Reply to Reviewer 1

2023-11-09

Dear Reviewer 1,

Thank you for your thoughtful feedback regarding this manuscript. Many suggestions have been thought provoking and this has been very helpful. I have taken care to address each of your comments and suggestions below. Your comments have been copied to this document in plain text, and my replies are below in italics.

There is no question from a statistical perspective that there is an issue here, and to the point of the author, this should be addressed. But somewhat distinctly, as you will see below, I would rather argue that the issue is with applying frequentist analyses at all since even with greater power a null result, even with equivalence testing, cannot truly be interpreted in the way the author seems to suggest.

I agree that applying a frequentist approach may not always be the best approach to statistical analysis in L3 acquisition (or providing evidence for equivalence). However, I argue that equivalence testing can provide evidence of performance within certain, theoretically guided upper and lower bounds, not necessarily that two groups are exactly alike. One point that I also mean to accentuate here is that when there are two theoretical outcomes being tested, they should both have clear and as objective as possible criteria for evaluation. In this case, the question was whether both groups showed evidence of access Spanish. In this case, support for the two theoretical predictions (that they do both or do not both show evidence of access to Spanish) come from the same statistical test. Specifically, as I argue in the body of the paper, it had been assumed that the lack of a main effect (of group) were support that both groups did indeed have access to Spanish, and that only the presence of evidence of a statistical difference would falsify this prediction. I argue that this particular paradigm alone is not appropriate for this conclusion. In the event that theory predicts two specific outcomes, evidence for each of these should be considered individually. In this example, a test of equivalence with specified equivalence bounds is just one way to provide evidence for both potential outcomes. That is not to say that other methods are not appropriate for a other situations. I have added a paragraph to the discussion mentioning that Bayesian approaches are also possible and likely more appropriate for other types of designs, as you mentioned.

In other words, it is not JUST a number=power issue (more below).
Let us consider a bit more deeply where I am going here. The models do not assume or need to that any two groups are exactly equivalent, they are about influence from a source language that given the methodology both groups have access to. Keep in mind that well designed studies test the specific property to be examined in the L3 in both other languages, including only those showing evidence of the target grammatical representations in each language is available for transfer/influence (matching the descriptive literature of the target).

My point here is not that groups need to be exactly equivalent, but that we need to specify how close is "close enough" to conclude that both groups behave similarly, just like we should consider how different is "different enough."

Also in well-designed studies, the property is important in that there is a clear distinction in the representational space. Because, at least to properly test the most restrictive model in terms of timing of its predictions, the TPM, task design is relatively simplistic for a relatively novice learner to do—to juxtapose the models properly one must be testing at the earliest stages of L3 exposure (of course this is not to suggest that testing at other stages of L3 development is irrelevant, but for testing models that includes the TPM there are issues beyond early exposure (see Puig-Mayenco & Rothman, 2020; Rothman, González-Alonso & Puig-Mayenco, 2019). And so, in principle, all participants have access to both grammatical representations. As a result, failure to show that either—irrespective of order of acquisition—shows signs of influence of one of the available languages is itself significant in the theoretical space, that the language not showing influence when it was available was the same one already tells you something about any default role order of acquisition likely plays (that is, not) and so on. In other words, it's not just that the two groups seem to "use" the same source resulting in a null effect, they also fail to show influence from the other available language at all.

I specified more about the implications for L3 models. My idea was more that, while the TPM and LPM both suggest that order of acquisition is not a determining factor, that the LPM suggests that influence from English would be seen in relatively performance on a single structure compared to an L2 group. This, of course, was outside the scope of this replication. I used this example to talk about how models should specify the magnitude of differences they expect to find and whether this is evidence of full transfer.

Major comments: I am very confused about the Results section. Firstly, for statistical reports, when a main effect is claimed (significant), p values, sometimes, are reported to be above .05 (to name a few, line 271, line 273, line 274, line 282, among others). I would assume the authors meant to use "<" instead of ">".

These were indeed in the wrong direction. They have been corrected.

Secondly, the authors conducted an a priori power analysis, showing a total sample of 214 (assuming a balanced design) is required to reach a power of .8. However, the final sample is 211 (not a huge difference I understand) and the groups do not have a balanced number. I am wondering if the authors have adjusted the equivalence bounds for this. If not (or even if so), the power issue that is central to the paper is still there. I suggest that the authors estimate the minimum detectable effect, i.e., a sensitivity analysis considering the final sample size was conducted.

A sensitivity analysis has been added to the "sample size justification" section, which reveals that the current sample size had the power to detect a minimum effect size of d = .39 (at level .8 and alpha of .05). For equivalence, the equivalence bounds at the current sample size were -.40388 and .40388. Thus, the difference between .4 and .40388 is a difference in .0388 standard deviations which corresponds to .09% difference in correctness in raw units. I assume that this difference is negligible and have added this information to the manuscript.

Thirdly, I understand the rationale of conducting t-tests along with equivalence tests. I am not sure why t-tests were needed when no interaction terms were found for the collocation task, although it does not matter in the end because these two t-tests gave the same interpretations. However, multiple comparisons should be corrected at least.

T-tests were chosen because they can directly be compared to the Tests of Equivalence - I added this information to the statistical analysis section. As far as I know, corrections for multiple comparisons should only be used if the model itself changes (adding variables), given that the t-test was used on the same data as the ANOVA and performs an analogous purpose, I do not think that this is a correct use of a correction for multiple comparisons.

Now, assuming every step was corrected, power is not of concern, and the writing is corrected, when the authors claimed "not equivalent and not different", I suggest the authors to further specify that even under TOST, this finding should be interpreted as: the current study cannot exclude meaningful effect sizes and that the result is interpreted as not equivalent and not different (i.e., insufficient data to draw a conclusion).

Thank you, I added this information to the manuscript.

Lastly, although I agree with the authors that TOST methods should be in the toolkits for us, it should be of a post-hoc nature, which I think the authors needs to point out. At the same time, what is, to me, one the of central problems the authors have with the original study, i.e., sample size and statistical power, remains for TOST methods. TOST methods also can be more restrictive than some other statistical modelling, e.g., what about repeated measures?

I added information to clarify: the use of a TOST here was specifically chosen so for the purpose of comparability to the original statistical analysis. Bayesian methods (using a specified ROPE) can and should also be used to determine practical equivalence. I disagree that this needs to be post-hoc. If we design a study with two outcomes, it is a fallacy to say that lack of evidence for one outcome is evidence for the other. Each *of the two possibilities needs its own set of criteria for evaluation, and good evidence for practical equivalence should entail both the absence of evidence of a difference (such a non-significant main effect) and specific, well defined evidence for equivalence. Both in a Bayesian Framework and a frequentist TOST, equivalence bounds need to be specified.*

Other comments: Page 5, line 95 to 98. Although I agree with the authors that the lack of a main effect does not entail practical equivalence, simplifying its reason to low power and high uncertainty is not optimal. Under the NHST if the null hypothesis is true, then the probability of ANY p-value is equally likely. That is to say, even with large sample size and lower population uncertainty, NHST doesn't "approve" the null hypothesis. It can only provide evidence against the null hypothesis if the p-value is below the chosen significance level. If the p-value is above the significance level, it doesn't prove that the null hypothesis is true; it just indicates that there isn't enough evidence to reject it.

I added specifically that NHST cannot provide evidence for the null, only against it, and moved the low power high uncertainty portion to a difference place in the manuscript

Page 8, line 166. Change filters"Country of Birth" to filters "Country of Birth".

Edited

Page 8, line 170 to line 173. Any statistical differences between these two groups?

I added detailed reporting of additional tets for each narrative comparison.

Page 9, line 195 to line 198. Are these differences statistically significant?

I added detailed reporting of additional tets for each narrative comparison.

Page 12, line 246 to line 247. Could you please specify more about the randomisation? Do you mean 1. participants finished LexTALE in one or more randomly selected language(s); or (2) participants finished LexTALE in all languages but the sequence of the test for each language is randomised?

The order was randomized and each participant completed the LexTALE in all 3 languages. I have added this information.

Page 12, line 250. Could you actually anonymise the OSF project and share the OSF link for review? I assume your R scripts are there as well, which would address some of my concerns.

Here is the OSF link: https://osf.io/t2vkp/?view_only=ec60116dde2846f1ade332e363452314

Page 19, line 370. I wish to caution the authors that under the frequentist approach, any results from a single study (regardless of the statistical power), are inconclusive, more so when the null effect is not falsified. Each individual study should be taken as a coin flip on if you found something or not. The purpose of NHST is to ensure that if you keep repeating the study that way and examine all the results together, they will produce false positives 5% of the time and false negatives 20% of the time (assuming alpha as .05). You will never know if any single one of the studies constitutes the errors. That being said, I completely agree and applaud the authors to promote publishing null results (and include all data information) for future meta-analyses. Also given related concerns articulated above, it would be nice to acknowledge that a non-frequentist approach might be the best way to go in future if equivalence must be pushed, somewhat address the real world issues I mentioned with this in the beginning part of the review.

Thank you for this point. I added a summary of this point in the discussion.