**Peer Review and Communication History**

**MS Title:** Obtaining Evidence for No Effect

**Author Name:** Zoltan Dienes

**Submitted:** June 18, 2020

**Editor First Decision: Revise & Resubmit**

Sep 9, 2020

Dear Zoltan Dienes,

I have now received all reviews of your manuscript, “Obtaining evidence for no effect” from qualified researchers. I also independently read the manuscript before consulting these reviews. Your paper addresses a topical aspect of statistical inference: How to specify models so that evidence can be stated for and against an effect. The contribution of the paper is a practical approach to this specification. However, the current version of the manuscript lacks clarity and concrete guidance. I therefore reject the current version of the manuscript but encourage you to submit a revised version for further consideration at Collabra: Psychology. Given the very serious critiques raised by the referees I do not want to sound overly encouraging about the prospects of a revised version. That said, I can certainly imagine a successful revision of this paper, but it would likely be quite different from the draft you submitted.

I encourage you to read the excellent comments by all three reviewers carefully. All of them have provided serious criticism of the manuscript in its current form but also meaningful suggestions for improvement. I will highlight issues I think are particularly salient here starting with issues that all reviewers agreed on. If you decide to resubmit, please include a document with a point-by-point response to the points I highlight here and the reviewers’ comments, outlining each change made in your manuscript or providing a suitable rebuttal.

1. All reviewers note that the depth of discussion for the two main frameworks discussed here (equivalence testing and Bayes factor) is not the same. While Bayes factor is explained at a more basic level equivalence testing is not introduced as a method in any detail. While the reviewers somewhat disagree on whether the paper should introduce both frameworks in more or less detail, I agree with Reviewer 3 that this is a matter of better defining the audience of the paper. Given that Collabra:Psychology has a wide readership beyond methodologists and statisticians I assume the intended audience is psychologists from diverse areas of expertise. Therefore, I would suggest to introduce both frameworks to some degree.
2. All reviewers highlight that the use of examples for determining the smallest effect of interest or a rough scale of expected effects could be improved. The practical examples are a strength of the manuscript. However, these examples could be more structured and more worked out. In some sections (See R2 point 6) the number of examples could be reduced to allow more detailed discussion of the most meaningful example.
3. The paper implicitly and explicitly highlights differences and similarities between the two approaches of defining the smallest effect of interest and the rough scale of expected (predicted?) effects. Both Reviewer 3 and I would like to see more explicit discussion of these two approaches. What are the strengths and weaknesses? What is the effect if these two are misspecified? Particularly with regards to the first I also assume the bounds have to be quite wide to ever get evidence for the null (otherwise the CI will never be within the bounds). How can one find the right balance? In a similar vein, Reviewer 1 notes the difference between practical interest and theoretical interest. This distinction is relevant for section 1 Smallest effect size of interest. Many of the most meaningful examples in the section provide guidance from a practical perspective, not necessarily a theoretical one. Is this a purposeful distinction between the frameworks or just by accident?
4. Additionally, given the suggested methods of determining scales and bounds are empirical, I think a discussion of noise variability is needed. Especially the calibration method seemingly might be affected by noise in measuring the relationship between the variables. My suspicion is that the effect of sample noise might be much more dramatic for setting fixed bounds on smallest effects of interest than for setting a rough scale.
5. Some of the reviewers and I are a bit uneasy about the distinction between “objective reasons” (I assume this means data-driven) and other reasons (I assume this means subjective, expert-driven; though I might also include default Bayes factors in this category) that are the basis of specifying effect sizes of interest. I think the heuristics presented here show that there is quite a bit of subjectivity on these so-called objective approaches. Maybe data-driven or empirical would be the better term. Additionally, the more important distinction may be whether the reasons are transparently stated (see also the comments by Reviewer 2) or not.

Reviewer 2 summarizes these points of critique as the following question: Can the paper provide helpful (practical) advice with regards to the specification of hypotheses? I would add that the manuscript would prove even more useful if it provides guidance on when to choose either of the two frameworks. In summary, while I think this is a promising manuscript it does lack some depth and clarity in its current form. I hope you will consider revising it and resubmitting for further consideration at Collabra: Psychology.

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. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission. **I also suggest that the revised version of the manuscript has page and line numbers. Please also check for typos and grammar.**

If you have any questions or difficulties during this process, please contact the editorial office at 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,
Julia Haaf

# Reviewer 1

##### 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.

The paper gives an overview of how to set minimally interesting effect sizes and how to set informative priors. I did something similar recently in the context of cognitive modelling (<https://ppw.kuleuven.be/okp/_pdf/Lee2018DIPFCM.pdf>) so the topic is, at least to me, interesting and timely. I do sense there is a better paper hiding underneath the current version. My comments aim to be helpful to uncover this gem.

Overall, I think the paper could benefit from having more worked examples, rather than abstract descriptions and references. The paper is at its strongest when it does so, but at its weakest when it does not.

I am wondering what exactly the author means with theory. Everything in the paper makes perfect sense if I replace “theory” with “statistical hypothesis”. But that doesn’t align well with my own rough definition of theory. This is not to say my definition is superior, but I would like to encourage the author to make his own rough definition clear. As an example, take the weight example. I realize this example is hedged using “or practical reasons”, but its practical nature seems to be strangely at odds with the rest of the paper, which talks about theory most of the time. The same seems to hold for the psychiatry example.

Relatedly, I couldn’t get my head around this sentence: “The model of H1 is a probability density function indicating how plausible different possible population effects are given the theory” This one sentence contains theory, H(ypothesis) and model; but I can’t figure out which entity is supposed to play which role, and how they relate.

I have a similar confusion regarding the use of the word “predict”. For me, prediction is something that models do in data space. Here, it seems that the word is used differently (e.g., “point prediction” refers to something in parameter space). Again, I am not arguing that my perspective on prediction is superior, but given there seem to be different reasonable interpretations of the term, it might be good to make the perspective used in the paper more explicit.

“Therefore a p-value can never provide evidence for an effect not being there” I do agree, on some level, but in this light I was surprised to read about equivalence testing as a way to provide evidence for no effect, which is of course just two p-values, associated with two one sided tests. Maybe the cited phrase should be “Therefore a SINGLE p-value can never provide evidence for an effect not being there”?

I don’t get why “(or at least in the range of, say, 0 to 6kg).” is between parentheses. Isn’t the fact that the hypothesis is expressed as a range exactly what differentiates this type of the previous type?
I am not sure another explanation of what a Bayes factor is is needed (especially, since p-values and equivalence testing, and inference by intervals are not explained either). There are ample references for readers needing to be introduced to Bayes factors.

For what it’s worth: I don’t agree that specifying a prior is enough to make a theory testable, and I don’t think Lakatos, mentioned in the introduction, would either, so “in order to severely test it” and “in order to test our theories severely” seems misguided, at least to me. I recently argued how we need something more to do that: a data prior (<https://psycnet.apa.org/record/2019-80330-001>).

“If there is no basis for justifying an effect size, consider if one only need estimate (e.g. derive means and confidence or credibility intervals) to draw the inferences needed”. But even for “just” estimation using H1, a prior is needed, so what does switching to estimation solve?

Figure 1: Is it really necessary that the upperbound and lowerbound are identical in absolute value (as panel A seems to suggest)?
I don’t know which is the best way is to cite <http://pcl.missouri.edu/bayesfactor>, but any reference omitting Morey seems wrong. Maybe contact Rouder or Morey to find out their preference?

Signed,
wolf vanpaemel

##### 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.

Review: Obtaining evidence for no effect

Summary:
The manuscript argues that in order to obtain evidence for the null hypothesis, it is necessary to specify an effect size of interest. It continues by outlining how an effect size of interest can be specified as a smallest effect size of interest, or as a rough scale of effect size.

Review:
I agree with the general claim of the authors that obtaining meaningful evidence for the non-existence of an effect requires the careful specification of hypotheses. However, the fact that statistical answers depend on statistical questions (i.e., Jeffreys platitude) has long been known, and thus, the argument that evidence for the null hypothesis depends on the specification of statistical models is hardly new. It remains to ask whether the manuscript can provide researchers with helpful advice regarding the specification of hypotheses. However, I get the impression that the manuscript will be more likely to create confusion than to provide actionable insights for applied researchers. Although it covers lots of different approaches to model specification, it fails to introduce any of these in sufficient depth to make them actionable. Moreover, the manuscript implicitly switches between many different statistical frameworks (Bayes factors, equivalence testing, severity testing, ROPE, …), which makes it difficult to match recommended approaches to any particular framework. Therefore, the unique contribution of the paper is unclear to me, and I do not believe that it is suitable for publication in its current form. I will give more detailed comments on the manuscript below.

Major:

1. Reading the manuscript, I have trouble telling what it adds to the current literature. As outlined before, the main claim (evidence for H0 depends on the specification of the models) is not new, and all recommendations about specifying effect sizes of interest have been described in greater detail in Lakens, Scheel, & Isager (2018), Etz et al. (2018), Dienes (2019), and Mayo (2018).
2. The manuscript implicitly touches on many different statistical frameworks (null hypothesis significance testing, equivalence testing, Bayes factors, severity testing, ROPE, …). However, the only framework the authors venture to explain in at least some detail are Bayes factors. Therefore, the reader is required to have at least a basic knowledge of almost all mentioned statistical frameworks to understand the manuscript, and judge the adequacy of the proposed methods for the specification of the effect size of interest in a particular method. I do not think that this is a realistic expectation, and even if it was, it would still be kind to the reader to explain the statistical concepts carefully instead of just hinting to background knowledge.
3. The title of the manuscript is: Obtaining evidence for no effect. This makes me wonder whether all described methods can actually provide evidence for no effect. The authors themselves claim that “the use of non-significance to indicate no effect is misleading”. However, later, they rather explicitly refer to equivalence testing and severity testing, both methods deeply rooted in significance testing, and even null hypothesis significance testing with an a-priori power analysis itself (see p. 5) as methods to obtain evidence in favor of the null hypothesis (see p. 4-6 of the manuscript). A similar issue applies to the ROPE method, which investigates the posterior probability that a parameter lies within a certain interval. Interestingly, with the right constellation of prior distributions and data, the posterior distribution could lie almost entirely within the ROPE interval, but there would be evidence for the alternative in the Bayes factor. With a narrow definition of statistical evidence, I would therefore argue that only the Bayes factor can measure the statistical evidence for the null hypothesis. This makes me wonder whether the focus of the manuscript is actually providing advice on how to “obtain evidence for no effect” or whether it is to provide advice on how to specify effect sizes of interest.
4. The manuscript separates two ways to specify an effect size of interest: Specifying a smallest effect size of interest or specifying a range of plausible values. In my opinion, presenting these approaches as independent of each other is misleading (see e.g., claims in the manuscript that they are “different models”, “two ways”). In fact, what the authors describe, is the specification of an interval null hypothesis, and the specification of an alternative hypothesis that is represented by a model with a probability distribution on parameters. However, none of these hypotheses is ever specified alone. The competing hypothesis is either implied by the method (e.g., in equivalence testing) or needs to be specified separately (e.g., in Bayes factors). The assumption that researchers can decide to specify only one hypothesis in a hypothesis test, is incorrect. Furthermore, the implicit switch between the specification of H0 and H1 will most likely confuse readers, especially considering the fact that the introduction of the manuscript only strongly points towards the specification of the alternative hypothesis (see the snow leopard and the “animal in the room” example).
5. In a related note, although the manuscript discusses the importance of specifying H1 in order to obtain meaningful evidence for H0, it misses the point that the inverse is also true. The specification of H0 (e.g., is it a point null or an interval?) also influences the evidence obtained for H1. With the current structure of strictly separating the specification of H0 and H1, I think it is difficult to bring this point across.
6. As a reader, I was somewhat overwhelmed by the number of different examples. For example, just take section i) of the SESOI section (p. 5). Here, we have: depression treatment, change in pain, positive/negative affect rating, redness of faces, and effectiveness of cognitive training within half a page of text. However, none of these examples were really helpful because they were too superficial to actually understand the method in depth (e.g., How many end users were asked? Was it an interview or a survey? How did the original authors arrive at their cut-off value for an interesting effect size?, …). A reader who would want to apply the method, or even find out whether it is applicable to their research, would need to go back to the original literature to figure out the details. As a reader, I would therefore really appreciate if the authors could drastically reduce the number of their examples in exchange for more depth (e.g., show how the SESOI is calculated with the different methods and applied to a certain statistical framework).
7. Throughout the manuscript, the authors make many implicit assumptions regarding the hypothesis testing scenarios and the model specification that are not always obvious for the reader. For example, the claim that with Bayes factors, researchers need to (only) specify the scale of an effect, makes at least the assumption that (1) this effect is defined under H1, (2) the prior on the effect under H1 follows a continuous prior distribution, (3) this continuous prior distribution has a scale parameter, (4) only the scale parameter needs to be adjusted, (5) the complementary H0 is pre-defined, for example as a point prior. This is a very limited view on Bayes factors because, as the authors write themselves, “a Bayes factor can use any models that seem worthwhile to compare”. These implicit assumptions make it difficult for the reader to see the generalizability of claims. I would therefore recommend to the authors to clarify the assumptions made throughout the manuscript, and maybe find an application example that makes it clear to what applied scenarios (e.g., t-test, regression) the methods can be applied to.
8. Overall, the manuscript argues in favor of more transparency in model specification. This becomes especially clear when the authors demand that the “numerical value for the effect should come from a public place one can point to (for example data), so that other people can criticize the reason for choosing that value” (p. 4). I absolutely agree with this sentiment. However, it is difficult to entirely remove arbitrary decisions from the research process (see also the current debates around multiverse analyses), and it is important to note that the proposed methods for effect size definition are no exception. For example, the degree of violation in assumptions (see section iv p. 6) requires the specification of a cut-off for assumption violation, or the confidence interval method requires the specification of the width of the confidence interval, as well as the determination of “interestingness” (p. 6). The question is, what makes these methods better than committee decisions or benchmarks (see p. 4), which the authors criticize for their intransparency? To me, it seems like the latter methods can be reported equally transparently as the recommended methods, and all methods entail a certain degree of arbitrariness. I wonder what the authors think about this and would appreciate if they could comment some more on the issues of transparency and sensitivity to arbitrary decisions.

Minor

1. In several places, the manuscript displays a very strict falsificationist attitude that is, in my view, not consistent with research reality in most disciplines. For example, the manuscript starts with the sentence “If there is no way for a theory to be shown wrong by a study, the putative test of the theory is no test at all”. However, how often does it happen that a theory can actually be “shown wrong”, i.e., finally disproved, by data? Even the most elegant theories are most often formulated as probabilistic models, which means that there is always some (small) likelihood that the data did occur under the model. How could a result of one single study therefore “show a theory wrong”? I would argue that it is, in fact, impossible. Additionally, most scientists would agree with George Box’s statement that “all models are wrong”. What would be the value of showing that a model is wrong in this case? My point is that although I agree that it is beneficial to be able to obtain evidence in favor of the null hypothesis, this does not warrant a strict all-or-nothing falsificationist view which is not in line with research reality. I would suggest that the authors tone down these statements and instead argue for the many benefits of finding (gradual) evidence for the null.
2. The short discussion of Bayes factor design analysis on p. 9 seems somewhat out of place. BFDA is not a method to find an effect size of interest for “when there are no prior studies”, as the section title suggests. However, I think a short discussion of the implications of model specification (e.g., specifying H0 as an interval or as a point, specifying a certain width of the prior distribution in H1) on the required sample size would be an interesting addition to the manuscript.
3. On p. 3, the authors argue that “neither can evidence for no theoretically interesting effect be provided by […] a default Bayes factor”. In my opinion, this is not true. A Bayes factor is always a measure of relative evidence in favor of one model over another. A default Bayes factor simply displays a lot of uncertainty about the effect size under the H1 model. If a comparison of this H1 model to a (typically point) H0 model is not of interest to a researcher, then this means that the models were misspecified for the specific application, but the Bayes factor can still display evidence in favor of H0.
4. Reading the conclusion of the manuscript (last paragraph p. 9), I find it way too strongly formulated. Since the manuscript describes methods and ideas that already (long) exist, I don’t see how it can argue for “radical change”. Also, it is not clear to me what the authors mean by “scientific hallucinations” in the context of the paragraph. It sounds like they argue that without a proper specification of a theory, we can neither find evidence for or against it. However, nothing of this seems to require any delusion on the side of the scientist.

##### 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 3

##### 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.

The manuscript under review considers the question how to investigate the absence of an effect in the context of frequentist and Bayesian hypothesis testing. The critical requirement in both instances is the definition of the smallest, theoretically relevant effect size. The manuscript aims to provide guidelines, principles, and practical advice.

In my view, the paper collects several approaches to this question, but cannot communicate these ideas in a clear and accessible manner. I am wondering who is the intended audience of the paper: Statistical experts who are looking for best practices? Or practitioners who want to improve their statistical inferences and draw better conclusions?
In both cases, the paper would benefit from more clarity in the way it presents its ideas and methods as well as more illustrative examples that guide researchers who want to adopt the strategies outlined here. In its current form, the paper was sometimes difficult to follow, seems to be unfinished in some parts, and had a couple of typos and grammatical mistakes.

Overall, psychologists can benefit from papers that provide a clear guidance on how to analyze experimental and observational data, especially in the context of evaluating null effects. The paper contains several useful ideas — it should be reworked to present these ideas in a more approachable manner so readers can benefit from it in their actual practice.

I have collected more specific comments below.
(A note on the manuscript formatting and/or Collabra‘s editorial system: The manuscript does not have page numbers nor line numbers, making commenting on particular sections very difficult. In my comments I will refer to the pages as they appear in the PDF, i.e. page 1 is the title page, page 2 contains the abstract, and page 3 is the beginning of the actual manuscript.)

Page 3, paragraph 3: „A p-value is calculated without reference to the possible size of an interesting effect. Therefore a p-value can never provide evidence for an effect not being there.“
The first sentence is technically true, but the context of performing a statistical test is relevant: For example, in an equivalence test the p-value is very much to be considered in the context of the absence of an effect. In a paper looking at the interpretation of statistical results in both significance testing and Bayes factors a more nuanced representation of the statistical methods (e.g. in context of either Fisherian or Neyman-Pearsonian testing) would be favorable.

Page 3, last paragraph + page 4, Figure 1:
The purpose of this figure is not really clear as the text does not match the examples provided. While the first textual example matches illustration B (an empirical effect size larger than the smallest effect of interest is consistent with the theory and thus corroboration for the theory), the second example in the text refers to the case where the theoretically plausible effect size is within an interval and any effect size estimate outside of this interval would be refutation of the theory. The illustration refers to this second equivalence by providing „models of theories for Bayes factors“, but it does not seem necessary at this point to consider this problem only in terms of probability densities. If an illustration to the examples is desired, it should be presented - at this point in the manuscript - in the same contexts for both examples so readers can better understand the illustration without extensive prior experience with both Bayes factors and significance tests.

Page 4, paragraph 1:
The author advocates for the use of „objective reasons“ when specifying smallest effect sizes of interest, but the negative examples are not convincing to me. For example, using common boundaries such as „a Cohen‘s d of 0.5“ can be one approach when the theory is vague and unspecific (I concur that such a theory might not be strictly testable and researchers should refrain from trying to use significance testing in such circumstances, but this seems to be not the focus of the present paper). Furthermore, expert committees can, if conducted transparently, also provide a solid justification for the present case. Providing examples of „ad hoc justifications“ for effect sizes, criticising weak theories that do not predict a minimally interesting effect size would be more convincing at this point.
The sentence „One can evade the problem by only reporting p-values and not interpreting non-significant results.“ is for me an example where the paper would benefit from further polishing and added clarity: An idea is presented that is technically true, but a very limited perspective. For example, non-significant results from both traditional significance tests and equivalence tests paint a different picture. This and the follow sentences also feel somewhat disconnected to the previous discussion on objective reasons for possible effect sizes.

Page 6, paragraph 2:
The method of using „calibration“ is interesting and a very valuable addition to the paper. Unfortunately, this section is very brief. The example could be extended and described in more detail, so researchers can adopt this method more easily.

Pages 4-6:
A more general point on finding smallest effect sizes of interest: The paper could also benefit from a discussion on whether the smallest effect size of interest is an useful tool. Why, for example, is a change of 0.30 relevant, but a change of 0.29 is not? Because on average 0.30 corresponds to a noticeable change? Depending on the context, it might as well have been 0.28 – measurement error or uncertainty in general might make this boundary arbitrary. Significance testing does require a single cut-off point, but a brief discussion of this issue seems to be fitting to the general context of the paper (especially considering that Bayes factors are discussed later).

Page 7, paragraph 2:
It seems to be limited to consider null hypotheses in Bayes factors only as point null hypotheses. In practice it might be most common, but as the author acknowledges there are alternatives available (see also my previous comment)

Page 8f:
The paper only briefly explains why now a scale of effect and not a smallest effect size is considered. Readers not as acquainted with the technicalities of Bayes factors might not be to follow as easily.
Furthermore, a more detailed discussion on how the two approaches compare would be helpful: In what ways are the approaches similar (e.g. calibration) and what differences exist? Why, for example is the „basic effect heuristic“ not useful for determining a smallest effect size?

Page 8, section „(ii) Basic effect heuristic“:
This is another example where a more verbose explanation would be helpful to guide readers in using this method.

Page 8, section „(iii) Calibration“:
While the previous example for calibration was straight forward (even when it could be expanded further), this example is more difficult to follow. Furthermore I did not find the section convincing in how participants‘ expectations should be a valid calibration for an actual change in outcome. This rests on some assumptions that should at least be mentioned here.

Page 9, last paragraph:
I am not sure if the paper actually argues for that and if this is really a „radical change“. At least it seems far less radical compared to initiatives to ban statistical significance.

##### 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**
Mar 12, 2021

EDITOR: *All reviewers note that the depth of discussion for the two main frameworks discussed here (equivalence testing and Bayes factor) is not the same. While Bayes factor is explained at a more basic level equivalence testing is not introduced as a method in any detail. While the reviewers somewhat disagree on whether the paper should introduce both frameworks in more or less detail, I agree with Reviewer 3 that this is a matter of better defining the audience of the paper. Given that Collabra:Psychology has a wide readership beyond methodologists and statisticians I assume the intended audience is psychologists from diverse areas of expertise. Therefore, I would suggest to introduce both frameworks to some degree.*

RESPONSE: A paragraph has been added further explaining inference by intervals, including equivalence testing.

EDITOR: *All reviewers highlight that the use of examples for determining the smallest effect of interest or a rough scale of expected effects could be improved. The practical examples are a strength of the manuscript. However, these examples could be more structured and more worked out. In some sections (See R2 point 6) the number of examples could be reduced to allow more detailed discussion of the most meaningful example.*

RESPONSE: I have removed almost all of the smallest effect of interest examples where they had been just referred to and not spelt out. Instead I have introduced a new real example I describe in some detail. Other examples in both sections have also been expanded.

EDITOR: *The paper implicitly and explicitly highlights differences and similarities between the two approaches of defining the smallest effect of interest and the rough scale of expected (predicted?) effects. Both Reviewer 3 and I would like to see more explicit discussion of these two approaches. What are the strengths and weaknesses? What is the effect if these two are misspecified?*

RESPONSE: I have now added a paragraph to the discussion about the differences between them, especially the fact that they answer different questions. I have also added discussion of how noise can lead to misspecification and how to deal with noise.

*EDITOR: Particularly with regards to the first I also assume the bounds have to be quite wide to ever get evidence for the null (otherwise the CI will never be within the bounds).*

RESPONSE: Yes, I agree. That is something I have long noticed; I refer to Linde et al (2020) for a reference on equivalence testing finding it particularly difficult to get support for H0 (and then I remembered I had said the same in Dienes 2016 and added that).

*EDITOR: How can one find the right balance?*

RESPONSE: I emphasize severity testing more now, which entails one must run enough subjects to potentially get support for H0 in order to severely test a theory that claims a difference. Thus, one should check one has the resources to run enough subjects to get support for H0 if one is to proceed with inference by intervals. (But what the minimal effect is should not be determined by such practical constraints but actual scientific context.)

*EDITOR: In a similar vein, Reviewer 1 notes the difference between practical interest and theoretical interest. This distinction is relevant for section 1 Smallest effect size of interest. Many of the most meaningful examples in the section provide guidance from a practical perspective, not necessarily a theoretical one. Is this a purposeful distinction between the frameworks or just by accident?*

RESPONSE: It is true as you say that most examples for getting a smallest effect size of interest are practical rather than theoretical. While I think it is easier in a practical context, it is not intrinsic to the method, and as you say I do give theoretical examples as well. I do not think I have anything more insightful to say about this, so I have not added anything to the manuscript.

*EDITOR: Additionally, given the suggested methods of determining scales and bounds are empirical, I think a discussion of noise variability is needed. Especially the calibration method seemingly might be affected by noise in measuring the relationship between the variables. My suspicion is that the effect of sample noise might be much more dramatic for setting fixed bounds on smallest effects of interest than for setting a rough scale.*

RESPONSE: I now discuss the problem of noise and regression, as it is one I (well really my post doc) have been working on the past few years. I also discuss variability in the estimate of the minimal interesting effect size and how one may therefore introduce a grey area, as suggested by Speigelhalter.

*EDITOR: Some of the reviewers and I are a bit uneasy about the distinction between "objective reasons" (I assume this means data-driven) and other reasons (I assume this means subjective, expert-driven; though I might also include default Bayes factors in this category) that are the basis of specifying effect sizes of interest. I think the heuristics presented here show that there is quite a bit of subjectivity on these so-called objective approaches. Maybe data-driven or empirical would be the better term. Additionally, the more important distinction may be whether the reasons are transparently stated (see also the comments by Reviewer 2) or not.*

RESPONSE: “Data-driven” matches many of the examples, but not all (e.g. the “ratio of scales heuristic” in Dienes, 2019), and also misses the more overarching point: There should be publicly available reasons, as you say; this is also the claim in the manuscript. Ultimately one needs to bottom out and hence, as you say, make a subjective judgment that one has gotten as far as one can at this point with the reasons available. But this subjectivity is just to say: Conclusions so far are provisional on the reasons available to date for drawing the conclusions. All those reasons can be publicly criticized and hence conclusions updated. If a parameter value is derived purely subjectively (“I feel in the depths of my soul the number is about 0.3”), two people disagreeing cannot with what is on the table criticize each other’s’ reasons, and hence cannot come to a better conclusion because of constraints intrinsic to the problem. Therefore, objective is a good description of what is needed.

*EDITOR: Reviewer 2 summarizes these points of critique as the following question: Can the paper provide helpful (practical) advice with regards to the specification of hypotheses? I would add that the manuscript would prove even more useful if it provides guidance on when to choose either of the two frameworks.*

RESPONSE: I have added a paragraph comparing the frameworks in the discussion.

Reviewer 1

*REVIEWER: The paper gives an overview of how to set minimally interesting effect sizes and how to set informative priors. I did something similar recently in the context of cognitive modelling (*[*https://ppw.kuleuven.be/okp/\_pdf/Lee2018DIPFCM.pdf*](https://ppw.kuleuven.be/okp/_pdf/Lee2018DIPFCM.pdf)*) so the topic is, at least to me, interesting and timely. I do sense there is a better paper hiding underneath the current version. My comments aim to be helpful to uncover this gem.*

RESPONSE: The positive comments are appreciated. The referenced paper is relevant and has been added to the discussion.

*REVIEWER: Overall, I think the paper could benefit from having more worked examples, rather than abstract descriptions and references. The paper is at its strongest when it does so, but at its weakest when it does not.*

RESPONSE: Most examples have been further fleshed out.

*REVIEWER: I am wondering what exactly the author means with theory. Everything in the paper makes perfect sense if I replace “theory” with “statistical hypothesis”. But that doesn’t align well with my own rough definition of theory. This is not to say my definition is superior, but I would like to encourage the author to make his own rough definition clear. As an example, take the weight example. I realize this example is hedged using “or practical reasons”, but its practical nature seems to be strangely at odds with the rest of the paper, which talks about theory most of the time. The same seems to hold for the psychiatry example.*

RESPONSE: I now define the terms substantial theory, statistical hypothesis and model in the first section on severe testing. A test of the statistical hypothesis is typically a test of the substantial theory. But that does not mean we should do away with the use of the term theory in the paper; the only reason testing the statistical hypothesis has interest is because it also tests a broader theory that is the thing truly of interest. The psychiatrist case happens to be deleted, but here there is a theory involved, namely that e.g. the application of a set of therapeutic principles improves depression. This is translated into a way of testing it, namely with specific DVs and IVs. The statistical hypothesis is a claim about that specific case. A model is mathematical representation of the statistical hypothesis. The usefulness of having a separate term for “model” as distinct from hypothesis is apparent for Bayes factors (should one model H1 as a normal or a Cauchy with scale factor s?). One can still apply it in the frequentist case, although it depends on what one takes the statistical hypothesis to be exactly (e.g. when people calculate power using an effect size s, they imply a model of H1 which is not exactly the motivating hypothesis that “Difference in RTs is different from 0” but rather how that translates into a case where s is the minimally interesting effect size).

*REVIEWER: Relatedly, I couldn’t get my head around this sentence: “The model of H1 is a probability density function indicating how plausible different possible population effects are given the theory” This one sentence contains theory, H(ypothesis) and model; but I can’t figure out which entity is supposed to play which role, and how they relate.*

RESPONSE: Hopefully it is now clear given the definition. The “Model of H1” is a mathematical representation of the hypothesis. We can construct the model only with a theory to guide us about what are relevant effect sizes in this case. The model is not the theory though; it is guided by the theory. So we need all terms used in that sentence, they play different roles.

*REVIEWER: I have a similar confusion regarding the use of the word “predict”. For me, prediction is something that models do in data space. Here, it seems that the word is used differently (e.g., “point prediction” refers to something in parameter space). Again, I am not arguing that my perspective on prediction is superior, but given there seem to be different reasonable interpretations of the term, it might be good to make the perspective used in the paper more explicit.*

RESPONSE: I use the term prediction in its everyday sense of a claim about what will happen (under certain circumstances). The reviewer is right that this must at some point be translated into a prediction about data. But predictions can occur at all levels of generality. General relativity predicts that light can bend. That’s a general prediction, and not yet translated into data space. Eddington did translate it into data space to test the theory in a very specific context, but that translation cannot be thought of as the only way the theory predicts. It is important the theory can predict generally so that the same conceptual prediction can be translated into different data spaces for different studies.

*REVIEWER: “Therefore a p-value can never provide evidence for an effect not being there” I do agree, on some level, but in this light I was surprised to read about equivalence testing as a way to provide evidence for no effect, which is of course just two p-values, associated with two one sided tests. Maybe the cited phrase should be “Therefore a SINGLE p-value can never provide evidence for an effect not being there”?*

RESPONSE: I had thought about this issue, which is why I used the singular. But the reviewer is right, it is not clear enough (as noted also by other reviewers). In the places where this claim is made I now make explicit that that it applies when the H0 is the absence of an effect.

*REVIEWER: I don’t get why “(or at least in the range of, say, 0 to 6kg).” is between parentheses. Isn’t the fact that the hypothesis is expressed as a range exactly what differentiates this type of the previous type? I am not sure another explanation of what a Bayes factor is is needed (especially, since p-values and equivalence testing, and inference by intervals are not explained either). There are ample references for readers needing to be introduced to Bayes factors.*

RESPONSE: In this section, each type of effect size is specified – a minimal one or a rough scale of effect, and in each case, the corresponding inferential approach is named – inference by intervals or Bayes factors. Neither is explained, that is left for later. The point is to introduce the two types of effects and the inferential approaches that goes with each.

*REVIEWER: For what it’s worth: I don’t agree that specifying a prior is enough to make a theory testable, and I don’t think Lakatos, mentioned in the introduction, would either, so “in order to severely test it” and “in order to test our theories severely” seems misguided, at least to me. I recently argued how we need something more to do that: a data prior (*[*https://psycnet.apa.org/record/2019-80330-001*](https://psycnet.apa.org/record/2019-80330-001)*).*

RESPONSE: The “in order to” phrasing implies a condition but it does not imply that the condition is sufficient. But the reviewer is right that it is good to specify sufficient conditions. I now emphasize test severity with its own sub-sections, and refer to the reviewer’s paper, which is indeed very relevant and forces researchers to think about genuine predictions.

*REVIEWER: “If there is no basis for justifying an effect size, consider if one only need estimate (e.g. derive means and confidence or credibility intervals) to draw the inferences needed”. But even for “just” estimation using H1, a prior is needed, so what does switching to estimation solve?*

RESPONSE: The prior used in Bayes factors serves a different purpose from that used in estimation and therefore will rarely be the same. The prior distribution used in estimation has the purpose of allowing the most accurate estimate of parameters. The purpose of the model of H1 is to represent the predictions of a theory. So one may find it hard to say what the prediction of a relevant theory is, but still use a vague prior to estimate parameters from data. I now spell this out in the discussion.

*REVIEWER: Figure 1: Is it really necessary that the upperbound and lowerbound are identical in absolute value (as panel A seems to suggest)?*

RESPONSE: Right; it shouldn’t be necessary though it is very typical. I don’t have a good practical example where they differ, except for the case in B, though presumably there could be examples in between. I have added to the Figure caption: “The equivalence region in A is typically symmetric though need not be.” I also now use a numerical example in the smallest effect size of interest section that matches the case in B.

*I don’t know which is the best way is to cite* [*http://pcl.missouri.edu/bayesfactor*](http://pcl.missouri.edu/bayesfactor)*, but any reference omitting Morey seems wrong. Maybe contact Rouder or Morey to find out their preference?*

I now reference Morey’s R package at the same time.

**Reviewer 2**

**REVIEWER:**

*Summary: The manuscript argues that in order to obtain evidence for the null hypothesis, it is necessary to specify an effect size of interest. It continues by outlining how an effect size of interest can be specified as a smallest effect size of interest, or as a rough scale of effect size.*

*Review: I agree with the general claim of the authors that obtaining meaningful evidence for the non-existence of an effect requires the careful specification of hypotheses. However, the fact that statistical answers depend on statistical questions (i.e., Jeffreys platitude) has long been known, and thus, the argument that evidence for the null hypothesis depends on the specification of statistical models is hardly new. It remains to ask whether the manuscript can provide researchers with helpful advice regarding the specification of hypotheses. However, I get the impression that the manuscript will be more likely to create confusion than to provide actionable insights for applied researchers. Although it covers lots of different approaches to model specification, it fails to introduce any of these in sufficient depth to make them actionable. Moreover, the manuscript implicitly switches between many different statistical frameworks (Bayes factors, equivalence testing, severity testing, ROPE, …), which makes it difficult to match recommended approaches to any particular framework. Therefore, the unique contribution of the paper is unclear to me, and I do not believe that it is suitable for publication in its current form. I will give more detailed comments on the manuscript below.*

*Major:*

1. *Reading the manuscript, I have trouble telling what it adds to the current literature. As outlined before, the main claim (evidence for H0 depends on the specification of the models) is not new, and all recommendations about specifying effect sizes of interest have been described in greater detail in Lakens, Scheel, & Isager (2018), Etz et al. (2018), Dienes (2019), and Mayo (2018).*

RESPONSE: These previous works offer advice about specific issues. On the pragmatics of how one connects theory to statistics: Lakens et al only for smallest effect size of interest, and then the advice is not in general about connecting theory to statistics and therefore includes advice I argue against in this paper; Dienes 2019 is only about Bayes factors and only in the case where there are no relevant prior studies; and Mayo urges people to make the connection but does not say how. This paper brings together advice relevant to both Bayesians and frequentists specifically emphasizing the role of theory testing.

1. *REVIEWER: The manuscript implicitly touches on many different statistical frameworks (null hypothesis significance testing, equivalence testing, Bayes factors, severity testing, ROPE, …). However, the only framework the authors venture to explain in at least some detail are Bayes factors. Therefore, the reader is required to have at least a basic knowledge of almost all mentioned statistical frameworks to understand the manuscript, and judge the adequacy of the proposed methods for the specification of the effect size of interest in a particular method. I do not think that this is a realistic expectation, and even if it was, it would still be kind to the reader to explain the statistical concepts carefully instead of just hinting to background knowledge.*

RESPONSE: Inference by intervals/equivalence testing is now described in more detail.

1. *REVIEWER: The title of the manuscript is: Obtaining evidence for no effect. This makes me wonder whether all described methods can actually provide evidence for no effect. The authors themselves claim that “the use of non-significance to indicate no effect is misleading”.*

RESPONSE: This sentence has been revised to indicate a significance test using the H0 of no effect cannot indicate no effect.

On evidence – I think Bayes factors are the most justified in providing a measure of evidence, in agreement with the reviewer - and it is the method I personally routinely use. On the other hand, Fisher thought in terms of evidence, Mayo also does. Neyman Pearson is not framed in terms of evidence, as the reviewer goes on to point out. Indeed, I have published examples in several places that show how Neyman Pearson conclusions can be against the evidence. Nonetheless, it is striking how similar the procedures are in inference by intervals, whether one uses Mayo’s error statistics, Neyman Pearson CIs or Bayesian CIs. The outcomes would be very similar in all cases and correlate with which way the evidence points.

*REVIEWER: However, later, they rather explicitly refer to equivalence testing and severity testing, both methods deeply rooted in significance testing, and even null hypothesis significance testing with an a-priori power analysis itself (see p. 5) as methods to obtain evidence in favor of the null hypothesis (see p. 4-6 of the manuscript). A similar issue applies to the ROPE method, which investigates the posterior probability that a parameter lies within a certain interval. Interestingly, with the right constellation of prior distributions and data, the posterior distribution could lie almost entirely within the ROPE interval, but there would be evidence for the alternative in the Bayes factor. With a narrow definition of statistical evidence, I would therefore argue that only the Bayes factor can measure the statistical evidence for the null hypothesis. This makes me wonder whether the focus of the manuscript is actually providing advice on how to “obtain evidence for no effect” or whether it is to provide advice on how to specify effect sizes of interest.*

RESPONSE: The point of the paper is to indicate that the only way to obtain evidence for no effect is by having scientifically relevant effect sizes of interest. I take it the reader sometimes wants to seek the former – and that’s why they need the latter. I doubt readers will seek out the latter for its own sake. What I felt was missing was an account that gave advice on how to do the latter, advice not wedded to any particular school of statistics. As advice on how to do the statistical part is already plentiful, what was missing was what one might call the pragmatics of statistical inference, namely, how one connects theory to particular tests. Although the reviewer starts by saying there is nothing new in this, I see these pragmatics as something almost entirely missing from substantive published papers in disciplines that use statistical inference. So my thought was that a paper concentrating on these pragmatics, and showing they are broader than any particular school of inference, was needed. I have added a few sentences to the first paragraph about this. As the reviewer says, there have been papers on specifying effect sizes of interest, but they have always been of limited scope.

1. *REVIEWER: The manuscript separates two ways to specify an effect size of interest: Specifying a smallest effect size of interest or specifying a range of plausible values. In my opinion, presenting these approaches as independent of each other is misleading (see e.g., claims in the manuscript that they are “different models”, “two ways”).*

RESPONSE: When one views approaches to getting evidence or support for no effect they fall into these two camps, with what I call inference by intervals cutting across schools of statistics. Further, specifying minimal effect sizes and rough scales of effect are practically very different enterprises, that can’t be assimilated into the same approaches. I now make this clearer, e.g. indicating a couple of the heuristics for obtaining a rough scale of effect cannot be used for minimally interesting effect sizes, as suggested by another reviewer.

*REVIEWER: In fact, what the authors describe, is the specification of an interval null hypothesis, and the specification of an alternative hypothesis that is represented by a model with a probability distribution on parameters. However, none of these hypotheses is ever specified alone. The competing hypothesis is either implied by the method (e.g., in equivalence testing) or needs to be specified separately (e.g., in Bayes factors). The assumption that researchers can decide to specify only one hypothesis in a hypothesis test, is incorrect.*

RESPONSE: Quite right. To make this clearer I have added to Figure 1 “Note for every test two models are specified, one for H0 and one for H1 (in the diagram, an interval H0 for inference by intervals and a point H0 for the Bayes factors).”

*REVIEWER: Furthermore, the implicit switch between the specification of H0 and H1 will most likely confuse readers, especially considering the fact that the introduction of the manuscript only strongly points towards the specification of the alternative hypothesis (see the snow leopard and the “animal in the room” example).*

RESPONSE: Right, you could say that the point of the manuscript is what to do with H1, though H0 is simultaneously specified for inference by intervals.

1. *REVIEWER: In a related note, although the manuscript discusses the importance of specifying H1 in order to obtain meaningful evidence for H0, it misses the point that the inverse is also true. The specification of H0 (e.g., is it a point null or an interval?) also influences the evidence obtained for H1. With the current structure of strictly separating the specification of H0 and H1, I think it is difficult to bring this point across.*

RESPONSE: The second paragraph introducing Bayes factors says “Typically, H0 is represented as a point prediction; for example, that of no effect. In that case, no minimal interesting effect need be specified; this can be approximated as close to zero, which will be a good enough approximation if the true minimal interesting effect is smaller than the standard error of the effect (otherwise a null interval can be specified: Morey & Rouder, 2011; Palfi & Dienes, 2019; see Skora et al., 2020, for actual use of a null interval hypothesis).“ The last reference has been added as an actual practical use of null intervals.

1. *REVIEWER: As a reader, I was somewhat overwhelmed by the number of different examples. For example, just take section i) of the SESOI section (p. 5). Here, we have: depression treatment, change in pain, positive/negative affect rating, redness of faces, and effectiveness of cognitive training within half a page of text. However, none of these examples were really helpful because they were too superficial to actually understand the method in depth (e.g., How many end users were asked? Was it an interview or a survey? How did the original authors arrive at their cut-off value for an interesting effect size?, …). A reader who would want to apply the method, or even find out whether it is applicable to their research, would need to go back to the original literature to figure out the details. As a reader, I would therefore really appreciate if the authors could drastically reduce the number of their examples in exchange for more depth (e.g., show how the SESOI is calculated with the different methods and applied to a certain statistical framework).*

RESPONSE: The number of examples for minimally interesting effect size has been dramatically cut and the reported examples fleshed out.

1. *REVIEWER: Throughout the manuscript, the authors make many implicit assumptions regarding the hypothesis testing scenarios and the model specification that are not always obvious for the reader. For example, the claim that with Bayes factors, researchers need to (only) specify the scale of an effect, makes at least the assumption that (1) this effect is defined under H1, (2) the prior on the effect under H1 follows a continuous prior distribution, (3) this continuous prior distribution has a scale parameter, (4) only the scale parameter needs to be adjusted, (5) the complementary H0 is pre-defined, for example as a point prior. This is a very limited view on Bayes factors because, as the authors write themselves, “a Bayes factor can use any models that seem worthwhile to compare”. These implicit assumptions make it difficult for the reader to see the generalizability of claims. I would therefore recommend to the authors to clarify the assumptions made throughout the manuscript, and maybe find an application example that makes it clear to what applied scenarios (e.g., t-test, regression) the methods can be applied to.*

RESPONSE: The introduction to Bayes factors says: “for the sake of argument consider a normal distribution centred on zero (Figure 1 C). If the theory predicts a direction, the half of the distribution below zero can be removed (thus, by convention, a positive effect is defined as in the direction predicted by theory) (Jeffreys, 1948; Wagenmakers, 2020) (Figure 1 D). The half-normal distribution requires its standard deviation being set; the standard deviation defines how steeply the curve drops off and thus sets the scale of effect predicted. In sum, the half-normal model of H1 (introduced by Dickey, 1973) assumes that the rough scale of effect is that given by its standard deviation, that smaller effects are more likely than larger ones, and that a rough maximum effect expected is about twice the standard deviation (see Dienes & McLatchie, 2018, for justification of why these assumptions may widely hold). If these assumptions appear to represent one’s theory adequately, then the task simplifies to specifying the rough scale of effect.”

1. *REVIEWER: Overall, the manuscript argues in favor of more transparency in model specification. This becomes especially clear when the authors demand that the “numerical value for the effect should come from a public place one can point to (for example data), so that other people can criticize the reason for choosing that value” (p. 4). I absolutely agree with this sentiment. However, it is difficult to entirely remove arbitrary decisions from the research process (see also the current debates around multiverse analyses), and it is important to note that the proposed methods for effect size definition are no exception. For example, the degree of violation in assumptions (see section iv p. 6) requires the specification of a cut-off for assumption violation, or the confidence interval method requires the specification of the width of the confidence interval, as well as the determination of “interestingness” (p. 6). The question is, what makes these methods better than committee decisions or benchmarks (see p. 4), which the authors criticize for their intransparency? To me, it seems like the latter methods can be reported equally transparently as the recommended methods, and all methods entail a certain degree of arbitrariness. I wonder what the authors think about this and would appreciate if they could comment some more on the issues of transparency and sensitivity to arbitrary decisions.*

RESPONSE: The reviewer is right that in all cases a judgment has to be made about its suitability, and do we stop bottoming out here. Yet there is a difference between specifying the predictions of the theory (minimally interesting effect size, scale factor), on the one hand, and the strength of evidence/degree of error control acceptable in the current case for drawing conclusions, on the other (the latter being a practical matter concerning how we move forward – e.g. was a confound controlled enough that we can get on with applying the theory?). Predictions of the theory should refer to matters relevant to theory as much as possible. Predictions of theories should not depend on committee votes per se. That is not to say an expert’s opinion is not valuable; but in itself a prediction of a theory formed by intuition is a promissory note that objective reasons are there to be found if we now look.

This is incidentally one way the paper differs from many others (in terms of the reviewer’s first point about what does the paper add). Most papers on Bayesian statistics phrase the issue as objective defaults vs subjective opinion. Objective defaults do not refer to the predictions of any theory. Subjective personal probabilities are also about what one feels, not what a theory predicts. My view is: It is not exactly either. Rather, I apply the critical rationalism of the Popperian tradition to statistics.

Minor

1. *REVIEWER: In several places, the manuscript displays a very strict falsificationist attitude that is, in my view, not consistent with research reality in most disciplines. For example, the manuscript starts with the sentence “If there is no way for a theory to be shown wrong by a study, the putative test of the theory is no test at all”. However, how often does it happen that a theory can actually be “shown wrong”, i.e., finally disproved, by data? Even the most elegant theories are most often formulated as probabilistic models, which means that there is always some (small) likelihood that the data did occur under the model. How could a result of one single study therefore “show a theory wrong”? I would argue that it is, in fact, impossible. Additionally, most scientists would agree with George Box’s statement that “all models are wrong”. What would be the value of showing that a model is wrong in this case? My point is that although I agree that it is beneficial to be able to obtain evidence in favor of the null hypothesis, this does not warrant a strict all-or-nothing falsificationist view which is not in line with research reality. I would suggest that the authors tone down these statements and instead argue for the many benefits of finding (gradual) evidence for the null.*

RESPONSE: In the first sentence, the phrase “shown wrong” has been replaced by the notion of data counting against a theory.

How frequently one takes theories to be falsified by a study (and its replications) depends on the generality of the theory considered – scientists seem to take low level theories as routinely falsified. Higher level ones only sometimes are (I could give several examples from the fields I work in), and are more likely worn down as the reviewer suggests. Popper’s notion of falsifying theories was not a black-and-white one-study-does-it – in his 1959 he pointed out that a falsifying result is actually a low level theory of the conditions under which the (falsifying) outcome is obtained, and this needs corroboration by replication. Further, on Popper’s mature account, any falsification is in principle provisional, as all claims are open to criticism. Still, the ideal of a study is that it potentially falsifies (it is precisely how a study may fail in this regard that takes up much of the criticism in the discussion about the weakness of auxiliary assumptions), and so this is what we strive for.

1. *REVIEWER: The short discussion of Bayes factor design analysis on p. 9 seems somewhat out of place. BFDA is not a method to find an effect size of interest for “when there are no prior studies”, as the section title suggests. However, I think a short discussion of the implications of model specification (e.g., specifying H0 as an interval or as a point, specifying a certain width of the prior distribution in H1) on the required sample size would be an interesting addition to the manuscript.*

RESPONSE: The sections are now better labelled with a new heading structure.

1. *REVIEWER: On p. 3, the authors argue that “neither can evidence for no theoretically interesting effect be provided by […] a default Bayes factor”. In my opinion, this is not true. A Bayes factor is always a measure of relative evidence in favor of one model over another. A default Bayes factor simply displays a lot of uncertainty about the effect size under the H1 model. If a comparison of this H1 model to a (typically point) H0 model is not of interest to a researcher, then this means that the models were misspecified for the specific application, but the Bayes factor can still display evidence in favor of H0.*

RESPONSE: Right, any Bayes factor provides a measure of evidence for the specified models. But if the model of H1 is not a good representation of the theory then the measure is irrelevant. Hence my phrasing in terms of “a theoretically interesting effect”.

1. *REVIEWER: Reading the conclusion of the manuscript (last paragraph p. 9), I find it way too strongly formulated. Since the manuscript describes methods and ideas that already (long) exist, I don’t see how it can argue for “radical change”. Also, it is not clear to me what the authors mean by “scientific hallucinations” in the context of the paragraph. It sounds like they argue that without a proper specification of a theory, we can neither find evidence for or against it. However, nothing of this seems to require any delusion on the side of the scientist.*

RESPONSE: The term “radical” has been removed as well as the term “hallucination”. I have now tried to express more precisely how I see what needs to be changed.

**Reviewer 3**

*REVIEWER: In my view, the paper collects several approaches to this question, but cannot communicate these ideas in a clear and accessible manner. I am wondering who is the intended audience of the paper: Statistical experts who are looking for best practices? Or practitioners who want to improve their statistical inferences and draw better conclusions? In both cases, the paper would benefit from more clarity in the way it presents its ideas and methods as well as more illustrative examples that guide researchers who want to adopt the strategies outlined here. In its current form, the paper was sometimes difficult to follow, seems to be unfinished in some parts, and had a couple of typos and grammatical mistakes.*

*Overall, psychologists can benefit from papers that provide a clear guidance on how to analyze experimental and observational data, especially in the context of evaluating null effects. The paper contains several useful ideas — it should be reworked to present these ideas in a more approachable manner so readers can benefit from it in their actual practice.*

*I have collected more specific comments below. (A note on the manuscript formatting and/or Collabra‘s editorial system: The manuscript does not have page numbers nor line numbers, making commenting on particular sections very difficult. In my comments I will refer to the pages as they appear in the PDF, i.e. page 1 is the title page, page 2 contains the abstract, and page 3 is the beginning of the actual manuscript.)*

*Page 3, paragraph 3: „A p-value is calculated without reference to the possible size of an interesting effect. Therefore a p-value can never provide evidence for an effect not being there.“ The first sentence is technically true, but the context of performing a statistical test is relevant: For example, in an equivalence test the p-value is very much to be considered in the context of the absence of an effect. In a paper looking at the interpretation of statistical results in both significance testing and Bayes factors a more nuanced representation of the statistical methods (e.g. in context of either Fisherian or Neyman-Pearsonian testing) would be favorable.*

RESPONSE: I have now changed the quoted sentence to refer to a p value where the H0 is no effect.

REVIEWER: P*age 3, last paragraph + page 4, Figure 1: The purpose of this figure is not really clear as the text does not match the examples provided. While the first textual example matches illustration B (an empirical effect size larger than the smallest effect of interest is consistent with the theory and thus corroboration for the theory), the second example in the text refers to the case where the theoretically plausible effect size is within an interval and any effect size estimate outside of this interval would be refutation of the theory. The illustration refers to this second equivalence by providing „models of theories for Bayes factors“, but it does not seem necessary at this point to consider this problem only in terms of probability densities. If an illustration to the examples is desired, it should be presented - at this point in the manuscript - in the same contexts for both examples so readers can better understand the illustration without extensive prior experience with both Bayes factors and significance tests.*

RESPONSE: I now indicate in the legend to the Figure what rough range of effects is implied by the models used for Bayes factors. Note that the text emphasized the scale factor rather than the range.

*REVIEWER: Page 4, paragraph 1: The author advocates for the use of „objective reasons“ when specifying smallest effect sizes of interest, but the negative examples are not convincing to me. For example, using common boundaries such as „a Cohen‘s d of 0.5“ can be one approach when the theory is vague and unspecific (I concur that such a theory might not be strictly testable and researchers should refrain from trying to use significance testing in such circumstances, but this seems to be not the focus of the present paper).*

RESPONSE: If the theory is vague why would one chose any particular value? Why 0.5?

*REVIEWER: Furthermore, expert committees can, if conducted transparently, also provide a solid justification for the present case.*

RESPONSE: And transparent must mean that the reasons for the decisions are available. Then what matters are simply the reasons.

*REVIEWER: Providing examples of „ad hoc justifications“ for effect sizes, criticising weak theories that do not predict a minimally interesting effect size would be more convincing at this point.*

RESPONSE: Though I think most theories do not predict minimally interesting effect sizes, many can nonetheless predict the sort of effect size that should occur, thus making them falsifiable.

*REVIEWER: The sentence „One can evade the problem by only reporting p-values and not interpreting non-significant results.“ is for me an example where the paper would benefit from further polishing and added clarity: An idea is presented that is technically true, but a very limited perspective. For example, non-significant results from both traditional significance tests and equivalence tests paint a different picture. This and the follow sentences also feel somewhat disconnected to the previous discussion on objective reasons for possible effect sizes.*

RESPONSE: This sentence and the following have been slightly rephrased to emphasize the point that non-significance does not provide evidence for H0, thereby not inadvertently impugning equivalence testing.

*REVIEWER: Page 6, paragraph 2: The method of using „calibration“ is interesting and a very valuable addition to the paper. Unfortunately, this section is very brief. The example could be extended and described in more detail, so researchers can adopt this method more easily.*

RESPONSE: The reviewer’s positive comment is appreciated. The calibration examples have been extended.

*REVIEWER: Pages 4-6: A more general point on finding smallest effect sizes of interest: The paper could also benefit from a discussion on whether the smallest effect size of interest is an useful tool. Why, for example, is a change of 0.30 relevant, but a change of 0.29 is not? Because on average 0.30 corresponds to a noticeable change? Depending on the context, it might as well have been 0.28 – measurement error or uncertainty in general might make this boundary arbitrary. Significance testing does require a single cut-off point, but a brief discussion of this issue seems to be fitting to the general context of the paper (especially considering that Bayes factors are discussed later).*

RESPONSE: This is a good point and the discussion now phrases this point in terms of how Bayes factors are often criticized for arbitrariness because they need a model of H1; yet such critics seem to fail to realize that if they use inference by intervals they also need to specify a model – and I personally believe they have a harder and more arbitrary task, as the reviewer implies. I also discuss the notion of a grey area in inference by intervals in order to smear the point minimally interesting effect size.

*REVIEWER: Page 7, paragraph 2: It seems to be limited to consider null hypotheses in Bayes factors only as point null hypotheses. In practice it might be most common, but as the author acknowledges there are alternatives available (see also my previous comment)*

RESPONSE: I quite agree. As the reviewer points out, I acknowledge alternatives. In fact, I think some interesting problems can non-obviously be solved by a null interval in Bayes factors, but this requires a different paper to go through, one on my list to write. I have added a reference to a paper that makes good use of a null interval hypothesis in its substantial domain.

*REVIEWER: Page 8f: The paper only briefly explains why now a scale of effect and not a smallest effect size is considered. Readers not as acquainted with the technicalities of Bayes factors might not be to follow as easily. Furthermore, a more detailed discussion on how the two approaches compare would be helpful: In what ways are the approaches similar (e.g. calibration) and what differences exist? Why, for example is the „basic effect heuristic“ not useful for determining a smallest effect size?*

RESPONSE: I have added a paragraph in the discussion comparing the approaches, and also highlighted which of the heuristics do not apply to a minimally interesting effect size.

*REVIEWER: Page 8, section „(ii) Basic effect heuristic“: This is another example where a more verbose explanation would be helpful to guide readers in using this method.*

RESPONSE: Verbosity has been greatly increased, as helpfully recommended, by describing a published study that used the heuristic.

*REVIEWER: Page 8, section „(iii) Calibration“: While the previous example for calibration was straight forward (even when it could be expanded further), this example is more difficult to follow. Furthermore I did not find the section convincing in how participants‘ expectations should be a valid calibration for an actual change in outcome. This rests on some assumptions that should at least be mentioned here.*

RESPONSE: The calibration example ahs been changed so as to be based on real data, and it has been fleshed out, with discussion of assumptions.

*REVIEWER: Page 9, last paragraph: I am not sure if the paper actually argues for that and if this is really a „radical change“. At least it seems far less radical compared to initiatives to ban statistical significance.*

RESPONSE: The term “radical” has been removed.

**Editor Final Decision: Revise & Resubmit**

May 26, 2021

Dear Zoltan Dienes,

I sent out your revised manuscript to two of the three original reviewers, and have now received the reviews of your manuscript, “Obtaining evidence for no effect”. I also independently read the response letter and manuscript before consulting these reviews. I agree with both reviewers that the manuscript has notably improved. I particularly like the changes to the structure and the examples throughout the manuscript. My reading of your responses and the current form of the manuscript are more in line with reviewer 2 than reviewer 1. Therefore, I encourage you to submit a revised version for further consideration at Collabra: Psychology.

I encourage you to carefully read and respond to both reviewers’ excellent comments. For example, the first three points by reviewer 1 on simplifying the examples could very much improve the manuscript in its current form. I also think adding a figure for the pain example is an excellent idea as I myself struggled through that section. In addition, both reviewers had opinions on the distinction between estimation and testing. Even if you do not agree with them, I hope you expand on your rationale in the manuscript itself to clarify the position in the paper for the readers.

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. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at 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,

Julia Haaf

# Reviewer 1

##### 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.

I think the revised version of the manuscript has visibly improved over the original version. The structure of the article is now much more comprehensible, and I appreciate that more room has been given to a smaller number of examples.

This being said, I still have doubts about the contribution that the article provides (see my previous points 1 and 6). The theoretical argument (uniting different statistical frameworks from a theory testing perspective) is certainly interesting, but I am not sure whether it has been sufficiently fleshed out to be picked up by the readers. For me, the manuscript still reads a lot like “in framework X you do A, in framework Y you do B”, and this information can be found elsewhere in greater depth. However, I believe that, in the end, this is an editorial decision, so I won’t comment on this anymore should there be further rounds of revisions.

There are a few more specific comments that I would like to make on the revised manuscript. You can find these below:

(1) It seems to me that some of the examples are overly complex, or may not be representative of typical analyses that substantive researchers perform. In particular, in the replication example, a half-normal prior is put on an effect that is expressed in percentages above chance level. Apart from the confusion this creates because there are many percentage values in the explanation (e.g., 95% CI does not refer to the effect scale but to the width of the CI), it is simply not clear to me what test was calculated here. The typical test for testing whether someone is correct above chance level would be a binomial test with a beta prior or, if we want to model the decision, a signal detection model. Placing a half-normal prior on differences in percentage that are necessarily restricted to 100%-chance level seems very odd to me because the support of the prior is larger than the possible values.

(2) The example for the basic effect heuristic also seems overly complicated to me. The chosen model is a logistic mixed-effects model. This is a fairly complex model – I am not sure how many readers will be familiar with it. I wonder if a simpler model wouldn’t be suited better to explain the heuristic. Additionally, it is not clear for which parameter the prior distribution is formulated, and what models are exactly compared (do the authors use a Savage-Dickey density ratio on the parameter or do they compare two full models? How is the null model defined? What are the priors on the other parameters?).

(3) I did not understand what was regressed on what in the pain example, and what the goal of the study (and therefore the calibration) was. There are a lot of variables (pain, expected pain, mindfulness condition, different (?) placebo conditions) and all of them measured in different studies – I tried to follow but I got very confused. Assuming that I am not the only one struggling with this paragraph, I would recommend to revise the structure of the paragraph and perhaps include a figure that makes the relationships between the measured quantities clear.

(4) It has to be made clear that the list of strategies to formulate prior distributions or a null interval is not exhaustive.

(5) In several places, the article equates the H0/H1 model with the interval or prior distribution specified. It needs to be made clear that this is a simplification. There is always a likelihood function involved, and any prior distribution or decision interval is defined in the context of this likelihood. In the Bayesian models, there are additionally other prior distributions involved (see my comment #2), and the model fit also depends on the definition of these priors (e.g., the prior on the scale of the effect in the logistic mixed model).

(6) I strongly disagree with the sentence “If there is no real theory to test, one can still estimate”. Estimation is only sensible in the context of a theory, i.e., a model specified by a likelihood function. How else would we know what to estimate? As far as I know, this is true for any statistical framework. Jeffreys even argues that hypothesis testing should always precede estimation, arguing that we first need to establish the existence of an effect before we can estimate its size.

(7) In the “(i) Replication” section of the article, the authors state that Verhagen and Wagenmakers (2014) and Ly et al. (2019) present similar methods to the one described by the authors. I would argue that their methods are too different from what is presented here to warrant this statement. Verhagen’s and Ly’s methods use the posterior of the original study (or an approximation thereof) as a prior for the replication. This means that they arrive at a noncentral prior distribution whose scale is based on the uncertainty about the parameter in the original study (which will usually be smaller if the original study was large). In contrast, the proposed method to formulate a prior for replication studies in this article assumes a half-normal distribution centered on zero (i.e., central), where the scale factor is determined by the point estimate of the effect size in the original study (i.e., not influenced by the sample size of the original study).

(8) Figure 1: The x- and y-axes need to be in the proper place and labeled. In panel C, the y axis seems to be hovering above ground, the y axis in panels A, B, and D is not labeled, the x axis is not properly labeled in any of the panels (I assume “effect size - >” in the middle of the plot is supposed to be the axis label?). The figure caption also needs some work: It sounds as if panels C and D show inference by intervals, which is incorrect.

##### 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.

The manuscript has much improved. I have a few remaining (small) comments.

In fig 1, I think writing “S=rough (etc)” is more clear than the current “rough (etc)=S”

“reasons for specifying a smallest effect size of interest” and “reasons for specifying a rough expected scale of effect.” I am not a native English speaker, so might have the wrong connotations, but to me, “reasons” reads like “why it is necessary to specify a minimum or a scale”, whereas I think the author refers to the basis/heuristics/sources/justifications/rationales for setting a minimum or a scale.

“Which approach one chooses depends on whether a minimally interesting effect size or the rough scale of effect is easier to justify using publicly available reasons” is wrong, in my opinion. There are other reasons than the availability of justifiable prior info to prefer, say, Bayes factor over equivalence testing. Maybe the author means “could depend” or “should depend” or “in part depends” instead of “depends”.

“Determine whether the predictions of the model (represented by what he calls a “data prior”) are implausible in the light of background knowledge.” is not entirely accurate (the data prior is not the predictions). I would phrase it like this: “Determine whether the predictions of the model are implausible, using a “data prior”, representing which outcomes can be considered plausible in the light of background knowledge.”

“So the problem of specifying a “prior” has to be confronted and cannot be avoided if one wishes to potentially be able to obtain evidence for no effect.” Vey nice. Just banning a term from one’s vocabulary (like frequentists do) does not make it go away.

While I agree that in the absence of information, estimation is safer than computing Bayes factors, I don’t agree with the argumentation used by the author: “The prior distribution used in estimation has the purpose of allowing the most accurate estimate of parameters.”. I would argue that, when prior are vague, the exact implementation of the vague prior has less impact on the parameter posterior than on the Bayes factor, which seems a good rationale for preferring estimation in such a case. But I don’t see how priors serve a different purpose. (I am just sharing my thoughts; I realize it is the author’s paper, not mine, so it is up to the author to decide whether he wants to expand or change the argumentation.)

“In sum, this article argues a change in approaching statistical testing is needed in order to test our theories severely:” This is not the message I got away with. Is the intended change the move from NHST to testing by intervals and/or using Bayes factors? Both approaches have had their share of advocates and practicioners, so I am not sure how advocating these methods constitutes “arguing for a change” in statistical testing. Maybe the intended change is something else (possibly relating to what counts as justifiable sources of priors, though I am not sure). In that case, I would recommend spelling out more clearly exactly what the author argue that should be changed. (That being said, I think the paper has value without arguing for a change: In my reading, it provides a timely and convenient overview of sources for setting priors when computing Bayes factors or doing inference by intervals, which is valuable in itself).

signed,
wolf vanpaemel

##### 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**

Aug 13, 2021

EDITOR: *I encourage you to carefully read and respond to both reviewers' excellent comments. For example, the first three points by reviewer 1 on simplifying the examples could very much improve the manuscript in its current form. I also think adding a figure for the pain example is an excellent idea as I myself struggled through that section. In addition, both reviewers had opinions on the distinction between estimation and testing. Even if you do not agree with them, I hope you expand on your rationale in the manuscript itself to clarify the position in the paper for the readers.*

The reviewers make some valuable points which I do indeed address below. I agree the pain example had become complex; I have now substantially simplified by not referring directly to real data. I haven’t added a figure of what is regressed on what as I thought this was more simply expressed in words (but easily added if you wish to give the nod given how it is now). Without the added complexities I hope the example is now clearer. I have also added more in the discussion on the distinction between estimation and hypothesis testing.

Reviewer 1

*REVIEWER: I think the revised version of the manuscript has visibly improved over the original version. The structure of the article is now much more comprehensible, and I appreciate that more room has been given to a smaller number of examples.*

*This being said, I still have doubts about the contribution that the article provides (see my previous points 1 and 6). The theoretical argument (uniting different statistical frameworks from a theory testing perspective) is certainly interesting, but I am not sure whether it has been sufficiently fleshed out to be picked up by the readers. For me, the manuscript still reads a lot like “in framework X you do A, in framework Y you do B”, and this information can be found elsewhere in greater depth. However, I believe that, in the end, this is an editorial decision, so I won’t comment on this anymore should there be further rounds of revisions.*

RESPONSE: I think the reviewer is right that “The theoretical argument (uniting different statistical frameworks from a theory testing perspective) is certainly interesting, but I am not sure whether it has been sufficiently fleshed out to be picked up by the readers.” While I see the framework permeate every paragraph, the underlying philosophy needs to be made more explicit. I have added two paragraphs to the discussion to highlight this point.

*REVIEWER: There are a few more specific comments that I would like to make on the revised manuscript. You can find these below:*

1. *It seems to me that some of the examples are overly complex, or may not be representative of typical analyses that substantive researchers perform. In particular, in the replication example, a half-normal prior is put on an effect that is expressed in percentages above chance level. Apart from the confusion this creates because there are many percentage values in the explanation (e.g., 95% CI does not refer to the effect scale but to the width of the CI), it is simply not clear to me what test was calculated here. The typical test for testing whether someone is correct above chance level would be a binomial test with a beta prior or, if we want to model the decision, a signal detection model. Placing a half-normal prior on differences in percentage that are necessarily restricted to 100%-chance level seems very odd to me because the support of the prior is larger than the possible values.*

RESPONSE: To avoid confusion with so many percentages, I have changed the units to milliseconds in this example. (Note, in terms of the original example, a percentage above chance, with a normal likelihood, may be a good approximate model of accuracy when dealing with a group of subjects; the reviewer is right about fitting a binomial for a single subject.)

1. *REVIEWER: The example for the basic effect heuristic also seems overly complicated to me. The chosen model is a logistic mixed-effects model. This is a fairly complex model – I am not sure how many readers will be familiar with it. I wonder if a simpler model wouldn’t be suited better to explain the heuristic. Additionally, it is not clear for which parameter the prior distribution is formulated, and what models are exactly compared (do the authors use a Savage-Dickey density ratio on the parameter or do they compare two full models? How is the null model defined? What are the priors on the other parameters?).*

RESPONSE: The reviewer was right, there was some needless ambiguity in the original description. I have clarified a few points, but not all the reviewer’s questions. The details of the full analysis are available in the original paper (an advantage of referring to a real example); the point of the example for the purposes of this paper is how one gets a rough expected effect size to model H1, so that is the issue that I have attempted to clarify.

1. *REVIEWER: I did not understand what was regressed on what in the pain example, and what the goal of the study (and therefore the calibration) was. There are a lot of variables (pain, expected pain, mindfulness condition, different (?) placebo conditions) and all of them measured in different studies – I tried to follow but I got very confused. Assuming that I am not the only one struggling with this paragraph, I would recommend to revise the structure of the paragraph and perhaps include a figure that makes the relationships between the measured quantities clear.*

RESPONSE: I have now considerably simplified the pain example by not using real data. I agree that, by virtue of using real data, the example introduced more complexity than necessary just to illustrate the main point. (I often like real data for examples when they are available, as a heuristic is only as good as it applies in reality, warts and all. But I can now refer people to the arXived version of the previous submission, which I now do in a footnote.)

1. *REVIEWER: It has to be made clear that the list of strategies to formulate prior distributions or a null interval is not exhaustive.*

RESPONSE: Both lists are now introduced by being described as non-exhaustive.

1. *REVIEWER: In several places, the article equates the H0/H1 model with the interval or prior distribution specified. It needs to be made clear that this is a simplification. There is always a likelihood function involved, and any prior distribution or decision interval is defined in the context of this likelihood. In the Bayesian models, there are additionally other prior distributions involved (see my comment #2), and the model fit also depends on the definition of these priors (e.g., the prior on the scale of the effect in the logistic mixed model).*

RESPONSE: I have added a new paragraph, the penultimate in the discussion, to point out that many more important choices need to be made in statistical inference than just the issue considered in this paper.

1. *REVIEWER: I strongly disagree with the sentence “If there is no real theory to test, one can still estimate”. Estimation is only sensible in the context of a theory, i.e., a model specified by a likelihood function. How else would we know what to estimate? As far as I know, this is true for any statistical framework. Jeffreys even argues that hypothesis testing should always precede estimation, arguing that we first need to establish the existence of an effect before we can estimate its size.*

RESPONSE: I have changed that specific sentence to “If there is only a sense of credibility that a parameter is relevant, one can still estimate.” In terms of the second half of the reveiwer’s paragraph, quoting from a new paragraph in the Discussion, “The problem of obtaining evidence for something not being there is not solved by not using a model of H1, … nor by using someone's default model of H1. That is because the evidence for something not being there is only as good as the grounds for claiming the effect, should it be there, is of the size modelled in the model of H1.” So Jeffreys testing using a e.g. default prior with a scale factor of a Cohen’s d of 1, and obtaining evidence for H0, may be evidence for H0 against THAT model – but why is that of any relevance to a particular scientific problem? One can answer the last question only if one has a theory. So given any credibility that the parameter is relevant, one can estimate - but not thereby obtain evidence for H0 relative to a model that is justifiably relevant to the context.

I have just checked Jeffreys book; see first paragraph of Chapter V and his take is fairly nuanced, and not an injunction to always hypothesis test before estimating. See also first few paragraphs of his estimation chapter, where he estimates without hypothesis testing, because he takes the general nature of the phenomenon for granted. Thus, in the current article, I have now presented an argument for the case for the position contrary to me, in order to refute it: “….One recommendation might be to find evidence for an effect before estimating it, to make sure there is something to estimate. On this approach, if there evidence for H0, one would not estimate. In fact, one can only obtain evidence for nothing being there given grounds for claiming what size it would be were it to be there. So a theory, however minimal, is always needed for hypothesis testing. One always estimates parameters if they are reported at all; in addition, one may also hypothesis test given a theory to test. Both estimation and hypothesis testing may use priors. [then the paragraph continues as before] The prior used in Bayes factors (i.e. the model of H1) serves a different purpose from that used in estimation and therefore will rarely be the same. …” The sight I have often seen of no estimates given when a result is non-significant is in my opinion one we must eradicate from the literature. (And likewise, though I rarely see it, for when B < 1/3, etc.)

1. *REVIEWER: In the “(i) Replication” section of the article, the authors state that Verhagen and Wagenmakers (2014) and Ly et al. (2019) present similar methods to the one described by the authors. I would argue that their methods are too different from what is presented here to warrant this statement. Verhagen’s and Ly’s methods use the posterior of the original study (or an approximation thereof) as a prior for the replication. This means that they arrive at a noncentral prior distribution whose scale is based on the uncertainty about the parameter in the original study (which will usually be smaller if the original study was large). In contrast, the proposed method to formulate a prior for replication studies in this article assumes a half-normal distribution centered on zero (i.e., central), where the scale factor is determined by the point estimate of the effect size in the original study (i.e., not influenced by the sample size of the original study).*

RESPONSE: The sentence reads “Verhagen and Wagenmakers (2014) and Ly, Etz, Marsman, and Wagenmakers (2019) present related methods.” The following sentence indicates how they are related: “The theory tested is that implied by any empirical paper: That the methods of the original study (as given in the Methods section) describe a procedure for obtaining the sort of effect obtained (as reported in the Results section).“

1. *REVIEWER: Figure 1: The x- and y-axes need to be in the proper place and labeled. In panel C, the y axis seems to be hovering above ground, the y axis in panels A, B, and D is not labeled, the x axis is not properly labeled in any of the panels (I assume “effect size - >” in the middle of the plot is supposed to be the axis label?). The figure caption also needs some work: It sounds as if panels C and D show inference by intervals, which is incorrect.*

RESPONSE: x and y axes labelled for the Bayes factors cases and x-axis labelled for inference by intervals; the meaning of the blocked spaces for inference by intervals was already indicated. A semi-colon has been put in the figure caption to avoid the misreading the reviewer mentions.

*Reviewer 2*

*The manuscript has much improved. I have a few remaining (small) comments.*

*In fig 1, I think writing “S=rough (etc)” is more clear than the current “rough (etc)=S”*

RESPONSE: Now changed as suggested.

*REVIEWER: “reasons for specifying a smallest effect size of interest” and “reasons for specifying a rough expected scale of effect.” I am not a native English speaker, so might have the wrong connotations, but to me, “reasons” reads like “why it is necessary to specify a minimum or a scale”, whereas I think the author refers to the basis/heuristics/sources/justifications/rationales for setting a minimum or a scale.*

RESPONSE: ”reasons” changed to “heuristics”.

REVEIWER: “Which approach one chooses depends on whether a minimally interesting effect size or the rough scale of effect is easier to justify using publicly available reasons” is wrong, in my opinion. There are other reasons than the availability of justifiable prior info to prefer, say, Bayes factor over equivalence testing. Maybe the author means “could depend” or “should depend” or “in part depends” instead of “depends”.

RESPONSE: “in part depends” now used, as suggested.

*REVIEWER: “Determine whether the predictions of the model (represented by what he calls a “data prior”) are implausible in the light of background knowledge.” is not entirely accurate (the data prior is not the predictions). I would phrase it like this: “Determine whether the predictions of the model are implausible, using a “data prior”, representing which outcomes can be considered plausible in the light of background knowledge.”*

RESPONSE: The sentence suggested by the reviewer has now been used.

*REVIEWER: “So the problem of specifying a “prior” has to be confronted and cannot be avoided if one wishes to potentially be able to obtain evidence for no effect.” Vey nice. Just banning a term from one’s vocabulary (like frequentists do) does not make it go away.*

RESPONSE: Thanks!

*REVEIWER: While I agree that in the absence of information, estimation is safer than computing Bayes factors, I don’t agree with the argumentation used by the author: “The prior distribution used in estimation has the purpose of allowing the most accurate estimate of parameters.”. I would argue that, when prior are vague, the exact implementation of the vague prior has less impact on the parameter posterior than on the Bayes factor, which seems a good rationale for preferring estimation in such a case. But I don’t see how priors serve a different purpose. (I am just sharing my thoughts; I realize it is the author’s paper, not mine, so it is up to the author to decide whether he wants to expand or change the argumentation.)*

RESPONSE: As I say in the article, the prior for a BF serves to represent the predictions of a theory to be tested; and this in no way is the function of the prior for estimation. I don’t make the argument for the latter in this article though; I now refer to a paper where I do. What I take to be the key argument is that if the prior had the function of representing the predictions of a theory, the theory could not be severely tested against the data.

*REVIEWER: “In sum, this article argues a change in approaching statistical testing is needed in order to test our theories severely:” This is not the message I got away with. Is the intended change the move from NHST to testing by intervals and/or using Bayes factors? Both approaches have had their share of advocates and practicioners, so I am not sure how advocating these methods constitutes “arguing for a change” in statistical testing. Maybe the intended change is something else (possibly relating to what counts as justifiable sources of priors, though I am not sure). In that case, I would recommend spelling out more clearly exactly what the author argue that should be changed. (That being said, I think the paper has value without arguing for a change: In my reading, it provides a timely and convenient overview of sources for setting priors when computing Bayes factors or doing inference by intervals, which is valuable in itself).*

RESPONSE: I have now added that a change is needed “, compared to typical practice.“ That is, it is true there are exceptions, and Vanpaemel’s papers are indeed exceptions as I now note. The colon then indicates what the change is, namely, “a range of predicted effect sizes must be specified that follow from the theory itself in its scientific context.”

**Editor Final Decision: Accept**

Sep 3, 2021

Dear Zoltan Dienes,

I have now had a chance to read over your manuscript “Obtaining evidence for no effect”, along with the letter describing the changes you made. Thank you for your responsiveness to the concerns that the reviewers and the previous action editor Julia Haaf 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,
Don van Ravenzwaaij