November 14, 2024

Dear Dr. Dienes,

We appreciate the additional opportunity to revise our Stage 1 submission titled "How Does Model (Mis)Specification Impact Statistical Power, Type I Error Rate, and Parameter Bias in Moderated Mediation? A Registered Report". We thank you and the reviewers for such thoughtful, helpful feedback. We appreciate the opportunity to revise our manuscript and resubmit this updated version. We were very grateful to have two of the same reviewers from the original submission, and we were glad to see that reviewer Mijke Rhemtulla was satisfied with this revision.

For this submission, we have made several changes to the manuscript: 1) including the Study Design Table at the end of the main manuscript in addition to the OSF project page, 2) including specific criteria for full, partial, and non-support of our hypotheses, and 3) improving the clarity of the analysis plan.

We believe that these revisions have met the requests of the reviewers and editor and demonstrate a clear step toward an in-principle acceptance as a Stage 1 Registered Report; however, we are open to additional feedback and rounds of revision.

On the following pages are responses to the revision requests, which are numbered for ease of reference. The reviewer comments are in normal text, and the responses are in bold text. We look forward to your response.

Thank you,

Jessica L. Fossum

Seattle Pacific University
3307 Third Ave. W., Suite 107
Seattle, WA 98119-1922

email: fossumj@spu.edu
phone: 206-281-2252

Editor: Zoltan Dienes

One major point: You appear to have lost your study design template.

We had included the study design template with the rest of the supplemental material on our OSF page (https://osf.io/vgkdt/?view_only=), but mistakenly did not include it at the end of the main manuscript document. We now include it in both places for this revision.

One crucial point of inference: You decide not to use equivalence testing because your hypothesis does not predict equivalence. But a test of a hypothesis is not severe unless it could show the hypothesis wrong; so it is precisely because the hypothesis predicts a difference that is inferentially desirable equivalence can be concluded (that is one reason why I prefer to call the inferential procedure using null interval hypotheses and confidence intervals "inference by intervals"). Otherwise how do results show the hypothesis wrong? Non-significance in itself does not do that. You say "We also report 99.9% confidence intervals and odds ratios to contextualize the results further." The motivation for this is of course that non-significance by itself does not do the job; indeed, for high N simulations, significance may not do the job either (if an irrelevant effect size is found significant). The problem in terms of a Registered Report is that statement that results will be contextualized allows inferential flexibility in. Just how will you do the contextualizing?

This is where the Design Table comes in: One is asked to nail down the whole inferential chain - including how one may draw different conclusions dependent on different patterns of data.

I am not saying you should use inference by intervals; just that if you test a hypothesis, its severity should be at least known. One alternative is just to estimate with CIs and drop p-values; then the hypotheses (Still listed in the DesignTable) are not about what effect exists, just a question about how big each effect is. (Then one must be scrupulous in not drawing any existential conclusions.)

You have done a good job of specifying how you will draw inferences in many cases; you just need to make sure you now make everything watertight.

We appreciate these points. We have taken steps to (1) ensure severe testing of our hypotheses and (2) specify criteria that result in clear conclusions as to whether the hypotheses are supported or refuted. We have carefully addressed each of our hypotheses to provide explicit criteria that need to be met for the hypothesis to be supported, refuted, and, in some cases, partially supported. These criteria have been added to the manuscript as well as the Study Design Table (final column). We present an example for Hypothesis 1a below. The adjustments we have made to our decision criteria not only create a watertight template for future interpretation of our results in light of our hypotheses but also address the problem that the p-values alone from the statistical models may not yield a severe test (e.g., two conditions are significantly, but not practically different). We do not directly use interval-based equivalence testing on the

coefficients from the statistical models given that we want to define the equivalence region on the most interpretable metric, which here is the raw scale of the outcome (e.g., statistical power) rather than the model-based coefficient values or effect sizes. However, our revised criteria, which focus on the raw statistical power, Type I error rates, and parameter bias values, are in the spirit of severe testing principles and allow us to clearly support or refute each hypothesis.

Example for Hypothesis 1a Interpretations in the Study Design Table:
Criterion A: more than 20% of conditions show greater than a 3% difference in power such that power is lower when models are over-specified compared to correctly specified for both continuous and dichotomous moderators
Criterion B: both coefficients for model specification are significant (p < .001)

H1a supported: A and B
H1a partial support: A not B
H1a refuted: not A (regardless of B)

Reviewer 3: Reny Baykova

1. Including a statement which specifies which parts of the study have been completed and which have not is good idea. However, this should also already be clear as one is reading the stage 1 report - everything that has already been done should be in past tense, and everything that has not been done should be in future tense. Verb tenses can be changed in the stage 2 registered report.

   We apologize for the lack of clarity. We had aimed to write everything in past tense to avoid having to change language for Stage 2, but it is clear now that this made it difficult to evaluate which parts of the research have been completed and which have not. All language describing studies that have not yet completed are now described in future tense.

2. On page 6, the first sentence of the last paragraph is unclear - "When the indirect effect is moderated...". Here, it is not clear what "indirect effect" is being discussed.

   We have added in a parenthetical to that sentence indicating that the indirect effect is the effect of X on Y through M, which is the mediated effect in all the moderated mediation models compared in this paper.

   "When either path that makes up the indirect effect (i.e., the effect of X on M or the effect of M on Y) is moderated, the conditional indirect effect quantifies the indirect effect at a specific value of the moderator." (page 6)

3. On page 11, hypothesis H1c states: "we hypothesized that parameter bias for over-specified models would be acceptable (<10%) in each condition". Up to this point,

no conditions have been discussed explicitly, so it is not clear what is meant by "each condition".

The sentence is revised to mirror the hypothesis in the same section about parameter bias for under-specified models, moving all mention of conditions to later in the manuscript where the table of simulation conditions is referenced.

"We hypothesize that parameter bias will be acceptable (<10%) for over-specified models (H1c)." (page 11)

4. On page 12, hypothesis H3a states: "We hypothesized that the type I error rate would be too high (liberal) in completely misspecified models". It is not clear what would be considered"too high". Hypothesis H3b should also be described in more detail - it is not clear what amount of bias would be considered "unacceptably high".

    The criterion for error rate being too high has been added to H3a, and equivalent criterion for unacceptably high parameter bias has been added to H3b.

    "We hypothesize that the type I error rate will be too high (liberal) according to the criterion set by Bradley (1978, > 0.075) in completely misspecified models (H3a). Additionally, we hypothesize that raw bias will be unacceptably high (greater than .00286, which is the raw bias equivalent of 10% relative bias) for completely misspecified models (H3b)." (pages 11-12)

5. On page 17, the conditions under which H1b would be supported are unclear. The text states: "H1b would be fully supported if all four coefficients for the number of moderated paths are significant... " . Here it is not clear what the "four coefficients for the number of moderated paths" refer to. If there is one model for dichotomous W and a second model for continuous W, then there will be 2 coefficients related to the number of moderated paths. It is not clear where the other 2 coeffients are coming from.

    There are two models (one for dichotomous *W* and one for continuous *W*), then for each model there are two coefficients since this is a sequentially coded categorical variable with three levels (1, 2, or 3 moderated paths). We have added in a description of the categorical variable coding strategy to indicate that it is two coefficients from both models, along with our additional criteria for hypothesis support reflected in the Study Design Table.

    "H1b would be fully supported if more than 20% of conditions show greater than a 3% difference in power such that power is lower when over-specified models have more moderated paths and all four coefficients (two for dichotomous moderators and two for continuous moderators) for the sequentially-coded variable representing number of moderated paths are significant." (page 17)

6. On page 17, the conditions under which hypothesis H1b would be partially supported need to be specified in more detail. The text says "If only some of the coefficients are significant in the predicted direction, H1b would be partially supported". "Some" is not specific enough.

   We have revised this section to address the ambiguity, in combination with addressing the editors concerns about setting stricter tests. This section now reads:

   "H1b would be partially supported if more than 20% of conditions show greater than a 3% difference in power but the coefficients are not all statistically significant." (page 17)

7. On page 17, it is not clear which models will be used to assess hypothesis H1c. It is also not clear what is meant by "conditions".

   H1c focuses on over-specified models. We have changed the ambiguous word "these" in that sentence to "the over-specified" in hopes that it improves clarity.

   "To test H1c (acceptable parameter bias for over-specified models), we interpreted the parameter bias resulting from the over-specified models." (page 17)

8. On page 17, it is not clear how hypothesis H2b will be assessed. Same as above, it is not clear what is meant by "conditions". As a general comment, the analysis plan will be much clearer and eaiser to followig if the tests corresponding to each hypothesis are described in detail.

   Instead of referencing H1c for comparison, the exact plan for H2b is now stated, clarifying that only under-specified models will be included.

   "To test H2b (unacceptable parameter bias for under-specified models), we interpret the parameter bias resulting from the under-specified models. H2b would be fully supported if many (>20%) of the simulation conditions result in a relative bias value of over 10%, and we would interpret this as a particularly high risk for the minimalist approach. Partial support would be if between 10%-20% of the simulation conditions resulted in a relative bias value of over 10%. If H2b is fully or partially supported, we will examine patterns among unacceptable bias values. If we see relative bias over 10% in over 20% of simulation conditions, H2b is refuted." (page 18)

   To address the broader concern that "conditions" are unclear, all references have been changed to "simulation conditions" specifically, which are described in more detail in the section with that same heading. Table 2 gives the details for each simulation condition. Additionally, we hope that the inclusion of the study design

table with the main manuscript will make the analysis plan easier to follow. Due to our target journal having a word limit of 5,000 words, we are hesitant to include any additional word count in the main manuscript.