**Peer Review Correspondence**

Ms Title: Bayesian frequentists: Examining the paradox between what researchers can conclude versus what they want to conclude from statistical results

Author names: Matthias Haucke

Manuscript submitted April 10, 2020

**Editor First Decision—Revise & Resubmit**

July 30, 2020

Dear Matthias Haucke,

I am writing in regard to your manuscript “Bayesian frequentists: Examining the paradox between what researchers can conclude versus what they want to conclude from statistical results” (MS 1108604), which you submitted for consideration by Collabra: Psychology.

I first wanted to apologize for the extremely long time it has taken to get you a decision on your paper. As you may know, there were some transitions since the time you submitted your paper, both in the editorial team and the editorial system, and this was partially responsible for the delays. In addition, I think that everyone has been a bit overwhelmed with additional responsibilities and challenges during the pandemic, and this also contributed to the delays. However, even taking these factors into account, the time it has taken to process this manuscript was too long, and for that I apologize. The team here at Collabra: Psychology will definitely work hard to make sure that processing time is improved in the future.

In any case, I was able to get two reviews for this manuscript, and as you will see, these reviewers provided comments that were thoughtful and constructive. Both reviewers appreciated the goals of the paper, and both thought that the data and your analyses could make a nice contribution to the literature. At the same time, the reviewers did raise some concerns about the clarity of the manuscript, and each raised some questions about the clarity (and sometimes accuracy) of your interpretation of your analyses and prior work. In addition to these issues, I had some concerns of my own. The issues that the reviewers and I raised are serious enough to prevent me from accepting this version of the paper, but I would be happy to consider a revised version that addresses the issues that the reviewers and I raised.

As you will see, the reviewers would like to see a bit more detail about your procedures and a bit more clarity regarding your discussion and interpretation of these procedures. For instance, Reviewer 1 would like to see the full materials in the supplemental material presented on OSF. In addition, this reviewer thought that some of the discussion of the specific questions that were asked was not clear.

Both reviewers noted a lack of clarity regarding the way you are using subjective and objective information. I agree that there is some confusion here about what information “counts” as subjective versus objective, how this information is actually used, how participants interpret “subjective beliefs” and “prior study results,” and how your questions reflect these three factors. Please see the reviewers’ comments for suggestions on how to address this lack of clarity.

Reviewer 2 also raised a number of concerns about the clarity of the paper. These concerns related to the structure of the paper itself, along with some more specific issues about the clarity of specific statements in the paper and concerns about the accuracy of certain interpretations given the data. I think that all of these issues are worth addressing, and if you believe that you cannot or prefer not to address these in the paper itself, please explain why in your response letter to me.

A critical issue that I had concerns the fit between the goals of the study and the specific tests you conducted. In the first paragraph, you say: “Frequentist techniques are often claimed to be incorrectly interpreted as Bayesian outcomes, which suggests that a Bayesian framework may fit better to inferences researchers frequently want to make (Briggs, 2012). The current study set out to test this proposition.” I think this is a worthy goal, but I’m not sure that the primary test you conducted really addresses this question. Specifically, one of the three main tests you conducted (in addition to assessing false beliefs about frequentist tests and preference for subjective beliefs versus prior study results) was about the discrepancies between what people think that frequentist tests can do and what those same researchers *want* their tests to do. But you never really explain why comparing what people think they are doing to what they want to do is relevant for the issues you describe in the introduction. I completely understand why it is useful to compare what researchers think frequentist tests do to what they actually do (this is the traditional investigation into misunderstandings of these tests), and I can understand why you might want to assess discrepancies between what these tests do and what researchers *want* to do. But I don’t think you clarified what investigating the discrepancy between what researchers think these tests do and what they want to do really adds.

Another way to think about this is to consider the possible outcomes. You emphasize that what researchers want to do does not match what they think they are doing with frequentist tests; but if participants in this study had been worse in their evaluations of the correctness of the six statements (i.e., if they thought that frequentist tests could do more than they really can do), the beliefs and desires would match, but it would still be the case that the frequentist tests aren’t doing what researchers want them to do.

So, to me, the main contribution of the data on desires is just the documentation of these desires and not the comparison with the beliefs. If I am missing something here, I think you need to clarify what the contribution of this comparison is.

I should note that this is also true for the prior beliefs and prior studies questions. You need to do more to explain why the relevant issue is the discrepancy between beliefs and desires rather than the desires themselves.

Minor Issues:

Daniël Lakens has a recent paper in Perspectives on Psychological Science on what he calls the statistician’s fallacy. He makes the case that statisticians often make assertions about what researchers “really want to know” with little evidence about what it is that researchers want to know. It seems that your paper provides some evidence about what researchers want to know. You might want to refer to this paper, as your own work provides somewhat of a response to it.

On p. 2, you state that “Stone (1969) argued that researchers are likely to collect more data after p-value snooping, compared to when they think of p-values as measures of cumulative evidence.” However, I think you need to do more to link this to the sentence that precedes it. Specifically, I think that you need to spell out how a lack of understanding of statistical methods leads to “p-value snooping,” you should explicitly describe what p-value snooping is, and you should explain in more detail what it means to think about p-values as measures of cumulative evidence.

I think it might be useful when you first mention the six statements (on p. 3) to point readers to the table that presents them. As Reviewer 2 noted, it would also be useful to remind readers precisely why each statement is wrong and to introduce earlier precisely how a Bayesian approach addresses the beliefs that are reflected in these statements.

I think it would be helpful to provide (probably on your OSF page) the exact wording of the invitation email that was sent to prospective participants (see Reviewer 1’s comment about this issue, too). As you note, the sample in this study appear to have a better understanding of these issues than participants in prior studies, and you argue that this was not due to selection based on content mentioned in the invitation. However, it would be helpful to make this invitation available for others to evaluate.

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained.

If you have any questions or difficulties during this process, please contact the editorial office at [editorialoffice@collabra.org](mailto:editorialoffice@collabra.org).

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Richard Lucas

**Reviewer 1**

**Open response questions**

Your name will not be revealed unless you wish to sign your review. If you chose to show your name, please type it in the text field below. If you prefer your comments to be anonymous, leave the field blank.

**Review of 1108604-bayesian-frequentist**

**Summary:** The paper “Bayesian frequentists: Examining the paradox between what researchers can conclude versus what they want to conclude from statistical results” offers an empirical investigation of the claim that there exists “a discrepancy between what researchers think they can conclude and what they want to be able to conclude” (p.1). Such a proposed discrepancy has typically emerged out of literatures on the theoretical and practical shortcomings of the frequentist approach to statistical inference, including widespread rates of p-value misconceptions. As a result, the Bayesian alternative to statistical inference has been commonly proposed as a solution to address this observed discrepancy, as its framework is believed to theoretically align with the inference goals that researchers are presumed to hold. The current paper’s method included directly surveying researchers’ confidence and interest ratings (respectively, “C” and “W” scores) for statements about a statistically significant result, in order to provide a direct comparison between what one believes they can conclude versus what they want to conclude from a statistical finding (task 1). Moreover, the authors also surveyed the extent to which researchers typically do (vs. wish to) account for prior beliefs and/or prior study findings when drawing statistical inferences (task 2).

**Overall impression:** Overall, I felt the authors had a clear set of research goals which are pertinent and valuable to be investigated; that said, some lack of clarity about the survey details left me questioning whether their (pre-registered and carried-out) approach was optimally suited to tackle these goals (see **major comment 1** below); related, I am confused on how ‘subjective’ vs. ‘objective’ priors are being conceptualized, and/or whether it is fair (or the authors’ intentions) to operationalize them, respectively, as beliefs vs. prior study outcomes (**major comment 2**). Finally, I think the paper (beyond presenting empirical findings) served in part as a nice tool (or short tutorial) to educate readers on the Bayesian perspective (which in my estimation aligns with the authors’ overall goals), however I feel that some key pieces of information could be added to provide a more comprehensive explanation (**major comment 3**).

**Major comments:**

**1)** I am specifically confused about the details of task 2: On the one hand, it is clear that participants are presented a specific study scenario, namely the ‘drug against depression’ study, thus I am left to assume that the follow-up survey questions (i.e. Can Prior Study… etc. scores) are directly related to this specific study scenario. This seems to be reaffirmed when the authors state “In a second task, participants were presented with a hypothetical scenario […] whether they would like to be able to incorporate previous findings, and their subjective beliefs into the analyses of the study results” (p. 17). On the other hand, it seems that everywhere else in the manuscript, these task 2 questions are described as probing participants about what they actually (or “typically”) do (in their own research?), e.g.: “we examined to what extent researchers think they typically incorporate…” (p. 6); “Afterwards, participants were asked to indicate to what extent they typically take into account…” (p. 9); “Our analysis resulted in two Bayes factors, each of which quantifies whether there is a discrepancy between what respondents typically take into account and what they would like to take into account” (p. 12); “…a difference in the extent to which people take into account prior study results and the extent they wish to do so” (p. 14).

My impression is that probing individuals on what they actually do vs. wish to do would be more relevant to the authors’ goals, insofar as assessing whether the majority of researchers’ current approach to statistical inference already aligns more readily with a Bayesian vs. frequentist approach. Therefore, I’m not certain that asking individuals on how they *would* act in this hypothetical scenario is an optimal measure, unless it was meant to act as a proxy for what researchers’ typically do? In this case, I would communicate this explicitly; that said, the authors’ themselves discuss that the clinical nature of the task may have influenced the respondents’ answers, thus begging the question as to whether this would be a suitable proxy.

Perhaps a bit nit-picky, but in task 2, when asking respondents about incorporating beliefs and prior study outcomes, sometimes the authors talk about incorporating these “in a statistical analysis” (p. 9), “into the analyses of the study results” (p. 17), or “into the prior” (p. 11), whereas other times they talk about incorporating them “in/into their statistical inference” (p. 1 & 17). To me, the process of setting up the analysis constraints (e.g., prior) versus the process of drawing inferences from the analysis outcomes are not the same thing. The former may involve, for instance, specifying a priori a range of plausible values for a parameter (which may strike a frequentist as biasing the outcome), whereas the latter may involve simply interpreting a given outcome in light of prior beliefs/knowledge (and in turn calibrating one’s belief or skepticism accordingly a posteriori). To me, only this latter case would be consistent with both a frequentist’s and Bayesian’s perspective of cumulative evidence; thus, I would speculate that how the question was phrased may importantly influence respondents’ answers.

**Suggestion major comment 1:** Please provide somewhere the actual full task details as they were presented verbatim to the respondents (e.g., listing / quoting the full survey questions within the body of the manuscript itself, or provide the full survey within osf as supplementary materials). Also, edit the sections of the text accordingly such that the listed confusions about task 2 are clear to a reader (even without accessing say the online supplementary materials).

**2)** The distinction between what constitutes a ‘subjective’ vs. ‘objective’ prior I find is unclear or inconsistent throughout the text. The authors introduce the concept of priors on p.6 citing the Wagenmakers (2007) article, in which ‘subjective’ priors are characterized as a “state of knowledge” which “quantifies a personal degree of belief”, and ‘objective’ priors as “predefined rules” which are “independent of the person who performs the analysis” and “are generally vague or uninformative” (p. 791). According to this perspective, I would assume that prior study results would contribute to a ‘subjective’ rather than ‘objective’ prior. In fact, this seems to be consistent with the statement in the current article: “subjective beliefs may well be informed by knowledge about previous study results” (p. 17). Given this perspective, I would *not* expect that “researchers [would] feel uncomfortable with using a subjective Bayesian approach” (p. 6); what I *would* expect is researchers being uncomfortable with introducing subjective bias into their analyses or interpretations of results, but here, the meaning of ‘subjective’ does not encompass the same connotations as it does from a Bayesian point of view.

I believe the crux of my confusion stems from the fact that the authors seem to be using ‘subjective’ and ‘objective’ sometimes from a Bayesian perspective, and other times from a formal language perspective. From a language perspective, it seems acceptable to dichotomize beliefs as ‘subjective’ versus study results as ‘objective’ (e.g., see p. 6-7 where authors repeat the phrase “subjective beliefs and objective prior study results”). However, this usage of the terms shouldn’t hold in the Bayesian sense, if and when (according to my understanding) both beliefs and prior study outcomes can contribute to the formation of a ‘subjective’ prior. This confusion, or misleading dichotomy, is further reinforced when authors explain that researchers’ desire to incorporate prior study results, but reluctance to incorporate subjective beliefs, “suggests that they might be more comfortable with a rather objective Bayesian approach” (p. 17)… How do these results reflect the objective Bayesian approach which (again, referring back to p. 6) was described as predefined, and comparatively uninformative, such as a default prior distribution? Moreover, later on p. 18, the authors liken a “prior distribution [that] is well founded by previous studies” to “the Bayesian notion of subjectivity”. Finally, their concluding remarks state that “researchers sympathize with some elements native to the philosophy of objective Bayesianism” and “a subjective Bayesian approach is deemed more acceptable if a prior distribution is built upon previous research” (p. 20)… At this stage, I’m not certain what is meant by ‘objective’ vs. ‘subjective’, especially since researchers’ desire to incorporate prior study findings was interpreted as both support for the objective Bayesian approach (quote p. 17) but also as aligning with the subjective Bayesian approach (quotes p.18 & 20).

**Suggestion major comment 2:** Overall, the authors need to more clearly delineate what is meant by ‘subjective’ vs. ‘objective’ in a Bayesian sense, and I think some key statements, such as “subjective beliefs may well be informed by knowledge about previous study results”, are worth mentioning up front when introducing the concept of priors. Currently, the manuscript reads as if the authors’ intentions (especially in the methods) were to (almost mutually exclusively) map beliefs onto subjective priors, and prior study outcomes onto objective priors; but come the discussion these concepts become muddled or intertwined. I think it would also benefit a reader to know up front in the methods that “nothing in the invitation email nor in the survey mentioned Bayesian statistics” (p. 18) such that *should* the survey have included use of the terms (‘subjective’ or ‘objective’), respondents would not be expected to interpret them in a Bayesian sense. Here, again, a reader would profit from seeing the survey questions verbatim.

**3)** When flushing out the A Bayesian Perspective section, I think the explanation of priors, posteriors, and BFs, would benefit from an equation (same as Eq. 7; Wagenmakers, 2007):

P(H1|D) / P(H0|D) = P(D|H1) / P(D|H0) x P(H1) / P(H0)

posterior odds = BF10 x prior odds

Especially for a novice reader, the following statement might be ambiguous: “A Bayes factor (BF) expresses how much our observed data shifts the balance of evidence from one hypothesis (e.g., the null hypothesis H0) to another (e.g., the alternative hypothesis Ha; for more details see Dienes, 2011)” (p. 5). My assumption is that this was meant to express how a BF can be taken to represent / quantify the shift from prior to posterior odds (a link which I find is made more explicit in the equation above; note: in the referenced Dienes (2011) article, this equation is only contained within the article notes). I also think the equation helps clarify the “BF10” subscript.

Within the article, the threshold for strong evidence (i.e. BF ≥ 10) is specified; what is missing is a specified range for inconclusive evidence (e.g., 1/3 ≤ BF ≤ 3). Without this additional information, statements such as “Bayes factors pointing towards the null hypothesis are indicative of no discrepancy,..” (p. 12) may suggest that *any* BF value less than 1 should constitute evidence for the null (which in and of itself would potentially promote erroneous inferences); it also would help a novice reader understand why the outcome in Table 4 (BF = 2.58) is interpreted as “almost no evidence for a difference” (p. 14).

Finally, sequential testing is listed among the advantages of using Bayes factors. As it can also be applied within the frequentist framework, it seems misleading to list specifically as an advantage to BFs. Moreover, in the Wagenmakers et al. (2018) cited reference, not only do the authors acknowledge that “the standard framework of p value NHST can be adjusted so that it can accommodate sequential testing”, but they also state that “From a Bayesian perspective, however, corrections for sequential monitoring are an anathema” (in other words, irrelevant). Perhaps it is worth clarifying that the advantage being expressed here is more accurately sequential testing without need for corrections or without the same concerns of inflating type-1 errors in the frequentist sense?

**Minor comments:**

**4) Concerning The misinterpretation of frequentist statistics section:** I would add in “statistically significant” (on p. 2) when providing the examples for an alternative vs. null hypothesis, specifically: “there is a statistically significant mean difference between…” and “there is no statistically significant mean difference between…”; I feel that describing the null hypothesis as indicating “no mean difference” could imply that evidence for the null should be inferred as evidence of no effect (which in and of itself is a misinterpretation of a non-significant p-value; see e.g., Hoekstra, Finch, Kiers, & Johnson, 2006).

The most common misconception in the Haller & Krauss (2002) study is said to be endorsed by “80% of statistics teachers as well as students” (p. 3); this is inconsistent with Table 1 of the Haller & Krauss (2002) article (p. 8) which lists 73% of methodology instructors and 68% of psychology students as falsely endorsing Statement 5.

While false endorsement of Statement 5 could be indirectly attributed to the misconception that you can assign a probability to a hypothesis (specifically if one failed to recognize that “the wrong decision” cannot be inferred from a p-value), I would argue that this item was primarily meant to pick up on the misconception that a single p-value tells one something about the Type-1 error rate. In contrast, Statements 1 through 4 are directly relevant to the core misconception highlighted within the current manuscript, namely the section beginning with “Thus, a majority of statistically educated researchers misinterpret frequentist statistics as the probabilities of one’s hypotheses being true” to the end of that paragraph (p. 4). I’m curious why the authors would isolate Statement 5, rather than appeal to the other direct measures of this misconception.. Moreover, as Statement 5 misconception rates in Haller & Krauss’s (2002) study were both the highest (67 - 73% incorrect) and the only specific values referenced in the text, one might be more inclined to accept the authors’ interpretation that “In comparison to the results of Krauss and Haller (2002), the performance of correctly interpreting p-values appears to have increased. This indicates that people have improved on their correct interpretation of t-statistics compared to the beginning of this century” (p. 13). Would these statements still hold for Statement 1 (10 – 34% incorrect), Statement 2 (17 – 32% incorrect), Statement 3 (10 – 20% incorrect), etc.? Arguably, these misconception rates would suggest that even at the beginning of this century, the vast majority of respondents correctly identified these as erroneous interpretations of a p-value.

Finally, given that this section is *not* aiming to outline in detail different types of p-value misconceptions, yet nevertheless serves to highlight some of the drawbacks of the frequentist approach to statistical inference, it may be worth including a reference or two (beyond the Haller & Krauss, 2002, text) on where a reader might access more information on specific types of p-value misconceptions to avoid, e.g., Badenes-Ribera et al. (2015), which also highlights the fallacy of conflating statistical significance with effect size or clinical/practical meaning.

**5) Concerning the Materials and Procedure section:** Table 1 typos: Please check statements 1 and 3; looks like the word “means” is missing after “..between the population”. In the Note, the statement “has been repeated by several authors” should be supported by reference(s).

Unless I misunderstand, although the intended sample to be collected was closer to 600 participants (i.e. 30% of 2,000), but the final sample collected was 117 (i.e. 9% of 1,282), in terms of drawing meaningful conclusions from the data, a sample of 117 should nevertheless allow one sufficient power to find strong support for H1 (assuming a moderate effect size), whereas a larger N falls would be required to find either strong support for H1 (assuming a small ES) or strong support for H0, correct? I’m referring the pre-registered sample sizes (i.e. assuming moderate ES, d = 0.5, required N = 31; assuming small ES, d = 0.2, required N = 168; assuming d = 0.0, required N = 132). If so, I think it would be worth mentioning in the Participants section, and/or in the Discussion, as it should indicate to a reader that the BF = 2.58 should specifically be interpreted cautiously (whereas the rest of the effects which are much greater than BF = 10 would have been sufficiently powered to be detected). Regarding academic position, 55% + 18% + 12% + 12% + 4% = 101%.

**5) Concerning the Results section:** Histograms: Is the y-axis frequency expressed in terms of raw numbers? Or a percentage? In terms of relating to past findings (i.e. Haller & Krauss, 2002 rates), percentages would be more understandable. X-axis should include “1” and “9” labels. As “C” and “W” scores are being directly compared, I would also ensure that the range of y-axis values match for Figures 1 and 2.

I am still not wholly convinced that the data supports the following interpretation: “In comparison to the results of […] appears to have increased” (p. 13). For instance, looking at Figure 1 Scores C1 frequencies (assuming raw values), I’d say scores 1 to 4 total roughly 80 respondents, a score of 5 (i.e. neutral?) approx. 10 respondents, and scores 6 to 9 the remaining cases, approx. 27 individuals. This should translate loosely to 68% of the 117 respondents who were some level of confident that this item was false, vs. 23% of the sample with some level of confidence that this item was true. Does this not loosely compare to the Haller & Krauss (2002) rates which observed 10 – 34% of respondents falsely endorsing the item as true? My sense is that the extremely skewed nature of the distributions seems to impart this comparatively greater awareness of about what constitutes an erroneous vs. accurate p-value interpretation, but I’m not yet convinced that this interpretation can be concluded from the data. Perhaps the rationale behind the authors’ interpretation needs to be more strongly argued or the interpretation should not be generalized across all items.

I’m confused about the following line: “In line with our hypothesis we can see that the means for Wish statements are generally higher than for Can statements” (p. 13). When was this hypothesis stated? The only explicit hypothesis I found stated in the text was on p. 6: “We hypothesize that researchers feel uncomfortable with using a subjective Bayesian approach because…” which clearly relates to task 2. Yet, the statement above was embedded within a discussion on task 1 results. If there are explicit hypotheses regarding task 1 and 2, these should be stated more explicitly in the intro, and accordingly followed up in the results / discussion.

Authors make reference to the score means in the text, e.g. “Inspection of the means indicate that scores…” (p. 14). It would help to have a table with the means or indicated below each plot. Additionally, listing the median scores would be relevant given the skewed nature of the data.

**5) Concerning the Discussion section:** On p. 16 – 17, concerning the explanation on why “Statement 6 is not applicable to a Bayesian framework,” I think it is interesting and informative to include the explanation on how – given a great number of experiments – there is a link between the size of the BF and the expected proportion of BFs one should expect to represent the correct model (put simply). That said, what I think is critically missing is the idea that (like the frequentist approach), no single p-value or BF should tell us anything about probability of replicability in and of itself, correct?

Finally, I’m surprised respondents were not probed for whether they employed frequentists vs. Bayesian techniques in their own research. While I expect that a random sample of psychology researchers should produce a group of respondents who primarily apply frequentist inference, and thus should still be appropriate to address the authors’ goals, I still think it would be fair to also recommend that future research probe for respondents’ typical analysis approach and/or familiarity with each framework (frequentist and Bayesian), in the same way that the authors’ recommended that future studies obtain demographics on individuals’ area of expertise (p. 19).

**Other very minor comments:** Attached alongside the review is a commented .pdf of the original submission. I found it was easier this way to flag some minor typos and mere suggestions for edits.

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  | ✔ |  |  |

**Reviewer 2**

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

In the manuscript *“Bayesian frequentists: Examining the paradox between what researchers can conclude versus what they want to conclude from statistical results”* the authors assessed whether researchers **1.** still showed common misconceptions about frequentist NHST, **2.** wished to make statements about their statistical analysis that are incompatible with frequentist NHST, and **3.** whether researchers would be willing to pay the price of constructing subjective priors that is needed to conduct Bayesian hypothesis testing. They conclude that the answers to these questions are 1. No, 2. Yes, and 3. Maybe. I find the study potentially quite interesting, and the results completely plausible. I also appreciate that the study was preregistered, the data are open, and the analysis script was made available. Overall, I think the paper would benefit from a bit of work both in structuring the manuscript as well as motivating the study and the conclusions. Here are some comments:

* Structure of the manuscript: The introduction is about evidence for misconceptions of frequentist NHST and then discussed the basis of Bayesian hypothesis testing. The authors discuss what they call *task 1* and *task 2* together throughout the manuscript (in the methods, result, and discussion sections). I was at times somewhat lost while reading the manuscript, but it took me a while to understand why. I think I could not quite make the connection between these parts. Task 1 is an attempt to assess common misconceptions of frequentist NHST and probably also why these misconceptions exist (because we want to make these kinds of judgments). Task 2 deals with the question why researchers may not use Bayesian hypothesis testing. The connection - that many of the misconceptions are actually legitimate statements when using Bayesian inference, is only made in the discussion section, but not earlier in the manuscript. There are two potential ways of strengthening the narrative of the paper: 1. by introducing this connection in the introduction, and including more interconnections between the discussion of Bayesian and frequentist inference throughout the manuscript; or 2. by splitting the study reporting into task 1 with discussion, then introducing Bayesian inference, and then reporting and discussion of task 2.
* P. 7, first sentence: “… to what extent they want to incorporate both subjective beliefs and objective prior study results” I would note that it is questionable that prior study results are in any way more objective. I would say they are not - neither in terms of scientific decisions that are made (how to combine these results, but also the methods chosen for the prior studies), nor in the statistical sense (they tend to be more narrow than priors based on subjective belief). I think the authors actually agree with me from what I read in the discussion section. I would be careful with the language here.
* P. 8, Table 1. I was missing some sort of discussion of these statements (except maybe there is a bit of it in the discussion section). Why are these misconceptions, why are they so prevalent, can we cluster them into groups of misconceptions, etc. The best place for this would probably be the intro. I feel readers must have read some of the previous papers on the topic to understand what is done and what the problem is. I also think the statements themselves are only meaningful if it is discussed what the statements attempt to express. For example, statement 6 seems relatively useless when one wants to do inference for the data at hand, but if we interpret it as a proxy for robustness and replicability then of course it can be meaningful.
* P. 9, “on a scale form 1 (not all at interested) to 9 (very interested).” *form* should be *from*
* Figures 1 and 2: If data are so skewed, then it might be better to use the median here? 😃 Also, I think the information from the two Figures could be shown more concisely and informatively by plotting each person’s truthfulness and wanting rating in the same plot. For example, this can be done with violin or boxplots and lines (like <https://github.com/jorvlan/open-visualizations>). I feel like proposing this is a bit out of line for a reviewer, but I think useful to see whether there are any participants who do not want to use these statements but still think they are appropriate (i.e. who go opposite to the overall trend).
* P.12: Why use a two-sided test? I suspect the hypothesis was always that researchers would wish to make these kinds of statements.
* P. 12: The authors write “According to the classification scheme suggested by Jeffreys (1998), we considered a Bayes factor of 10 (in favor or against the alternative hypothesis) as strong evidence.” It is odd that they use the cutoff given their previous statement “The BF is directly interpretable in relation to one’s hypothesis.” If the BF is immediately interpretable, then why use a cutoff. Also, given the vast evidence for most of their tests I am not sure this specific cutoff makes any sense in the context of the study and they do not refer to it afterwards. So, maybe just take it out? Just as a side note, I did not understand what the critical comparison for the t-test was until the very end of the paragraph. Maybe it makes sense to lead with this information.
* P. 13: " This indicates that people have improved on their correct interpretation of t-statistics compared to the beginning of this century." I wonder whether that is true without limitations. I guess researchers have seen a considerable number of papers on the issue in the recent years. They also know that psych methods researcher want to trick them in their surveys and are quite suspicious. Also, maybe saying “do you think this is correct” and “would you wish to make such a statement” in the same questionnaire is a bit leading.
* P. 15: “However, the results show contradicting evidence for the notion that a Bayesian framework better describes how researchers want to use inferential statistics.” This notion is transported through the entire discussion section, that researchers want to interpret results as if they used Bayesian analysis but do not want to incorporate their subjective belief. I do not think the data support this claim. Participants in the study *did not* say they would not like to use subjective belief. Actually, Fig 3 bottom right is all over the place. Some people seem to wan to incorporate their beliefs, others don’t. The BF is in the direction of the alternative hypothesis, granted it is a small BF, but > 1. So, the conclusion “researchers seems to like being able to incorporate prior results but not personal beliefs into statistical inference” does not seem appropriate given the data.
* P. 16: “Bayes factors enable drawing the conclusions from […] statement 5 (we can find the probability of making a wrong decision, when rejecting the null hypothesis, given a prior belief).” But we do not make decisions in Bayesian inference based on Bayes factor. We just report evidence. The proposed Bayesian inference in the manuscript does not even mention decision making. Therefore, statement 5 does not apply to Bayesian inference as it is used in psychology.
* P. 16: “then they would hit the Bayes factor criterion associated with the correct model 10 times more often than the Bayes factor associated with the incorrect model”. Is that so? I only briefly read Tendeiro, Kiers, & van Ravenzwaaij (2019), but I am quite certain that such a statement does not appear in Rouder (2014). The analysis in that paper is not concerned with correct or incorrect models.
* I think this study is a missed opportunity for some qualitative assessment. Given that the target sample is experts who are asked about analyses related to their work, I think it could be very useful to give them some room for explanations and qualifications of their responses. This is especially true for task 2. Under which circumstances and what information would they like to incorporate? There seem to be huge individual differences - what are the different rationales? The expertise and education of these researchers is probably quite variable, so this information could help interpreting the data patterns. Additionally, it would be interesting to ask researchers if, given the opportunity to use a statistical approach that allows them to make statements like statements 2 and 4 in the questionnaire, they would be willing to pay the price of setting priors.

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  | ✔ |  |  |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  | ✔ |  |

**Author Response to Reviewers**

October 29, 2020

Dear Dr. Lucas,

We would like to submit the revised manuscript “Bayesian frequentists: Examining the paradox between what researchers can conclude versus what they want to conclude from statistical results” (MS 1108604), for publication in Collabra: Psychology. Below is a detailed description of the ways in which we have incorporated the suggestions of yourself and the reviewers. We reproduced the relevant sections of the decision letter in standard font and added our response in bold. We added a document with all tracked changes in comparison to the original submission. Moreover, major changes in the manuscript are highlighted in blue. We hope that this revision is satisfactory. We look forward to your comments.

Kind regards,

Matthias Haucke

Jonas Miosga

Rink Hoekstra

Don van Ravenzwaaij

**Editor**

As you will see, the reviewers would like to see a bit more detail about your procedures and a bit more clarity regarding your discussion and interpretation of these procedures. For instance, Reviewer 1 would like to see the full materials in the supplemental material presented on OSF.

**E1: We have made all the materials available on https://osf.io/r75qd/.**

In addition, this reviewer thought that some of the discussion of the specific questions that were asked was not clear.

**E2: We have clarified for each of the statements why they are false in the main manuscript in a footnote on p. 3.**

Both reviewers noted a lack of clarity regarding the way you are using subjective and objective information. I agree that there is some confusion here about what information "counts" as subjective versus objective, how this information is actually used, how participants interpret "subjective beliefs" and "prior study results," and how your questions reflect these three factors. Please see the reviewers' comments for suggestions on how to address this lack of clarity.

**E3: We have replaced ‘objective’ and ‘subjective’ with ‘prior study results’ and ‘your own beliefs’ in the manuscript, as that is directly in line with our study materials (see** [**https://osf.io/r75qd/**](https://osf.io/r75qd/) **for materials).**

Reviewer 2 also raised a number of concerns about the clarity of the paper. These concerns related to the structure of the paper itself, along with some more specific issues about the clarity of specific statements in the paper and concerns about the accuracy of certain interpretations given the data.

**E4: We have rewritten the introduction in order to better signpost the goals of our paper, the rationale of our two studies, and the way the two studies (in our opinion) flow logically from one another.**

A critical issue that I had concerns the fit between the goals of the study and the specific tests you conducted. In the first paragraph, you say: "Frequentist techniques are often claimed to be incorrectly interpreted as Bayesian outcomes, which suggests that a Bayesian framework may fit better to inferences researchers frequently want to make (Briggs, 2012). The current study set out to test this proposition." I think this is a worthy goal, but I'm not sure that the primary test you conducted really addresses this question. Specifically, one of the three main tests you conducted (in addition to assessing false beliefs about frequentist tests and preference for subjective beliefs versus prior study results) was about the discrepancies between what people think that frequentist tests can do and what those same researchers want their tests to do. But you never really explain why comparing what people think they are doing to what they want to do is relevant for the issues you describe in the introduction. I completely understand why it is useful to compare what researchers think frequentist tests do to what they actually do (this is the traditional investigation into misunderstandings of these tests), and I can understand why you might want to assess discrepancies between what these tests do and what researchers want to do. But I don't think you clarified what investigating the discrepancy between what researchers think these tests do and what they want to do really adds.

Another way to think about this is to consider the possible outcomes. You emphasize that what researchers want to do does not match what they think they are doing with frequentist tests; but if participants in this study had been worse in their evaluations of the correctness of the six statements (i.e., if they thought that frequentist tests could do more than they really can do), the beliefs and desires would match, but it would still be the case that the frequentist tests aren't doing what researchers want them to do.

So, to me, the main contribution of the data on desires is just the documentation of these desires and not the comparison with the beliefs. If I am missing something here, I think you need to clarify what the contribution of this comparison is.

I should note that this is also true for the prior beliefs and prior studies questions. You need to do more to explain why the relevant issue is the discrepancy between beliefs and desires rather than the desires themselves.

**E5: We agree that demonstrating a discrepancy between what traditional tests do and what researchers want them to do shows that the standard tools for statistical inference do not (completely) match the researchers’ needs. We have added the following section on pages 6/7:**

*Demonstrating a discrepancy between what traditional tests do and what researchers want them to do shows that the standard tools for statistical inference do not (completely) match the researchers’ needs. Our study goes one step further and attempts to demonstrate not only the mismatch between what traditional tests do and what researchers want them to do, but also researchers’ awareness of this mismatch. We expect that researchers will score higher on the items indicating they want to be able to draw the conclusions in the six statements than on the items indicating they can draw the conclusions in the six statements.*

**In essence, we wanted to show that researchers (to an extent) actively prefer some of the features that Bayesian inference has to offer, rather than inadvertently looking for Bayesian tools without being aware of this.**

Minor Issues:

Daniël Lakens has a recent paper in Perspectives on Psychological Science on what he calls the statistician's fallacy. He makes the case that statisticians often make assertions about what researchers "really want to know" with little evidence about what it is that researchers want to know. It seems that your paper provides some evidence about what researchers want to know. You might want to refer to this paper, as your own work provides somewhat of a response to it.

**E6: We now cite Lakens’ paper in the introduction when we provide justification of our study (p. 6).**

On p. 2, you state that "Stone (1969) argued that researchers are likely to collect more data after p-value snooping, compared to when they think of p-values as measures of cumulative evidence." However, I think you need to do more to link this to the sentence that precedes it. Specifically, I think that you need to spell out how a lack of understanding of statistical methods leads to "p-value snooping," you should explicitly describe what p-value snooping is, and you should explain in more detail what it means to think about p-values as measures of cumulative evidence.

**E8: We agree that the connection was not very clear. In our extensive rewriting of the introduction, we have removed this paragraph.**

I think it might be useful when you first mention the six statements (on p. 3) to point readers to the table that presents them. As Reviewer 2 noted, it would also be useful to remind readers precisely why each statement is wrong and to introduce earlier precisely how a Bayesian approach addresses the beliefs that are reflected in these statements.

**E9: Agreed, we have moved our detailed examination of the statements and which are valid under both the frequentist and Bayesian frameworks to the introduction when we first discuss them (see p. 2/3).**

I think it would be helpful to provide (probably on your OSF page) the exact wording of the invitation email that was sent to prospective participants (see Reviewer 1's comment about this issue, too). As you note, the sample in this study appear to have a better understanding of these issues than participants in prior studies, and you argue that this was not due to selection based on content mentioned in the invitation. However, it would be helpful to make this invitation available for others to evaluate.

**E10: Agreed, we now provide the Email text on https://osf.io/r75qd/, see page 8.**

**Reviewer 1**

I am specifically confused about the details of task 2: On the one hand, it is clear that participants are presented a specific study scenario, namely the ‘drug against depression’ study, thus I am left to assume that the follow-up survey questions (i.e. Can Prior Study… etc. scores) are directly related to this specific study scenario. This seems to be reaffirmed when the authors state “In a second task, participants were presented with a hypothetical scenario […] whether they would like to be able to incorporate previous findings, and their subjective beliefs into the analyses of the study results” (p. 17). On the other hand, it seems that everywhere else in the manuscript, these task 2 questions are described as probing participants about what they actually (or “typically”) do (in their own research?), e.g.: “we examined to what extent researchers think they typically incorporate…” (p. 6); “Afterwards, participants were asked to indicate to what extent they typically take into account…” (p. 9); “Our analysis resulted in two Bayes factors, each of which quantifies whether there is a discrepancy between what respondents typically take into account and what they would like to take into account” (p. 12); “…a difference in the extent to which people take into account prior study results and the extent they wish to do so” (p. 14).

My impression is that probing individuals on what they actually do vs. wish to do would be more relevant to the authors’ goals, insofar as assessing whether the majority of researchers’ current approach to statistical inference already aligns more readily with a Bayesian vs. frequentist approach. Therefore, I’m not certain that asking individuals on how they would act in this hypothetical scenario is an optimal measure, unless it was meant to act as a proxy for what researchers’ typically do? In this case, I would communicate this explicitly; that said, the authors’ themselves discuss that the clinical nature of the task may have influenced the respondents’ answers, thus begging the question as to whether this would be a suitable proxy.

**R1-1: Thanks for this observation: we completely agree. The exact wording of the questions, that we have now made available on an OSF page makes clear we did mean for it to act as a proxy for what researchers typically do. We have included the relevant texts verbatim in the main manuscript (p. 9 – p. 11) and linked to the OSF page (**<https://osf.io/r75qd/>, **see page 9 for the complete materials.**

***In a second task, we asked researchers how they would act in a hypothetical scenario as a proxy for what researchers typically do.* *Participants were presented with the following text: “******Suppose you are about to study the effect of a new drug against depression. A former study with a drug similarly synthesized showed moderate to strong effects in the treatment of depression. However, due to your knowledge about the development of the drug you are suspicious and have strong beliefs against the drug's efficiency.”. Afterwards participants were asked to indicate to what extent they typically take into account (1) results from a previous study (“To what extent do you typically take into account the previous study in your statistical analysis?”) and (2) their own beliefs in a statistical analysis (“To what extent do you typically take into account your belief in a statistical analysis?”), on a scale from 1 (never) to 9 (always).* *Finally, participants were asked to indicate to what extent they would like to take into account (1) results from previous studies and (2) their own beliefs in a statistical analysis, on a scale form 1 (not all at interested) to 9 (very interested) (“These questions are related to the scenario before. Suppose you do know a way how to incorporate existing information in your analysis. Please indicate below to whether you want to take into account results from a previous study and your own beliefs into any statistical analysis. Note that there are no right or wrong answers”).* p.10**

Perhaps a bit nit-picky, but in task 2, when asking respondents about incorporating beliefs and prior study outcomes, sometimes the authors talk about incorporating these “in a statistical analysis” (p. 9), “into the analyses of the study results” (p. 17), or “into the prior” (p. 11), whereas other times they talk about incorporating them “in/into their statistical inference” (p. 1 & 17). To me, the process of setting up the analysis constraints (e.g., prior) versus the process of drawing inferences from the analysis outcomes are not the same thing. The former may involve, for instance, specifying a priori a range of plausible values for a parameter (which may strike a frequentist as biasing the outcome), whereas the latter may involve simply interpreting a given outcome in light of prior beliefs/knowledge (and in turn calibrating one’s belief or skepticism accordingly a posteriori).To me, only this latter case would be consistent with both a frequentist’s and Bayesian’s perspective of cumulative evidence; thus, I would speculate that how the question was phrased may importantly influence respondents’ answers.

***Suggestion major comment 1*:** Please provide somewhere the actual full task details as they were presented verbatim to the respondents (e.g., listing / quoting the full survey questions within the body of the manuscript itself, or provide the full survey within osf as supplementary materials). Also, edit the sections of the text accordingly such that the listed confusions about task 2 are clear to a reader (even without accessing say the online supplementary materials).

**R1-2: We agree we could have been clearer on this: the wording of our questions referred to incorporating personal beliefs and prior study results into the analysis. We do not quite see how this is different from incorporating them into one’s prior (which is subsequently used in the analysis), but we have changed the wording throughout the paper to refer to the analysis and not the prior.**

**We explicitly state the exact wording of each Question now on page 8 to 10. In addition, the full survey can now be found on** **https://osf.io/r75qd/.**

The distinction between what constitutes a ‘subjective’ vs. ‘objective’ prior I find is unclear or inconsistent throughout the text. The authors introduce the concept of priors on p.6 citing the Wagenmakers (2007) article, in which ‘subjective’ priors are characterized as a “state of knowledge” which “quantifies a personal degree of belief”, and ‘objective’ priors as “predefined rules” which are “independent of the person who performs the analysis” and “are generally vague or uninformative” (p. 791). According to this perspective, I would assume that prior study results would contribute to a ‘subjective’ rather than ‘objective’ prior. In fact, this seems to be consistent with the statement in the current article: “subjective beliefs may well be informed by knowledge about previous study results” (p. 17). Given this perspective, I would not expect that “researchers [would] feel uncomfortable with using a subjective Bayesian approach” (p. 6); what I would expect is researchers being uncomfortable with introducing subjective bias into their analyses or interpretations of results, but here, the meaning of ‘subjective’ does not encompass the same connotations as it does from a Bayesian point of view.

I believe the crux of my confusion stems from the fact that the authors seem to be using ‘subjective’ and ‘objective’ sometimes from a Bayesian perspective, and other times from a formal language perspective. From a language perspective, it seems acceptable to dichotomize beliefs as ‘subjective’ versus study results as ‘objective’ (e.g., see p. 6-7 where authors repeat the phrase “subjective beliefs and objective prior study results”). However, this usage of the terms shouldn’t hold in the Bayesian sense, if and when (according to my understanding) both beliefs and prior study outcomes can contribute to the formation of a ‘subjective’ prior. This confusion, or misleading dichotomy, is further reinforced when authors explain that researchers’ desire to incorporate prior study results, but reluctance to incorporate subjective beliefs, “suggests that they might be more comfortable with a rather objective Bayesian approach” (p. 17)… How do these results reflect the objective Bayesian approach which (again, referring back to p. 6) was described as predefined, and comparatively uninformative, such as a default prior distribution? Moreover, later on p. 18, the authors liken a “prior distribution [that] is well founded by previous studies” to “the Bayesian notion of subjectivity”. Finally, their concluding remarks state that “researchers sympathize with some elements native to the philosophy of objective Bayesianism” and “a subjective Bayesian approach is deemed more acceptable if a prior distribution is built upon previous research” (p. 20)… At this stage, I’m not certain what is meant by ‘objective’ vs. ‘subjective’, especially since researchers’ desire to incorporate prior study findings was interpreted as both support for the objective Bayesian approach (quote p. 17) but also as aligning with the subjective Bayesian approach (quotes p.18 & 20).

**Suggestion major comment 2:** Overall, the authors need to more clearly delineate what is meant by ‘subjective’ vs. ‘objective’ in a Bayesian sense, and I think some key statements, such as “subjective beliefs may well be informed by knowledge about previous study results”, are worth mentioning up front when introducing the concept of priors. Currently, the manuscript reads as if the authors’ intentions (especially in the methods) were to (almost mutually exclusively) map beliefs onto subjective priors, and prior study outcomes onto objective priors; but come the discussion these concepts become muddled or intertwined. I think it would also benefit a reader to know up front in the methods that “nothing in the invitation email nor in the survey mentioned Bayesian statistics” (p. 18) such that should the survey have included use of the terms (‘subjective’ or ‘objective’), respondents would not be expected to interpret them in a Bayesian sense. Here, again, a reader would profit from seeing the survey questions verbatim.

**R1-3: We agree completely that we should have been much clearer. We now stick to the Bayesian definition of subjectivity and objectivity throughout the text. Specifically, we now clarify upfront that subjective priors can be derived in two ways: based on one’s own belief and/or based on a previous study.**

When flushing out the A Bayesian Perspective section, I think the explanation of priors, posteriors, and BFs, would benefit from an equation (same as Eq. 7; Wagenmakers, 2007):

P(H1|D) / P(H0|D) = P(D|H1) / P(D|H0) x P(H1) / P(H0)

posterior odds = BF10 x prior odds

Especially for a novice reader, the following statement might be ambiguous: “A Bayes factor (BF) expresses how much our observed data shifts the balance of evidence from one hypothesis (e.g., the null hypothesis H0) to another (e.g., the alternative hypothesis Ha; for more details see Dienes, 2011)” (p. 5). My assumption is that this was meant to express how a BF can be taken to represent / quantify the shift from prior to posterior odds (a link which I find is made more explicit in the equation above; note: in the referenced Dienes (2011) article, this equation is only contained within the article notes). I also think the equation helps clarify the “BF10” subscript.

Within the article, the threshold for strong evidence (i.e. BF ≥ 10) is specified; what is missing is a specified range for inconclusive evidence (e.g., 1/3 ≤ BF ≤ 3). Without this additional information, statements such as “Bayes factors pointing towards the null hypothesis are indicative of no discrepancy,..” (p. 12) may suggest that *any* BF value less than 1 should constitute evidence for the null (which in and of itself would potentially promote erroneous inferences); it also would help a novice reader understand why the outcome in Table 4 (BF = 2.58) is interpreted as “almost no evidence for a difference” (p. 14).

Finally, sequential testing is listed among the advantages of using Bayes factors. As it can also be applied within the frequentist framework, it seems misleading to list specifically as an advantage to BFs. Moreover, in the Wagenmakers et al. (2018) cited reference, not only do the authors acknowledge that “the standard framework of p value NHST can be adjusted so that it can accommodate sequential testing”, but they also state that “From a Bayesian perspective, however, corrections for sequential monitoring are an anathema” (in other words, irrelevant). Perhaps it is worth clarifying that the advantage being expressed here is more accurately sequential testing without need for corrections or without the same concerns of inflating type-1 errors in the frequentist sense?

**R1-4: Again, we agree. We have corrected this phrasing and have included the suggested equation along with a little bit more explanation of the terms in the equation (page 5).**

Minor comments:

Concerning The misinterpretation of frequentist statistics section: I would add in “statistically significant” (on p. 2) when providing the examples for an alternative vs. null hypothesis, specifically: “there is a statistically significant mean difference between…” and “there is no statistically significant mean difference between…”; I feel that describing the null hypothesis as indicating “no mean difference” could imply that evidence for the null should be inferred as evidence of no effect (which in and of itself is a misinterpretation of a non-significant p-value; see e.g., Hoekstra, Finch, Kiers, & Johnson, 2006).

**R1-5: We agree with the reviewer that accepting the null hypothesis based on a non-significant effect only is highly problematic, and we would certainly not want to implicitly endorse such a practice. We do think, however, that in this particular case such a misinterpretation is unlikely. We are talking about hypotheses here, and they are per definition about the population. Statistical significance, on the other hand, only enters at the level of the sample (if we were to have all population data, a significant test would be unnecessary). For this reason, we prefer to keep the current formulation, despite our strong agreement about the undesirability of accepting the null.**

The most common misconception in the Haller & Krauss (2002) study is said to be endorsed by “80% of statistics teachers as well as students” (p. 3); this is inconsistent with Table 1 of the Haller & Krauss (2002) article (p. 8) which lists 73% of methodology instructors and 68% of psychology students as falsely endorsing Statement 5.

**R1-6: Astute catch! We have corrected this mistake (page 3).**

While false endorsement of Statement 5 could be indirectly attributed to the misconception that you can assign a probability to a hypothesis (specifically if one failed to recognize that “the wrong decision” cannot be inferred from a p-value), I would argue that this item was primarily meant to pick up on the misconception that a single p-value tells one something about the Type-1 error rate. In contrast, Statements 1 through 4 are directly relevant to the core misconception highlighted within the current manuscript, namely the section beginning with “Thus, a majority of statistically educated researchers misinterpret frequentist statistics as the probabilities of one’s hypotheses being true” to the end of that paragraph (p. 4). I’m curious why the authors would isolate Statement 5, rather than appeal to the other direct measures of this misconception.. Moreover, as Statement 5 misconception rates in Haller & Krauss’s (2002) study were both the highest (67 - 73% incorrect) and the only specific values referenced in the text, one might be more inclined to accept the authors’ interpretation that “In comparison to the results of Krauss and Haller (2002), the performance of correctly interpreting p-values appears to have increased. This indicates that people have improved on their correct interpretation of t-statistics compared to the beginning of this century” (p. 13). Would these statements still hold for Statement 1 (10 – 34% incorrect), Statement 2 (17 – 32% incorrect), Statement 3 (10 – 20% incorrect), etc.? Arguably, these misconception rates would suggest that even at the beginning of this century, the vast majority of respondents correctly identified these as erroneous interpretations of a p-value.

**R1-7: Thank you for this suggestion. We now added a more detailed interpretation of statements 1-3 in comparison to the other statements (see page 14-15).**

Finally, given that this section is *not* aiming to outline in detail different types of p-value misconceptions, yet nevertheless serves to highlight some of the drawbacks of the frequentist approach to statistical inference, it may be worth including a reference or two (beyond the Haller & Krauss, 2002, text) on where a reader might access more information on specific types of p-value misconceptions to avoid, e.g., Badenes-Ribera et al. (2015), which also highlights the fallacy of conflating statistical significance with effect size or clinical/practical meaning.

**R1-8: Thank you for this useful literature, we now added it to the introduction. (page 5)**

Concerning the Materials and Procedure section: Table 1 typos: Please check statements 1 and 3; looks like the word “means” is missing after “..between the population”. In the Note, the statement “has been repeated by several authors” should be supported by reference(s).

**R1-9: Correct, thank you for spotting this, we have corrected this in the Table. We have also included the reference in the manuscript (page 9-10).**

Unless I misunderstand, although the intended sample to be collected was closer to 600 participants (i.e. 30% of 2,000), but the final sample collected was 117 (i.e. 9% of 1,282), in terms of drawing meaningful conclusions from the data, a sample of 117 should nevertheless allow one sufficient power to find strong support for H1 (assuming a moderate effect size), whereas a larger N falls would be required to find either strong support for H1 (assuming a small ES) or strong support for H0, correct? I’m referring the pre-registered sample sizes (i.e. assuming moderate ES, d = 0.5, required N = 31; assuming small ES, d = 0.2, required N = 168; assuming d = 0.0, required N = 132). If so, I think it would be worth mentioning in the Participants section, and/or in the Discussion, as it should indicate to a reader that the BF = 2.58 should specifically be interpreted cautiously (whereas the rest of the effects which are much greater than BF = 10 would have been sufficiently powered to be detected). Regarding academic position, 55% + 18% + 12% + 12% + 4% = 101%.

**R1-10: We have added the requested information about the a-priori power calculation in the Participants section. We also elaborate on the magnitude of the BF=2.58. Furthermore, we have rectified what turned out to be a rounding error in the reported percentages of academic position.**

***The second Bayes factor indicates almost no evidence for a difference in the extent to which people take into account a-priori beliefs and the extent to which they want to be able to.*** ***This small BF might however result from the low obtained sample size and a possible low effect size.* p.15**

***Respondents indicated to be faculty member (55%), graduate student (18%), post-doctorate (12%), external researcher (12%), and other (3%).* p.8**

Concerning the Results section: Histograms: Is the y-axis frequency expressed in terms of raw numbers? Or a percentage? In terms of relating to past findings (i.e. Haller & Krauss, 2002 rates), percentages would be more understandable. X-axis should include “1” and “9” labels. As “C” and “W” scores are being directly compared, I would also ensure that the range of y-axis values match for Figures 1 and 2.

**R1-11: Thank you for this comment, we now changed the graphs so that the proportions are displayed. Moreover, the y-axes do match for Figure 1 and Figure 2. See p. 11-13**

I am still not wholly convinced that the data supports the following interpretation: “In comparison to the results of […] appears to have increased” (p. 13). For instance, looking at Figure 1 Scores C1 frequencies (assuming raw values), I’d say scores 1 to 4 total roughly 80 respondents, a score of 5 (i.e. neutral?) approx. 10 respondents, and scores 6 to 9 the remaining cases, approx. 27 individuals. This should translate loosely to 68% of the 117 respondents who were some level of confident that this item was false, vs. 23% of the sample with some level of confidence that this item was true. Does this not loosely compare to the Haller & Krauss (2002) rates which observed 10 – 34% of respondents falsely endorsing the item as true? My sense is that the extremely skewed nature of the distributions seems to impart this comparatively greater awareness of about what constitutes an erroneous vs. accurate p-value interpretation, but I’m not yet convinced that this interpretation can be concluded from the data. Perhaps the rationale behind the authors’ interpretation needs to be more strongly argued or the interpretation should not be generalized across all items.

**R1-12: Please see our earlier comment on R1-7.**

I’m confused about the following line: “In line with our hypothesis we can see that the means for Wish statements are generally higher than for Can statements” (p. 13). When was this hypothesis stated? The only explicit hypothesis I found stated in the text was on p. 6: “We hypothesize that researchers feel uncomfortable with using a subjective Bayesian approach because…” which clearly relates to task 2. Yet, the statement above was embedded within a discussion on task 1 results. If there are explicit hypotheses regarding task 1 and 2, these should be stated more explicitly in the intro, and accordingly followed up in the results / discussion.

**R1-13: We have thoroughly rewritten the introduction (including this line) and now clearly outline our original hypotheses/expectations for the first two tasks in the last two paragraphs of the introduction.**

Authors make reference to the score means in the text, e.g. “Inspection of the means indicate that scores…” (p. 14). It would help to have a table with the means or indicated below each plot. Additionally, listing the median scores would be relevant given the skewed nature of the data.

**R1-14: We have presented the mean scores in the plots next to the vertical dashed lines. We prefer not to print the medians in addition so as to not present too many numbers: all data are visible to the reader.**

Concerning the Discussion section: On p. 16 – 17, concerning the explanation on why “Statement 6 is not applicable to a Bayesian framework,” I think it is interesting and informative to include the explanation on how – given a great number of experiments – there is a link between the size of the BF and the expected proportion of BFs one should expect to represent the correct model (put simply). That said, what I think is critically missing is the idea that (like the frequentist approach), no single p-value or BF should tell us anything about probability of replicability in and of itself, correct?

**R1-15: We agree. This explanation was added to the section:**

***As such, Bayes factors allow one to draw conclusions about the probability of drawing the wrong conclusion in the long run based on the strength of evidence. Put simply – given a great number of experiments – there is a link between the size of the BF and the expected proportion of BFs that provide evidence in favor of the correct model. However, a single Bayes factor does not allow for predicting, say, the size of the Bayes factor in a replication attempt.* p.18**

Finally, I’m surprised respondents were not probed for whether they employed frequentists vs. Bayesian techniques in their own research. While I expect that a random sample of psychology researchers should produce a group of respondents who primarily apply frequentist inference, and thus should still be appropriate to address the authors’ goals, I still think it would be fair to also recommend that future research probe for respondents’ typical analysis approach and/or familiarity with each framework (frequentist and Bayesian), in the same way that the authors’ recommended that future studies obtain demographics on individuals’ area of expertise (p. 19).

**R1-16: This is a valid point, and we included this in the discussion section (page 22).**

***In our study, we did not ask participants to indicate their field of expertise, thus future studies would need to study whether being comfortable with including prior beliefs changes with the expertise a researcher has in the context of the research question. We also did not ask participants to report what kind of statistical inference they typically employ in their own work. As such, the results of the present study do not allow linking participants’ responses to their own practices, but we believe this to be a fruitful avenue to explore in future studies.***

***Finally, future studies should include options for open answer responses. It might be interesting to learn from researchers under which circumstances they would like to incorporate a prior and what information they would like to base such a prior on.* p.20**

Other very minor comments: Attached alongside the review is a commented .pdf of the original submission. I found it was easier this way to flag some minor typos and mere suggestions for edits.

**R1-17: Thank you, the suggestions were incorporated into the new document.**

**Reviewer 2**

Structure of the manuscript: The introduction is about evidence for misconceptions of frequentist NHST and then discussed the basis of Bayesian hypothesis testing. The authors discuss what they call *task 1* and *task 2* together throughout the manuscript (in the methods, result, and discussion sections). I was at times somewhat lost while reading the manuscript, but it took me a while to understand why. I think I could not quite make the connection between these parts. Task 1 is an attempt to assess common misconceptions of frequentist NHST and probably also why these misconceptions exist (because we want to make these kinds of judgments). Task 2 deals with the question why researchers may not use Bayesian hypothesis testing. The connection - that many of the misconceptions are actually legitimate statements when using Bayesian inference, is only made in the discussion section, but not earlier in the manuscript. There are two potential ways of strengthening the narrative of the paper: 1. by introducing this connection in the introduction, and including more interconnections between the discussion of Bayesian and frequentist inference throughout the manuscript; or 2. by splitting the study reporting into task 1 with discussion, then introducing Bayesian inference, and then reporting and discussion of task 2.

**R2-1: We have substantially rewritten the introduction in order to better signpost the goals of our paper, the rationale of our two studies, and the way the two studies (in our opinion) flow logically from one another (see also E4).**

P. 7, first sentence: “… to what extent they want to incorporate both subjective beliefs and objective prior study results” I would note that it is questionable that prior study results are in any way more objective. I would say they are not - neither in terms of scientific decisions that are made (how to combine these results, but also the methods chosen for the prior studies), nor in the statistical sense (they tend to be more narrow than priors based on subjective belief). I think the authors actually agree with me from what I read in the discussion section. I would be careful with the language here.

**R2-2: We agree and we have made changes throughout the text to be more consistent with using the term subjective in the Bayesian sense (see also E3).**

P. 8, Table 1**.** I was missing some sort of discussion of these statements (except maybe there is a bit of it in the discussion section). Why are these misconceptions, why are they so prevalent, can we cluster them into groups of misconceptions, etc. The best place for this would probably be the intro. I feel readers must have read some of the previous papers on the topic to understand what is done and what the problem is. I also think the statements themselves are only meaningful if it is discussed what the statements attempt to express. For example, statement 6 seems relatively useless when one wants to do inference for the data at hand, but if we interpret it as a proxy for robustness and replicability then of course it can be meaningful.

**R2-3: Agreed, we have elaborated on why each of the statements cannot be made with respect to a significant *p*-value in the introduction (see also E9).**

P. 9, “on a scale form 1 (not all at interested) to 9 (very interested).” form should be from

**R2-4: Thank you, we fixed this.**

Figures 1 and 2: If data are so skewed, then it might be better to use the median here? 😃 Also, I think the information from the two Figures could be shown more concisely and informatively by plotting each person’s truthfulness and wanting rating in the same plot. For example, this can be done with violin or boxplots and lines (like <https://github.com/jorvlan/open-visualizations>). I feel like proposing this is a bit out of line for a reviewer, but I think useful to see whether there are any participants who do not want to use these statements but still think they are appropriate (i.e. who go opposite to the overall trend).

**R2-5: We appreciate the suggestion, but believe the recommended visualizations are appropriate for more complicated data than those from our study. We essentially have a bunch of ones, twos, threes, …, and nines. For this, we believe a standard histogram to be appropriate and probably more intuitive for the typical reader (see also R1-14).**

P.12: Why use a two-sided test? I suspect the hypothesis was always that researchers would wish to make these kinds of statements.

**R2-6: When we preregistered our study, we decided that although we had some expectations as to the direction of the effect, we were not 100% sure the effect would not go in the opposite direction. For a Bayesian one-sided *t*-test, all mass of the prior on the other side is truncated, which we considered too severe given that we were not 100% certain. As such, we prefer to stick with our preregistered analysis strategy (also note that the Bayes factors are already of enormous magnitude and would only get larger, so nothing would change qualitatively).**

P. 12: The authors write “According to the classification scheme suggested by Jeffreys (1998), we considered a Bayes factor of 10 (in favor or against the alternative hypothesis) as strong evidence.” It is odd that they use the cutoff given their previous statement “The BF is directly interpretable in relation to one’s hypothesis.” If the BF is immediately interpretable, then why use a cutoff. Also, given the vast evidence for most of their tests I am not sure this specific cutoff makes any sense in the context of the study and they do not refer to it afterwards. So, maybe just take it out? Just as a side note, I did not understand what the critical comparison for the t-test was until the very end of the paragraph. Maybe it makes sense to lead with this information.

**R2-7: The Bayes factor threshold was preregistered to allow us to do an a-priori Bayesian power calculation (see** [**https://osf.io/uhskf/**](https://osf.io/uhskf/)**). Once the data had been collected, it became clear we had worried needlessly and the Bayes factors were all huge, but we believe in order to be transparent it is important to still report the Bayes factor threshold we had originally preregistered (see also R1-10).**

P. 13: " This indicates that people have improved on their correct interpretation of t-statistics compared to the beginning of this century." I wonder whether that is true without limitations. I guess researchers have seen a considerable number of papers on the issue in the recent years. They also know that psych methods researcher want to trick them in their surveys and are quite suspicious. Also, maybe saying “do you think this is correct” and “would you wish to make such a statement” in the same questionnaire is a bit leading.

**R2-8: We are not entirely sure what the reviewer is asking, but we agree we can make our statement more nuanced. We have replaced the word ‘indicates’ with ‘suggests’. Hopefully this addresses the issue at hand.**

P. 15: “However, the results show contradicting evidence for the notion that a Bayesian framework better describes how researchers want to use inferential statistics.” This notion is transported through the entire discussion section, that researchers want to interpret results as if they used Bayesian analysis but do not want to incorporate their subjective belief. I do not think the data support this claim. Participants in the study did not say they would not like to use subjective belief. Actually, Fig 3 bottom right is all over the place. Some people seem to wan to incorporate their beliefs, others don’t. The BF is in the direction of the alternative hypothesis, granted it is a small BF, but > 1. So, the conclusion “researchers seems to like being able to incorporate prior results but not personal beliefs into statistical inference” does not seem appropriate given the data.

**R2-9: Agreed, we have rewritten our conclusions in the Discussion throughout to reflect the fact that evidence for a discrepancy between researchers’ current practices and wishes with respect to incorporating their subjective beliefs is ambiguous.**

P. 16: “Bayes factors enable drawing the conclusions from […] statement 5 (we can find the probability of making a wrong decision, when rejecting the null hypothesis, given a prior belief).” But we do not make decisions in Bayesian inference based on Bayes factor. We just report evidence. The proposed Bayesian inference in the manuscript does not even mention decision making. Therefore, statement 5 does not apply to Bayesian inference as it is used in psychology.

**R2-10: Many experts on Bayesian inference do believe Bayes factors are suitable for decision making (see e.g., Aczel et al., 2020, full ref below). We now added a qualifier to the discussion about decision making, along with this reference.**

**Aczel, B., Hoekstra, R., Gelman, A., Wagenmakers, E.-J., Klugkist, I. G., Rouder, J., Vandekerckhove, J., Lee, M., Morey, R., Vanpaemel, W., Dienes, Z., & van Ravenzwaaij, D. (2020).**[**Discussion Points for Bayesian Inference**](https://www.nature.com/articles/s41562-019-0807-z)**. *Nature Human Behaviour, 4,* 561-563.**

P. 16: “then they would hit the Bayes factor criterion associated with the correct model 10 times more often than the Bayes factor associated with the incorrect model”. Is that so? I only briefly read Tendeiro, Kiers, & van Ravenzwaaij (2019), but I am quite certain that such a statement does not appear in Rouder (2014). The analysis in that paper is not concerned with correct or incorrect models.

**R2-11: We are not able to identify what in this sentence the reviewer objects to. Perhaps it would have been better if we would have used true model instead of correct model, Rouder uses “null serves as truth” and “alternative serves as truth” in his 2014 paper. We have rewritten the passage in the Discussion in an attempt to avoid confusion.**

I think this study is a missed opportunity for some qualitative assessment. Given that the target sample is experts who are asked about analyses related to their work, I think it could be very useful to give them some room for explanations and qualifications of their responses. This is especially true for task 2. Under which circumstances and what information would they like to incorporate? There seem to be huge individual differences - what are the different rationales? The expertise and education of these researchers is probably quite variable, so this information could help interpreting the data patterns. Additionally, it would be interesting to ask researchers if, given the opportunity to use a statistical approach that allows them to make statements like statements 2 and 4 in the questionnaire, they would be willing to pay the price of setting priors.

**R2-12: We agree that there are many more interesting questions to be asked on this front. We have added the following bit to the Discussion:**

***Finally, future studies should include options for open answer responses. It might be interesting to learn from researchers under which circumstances they would like to incorporate a prior and what information they would like to base such a prior on.*** **p.20**

**Final Editor decision—Accept**

Jan 20, 2021

Dear Dr. Haucke,

I have now had a chance to read over your manuscript “Bayesian frequentists: Examining the paradox between what researchers can conclude versus what they want to conclude from statistical results”, along with the letter describing the changes you made. Thank you for your responsiveness to the concerns that the reviewers and I raised. I am happy to say that your paper is now officially accepted for publication in Collabra: Psychology. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published! I hope your experiences with Collabra: Psychology have been positive and that you will continue to consider it as an outlet for your work.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point. Our managing editor will contact you in case there are any pre-prodution file related questions. You will have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely, Richard Lucas