Journal of Clinical and Translational Research
Peer review process file 03.2017S2.007

Making 'null effects' informative: statistical techniques and inferential frameworks

Christopher Harms, Daniël Lakens

Corresponding author
Christopher Harms, Bonn, Germany.

Handling editor:
Michal Heger
Department of Experimental Surgery, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands

Review timeline:

Received: 1 April, 2018
Editorial decision: 27 May, 2018
Revision received: 18 June, 2018
Editorial decision: 17 July, 2018
Revision received: 24 July, 2018
Editorial decision: 24 July, 2018
Published ahead of print: 30 July, 2018

1st Editorial response

Date: 27-May-2018

Ref.: Ms. No. JCTRes-D-18-00012
Making 'Null Effects' Informative: Statistical Techniques and Inferential Frameworks
Journal of Clinical and Translational Research

Dear authors,

Reviewers have now commented on your paper. You will see that they are advising that you revise your manuscript. If you are prepared to undertake the work required, I would be pleased to reconsider my decision.

For your guidance, reviewers' comments are appended below.

If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you resubmit your work.

Your revision is due by Jun 26, 2018.

To submit a revision, go to https://jctres.editorialmanager.com/ and log in as an Author. You will see a menu item called Submission Needing Revision. You will find your submission record there.

Yours sincerely,

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Reviewers' comments:

Reviewer #1: This is a very clear summary of the three main methods for providing support for a null hypothesis. I just have a few minor requests for revisions.

"Because the 95% HDI ([−4.24; 0.32]) lies well within those
bounds (as can be seen in Figure 3), we declare a difference of exactly zero to not be credible for practical purposes."

The consequent does not follow *because of* the prior clause. A difference of exactly zero is not credible because it was assigned zero probability in the prior. So a difference of exactly zero will never be credible no matter what the data.
The sentence goes on:
" We do not, however, reject any other specific value within the ROPE"
There is no basis for treating zero differently from any other specific value; the prior gives all exact values a zero probability.
On a related point, the Bayesian estimation procedure used allows us to give probabilities not to values but to the true value being in a certain interval.

"Kruschke and Liddell (2017) use the 95%
interval as a convention related to the 5% significance level, but the width of the HDI is arbitrary, should only be seen as a useful summary of the complete posterior distribution,"

But given the decision rule, the use of a certain percentage interval leads to black and white decisions that could change depending on the precise percentage used. A better way to live up to the claim just quoted is to ask what percentage of the posterior distribution is within the ROPE (as recommended by e.g. Greenwald 1975 Psych Bulletin). This is not what Kruschke recommends, but it is a more natural way of interpreting the full posterior without arbitrarily summarizing it as a specific HDI.

"A common criticism on Bayes factors is that they are much more sensitive to the specification of the prior than Bayesian model estimation. In a Bayesian estimation framework (such as the ROPE procedure) the data quickly overwhelm the prior, so the prior has a very limited effect on the final statistical inference. For Bayes factors, on the other hand, priors have much more weight and thus need to justified carefully before looking at the data."
This contrast in priors used in hypothesis testing vs estimation implies they have pretty much the same function but one method- estimation - is more robust. But this misconstrues the situation. In estimation the function of the prior is to allow the best estimate of a parameter; thus, for example, priors should be vague, and may slightly shrink estimates, and be overwhelmed by large amounts of data. In hypothesis testing the function of a prior is to indicate what a theory predicts. Naturally, decisions are sensitive to what a theory predicts. Ideally such priors are not vague, and , if the theory is wrong, they should be very different from the data even for large amounts. So this section should be spelled out a little differently,

making these points.

"In Bayesian estimation the prior can be used to regularise parameter estimates. Especially in small samples and in more complex models, this avoids overfitting the data and can lead to better estimates for out-of-sample inferences and predictions (Gelman et al., 2013, Chapter 14.6). While the perceived subjectivity of priors has been critised by frequentist analysts, with increasing amounts of data, the prior has less influence on the statistical inference."

I think these two sentences work against each other. If regularization is a good thing, why hope it disappears? Further, I doubt many readers will understand what regularization means. The flow should be: why shrinking parameter estimates in a relatively automatic way allows better generalization, and a positive aspect of this process is the shrinkage reduces the bigger the data set.

Reviewer #2: This article presents a relevant discussion of the limitations of NHST for evaluating support in favor of the null hypothesis, and provides an accessible overview of alternative statistical methods that can provide researchers with desired information about possible null effects. The style and presentation of the article is great, and very well suited for the intended audience. Overview papers such as these can be of great use for applied users of statistics who may not be overly familiar with the statistical and philosophical debates concerning hypothesis testing, and can be helpful for guiding them in a direction that best suits their needs. Also, the availability of all analysis scripts for each of the performed analyses will be greatly appreciated by the intended audience. As such, I feel that despite the lack of novelty (which cannot reasonably be expected of a paper that intends to provide an overview of relevant existing procedures) the paper can end up making a relevant contribution to the field. I do have some points that I would like the authors to address in a revision, which I believe should not be considered major but will nevertheless improve the paper. I will list these points (not ordered based on importance) below.

Lines 57-60 inform us that there are no statistical techniques that allow us to evaluate a statistical hypothesis in isolation. However, this is exactly what the Fisherian null hypothesis testing approach does: evaluate a null hypothesis without considering any alternative hypothesis. I would fully agree that this does not allow one to draw conclusions about whether or not that hypothesis is (likely to be) true, but as this is also not the intention of the procedure, its intended conclusions can be draw perfectly well without reference to alternative hypothesis. If the authors want to make the point that this inferential framework is not relevant for the evaluation of null effects, they will have to provide a more extensive critique of Fisherian null hypothesis testing (as opposed to standard NHST testing, which constitutes a hybrid of the Fisherian and Neyman/Pearson approach to hypothesis testing, see for example the work of Gigerenzer). Generally, the authors may want to consider devoting a bit of attention to this hybrid nature of NHST as it is currently practiced, as the exact nature of this hybrid form is of importance for making the claim that it does or does not allow one to draw inferences about the null hypothesis without these inferences depending on the specification of an alternative hypothesis.

Line 110: I am sure the readers would appreciate getting more information about how exactly these fictional data were obtained.

Page 7 and further: In many places in the text of the manuscript, the distinction between a sample-level difference (which can be observed and does not depend on any statistical procedure) and a population-level difference (which can never be said to be observed but only inferred, and where these inferences depend on the used procedure) is not consistently made. For example, on line 110, it is stated that 'a difference ... was not observed', which either refers to a sample-level difference (which was observed), or a population-level difference (which is never observed). Likewise, no statistical test is needed to evaluate whether a sample-level observed difference is considered meaningful (line 129/130). This is not just semantics, because for a correct understanding of the importance of the choice of methods it is quite relevant for the readers to realize that no procedure enables you to 'observe' differences in the population. Consistently making it clear whether a sample-level or population-level difference is referred to is needed to improve the clarity of the arguments. I have not listed the other occurrences, but they are not limited to the two cases mentioned.

Lines 166-168: It now states that the SESOI should both be determined before and after collecting the data.

Lines 219 and further: Up to this point, Bayesian analysis has been presented in a subjective way, with the prior distribution capturing prior beliefs. From this point on, most of the Bayesian procedures described make use of nonsubjective priors, while the interpretation of the posterior distribution (and the inferences based upon it) as representing posterior credibility is maintained. This can be considered misleading, as the interpretation is not warranted if a prior distribution is not picked based on prior beliefs but based on statistical considerations. At the very least the authors will need to clarify that the interpretation of the posterior as capturing posterior credibility hinges on whether the prior that is used indeed matches prior beliefs. This is partly discussed at the end of the manuscript, but it would be helpful to still briefly make this clear when the procedures are first presented.

General point: In a fully Bayesian framework, it is quite difficult to defend assigning nonzero probability to a point null hypothesis, as one would generally assume neighbouring values over the range of the parameter that is considered to not differ markedly (why would it a priori be much more probable that a group difference is exactly zero rather than for example 0.00000000000000001?). Such exact null hypotheses should in the social sciences hence often be assigned 0 probability, which makes the enterprise of evaluation point null hypotheses (rather than approximate null hypothesis that consider the range of no meaningful effects) a priori rather irrelevant. It would be relevant for the authors to reflect more on what kind of inferences one would generally hope to be able to make in the context of studying null effects in the social sciences (more than what is currently done on page 17/18), as this will have to be what guides the choice of methods of the applied researcher. The fact that an exact null hypothesis can often be rejected a priori (nicely illustrated by Cohen in his 1994 paper) would seem to be relevant to mention in the discussion, which has consequences for the desirability of testing exact versus approximate null hypotheses.

Line 333: The Bayes factor being close to 1 suggest that the data do not strongly support one hypothesis over the other, but that does not warrant the conclusion that there is no good reason to conclude in favor of either model, which is a statement that (in its current form) concerns the credibility of the two models, and hence depends on the (currently not discussed) prior probability assigned to the two models. The authors make this point on page 18, but the formulation in this sentence does not align with that.

Line 360/361: It is worth pointing out that the reader is only able to update their own priors based on the presented BF if they adopt proportionally identical priors as the researchers (i.e., work with prior distributions of the same shape). In the case of for example using earlier research or subjective beliefs to inform the choice of priors, not all readers would be willing to take this as their starting point, and those readers would not be able to update their beliefs based on the reported BF alone. It is in that sense misleading to claim that the BF contains all the necessary information to make an inference.

Line 387-389: This sentence should be improved. Any significant effect may constitute a Type I error and hence the distinction needs to be made more clear: Conditional on the significant effect not constituting a Type I error, under an equivalence test one still cannot conclude that the effect is exactly zero.

Line 451: Given that many of the alternatives to NHST that are considered in this paper still test a null hypothesis, I would suggest also using the term 'NHST' in this sentence instead of the more generic 'null-hypothesis tests', as the former specifically refers to the Fisher-Neyman/Pearson hybrid that is used in the social sciences while the latter term is more generic.

Noticed typos:
- line 29: 'testinyg'
- line 68: 'interpretat'
- line 342: 'amodel'

Reviewer #3:
This manuscript provides an accessible introduction to three statistical techniques for deciding to accept a null value or null hypothesis. I think the manuscript could make a useful contribution to special issue of JCTR, and the comments below are offered with the hope of enhancing the impact of the article.

p. 3: The manuscript lists three scenarios in which the goal is not to reject a point value. Missing from the list is one important situation: confirming a specific quantitative prediction (other than a null value). It may be rare in clinical research to have a theory that makes a specific quantitative prediction, but in science more generally this does happen. Physics is the paradigmatic example in which quantitative theories make specific predictions that are then confirmed. Thus, the list should have four scenarios.

p. 9 line 168: Typo in phrasing? Should this say "*before* looking at the collected data"?

p. 9 lines 168-169: The expression is confusingly stated: "An informative study should be designed such that it is well-powered to both detect and reject the [SESOI]." Instead perhaps this: "A study should have sufficient power (i) to detect an effect that exceeds the SESOI and (ii) to demonstrate equivalence to zero for a null effect."

p. 12, line 222, says "The goal of this approach is to arrive at an approximation of the posterior distribution..." No, in fact MCMC is arbitrarily accurate with processing time, and the goal is an accurate representation of the posterior, not an approximation. (This is unlike various approximations to p values, which are *inherently* approximations.) I think this

sentence can simply be deleted.

p. 12, footnote 5: "If the prior that is used for the model is not uniform (as in the BEST model) differences between an HDI and a confidence interval are to be expected." That is technically true but also misleading because it suggests that the prior in the BEST model is noticeably influencing the result. In fact, the prior is diffuse and is specifically designed to have minimal influence on the posterior distribution.

p. 13, line 243: Do not use the term "hypothesis test" in this context. The HDI+ROPE procedure is definitely *not* a hypothesis test. In the Bayesian realm, the term "hypothesis test" is strictly reserved for model comparison and Bayes factors.

p. 13, statement of HDI+ROPE decision rule. For the latest and clearest statement of the HDI+ROPE decision rule, please see the in-press article titled "Rejecting or accepting parameter values in Bayesian estimation" at https://osf.io/s5vdy/ to be published in Advances in Methods and Practices in Psychological Science. Presumably you'll also want to cite the companion articles in that issue, regarding equivalence testing and Bayes factors.

p. 13, line 257: Delete "not" from "we declare a difference of exactly zero to [not] be credible...". And it's not really accurate to say that zero is "credible" either. The "credible" values are summarized by the HDI. The decision rule says that the estimated value of the parameter is practically equivalent to zero (even in cases when zero is not inside the HDI).

p. 16, after 3rd paragraph that mentions a criticism of Bayes factors: In the article titled "Bayesian Data Analysis for Newcomers," at https://link.springer.com/article/10.3758/s13423-017-1272-1 or https://psyarxiv.com/nqfr5/, Kruschke & Liddell listed five main caveats about Bayes factors that are directly relevant to the present manuscript. In particular, consider caveats 4 and 5 which point out key differences in the information provided by Bayes factors as opposed to parameter estimation:
"4. The Bayes factor indicates nothing about the magnitude of the effect or the precision of the estimate of the magnitude. In this way, using a Bayes factor alone is analogous to using a p value alone without a point estimate or confidence interval."
"5. The Bayes factor can accept a null prior even when there is poor precision in the estimate of the magnitude of effect. In other words, the Bayes factor can accept the null prior even when an estimate of the magnitude indicates there is a wide range of credible non-null values for the effect."

p. 17, lines 320-327: It would be helpful to illustrate the alternative model prior, because Figure 4(D) does not show the alternative model prior described at line 320. And I'm dubious about the Bayes factor being only 2.95 --- I'd be less dubious if I could see the probability density at zero in the prior (to visually approximate the Savage-Dickey density ratio).

p. 18, line 361: But if the posterior odds are what determines a researcher's decision, then the BF should *not* be used to make a decision. This was Caveat 3 from "Bayesian for Newcomers" mentioned in the previous comment.

p. 21, line 392, parenthetical remark: It's confusing to talk about effects that reject the null hypothesis but are equivalent to zero. I understand what you mean, but I think it could be quite confusing to people who are not familiar with equivalence testing.

p. 21, line 403: It is correct to state that the HDI does not include parameter values that are true, but if this is worth stating then it's also worth immediately stating that the the 95% confidence interval also does not include parameter values that are true. In both frequentist and Bayesian estimation of parameters, we're only finding parameter values that are least bad in the context of the chosen model. The model itself might be a terrible model of the data, regardless of whether it's estimated using frequentist or Bayesian methods.

Throughout: The manuscript says repeatedly that null hypotheses cannot be proven, nor can they be disproven. This is definitely worth saying once, and maybe even worth stating a second time in the final discussion. But it doesn't need to be said more than twice.

Footnotes: Readers usually ignore footnotes. If readers search and find the footnotes, then the footnotes disrupt the flow of the text and argument. Either way, footnotes don't work well. If the content of a footnote is worth saying, put it in the main text, otherwise delete it.

In general I think it's good policy to have anonymous reviews, but in this case it's relevant for the authors to know my identity: John Kruschke. Thank you for the opportunity to review the manuscript.

Reviewer #4:
Although I disagree strongly with the null hypothesis significance testing procedure, including when used as the authors wish to use it to support null hypotheses, I nevertheless recommend in favor of publication. Here is why. In the first place, I believe that many effects really are small, and in science as currently practiced, one either needs to luck out and get a large sample effect size to pass the usual significance threshold, or one needs an extremely large sample size to overcome the small effect size. The authors show a way out of this that, though I disagree with it, still may benefit the field. In addition, I want to make sure that my own strong prejudices against null hypothesis significance testing influence my recommendation as little as possible.

A second benefit of the manuscript is the inclusion of Bayesian methods along with frequentist ones. A limitation is that the authors do not address the philosophical issues that go along with these different ways of thinking, but an argument could be made that this wouldn't be the type of paper where that would be appropriate. Even with the limitation, whether the limitation is justified or not, an advantage is that the reader can compare the different methods, including that they all deal with different questions.

Having supported publication, I suggest some minor changes. First, I think the title is misleading. The issue is not whether the effect is null, as practically no effects in the soft sciences are exactly 0.00000000000000000000000. (A physics exception may be the Michelson & Morley, 1887 effect where although the sample size was not equal to zero, the typical physics assumption is that the population effect size is zero. Interestingly, some physicists are now questioning this last.) Rather, the issue the authors bring out is whether the effect is small enough that it can be treated as unimportant. I think this should be reflected in the title.

A second possible change goes with the first. That is, I think some of the text throughout, including the abstract, could be changed to render clear that the authors don't really expect that any population effect sizes are EXACTLY 0. It seems to me that the authors are clear

about this in some places but not in others (such as the abstract and title).

A third possible change is particularly easy and pertains to citations. Consider the following sentence that I cut and pasted from the ms: "When researchers only publish scientific findings that statistically reject null effects, the scientific literature is biased, which hinders the accumulation of scientific knowledge." I agree with this statement but feel that the authors should provide supporting citations. For example, there was a recent discussion of this in BASP and the authors could cite that discussion though they could cite others in addition to, or instead of, the BASP authors. I list references below. For full disclosure, my name is David Trafimow and I am the editor of BASP.

Grice, J. W. (2017). Comment on Locascio's results blind manuscript evaluation proposal. Basic
and Applied Social Psychology, 39(5), 254-255.
https://doi.org/10.1080/01973533.2017.1352505
Hyman, M. (2017). Can 'results blind manuscript evaluation' assuage 'publication bias'? Basic and Applied Social Psychology, 39(5), 247-251.
https://doi.org/10.1080/01973533.2017.1350581
Locascio, J. (2017a). Results blind publishing. Basic and Applied Social Psychology. 39(5), 239-
246. https://doi.org/10.1080/01973533.2017.1336093
Locascio, J. (2017b). Rejoinder to responses to "results blind publishing." Basic and Applied Social Psychology. 39(5), 258-261. https://doi.org/10.1080/01973533.2017.1356305
Marks, M. J. (2017). Commentary on Locascio 2017. Basic and Applied Social Psychology. 39(5), 252-253. https://doi.org/10.1080/01973533.2017.1350580

Fourth, and this is extremely minor, in the version I received there were figures on the text and also at the end. I expect that one or the other is sufficient.

In summary, I recommend the ms be published but also feel that some minor changes would improve it. To ensure that my prejudices against significance testing did not cause me to recommend harmful changes to the ms, however, I would like to make a final recommendation that my suggested changes be left up to the authors rather than imposed on them.

Authors' rebuttal

Responses to the Reviewers

Reviewer #1
This is a very clear summary of the three main methods for providing support for a null hypothesis. I just have a few minor requests for revisions.

Authors' Response:
Thank you for your positive evaluation of our manuscript and your constructive requests to the text.

"Because the 95% HDI ([−4.24; 0.32]) lies well within those bounds (as can be seen in Figure 3), we declare a difference of exactly zero to not be credible for practical purposes."

*The consequent does not follow *because of* the prior clause. A difference of exactly zero is not credible because it was assigned zero probability in the prior. So a difference of exactly zero will never be credible no matter what the data.*

Authors' Response:
The original sentence has been rewritten with respect to the comments made by John Kruschke (Reviewer #3, see below). The sentence now states that the "difference of exactly zero is accepted for practical purposes" as it is more in line with the ROPE framework proposed by Kruschke.
The sentence goes on:

" We do not, however, reject any other specific value within the ROPE"

There is no basis for treating zero differently from any other specific value; the prior gives all exact values a zero probability.

On a related point, the Bayesian estimation procedure used allows us to give probabilities not to values but to the true value being in a certain interval.

Authors' Response:
We agree with the reviewer that the Bayesian framework allows for nuanced interpretations and is not limited to dichotomous decisions in general. The ROPE procedure, however, is aimed at providing a useful and consistent decision rule if a dichotomous decision about a specific value is required. We have stressed this point e.g. in the sentence on page 13, line 273ff.: "Importantly, even if only summaries are presented such as means, standard deviations, or credibility intervals, the whole posterior distribution is available to provide the statistical inference".
Rebuttal letter

"Kruschke and Liddell (2017) use the 95% interval as a convention related to the 5% significance level, but the width of the HDI is arbitrary, should only be seen as a useful summary of the complete posterior distribution," But given the decision rule, the use of a certain percentage interval leads to black and white decisions that could change depending on the precise percentage used. A better way to live up to the claim just quoted is to ask what percentage of the posterior distribution is within the ROPE (as recommended by e.g. Greenwald 1975 Psych Bulletin). This is not what Kruschke recommends, but it is a more natural way of interpreting the full posterior without arbitrarily summarizing it as a specific HDI.

Authors' Response:
In addition to the point raised before, we have added the recommendation by Greenwald (1975) to the manuscript: "An alternative way to investigate practical equivalence using a Bayesian posterior distribution is to examine the probability mass contained in the ROPE (for details, see Greenwald, 1975, p. 18)." (p. 15, line 315ff).

"A common criticism on Bayes factors is that they are much more sensitive to the specification of the prior than Bayesian model estimation. In a Bayesian estimation framework (such as the ROPE procedure) the data quickly overwhelm the prior, so the prior has a very limited effect on the final statistical inference. For Bayes factors, on the other hand, priors have much more weight and thus need to justified carefully before looking at thedata."

This contrast in priors used in hypothesis testing vs estimation implies they have pretty much the same function but one method- estimation - is more robust. But this misconstrues the situation. In estimation the function of the prior is to allow the best estimate of a parameter; thus, for example, priors should be vague, and may slightly shrink estimates, and be overwhelmed by large amounts of data. In hypothesis testing the function of a prior is to indicate what a theory predicts. Naturally, decisions are sensitive to what a theory predicts. Ideally such priors are not vague, and , if the theory is wrong, they should be very different from the data even for large amounts. So this section should be spelled out a little differently, making these points.

Authors' Response:
We have added a clarification on the different purposes of a prior in these two contexts: "It is important to note, however, that priors have different purposes in the two applications: In Bayesian models for estimation, the priors are used as a device for regularization and shrinkage of parameter estimates. This can be driven by subjective beliefs or statistical considerations (see discussion on subjective and objective use of priors above) .For Bayes factors, on the other hand, priors should represent the predictions of a theory." (p. 18, line 364ff).

"In Bayesian estimation the prior can be used to regularise parameter estimates. Especially in small samples and in more complex models, this avoids overfitting the data and can lead to better estimates for out-of-sample inferences and predictions (Gelman et al., 2013, Chapter 14.6). While the perceived subjectivity of priors has been critised by frequentist analysts, with increasing amounts of data, the prior has less influence on the statistical inference."

I think these two sentences work against each other. If regularization is a good thing, why hope it disappears? Further, I doubt many readers will understand what regularization means. The flow should be: why shrinking parameter estimates in a relatively automatic way allows better generalization, and a positive aspect of this process is the shrinkage reduces the bigger the data set.

Authors' Response:
The paragraph has been rephrased to better capture the benefits of using priors for regularizing estimates: "In Bayesian estimation the prior can be used to shrink or regularise parameter estimates. Through Bayes' theorem, priors provide an automatic way to implement shrinkage in a statistical model. Especially in small samples and more complex models, this avoids overfitting the data and can lead to better estimates for out-of-sample inferences and predictions (Gelman et al., 2013, Chapter 14.6). With more data parameter estimates become more precise and the prior has less influence on the posterior distribution, thus providing less shrinkage as is desirable in most models. Finally, the Bayesian approach

to statistical modelling is very versatile and can be used even in complex models such as
hierarchical generalized models. Bayesian hierarchical or multilevel models are particularly useful in clinical research, for example, when using clustered samples or repeated measurements" (p. 26, line 525ff.).

Reviewer #2
This article presents a relevant discussion of the limitations of NHST for evaluating support in favor of the null hypothesis, and provides an accessible overview of alternative statistical methods that can provide researchers with desired information about possible null effects. The style and presentation of the article is great, and very well suited for the intended audience.

Overview papers such as these can be of great use for applied users of statistics who may not be overly familiar with the statistical and philosophical debates concerning hypothesis testing, and can be helpful for guiding them in a direction that best suits their needs. Also, the availability of all analysis scripts for each of the performed analyses will be greatly appreciated by the intended audience. As such, I feel that despite the lack of novelty (which cannot reasonably be expected of a paper that intends to provide an overview of relevant existing procedures) the paper can end up making a relevant contribution to the field. I do have some points that I would like the authors to address in a revision, which I believe should not be considered major but will nevertheless improve the paper. I will list these points (not ordered based on importance) below.

Authors' Response:
Thank you for your positive comments and constructive remarks on our manuscript. We address your points in the following.

Lines 57-60 inform us that there are no statistical techniques that allow us to evaluate a statistical hypothesis in isolation. However, this is exactly what the Fisherian null hypothesis testing approach does: evaluate a null hypothesis without considering any alternative hypothesis. I would fully agree that this does not allow one to draw conclusions about whether or not that hypothesis is (likely to be) true, but as this is also not the intention of the procedure, its intended conclusions can be draw perfectly well without reference to alternative hypothesis. If the authors want to make the point that this inferential framework is not relevant for the evaluation of null effects, they will have to provide a more extensive critique of Fisherian null hypothesis testing (as opposed to standard NHST testing, which constitutes a hybrid of the Fisherian and Neyman/Pearson approach to hypothesis testing, see for example the work of Gigerenzer). Generally, the authors may want to consider devoting a bit of attention to this hybrid nature of NHST as it is currently practiced, as the exact nature of this hybrid form is of importance for making the claim that it does or does not allow one to draw inferences about the null hypothesis without these inferences depending on the specification of an alternative hypothesis.

Authors' Response:
Concerning the different perspectives on frequentist null-hypothesis testing, we added a section briefly outlining the differences and how we focus on the Neyman-Pearson paradigm for long-run error control (cf. p. 8, line 138ff.).

Line 110: I am sure the readers would appreciate getting more information about how exactly

these fictional data were obtained.

Authors' Response:
We have added a sentence on the simulation of the data: "For the imaginary study we simulated random samples using R from two independent normal distributions" (p. 6, line 117f.).
Page 7 and further: In many places in the text of the manuscript, the distinction between a sample-level difference (which can be observed and does not depend on any statistical procedure) and a population-level difference (which can never be said to be observed but only inferred, and where these inferences depend on the used procedure) is not consistently made. For example, on line 110, it is stated that 'a difference ... was not observed', which either refers to a sample-level difference (which was observed), or a population-level difference (which is never observed). Likewise, no statistical test is needed to evaluate whether a sample-level observed difference is considered meaningful (line 129/130). This is not just semantics, because for a correct understanding of the importance of the choice of methods it is quite relevant for the readers to realize that no procedure enables you to 'observe' differences in the population. Consistently making it clear whether a sample-level or population-level difference is referred to is needed to improve the clarity of the arguments. I have not listed the other occurrences, but they are not limited to the two cases mentioned.

Authors' Response:
Thank for noticing this important inconsistency. We have revised several instances of this distinction and write about "population effect sizes" versus "observed differences" more consistently, e.g. on page 7, line 132f: "This might be because there is no difference between the two populations from which the two groups are sampled, [...]".

Lines 166-168: It now states that the SESOI should both be determined before and after collecting the data.

Authors' Response:
The sentence was corrected and rephrased - also according to the comments by another reviewer: "Ideally, the SESOI should be informed by theory and previous research (such as meta-analyses or systematic reviews). The SESOI needs to be determined before collecting the data (similar to decisions about the sample size, the alpha level, and the desired statistical power). An informative study should be designed to have sufficient power both (i) to detect an effect that exceeds the SESOI and (ii) to demonstrate equivalence to zero or another specific value (thus rejecting the smallest effect size of interest)" (p. 9f, line 188ff).

Lines 219 and further: Up to this point, Bayesian analysis has been presented in a subjective way, with the prior distribution capturing prior beliefs. From this point on, most of the Bayesian procedures described make use of nonsubjective priors, while the interpretation of the posterior distribution (and the inferences based upon it) as representing posterior credibility is maintained.

This can be considered misleading, as the interpretation is not warranted if a prior distribution is not picked based on prior beliefs but based on statistical considerations. At the very least the authors will need to clarify that the interpretation of the posterior as capturing posterior credibility hinges on whether the prior that is used indeed matches prior beliefs.

This is partly discussed at the end of the manuscript, but it would be helpful to still briefly make this clear when the procedures are first presented.

Authors' Response:
We have added a sentence to the explanation of the posterior distribution highlighting the point raised by the reviewer: "If the prior distribution is accepted to represent a valid allocation of belief, the posterior distribution represents rationally updated belief through the observed data" (p. 13, line 258f.)

General point: In a fully Bayesian framework, it is quite difficult to defend assigning nonzero probability to a point null hypothesis, as one would generally assume neighbouring values over the range of the parameter that is considered to not differ markedly (why would it a priori be much more probable that a group difference is exactly zero rather than for example 0.00000000000000001?). Such exact null hypotheses should in the social sciences hence often be assigned 0 probability, which makes the enterprise of evaluation point null hypotheses (rather than approximate null hypothesis that consider the range of no meaningful effects) a priori rather irrelevant. It would be relevant for the authors to reflect more on what kind of inferences one would generally hope to be able to make in the context of studying null effects in the social sciences (more than what is currently done on page 17/18), as this will have to be what guides the choice of methods of the applied researcher. The fact that an exact null hypothesis can often be rejected a priori (nicely illustrated by Cohen in his 1994 paper) would seem to be relevant to mention in the discussion, which has consequences for the desirability of testing exact versus approximate null hypotheses.

Authors' Response:
The question whether point null hypothesis are realistic or at least useful is an important issue. We have added and extended a paragraph on this question and refer to Meehl's (1990) work on the "crud factor" (cf. p. 5, line 84ff.).

Line 333: The Bayes factor being close to 1 suggest that the data do not strongly support one hypothesis over the other, but that does not warrant the conclusion that there is no good reason to conclude in favor of either model, which is a statement that (in its current form) concerns the credibility of the two models, and hence depends on the (currently not discussed) prior probability assigned to the two models. The authors make this point on page 18, but the formulation in this sentence does not align with that.

Authors' Response:
We have added a sentence in this section and referred to the extended discussion below: "Thus, although the data support the null model over the alternative model, there is no good reason to conclude in favor of either model - at least if not either model is much more reasonable than the other a priori without respect to the data (this issue is extended on below)" (p. 20, line 404ff).

Line 360/361: It is worth pointing out that the reader is only able to update their own priors based on the presented BF if they adopt proportionally identical priors as the researchers (i.e., work with prior distributions of the same shape). In the case of for example using earlier research or subjective beliefs to inform the choice of priors, not all readers would be willing to take this as their starting point, and those readers would not be able to update their beliefs based on the reported BF alone. It is in that sense misleading to claim that the BF contains all the necessary information to make an inference.

Authors' Response:
We have rephrased the sentence to include the necessary agreement with the prior distributions for the parameters: "If a researcher accepts the prior distributions for the parameters in the models compared in the Bayes factor, the Bayes factor contains the necessary information to update their own prior odds and make an inference - but the Bayes factor is by itself not sufficient to reach a conclusion" (p. 21, line 434ff.).
Line 387-389: This sentence should be improved. Any significant effect may constitute a Type I error and hence the distinction needs to be made more clear: Conditional on the significant effect not constituting a Type I error, under an equivalence test one still cannot conclude that the effect is exactly zero.

Authors' Response:
The sentence has been rephrased for clarity: "If we conclude statistical equivalence, we can reject the presence of effect sizes more extreme than the smallest effect size of interest with a known error rate, but we can not conclude the true effect is exactly zero – there might be a true but small effect" (p. 24, line 476ff).

Line 451: Given that many of the alternatives to NHST that are considered in this paper still test a null hypothesis, I would suggest also using the term 'NHST' in this sentence instead of the more generic 'null-hypothesis tests', as the former specifically refers to the Fisher-Neyman/Pearson hybrid that is used in the social sciences while the latter term is more generic.

Authors' Response:
According to the reviewer's comment we have rephrased the sentence: "Researchers who only rely on null-hypothesis significance tests limit themselves in only asking the question whether the null-hypothesis can be rejected" (p. 26, line 543ff).

Noticed typos:
- line 29: 'testinyg'
- line 68: 'interpretat'
- line 342: 'amodel'

Authors' Response:
We have checked the manuscript for typos and have corrected them, including the three mentioned by the reviewer.

Reviewer #3
This manuscript provides an accessible introduction to three statistical techniques for deciding to accept a null value or null hypothesis. I think the manuscript could make a useful contribution to special issue of JCTR, and the comments below are offered with the hope of enhancing the impact of the article.

Authors' Response:
Thank you for your constructive remarks on our manuscript. We have incorporated many of your suggestions and will respond to them point-by-point in the following.
p. 3: The manuscript lists three scenarios in which the goal is not to reject a point value. Missing from the list is one important situation: confirming a specific quantitative prediction (other than a null value). It may be rare in clinical research to have a theory that makes a

specific quantitative prediction, but in science more generally this does happen. Physics is the paradigmatic example in which quantitative theories make specific predictions that are then confirmed. Thus, the list should have four scenarios.

Authors' Response:
Although we agree the fourth scenario exists, the special issue focuses exclusively on 'null effects'. It is true that all three approaches can be used to confirm specific quantitative predictions (e.g., by setting equivalence bounds to 0.54 and 0.56 when predicting an effect of 0.55) but this scenario falls outside of the topic of the special issue.
*p. 9 line 168: Typo in phrasing? Should this say "*before* looking at the collected data"?*

Authors' Response:
Thank you for catching this typo. The sentence has been corrected (see above).
p. 9 lines 168-169: The expression is confusingly stated: "An informative study should be designed such that it is well-powered to both detect and reject the [SESOI]." Instead perhaps this: "A study should have sufficient power (i) to detect an effect that exceeds the SESOI and (ii) to demonstrate equivalence to zero for a null effect."

Authors' Response:
Thank you for the suggested sentence, which have rephrased and integrated in the manuscript. The sentence now reads as follows: "An informative study should be designed to have sufficient power both (i) to detect an effect that exceeds the SESOI and (ii) to demonstrate equivalence to zero or another specific value (thus rejecting the smallest effect size of interest)." (p. 10, line 191ff).

*p. 12, line 222, says "The goal of this approach is to arrive at an approximation of the posterior distribution..." No, in fact MCMC is arbitrarily accurate with processing time, and the goal is an accurate representation of the posterior, not an approximation. (This is unlike various approximations to p values, which are *inherently* approximations.) I think this sentence can simply be deleted.*

Authors' Response:
We agree that the sentence was imprecise about the nature of the MCMC samples. We have rewritten replaced it to highlight that MCMC yields samples from the posterior, which in turn can be used to make inferences about the parameters. The new sentence reads: "When using a Bayesian statistical model, samples from the posterior distribution are generated which can be used to make inferences about the data" (p. 13, line 267ff).

p. 12, footnote 5: "If the prior that is used for the model is not uniform (as in the BEST model) differences between an HDI and a confidence interval are to be expected." That is technically true but also misleading because it suggests that the prior in the BEST model is noticeably influencing the result. In fact, the prior is diffuse and is specifically designed to have minimal influence on the posterior distribution.

Authors' Response:
The footnote has been moved into the main text and a sentence has been added to highlight the diffuse nature of the priors in the BEST model. We think, it is important to explain why HDI and CI are similar but not identical, because in our experience researchers often wonder why they are very similar but not exactly identical, and what this means for Bayesian inferences in general. However, we not explicitly say there are "small" differences.

*p. 13, line 243: Do not use the term "hypothesis test" in this context. The HDI+ROPE procedure is definitely *not* a hypothesis test. In the Bayesian realm, the term "hypothesis test" is strictly reserved for model comparison and Bayes factors.*

Authors' Response:
The term "hypothesis test" for the ROPE procedure has been replaced with "dichotomous decision". We have also added two sentences highlighting the distinction between the decision rule and Bayesian hypothesis testing: "In the vocabulary of Bayesian statistics, using a decision rule on a posterior distribution of a single model does not constitute "hypothesis testing". The term "Bayesian hypothesis testing" refers strictly to the use of Bayes factors for model selection, which will be considered in the section of the paper." (p. 15, line 314ff).

p. 13, statement of HDI+ROPE decision rule. For the latest and clearest statement of the HDI+ROPE decision rule, please see the in-press article titled "Rejecting or accepting parameter values in Bayesian estimation" at https://osf.io/s5vdy/ to be published in Advances in Methods and Practices in Psychological Science. Presumably you'll also want to cite the companion articles in that issue, regarding equivalence testing and Bayes factors.

Authors' Response:
We have updated the formulation of the decision rule to match the statement in Kruschke (2018). To be more in line with the vocabulary of our present manuscript, we made a change to the quote by talking about the "parameter's posterior distribution" rather than about the "parameter distribution". The article regarding equivalence testing and Bayes factors is cited as Lakens, Scheel & Isager (2018b).

p. 13, line 257: Delete "not" from "we declare a difference of exactly zero to [not] be credible...".

And it's not really accurate to say that zero is "credible" either. The "credible" values are summarized by the HDI. The decision rule says that the estimated value of the parameter is practically equivalent to zero (even in cases when zero is not inside the HDI).

Authors' Response:
The sentence has been rephrased and now reads: "[...] we declare a difference of exactly zero to be accepted for practical purposes based on the decision rule above. We do not, however, accept or reject any other specific value within the ROPE" (p. 15, 312f).
p. 16, after 3rd paragraph that mentions a criticism of Bayes factors: In the article titled

"Bayesian Data Analysis for Newcomers," at https://link.springer.com/article/10.3758/s13423-017-1272-1 or https://psyarxiv.com/nqfr5/, Kruschke & Liddell listed five main caveats about Bayes factors that are directly relevant to the present manuscript. In particular, consider caveats 4 and 5 which point out key differences inthe information provided by Bayes factors as opposed to parameter estimation:
"4. The Bayes factor indicates nothing about the magnitude of the effect or the precision of the estimate of the magnitude. In this way, using a Bayes factor alone is analogous to using a p value alone without a point estimate or confidence interval."
"5. The Bayes factor can accept a null prior even when there is poor precision in the estimate of the magnitude of effect. In other words, the Bayes factor can accept the null prior even

when an estimate of the magnitude indicates there is a wide range of credible non-null values for the effect."

Authors' Response:
Thank you for the reference to this very insightful discussion of Bayes factors. We have included the fourth issue to the main text as a valuable addition and referenced your paper: "Moreover, Bayes factors -- very much like p-values -- do not convey information about the magnitude of an effect or the uncertainty in its estimation. See Kruschke & Liddell (2018) for further criticism on Bayes factors" (p. 18, line 375f). The fifth issue raised in this paper is already mentioned in the discussion section: "A Bayes factor can indicate strong support for a null model relative to an alternative model, but both models can be wrong" (p. 25, line 498f).

p. 17, lines 320-327: It would be helpful to illustrate the alternative model prior, because Figure 4(D) does not show the alternative model prior described at line 320. And I'm dubious about the Bayes factor being only 2.95 --- I'd be less dubious if I could see the probability density at zero in the prior (to visually approximate the Savage-Dickey density ratio).

Authors' Response:
We have included a figure (Figure 4) showing the prior and posterior distributions for the alternative model to visually indicate the Savage-Dickey ratio for the Bayes factor. A brief explanation of the figure has been added to the main text and a reference to Wagenmakers et al. (2010) has been added.

*p. 18, line 361: But if the posterior odds are what determines a researcher's decision, then the BF should *not* be used to make a decision. This was Caveat 3 from "Bayesian for Newcomers" mentioned in the previous comment.*

Authors' Response:
We agree with this comment and this is the reason we stress the need for researchers to update their beliefs based on the Bayes factor (as summary of the data). We have added a sentence to further clarify this.

p. 21, line 392, parenthetical remark: It's confusing to talk about effects that reject the null hypothesis but are equivalent to zero. I understand what you mean, but I think it could be quite confusing to people who are not familiar with equivalence testing.

Authors' Response:
In the section on equivalence testing we have added a paragraph providing the necessary background for this sentence in the discussion, i.e. explaining the different possible outcomes when conducting an equivalence test (p. 10, line 212f).

p. 21, line 403: It is correct to state that the HDI does not include parameter values that are true, but if this is worth stating then it's also worth immediately stating that the the 95% confidence interval also does not include parameter values that are true. In both frequentist and Bayesian estimation of parameters, we're only finding parameter values that are least bad in the context of the chosen model. The model itself might be a terrible model of the data, regardless of whether it's estimated using frequentist or Bayesian methods.

Authors' Response:
The sentence has been extended to specifically mention confidence intervals and

highlighting that no statistical procedure can tell researchers which values are the "true"
values: "As with other measures of uncertainty such as confidence intervals, Bayesian credibility intervals are not guaranteed to contain true parameter values. The credible intervals contain values which are deemed credible based on the prior and the observed data with a specified posterior probability" (p. 25, line 493ff).

Throughout: The manuscript says repeatedly that null hypotheses cannot be proven, nor can they be disproven. This is definitely worth saying once, and maybe even worth stating a second time in the final discussion. But it doesn't need to be said more than twice.

Authors' Response:
Considering the title of our manuscript and the topic of this special issue, we feel that it is important to remind readers about this epistemological fact. While revising the manuscript (see comments above and below) some instances of this have been rephrased. We strongly feel, however, that it is necessary and worthwhile to stress this point as many researchers have not, yet, internalized this fact. Especially readers who do not read the paper in full.

Footnotes: Readers usually ignore footnotes. If readers search and find the footnotes, then the footnotes disrupt the flow of the text and argument. Either way, footnotes don't work well. If the content of a footnote is worth saying, put it in the main text, otherwise delete it.

Authors' Response:
We have moved some footnotes to the main text if they contain relevant information to the understanding of the manuscript. The remaining footnotes contain information that goes beyond the main topic of the article and are left for the interested reader to follow up the main text.

Reviewer #4

Although I disagree strongly with the null hypothesis significance testing procedure, including when used as the authors wish to use it to support null hypotheses, I nevertheless recommend in favor of publication. Here is why. In the first place, I believe that many effects really are small, and in science as currently practiced, one either needs to luck out and get a large sample effect size to pass the usual significance threshold, or one needs an extremely large sample size to overcome the small effect size. The authors show a way out of this that, though I disagree with it, still may benefit the field. In addition, I want to make sure that my own strong prejudices against null hypothesis significance testing influence my recommendation as little as possible.

Authors' Response:
Thank you for your balanced and positive evaluation of our work. We were hoping to build bridges between frequentist and Bayesian approaches, as well as discussing approaches not based on hypothesis testing (i.e., the ROPE procedure based on a Bayesian parameter estimation framework), than reiterating the divisive debate.

A second benefit of the manuscript is the inclusion of Bayesian methods along with frequentist ones. A limitation is that the authors do not address the philosophical issues that go along with these different ways of thinking, but an argument could be made that this wouldn't be the type of paper where that would be appropriate. Even with the limitation, whether the

limitation is justified or not, an advantage is that the reader can compare the different methods, including that they all deal with different questions.

Authors' Response:
We indeed aimed the article at practical researchers who - unfortunately - choose their methods much less on the grounds of deep philosophical considerations rather than on what was feasible and provides answers to their questions.

In order to direct interested readers to the more foundational issues of using frequentist versus Bayesian methodology, we have added references to books and articles covering this topic.

As noted by the reviewer, our main aim was to illustrate how these approaches deal with different (but related) questions. The importance lies in the correct interpretation of each method. Combining the outcomes of different methods to a general inference about the research question, deserves elaboration, but would fill several journal issues.

Having supported publication, I suggest some minor changes. First, I think the title is misleading. The issue is not whether the effect is null, as practically no effects in the soft sciences are exactly 0.00000000000000000000000. (A physics exception may be the Michelson & Morley, 1887 effect where although the sample size was not equal to zero, the typical physics assumption is that the population effect size is zero. Interestingly, some physicists are now questioning this last.) Rather, the issue the authors bring out is whether the effect is small enough that it can be treated as unimportant. I think this should be reflected in the title.

Authors' Response:
We actually agree that the title of our manuscript is not a correct way of thinking about 'null effects', and you touch a relevant concern about the use or adequacy of point null hypotheses. In our experience, researchers think in terms of 'null effects', and "proving the null", and we wanted to appeal to readers who think like this in the title, but then teach them the correct approach is the think about effect sizes too small to matter (even when not exactly null). We have dealt with the reviewers comments in two ways. First, we have added quotation marks around the "Proving the Null" to make it clearer this is not considered a formally correct phrasing, but more a well known figure of speech. Second, we have extended the section on 'null effects' to make it clearer that often, especially in non-experimental designs, the null might not be 0.00000, and we have added references to Meehl's writings about the 'crud factor' (see quoted changes in the next response).

A second possible change goes with the first. That is, I think some of the text throughout, including the abstract, could be changed to render clear that the authors don't really expect that any population effect sizes are EXACTLY 0. It seems to me that the authors are clear about this in some places but not in others (such as the abstract and title).

Authors' Response:
We have updated sections of paper that benefit from using more precise language regarding the size of the effect compared to a point null. We added, for example, the following paragraph to the introduction: "One can argue that in most studies without random assignment to conditions, and perhaps even in some studies with random assignment, it can be expected that the true (population) effect size is unequal to zero. Often an effect size of

exactly zero (as assumed in the null hypothesis) is implausible (see the theoretical work on
the "crud factor", Meehl, 1990)" (p. 5, 85ff).

A third possible change is particularly easy and pertains to citations. Consider the following sentence that I cut and pasted from the ms: "When researchers only publish scientific findings that statistically reject null e ects, the scienti c literature is biased, which hinders the accumulation of scientific knowledge." I agree with this statement but feel that the authors should provide supporting citations. For example, there was a recent discussion of this in BASP and the authors could cite that discussion though they could cite others in addition to, or instead of, the BASP authors. I list references below. For full disclosure, my name is David Trafimow and I am the editor of BASP.

Grice, J. W. (2017). Comment on Locascio's results blind manuscript evaluation proposal. Basic and Applied Social Psychology, 39(5), 254-255. https://doi.org/10.1080/01973533.2017.1352505
Hyman, M. (2017). Can 'results blind manuscript evaluation' assuage 'publication bias'? Basic and Applied Social Psychology, 39(5), 247-251. https://doi.org/10.1080/01973533.2017.1350581
Locascio, J. (2017a). Results blind publishing. Basic and Applied Social Psychology. 39(5), 239-246. https://doi.org/10.1080/01973533.2017.1336093
Locascio, J. (2017b). Rejoinder to responses to "results blind publishing." Basic and Applied Social Psychology. 39(5), 258-261. https://doi.org/10.1080/01973533.2017.1356305
Marks, M. J. (2017). Commentary on Locascio 2017. Basic and Applied Social Psychology. 39(5), 252-253. https://doi.org/10.1080/01973533.2017.1350580

Authors' Response:
Thank you for directing us to relevant references. We have added references to the Locascio (2017) article as a possible solution to publication bias and a reference to Kühberger et al. (2014) underlining the problems caused by publication bias (p. 3, line 50ff; p. 27, line 549ff)).

Fourth, and this is extremely minor, in the version I received there were figures on the text and also at the end. I expect that one or the other is sufficient.

Authors' Response:
Thank you, this is a result of the manuscript submission software appending the figures to the end.

In summary, I recommend the ms be published but also feel that some minor changes would improve it. To ensure that my prejudices against significance testing did not cause me to recommend harmful changes to the ms, however, I would like to make a final recommendation that my suggested changes be left up to the authors rather than imposed on them.

2nd editorial decision

Date: 17-Jul-2018

Ref.: Ms. No. JCTRes-D-18-00012R1
Making 'Null Effects' Informative: Statistical Techniques and Inferential Frameworks

Journal of Clinical and Translational Research

Dear author(s),

Reviewers have submitted their critical appraisal of your paper. The reviewers' comments are appended below. Based on their comments and evaluation by the editorial board, your work was FOUND SUITABLE FOR PUBLICATION AFTER MINOR REVISION.

If you decide to revise the work, please itemize the reviewers' comments and provide a point-by-point response to every comment. An exemplary rebuttal letter can be found on at http://www.jctres.com/en/author-guidelines/ under "Manuscript preparation." Also, please use the track changes function in the original document so that the reviewers can easily verify your responses.

Your revision is due by Aug 16, 2018.

To submit a revision, go to https://jctres.editorialmanager.com/ and log in as an Author. You will see a menu item call Submission Needing Revision. You will find your submission record there.

Yours sincerely,

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Reviewers' comments:

Reviewer #1: The authors have responded well to the points made last review. I give a few places below where I would recommend a change, but I don't need to see if and how these points have been addressed.

In discussing equivalence testing, and how it can be considered as whether the confidence interval is within the equivalence region, refer the reader to Richard Morey's criticism of confidence intervals.

p 14
"In the BEST
289 model, the priors are not uniform but chosen to have minimal impact on the inferences, so
290 even if the number of observations is relatively small, the prior should not have too much
291 influence on the results."

Refer the reader to later discussion in the paper for why the prior is actually desirable (e.g. shrinkage).

p 15

"we declare a difference of exactly zero to be accepted for practical
313 purposes based on the decision rule above. We do not, however, accept or reject any other
314 specific value within the ROPE."

There is no mathematical justification nor any practical need to draw these conclusions. The maths are quite clear: For the model used, every specific value has zero probability; and the value zero has no special status. But if we define a region of values as being no practical interest, we can say e.g. that with probability > 95%, the true value is one of no interest. Why should we want to say something more, when it is not justified? (So I phrased this procedure as defining a "null region" and then considering the support for the claim that the true value lies in the null region, Dienes, 2014).

"In the Bayesian approach we can make statements about
322 which values we believe are most credible, based on the data and the model, while in
323 frequentist statistics we make dichotomous decisions based on long-run error rates."

That's right; so why not stick to this and avoid making a dichotomous decision about a precise value?

Reviewer #2: While the authors have addressed some of the issues, there are some minor (but relevant) issues that I would argue remain to be addressed.

In the revised manuscript, the authors devote more space to discussing the differences between the approach of Fisher and of Neyman-Pearson in how hypotheses are evaluated. The authors claim (line 149-152) that they follow a Neyman-Pearson approach to hypothesis testing, but the approach that they present seems to match NHST, the standard hybrid of both of these approaches, rather than the Neyman-Pearson approach. Under the latter, it makes no sense to give special status to one of the hypotheses, as both are compared on equal footing: either hypothesis A is preferred/selected, or hypothesis B is preferred/selected. The issue of how to interpret 'null results' is not present under this framework as the goal is not to evaluate a null hypothesis but simply to determine which of two hypotheses should be preferred/selected. It is unclear why the authors attempt to present standard practice as matching the Neyman-Pearson approach, as NHST has strong Fisherian components in it as well that cannot be ignored and are also central to why NHST is a problematic tool for dealing with null hypotheses. I want to again refer the authors to the work of Gigerenzer, who extensively describes the hybrid nature of standard NHST and the fundamental problems this brings with it. I do not believe that attempting to characterize standard hypothesis-testing practice or even idealized standard practice as being in line with the Neyman-Pearson approach is correct or helpful for the purpose of the article: the focus on determining whether a null hypothesis should or should not be rejected is specifically Fisherian, and the issue of how to interpret null effects is not really present in a Neyman-Pearson approach that treats both hypotheses on equal footing. Effectively, there is no issue with null effects in the Neyman-Pearson approach, as there is no null hypothesis that gets a special status.

When discussing the simulated data, it is now only mentioned that the data were generated using two normal distributions. Given that the different methods that are discussed all attempt to evaluate whether or not there is a difference in the means of the two groups in the example, it would be helpful to give the reader information on the mean and variance of the two normal distributions that were used to generate the data.

The authors claim (line 89) that while null effects are practically implausible, testing an exact null hypothesis is still useful for purposes of model comparison. But NHST is in its standard

form not mainly used as a tool for model comparison, it is used as a tool for evaluating hypotheses. The point raised in the previous round of reviews therefore remains: If the null hypothesis can be rejected a priori, what then is the relevance of testing this hypothesis (rather than a 'no substantively relevant effect' hypothesis)? And if we should merely see it as a tool for model comparison, then this goes against what most applied users perceive NHST to be useful for, and this argument would also need to be made in much more detail.

Reviewer #3:
The authors responded well to the first round of reviews, for which I thank them. This revision is much improved. I think the manuscript will make a useful article in the special issue of JCTR. I have only a few remaining suggestions for minor changes.

p. 4, line 65: "certain level of certainty" is awkward. Consider rephrasing.

p. 5, line 78: "the observed data was not..." Change to, "the observed tendency of the data was not...", or "the observed effect size was not...".

p. 10, line 199-200: Change to, "... conclude the effect *size* is statistically equivalent *to zero* ..."

Fig 5, panel D: The alternative-hypothesis prior is centered on zero in this Figure 5, but the alternative-hypothesis prior is not centered on zero in the example of Fig 4. This could be confusing to readers who are not familiar with Bayesian hypothesis testing, so it's worth mentioning, perhaps in the caption of Fig 5, something about the placement of the alternative-hypothesis distribution.

p. 26, line 529 and elsewhere: Here it is mentioned again that "the prior can be used to shrink or regularise parameter estimates." But this use is never explained or exemplified in the article, so it remains mysterious to readers who don't already know what it means. Mentioning it can be misleading because readers may think that the Bayesian estimation or Bayes factor results in the manuscript are affected by this mysterious shrinkage or regularization, when in fact they are not. This needs to be clarified.

That's all. Again, I think this revision responded well to the first round of reviews. Thanks again for the opportunity to review this manuscript. John Kruschke

Reviewer #4: The authors did a conscientious job with the revision. I believe it can go to press now.

David Trafimow

Authors' response

Responses to the Reviewers (after Review Round 2)

Dear Editor,

Thank you for your editorial letter, and for the reviewers for providing excellent comments on our previous revision. Below we have responded to the remaining comments and questions by the reviewers. We indicate how we incorporated their suggestions for changes in the manuscript.

It is notable, that most of the remaining issues about the statistical details are down to fundamental differences in how statistics should be used to answer empirical questions. This is expected for a paper that tries to cover both frequentist and Bayesian statistics from both an estimation and hypothesis testing approach. The very balanced reviews and the fact that different reviews suggest changes in very different directions, convinces us that we were able to find a neutral compromise in how we presented these approaches to researchers reading your journal. It was our aim to give reasonable space for different perspectives and highlight both the benefits and limitations of the different approaches.

We hope to have addressed all remaining issues to your satisfaction and are looking forward to the special issue.

Warm regards

Christopher Harms & Daniel Lakens

Reviewer #1
The authors have responded well to the points made last review. I give a few places below where I would recommend a change, but I don't need to see if and how these points have been addressed.

In discussing equivalence testing, and how it can be considered as whether the confidence interval is within the equivalence region, refer the reader to Richard Morey's criticism of confidence intervals.

Authors' response:
The proper interpretation of confidence intervals can be difficult. However, it is important to understand the use of confidence intervals in equivalence testing relies on whether they include the value that is tested against, or not, in line with how they were proposed by Neyman and Pearson. To remind readers of the correct interpretation of confidence intervals in frequentist estimation we have added the reference to Morey et al. (2016). Note that Morey's criticism is on the exclusive use of CI as an approach to statistical inferences, as proposed by Cumming (2014), but not to their use in Neyman-Pearson hypothesis testing.
Rebuttal letter

p 14 "In the BEST model, the priors are not uniform but chosen to have minimal impact on the inferences, so even if the number of observations is relatively small, the prior should not have too much influence on the results."

Refer the reader to later discussion in the paper for why the prior is actually desirable (e.g. shrinkage).

Authors' response:
Although we understand the reviewer might have the previous knowledge to understand a

comment about shrinkage at this point in the manuscript, we think that a reference to
shrinkage is too early to be understood by most readers (see also comment by reviewer #3). Instead, we have chosen to refer to the discussion of the BEST model in Kruschke (2013).

p 15 "we declare a difference of exactly zero to be accepted for practical purposes based on the decision rule above. We do not, however, accept or reject any other specific value within the ROPE."

There is no mathematical justification nor any practical need to draw these conclusions. The maths are quite clear: For the model used, every specific value has zero probability; and the value zero has no special status. But if we define a region of values as being no practical interest, we can say e.g. that with probability > 95%, the true value is one of no interest. Why should we want to say something more, when it is not justified? (So I phrased this procedure as defining a "null region" and then considering the support for the claim that the true value lies in the null region, Dienes, 2014).

Authors' response:
This is a critique on the ROPE procedure itself. The procedure is one way to use the Bayesian posterior distribution for inferences and by no means the only one. Kruschke (2018) explains the rationale of the ROPE approach and considering his review, we are not far from the intended interpretation and thus would recommend to leave the section as it is. In the previous revision we have already added the reference to Greenwald (1975)'s proposed use of the posterior distribution. We have added another sentence to the paragraph highlighting that there are further ways to interpret the posterior distribution and that a dichotomous decision is not necessary if not desired, in contrast to significance testing (p. 15, line 320ff): "An alternative way to investigate practical equivalence using a Bayesian posterior distribution is to examine the probability mass contained in the ROPE (for details, see Greenwald, 1975, p. 18). It is important to highlight, that the basis for inferences is the full posterior distribution. It is up to the researcher to decide whether they want to make a dichotomous decision about a single parameter value or rather make a probability statement (cf. Discussion section on inferential frameworks)."

"In the Bayesian approach we can make statements about which values we believe are most credible, based on the data and the model, while in frequentist statistics we make dichotomous decisions based on long-run error rates." That's right; so why not stick to this and avoid making a dichotomous decision about a precise value?

Authors' response:
The comment refers to the fundamental question if dichotomous decisions should be the outcome of a statistical analysis or not. Even the authors of the paper disagree regarding this question. We have presented different approaches side-by-side, and our overarching philosophy is that in many different research areas different questions are worth asking, and require different statistical approaches. Our goal is that readers familiar with only significance testing (likely to be the majority of readers of the journal) can find an easy way to improve their inferences. Interested readers can follow-up equivalence tests by reading on two further approaches, highlighting the continuous interpretation of statistical outcomes. The added sentence (p. 15, line 320ff., see response to previous comment) should also clarify this further.

Reviewer #2
While the authors have addressed some of the issues, there are some minor (but relevant) issues that I would argue remain to be addressed.

In the revised manuscript, the authors devote more space to discussing the differences between the approach of Fisher and of Neyman-Pearson in how hypotheses are evaluated. The authors claim (line 149-152) that they follow a Neyman-Pearson approach to hypothesis testing, but the approach that they present seems to match NHST, the standard hybrid of both of these approaches, rather than the Neyman-Pearson approach. Under the latter, it makes no sense to give special status to one of the hypotheses, as both are compared on equal footing: either hypothesis A is preferred/selected, or hypothesis B is preferred/selected. The issue of how to interpret 'null results' is not present under this framework as the goal is not to evaluate a null hypothesis but simply to determine which of two hypotheses should be preferred/selected. It is unclear why the authors attempt to present standard practice as matching the Neyman-Pearson approach, as NHST has strong Fisherian components in it as well that cannot be ignored and are also central to why NHST is a problematic tool for dealing with null hypotheses. I want to again refer the authors to the work of Gigerenzer, who extensively describes the hybrid nature of standard NHST and the fundamental problems this brings with it. I do not believe that attempting to characterize standard hypothesis-testing practice or even idealized standard practice as being in line with the Neyman-Pearson approach is correct or helpful for the purpose of the article: the focus on determining whether a null hypothesis should or should not be rejected is specifically Fisherian, and the issue of how to interpret null effects is not really present in a Neyman-Pearson approach that treats both hypotheses on equal footing. Effectively, there is no issue with null effects in the Neyman-Pearson approach, as there is no null hypothesis that gets a special status.

Authors' response:
The incorrect use of statistical procedures, as is unfortunately common with significance testing is indeed a problem. And the so-called "hybrid NHST" is an example of a statistical method used incorrectly. The reviewer states that our approach matches the standard hybrid NHST - but in our view, this assessment is incorrect. Our descriptions follow directly from the writing of Neyman and Pearson (1933) which we have studied carefully. We want to explain this in three points.

(1) The Neyman-Pearson approach extends the Fisherian significance test by introducing an alternative hypothesis and the concepts of Type-I- and Type-II-error. The crucial distinction is, that N-P requires an analyst to make a binary decision based on an a prior chosen significance level. We were careful to present the testing approach in this paper in the correct and coherent N-P tradition.

(2) The hybrid between Neyman-Pearson testing and Fisherian p -values emerges, when p -values are interpreted without a fixed significance-level as measure of evidence (most commonly without power analysis). We, again, have taken care not to lend ourselves and the reader to this interpretation. Indeed, we carefully explain the goal of controlling error rates in the section on the equivalence testing procedure.

(3) In the Neyman-Pearson framework, there actually is a difference between the null and the alternative hypothesis. The null hypothesis is the hypothesis under scrutiny (for a point or range hypothesis) - the alternative hypothesis is less precise and follows from the a priori power analysis, where the effect size of interest can be set based on several different

criteria. The terminology of “true/false positive” and “true/false negative” can only arise under this interpretation. The reviewer is, furthermore, not correct in arguing that there is no issue with null effects. As Neyman and Pearson (1933) write: "Every test of a statistical hypothesis in the sense described above, consists in a rule of rejecting the hypothesis when a specified character, x, of the sample lies within certain critical limits, and accepting it or remaining in doubt in all other cases." Because we can remain in doubt, null results are an issue in a Neyman Pearson approach. The correct use of the equivalence testing approach using two one-sided tests (TOST procedure) follows from a Neyman-Pearson approach to statistics, as we have carefully described. Finally, because a null-hypothesis test in a Neyman-Pearson approach can still lead researchers to reject practically meaningless results, the TOST procedure is an important complement, because it requires researchers to specify a null-interval of values practically equivalent to zero, which can then be used to distinguish statistically significant *and* statistically equivalent results - and important benefit to traditional NHST. For further reading, we refer to the literature already referenced in the introduction of equivalence testing in our manuscript.

When discussing the simulated data, it is now only mentioned that the data were generated using two normal distributions. Given that the different methods that are discussed all attempt to evaluate whether or not there is a difference in the means of the two groups in the example, it would be helpful to give the reader information on the mean and variance of the two normal distributions that were used to generate the data.

Authors’ response:
The details for the simulation are available in the accompanying scripts. We have added a footnote to clarify that the scripts are available. We do not consider it necessary to explain the data generation in detail in the running text: The paper is not aimed as a technical validation of any method and in real-world research projects, the true data-generating process is also unknown. If readers want to follow up on our example, make changes to see how the methods behave under different scenarios or simply check the code the accompanying R scripts provide all information necessary. The present paper is fully reproducible and transparent.

The authors claim (line 89) that while null effects are practically implausible, testing an exact null hypothesis is still useful for purposes of model comparison. But NHST is in its standard form not mainly used as a tool for model comparison, it is used as a tool for evaluating hypotheses. The point raised in the previous round of reviews therefore remains: If the null hypothesis can be rejected a priori, what then is the relevance of testing this hypothesis (rather than a 'no substantively relevant effect' hypothesis)? And if we should merely see it as a tool for model comparison, then this goes against what most applied users perceive NHST to be useful for, and this argument would also need to be made in much more detail.

Authors’ response:
First of all, we would like to point out that our paper is not an argument for testing a point null in significance testing - if anything we aim to complement frequentist testing with equivalence testing, where the null-hypothesis is not a point value of exactly 0, but (if chosen well) a much more relevant hypothesis to reject. The question of the usefulness of a point-null hypothesis in significance testing is an important question - and one that is still debated in the literature. We do not aim to settle this debate with the present paper. Without going into too much detail, it should be noted that hypotheses are, in fact, models (not in the Bayesian sense of generative models, though) and hypothesis testing is a form of model

selection. This is true for both Bayesian hypothesis testing using Bayes factors and
null-hypothesis significance testing.

The more fundamental question of the null hypothesis as a point of reference for testing (as is regular practice in significance testing and equivalence testing) the reviewer raises is whether the null can be rejected a-priori. For the area of randomized control clinical trials, for example, the research question whether a treatment is identical to a placebo might be worthwhile to investigate. Other areas of research might pose different questions and in any case, the statistical method needs to be chosen based on the substantive question at hand. The present paper does not seek and cannot answer this question and we think a in-depth discussion is beyond the scope of this paper, but we see our explanation of three alternatives of traditional significance testing all as underlining the importance of testing
other hypotheses than a point null.

Reviewer #3
The authors responded well to the first round of reviews, for which I thank them. This revision is much improved. I think the manuscript will make a useful article in the special issue of JCTR. I have only a few remaining suggestions for minor changes.

p. 4, line 65: "certain level of certainty" is awkward. Consider rephrasing.

Authors' response:
The line has been rephrased to: "Rare events will happen, and thus the absence of an effect is always concluded based on a defined probability of making an error, or given a particular level of certainty." (p. 4, line 65ff.)

p. 5, line 78: "the observed data was not..." Change to, "the observed tendency of the data was not...", or "the observed effect size was not...".

Authors' response:
The line has been rephrased to: "[...], the observed effect size was not sufficiently different from zero [...]." (p. 5, line 78f.)

p. 10, line 199-200: Change to, "... conclude the effect ***size*** *is statistically equivalent* ***to zero****..."*

Authors' response:
The line has been rephrased to: "[...], and conclude that the group difference is statistically equivalent to zero, [...]" (p. 10, line 199f)
\
Fig 5, panel D: The alternative-hypothesis prior is centered on zero in this Figure 5, but the alternative-hypothesis prior is not centered on zero in the example of Fig 4. This could be confusing to readers who are not familiar with Bayesian hypothesis testing, so it's worth mentioning, perhaps in the caption of Fig 5, something about the placement of the alternative-hypothesis distribution.

Authors' response:
The caption for figure 5 has been extended: "[...] (D) For the Bayes
factor, two models are compared that differ in their prior distributions: The M0 prior is a

point mass of 1 at an effect size of 0, the alternative model M1 is here plotted as a Normal
distribution centered on 0 as an example. Note, that other alternative models can be used, e.g. centered on a value derived from theory or previous studies (cf. working example and Figure 4)."

p. 26, line 529 and elsewhere: Here it is mentioned again that "the prior can be used to shrink or regularise parameter estimates." But this use is never explained or exemplified in the article, so it remains mysterious to readers who don't already know what it means. Mentioning it can be misleading because readers may think that the Bayesian estimation or Bayes factor results in the manuscript are affected by this mysterious shrinkage or regularization, when in fact they are not. This needs to be clarified. That's all. Again, I think this revision responded well to the first round of reviews. Thanks again for the opportunity to review this manuscript. John Kruschke

Reviewer #4
The authors did a conscientious job with the revision. I believe it can go to press now.

David Trafimow

3rd Editorial response

Date: 24-Jul-2018

Ref.: Ms. No. JCTRes-D-18-00012R2
Making 'Null Effects' Informative: Statistical Techniques and Inferential Frameworks
Journal of Clinical and Translational Research

Dear authors,

I am pleased to inform you that your manuscript has been accepted for publication in the Journal of Clinical and Translational Research.

You will receive the proofs of your article shortly, which we kindly ask you to thoroughly review for any errors.

Thank you for submitting your work to JCTR.

Kindest regards,

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Comments from the editors and reviewers: