors could enhance their research.  
  
Second, ironically the paper also has little to say about possible disadvantages of Bayes factors. Although full discussion of the many different approaches to statistical inference is clearly unrealistic, nevertheless I believe it would be beneficial for the paper to provide a broader context for Bayes factors. For example, many Bayesians prefer interval estimation (often for effect size measures) instead of any form of hypothesis testing. Some acknowledgment of various perspectives, even within the Bayesian framework would strengthen the paper. In a similar vein, I believe that frequentists would raise at least two concerns about reliance on Bayes factors. One concern involves the risk of claiming an effect exists when in reality there is no effect. The paper states on Page 22 that "we should be conservative," but I would maintain that one of the most damaging critiques of Bayes factors is that there is no explicit control of the probability of claiming an effect exists when in reality there is no effect. In other words, if it is essential to be conservative, is it not better to use frequentist methods and explicitly control the Type I error rate at a desired alpha level? I do not expect this paper to resolve this debate, but it seems very important to me that the paper acknowledge this issue. Another concern I believe many frequentists would raise about Bayes factors as opposed to traditional p values relies on an argument that Bayes factors are at least loosely rescaled versions of p values. The manuscript dodges this issue by focusing almost exclusively on interpreting frequentist methods as intrinsically leading to a binary yes-no decision, which I believe is much too narrow. Instead, I would suggest that a more appropriate comparison is between Bayes factors and p values. Even if Bayes factors and p values tend to track one another in comparable situations, I would suggest that the authors can put much more emphasis on the fact that certain types of hypotheses lend themselves to Bayes factors but not to p values.  
  
Third, although the paper is generally very well written, I was disappointed in the paper's approach to the logical and conceptual underpinnings of Bayes factors. Figure 1 and the presentation accompanying this figure are apparently intended to provide readers with a conceptual understanding of the Bayes factor, but unfortunately in my judgment this was the weakest section of the entire paper. I fear that most readers will feel very confused after reading this material. One problem is that it is too abstract. Why not start with an example involving a research question and observed data? More generally, too much is happening in Figure 1 with little or even no explanation. For example, why is the distribution for Ha centered at zero? Why is it symmetric? Why is the variance either 1.0 or 0.5? More generally, exactly what does Ha mean? Is it that the null hypothesis is false or is it that the parameter(s) in question could have any theoretically possible value including zero? In NHST, Ha does not include zero as a possible value; if the same is not true for the Bayes factor, this needs to be emphasized. If it is true, why is the probability density at the null value not zero under the alternative hypothesis? In addition, we are told the value of the sample mean but unless I overlooked it, there is no mention of the sample variance or the sample size, which might lead some readers to infer that they are not relevant. Perhaps the authors' intentions for this example are different from mine, but in any event I do not think this example achieves what I believe is an essential goal of helping readers develop a conceptual understanding of the Bayes factor.  
I also have some other concerns about the paper's conceptual approach. For example, on Page 3 we are told that the version of Bayes factor implemented in Bain is the "approximate adjusted fractional Bayes factor." I fear that readers will have no idea what this means, and will inevitably be left wondering how many other versions of the Bayes factor exist. Is the version presented here the most common version in Bayesian statistics? For example, is the distribution for Ha centered at the null value in all versions? Is the version presented here the same as the JZS Bayes factor? In my experience, the JZS Bayes factor has received by far the most attention from psychologists, which entails quite a bit of discussion of the Cauchy distribution and accompanying tuning. Unless I overlooked it, the current manuscript seems silent on these issues, which at best will be puzzling to readers to as they explore the Bayesian literature. As another example, in several places the paper mentions the "best hypothesis" as opposed to a hypothesis that is true, but what does "best" mean? Even if "best" means best fit given model complexity, I suspect that many readers will say that the goal of science is to identify the model that is the actual population model. The statement on Page 30 that the Bayes factor may select the hypothesis that best describes the data, not the hypothesis that best describes the population from which the data were sampled seems antithetical to the goals of science. As yet another example, I was confused near the bottom of Page 13 why BF0a is still .001 (as it was on Page 12) although Ha itself is different from what it was on Pages 11 and 12. I would have thought that Ha on Page 13 corresponds to a hypothesis that all three population means are different from one another, unlike the statement of Ha at the top of Page 11. Is it just a coincidence that the value of BF0a remains the same (to three decimal places)? And why at the top of Page 11 are we told that Ha means that the three population means may have any combination of values? Is complete equality of the means included in Ha? The definition would seem to imply "yes," but if so Ha is entirely different here than it would be from a frequentist perspective.  
  
Fourth, I struggled in a number of places trying to understand how strict equality of means (or strict equality of parameters in general) fits into the picture. On the one hand, we are told on Page 8 that the entire tutorial assumes that H0 and Ha are equally likely. On the other hand, parts of the paper berate frequentist null hypothesis significance testing for testing a hypothesis that logically can never be true. Why should this be a fundamental flaw of frequentist methods but not a problem for Bayesian methods? On a related point, why should H0 and Ha be equally likely a priori?  
  
Fifth, I have reservations about the role of sample size as it affects the Bayes factor along with updating recommendations. The guidelines for when to update seem very subjective and thus subject to potential researcher "degrees of freedom" that could lead to bias, even with preregistration. I also believe that the suggestion of starting with an initial batch of (at least) 10 participants may be much too low. As best I can tell, there is no acknowledgment that the sample size should depend on the research design and the type of analysis to be conducted. Moreover, for virtually any type of psychological research, an n of 10 (even per cell) is likely to be much too small. Arguably even more important in this type of tutorial is a clear explanation of how sample size affects the value of the Bayes factor.  
  
Sixth, I was pleased to see a discussion of statistical assumptions, but disagree with some of the practical recommendations, such as discarding outliers. Even if rules for identifying outliers are developed a priori, it seems unwise to discard outliers (unless it is very clear they are invalid) instead of using more robust methods that are resistant to outliers.  
  
Seventh, I believe it would be useful for the authors to be clearer about whether the principles of the Bayes factor are applicable to all statistical methods or only a subset of methods independently of current software implementations.  
  
Minor Issues (in chronological order based on page number)  
  
(1) Page 3 (and also Page 35): It would be helpful to clarify the relationship between Bain and JASP.  
(2) Page 4 (and also Pages 35 and 39): It seems odd for an archival journal article to invite readers to contact an author. It clearly does not need to be stated 3 times.  
(3) Page 6, 3 lines from bottom: I believe it would be helpful to add a verbal explanation of how BFa0 relates to BF0a.  
(4) Pages 5-6: It seems premature to mention the Monin data because the study is not mentioned again until Page 10.  
(5) Page 7: Testing that the population mean equals zero could be motivated with an example involving a within-subjects comparison of means.  
(6) Page 20: I would suggest adding John, Loewenstein, and Prelec (2012) as a relevant reference for the prevalence of questionable research practices.  
(7) Page 30: It would be useful to explain why informative hypotheses need to be compatible.  
(8) Page 30: Some readers might infer that the best way to evaluate replicability is to compare values of the Bayes factor. Although this may be of some relevance, it hardly seems like the only way, or even the best way, to evaluate replicability.