July 9, 2024

Dear Dr. Dienes,

We appreciate the 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 apologize for the long delay between submissions and understand if this may impact the ability to receive feedback from the same reviewers. The lead author of the project was transitioning to a new position over the course of this academic year which severely impacted the ability to work on the revision, but we continue to be motivated to complete the project as proposed.

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 have made several major changes to the manuscript: 1) removing all mention of exploratory analyses, 2) revising hypotheses to include how we will determine if the hypothesis is supported, 3) recreating the figures to be easier to read, and 4) integrated in parameter bias throughout the study to ensure that we are providing a balanced account of the relative pros and cons of over vs. under-specifying models.

Additionally, we have moved the Systematic Review to the Appendix. This was motivated both by wanting to shift the focus to the proposed study and also by the very low word limit at our target journal. We are trying to stay well within the limit to ensure an easy transition to the journal.

I believe that these revisions have met the requests of the reviewers 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 short 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. I 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 additional point on your inferential procedure. You say that "only effects with an odds ratio greater than 1.68 will be considered meaningful," but also that you will reject a hypothesized effect on power if the effect is non-significant. In itself, non-significance is consistent with any population effect size, hence one over 1.68. So this leads to a possible inferential contradiction. Why not set up an equivalence region/null interval of +- ln 1.68 for ln OR (or make it one sided if you like), and if a suitable CI is wholly outside the null interval conclude power is meaningfully affected, wholly inside it conclude that power is not meaningfully affected, and otherwise suspend judgment? In any case, as you have provided no grounds for asserting non-significance is meaningful, some justified inferential procedure is needed to allow the conclusion that power is not affected by a variable/interaction.

**We appreciate your equivalence interval suggestion. Upon reflection of our goals, we opted to go with an alternative approach - our hypothesis is that power differs across model types, while an equivalence approach would be better suited if our hypothesis predicted no difference. In the spirit of your comment, however, we have revised our inferential procedure to avoid any contradictory inferences. For power comparisons across model types, we will conduct hypothesis testing with an alpha level of .001 as our cutoff level for statistical significance, reporting the odds ratio as a (descriptive) measure of effect size, along with the corresponding 99.9% confidence interval. We will also focus on descriptive trends, as suggested by Reviewer 1.**

On another point of inferential procedure, to clarify a point by Baykover, for a registered report, pre-register only analyses where the full inferential chain is nailed down; that is, to keep things clear, exploratory analyses are not mentioned in the Stage 1. They can of course be reported in the Stage 2 in a separate exploratory results section.

**We have removed any mention of exploratory analyses from this Stage 1 manuscript. We will plan to only include exploratory analyses in the Stage 2 in a separate exploratory results section.**

**Reviewer 1: Mijke Rhemtulla**

1. First, my only concern about the proposed simulation is how to interpret results in the under-specified condition. This concern stems from the fact that this paper focuses on power and not on the consequences of missing out on true moderation effects. So in the underspecified condition, suppose you find high power, would you conclude that under-specification is a good thing? Or no big deal? Or that a minimal strategy is good? I guess I'm a little worried that the study is set up to find that over-specification (which, to my mind, is not really a misspecification, because nothing is missed – we're simply estimating zeroes rather than fixing them at their true unknown population values) is worse than underspecification, in which the model is not just inefficient but actually

wrong. In the analysis section, the authors write that “We believe that in some cases underspecification may increase power and in other cases decrease power.” That’s my intuition as well – for this reason I think it’s important to clarify why you think it’s worthwhile to examine these conditions, and how the results will be interpreted or translated into recommendations for practice. It might be worthwhile to consider also reporting bias in addition to power, because bias may be able to reveal the consequence of the misspecification even when power is high. I think the paper could benefit from some discussion of whether missed moderation effects are something we should worry about – although the complete misspecification condition will at least shed some light on the possible consequences of missed moderation effects on type-I error rates for other effects.

**Thank you for your concern on this issue, and this comment has greatly impacted our thinking about the study and the outcomes that we will collect. In particular, we have integrated in parameter bias throughout the study to ensure that we are providing a balanced account of the relative pros and cons of over vs. under-specifying models. Your intuition about under-specification as highly risky is shared by us, and we agree that the previous design did not fully capture the potential for error (even in the presence of good power) when under-specifying a model. We believe that by incorporating parameter bias in the design, hypotheses, and conclusions we will provide a more balanced picture of the costs of under-specifying a model.**

2. I’m not clear on the total number of simulation conditions. The design is described as 6x9x6x2x2x6, but not every one of those cells exists, and some factors are not included in this design – like the 9 sample sizes. (There is a sentence reading “Effect size on the interaction term and sample size were varied”, which made me think that effect size on the interaction term is a different factor than the 2-level “effect size” factor that is described as a between-level factor in the simulation. At the bottom of page 17, the interaction is said to account for 1%, 3% or 5% of explained variance – so I guess this isn’t the 2-level effect size factor, because it has 3 levels. (Also at some later point it was mentioned that the interaction effect is sometimes negative but I’m not sure I saw that in the design). In short, I’m a bit confused about the setup. A table would be helpful, as well as a clearer description of what was varied and how it was varied, and what the total number of simulation conditions is.

**Thank you for this feedback. We now include Table 2 that lists each design factor, along with all the possible levels of each design factor. We have this aligning with the paragraph describing the design as 6x9x6x2x2x6 (revised to be 6 (Between: Generating Model) x 9 (Between: Sample Size) x 3 (Between: Effect Size Magnitude) x 2 (Between: Normal or Dichotomous *X*) x 2 (Between: Normal or Dichotomous *W*) x 6 (Within: Analysis Model) factorial design). The 9 is for the between subjects factor sample size. Not every cell exists, but the only exclusions are Models 58 or 59 when W is continuous. This is now described in the table**

**note. We have removed the effect size sign and choose to focus on positive effects only for simplicity and because it has been shown in previous simulation studies to not make a difference (Fossum & Montoya, 2023).**

3. "Interaction terms such as XW also had a variance of one" – how was this ensured? Was the product term rescaled to have variance of 1? More generally, I wanted to read about how the data were simulated – was it piecemeal (e.g., you first generated data on X and W and errors and then computed M and Y from those?) or did you use a joint distribution (in which case, how did the interaction terms work)? What was the correlation between the interaction terms and X and W? Etc.

   **We relied on the following equation to generate a product term with a variance of 1: Var(XW)=Var(X)Var(W) + Var(X)(E(W))^2 + Var(W)(E(X))^2 which applies if X and W are independent. We generated both W and X to have E(X) = E(W) = 0. We achieved this by generating balanced dichotomous variables (X and W) coded as -1 and 1, this also results in a variance of 1 for X and W. For normally distributed variables, the mean was set to 0 and the variance was set to 1. This sets the variance of the product term to be 1 in expectation, but is not fixed to be 1 in any given sample due to sampling variability. We now include this description along with more details about how the data were simulated (piecemeal) in the simulation procedure section.**

4. Is nonconvergence ever an issue with mixed effect logistic regression models? If it might be, please specify what you intend to do with nonconverged replications (will they be included in the analyses? will they be reported? will another estimator be used as a backup?)

   **We do not anticipate nonconvergence being an issue for these models. In a pilot version of this study with only 1000 replications in each condition, nonconvergence was never an issue.**

5. "We chose this effect size metric instead of statistical significance due to the large amount of data likely favoring statistical significance." I agree that significance tests are inappropriate as a way to interpret results of a simulation, and that effect sizes are more appropriate, though my personal preference would be for you to simply plot all the results (with confidence intervals) and describe the trends that you observe. But I'm confused here because this sentence suggests that significance tests will not be used, but the next sentence goes on to give a p-value threshold and the remainder of the analysis section describes how results will be interpreted on the basis of their statistical significance. At a minimum, the rationale for the analytic choices should be clarified. I would be very happy to see a strategy with no p-values involved.

   **Thank you for this suggestion. Given the issues with attempting to simultaneously use p-values and effect size thresholds, we e have decided to rely on significance**

**testing when comparing power across the different types of models (e.g. H1a). However, we (1) use a much more stringent alpha level of .001 due to the large number of replications and (2) report ORs and 99.9% confidence intervals to contextualize the results further. We also plan to observe trends descriptively, as you suggest. Finally, for hypotheses wherein the different types of models are not explicitly compared (e.g. H1b, H1c), we plan to focus on descriptive trends.**

6. I'm unclear on why the effect size variable is sometimes coded sequentially and sometimes as two variables (valence and absolute value). It would be helpful to see an explanation of why the coding of this factor changes.

   **Thank you for pointing out this discrepancy. We intended to test both the magnitude of the effect size on the interaction term (three possible values, sequentially coded in order of size), and also the valence (positive vs. negative). Upon further reflection, we have opted to just test the effect size magnitude (additional context given in response 2). We have gone through and edited the manuscript to remove negative effect sizes.**

7. Hypothesis 1b describes an interaction, which will be tested by including an interaction term in the analysis model, but then it sounds like only the main effect will be interpreted as evidence for the hypothesis ("a significant effect of number of moderated paths in the analysis model"), and no mention is made of the interaction effect.

   **We have now clarified this hypothesis. Because Hypothesis 1b is only concerned with how the number of moderated paths affects power for over-specified models, we propose this as a subgroup analysis. Specifically, we plan to test this by assessing the effect of (sequentially coded) number of moderated paths in a multilevel logistic regression model that removes the model specification main effect (given that the sample is now only over-specified models). If both coefficients (2 vs 1 path, 3 vs 2 paths) show a decrease in power, H1b is fully supported (partially supported if only one of the coefficients is significant). We believe this captures our intention more accurately than testing an interaction which would be sensitive to the pattern of effects in correctly specified models.**

8. For the analyses, I wanted to see more detail on how the models were fit (in R? using lmer or brms? etc.). Including equations for the linear mixed effects models, or lmer code (if that's what will be used to run them) would help to make the analyses totally transparent. I wasn't sure, for example, if "random intercepts only" means random intercept for condition, or if you intended a more complicated set of random intercepts.

   **Our analysis script is included in the supplemental material on OSF, and clarification on the random intercepts being random intercepts for each simulated dataset: "we will use multilevel logistic regression with random intercepts for the within-subjects factor of data analysis model to predict rejection".**

9. There seems to be a disconnect between using the percentile bootstrap confidence interval at 95%, as described, and using a Wald test (mentioned several times as a method for determining significance) and using a p-value threshold of .001.

   **After further refining our analyses, there is no longer any comparison of a set of coefficients, so we have removed the Wald test and focus entirely on significance tests at the p value threshold of .001 for the models and presenting results in tables and figures. The 95% confidence intervals apply to the simulated data, where that is the threshold for bootstrapped confidence intervals in calculating power or type I error rate..**

10. Figure 1: it took me a minute to find the conceptual vs. statistical models – the description "(right)" had me thinking that the right column of the figure would have the statistical models, but instead it's the right figure within each cell. It would be very helpful if these figures were re-drawn to have larger arrowheads and larger coefficient labels – I had to zoom in on the pdf quite a lot to decipher these.

    **Thank you for the helpful ideas for improving the figure. We have remade the figure to be larger, specifically with larger arrowheads and coefficient labels.**

11. Table 2: Consider using boxplots here to show the range of sample sizes in addition to the median for each model? I love this systematic review and think it would be neat to see a little bit more detail in these sample sizes that are presented.

    **Thank you for this suggestion. We think it's a great idea, and have included boxplots in Figure A1 to add in more detail. Due to several high outliers, those values are listed and only sample sizes under 3,000 are included in the figure (and the specific higher sample sizes are noted in the figure notes).**

12. I found it hard to keep straight the research questions and associated hypotheses – maybe a succinct table that lays these out would be valuable? I'm not sure to what extent these are just laid out like this for the stage 1 RR vs. the final paper.

    **We have revised the Study Design Table in the appendix to be more succinct, limited to just one page per hypothesis. The research question and associated hypotheses section in the main manuscript has been revised for conciseness and clarity.**

13. Equations: I found the verbal description of all possible equations for M and Y given in the text around Equations (1) – (6) not very easy to read. I wondered if putting these 6 equations in a table, or include them in the Figure that shows all 8 models, with the path

labels corresponding to the equation coefficients, would be both clearer and more succinct.

**We have incorporated the equations into the statistical diagram figure, and removed them from the in-text description in the manuscript. Thank you for this suggestion which made the manuscript clearer and more succinct.**

14. p. 2 “whether a proposed mediator” à “whereby a proposed mediator”

    **Addressed**

15. p.2 re: the WebofScience count increasing from 2020-2022, is there a denominator number that can be pulled from WebofScience to indicate whether moderated mediation models are increasing as a proportion of published articles vs. following the trend of increasing publications overall?

    **To our knowledge, it is not possible to get a denominator number from WebofScience. It appears to be possible with Scopus, but we do not have access to that content. We still feel that it is valuable to report that the number of academic articles being published on this subject is increasing, so the sentence now reads "Moderated mediation analyses are used across disciplines, with WebofScience counting 2,602 published articles using the analysis in 2020, 3,499 in 2021, and 3,815 in 2022."**

16. p. 2 “it’s” à its

    **Addressed**

17. p. 3 “detect a small effect” à should this be a small mediation effect? or is that effect size for a main effect? (or is power the same?)

    **Thank you for pointing this out. We have revised the comparison to read "a mediated effect when both paths involved in the indirect effect are small to medium (an effect size common in psychology)"..**

18. p. 3 “foul play” is pretty judgey – I think many proponents of science reform encourage the narrative that p-hacking is usually unintended, not malicious.

    **Thank you. We have removed "foul play" from the manuscript, and instead just left in p-hacking in that sentence.**

19. p. 4 “methods” à “method’s”

**Addressed**

20. p. 4 “types I”

    **Addressed**

21. p. 5 “estimated with [a] commonly used [SPSS?] macro”

    **Revised to include all available versions of the macro, "estimated with a commonly used macro available for SPSS, SAS, and R: PROCESS"**

22. p. 9 “model being estimates” à estimated

    **Addressed**

23. Table 1 note: “assume dichotomous moderated is with” à moderator has?

    **Addressed**

24. p. 10 “dichotomous vs. continuous predictor variables” does predictor variables include the moderator?

    **Correct, predictor variables included the moderator in that study. We had originally included the idea that dichotomous vs. continuous predictor variables affect power and sample size planning citing the McClelland and Judd (1993) paper, but in that paper the way the variables were generated impacted the results and it has been shown to not have an effect (Coutts, 2023). As such, we have removed that from the sample size planning section and instead consider referencing it in the discussion section.**

25. p. 10 “tools available to [do] sample size planning”
    **Addressed**

26. p. 11 “researcher[s] may need”

    **Addressed**

27. p. 11 “Statistical power [analysis] for moderated mediation…”

    **Addressed**

28. p. 12 “at least one additional path is moderated” à “allowed to be moderated”?

    **Addressed**

29. p. 12 “can introduce excessive collinearity, especially with the interactions” à Is this true? I’m not overly familiar with the debate, but I was under the impression that the thought the idea that not centering interaction terms can result in multicollinearity had been debunked?

    **The issue we are specifically referring to is collinearity between multiple interaction terms (e.g., XW and MW in a model for Y), rather than the very common misconception that centering variables as part of an interaction somehow changes the results due to reduced multicollinearity. This distinction is now made more clear in the manuscript itself.**

30. p. 12 “Under-specification … may also add unnecessary parameters” this sentence confused me – I thought it meant that by omitting one parameter, that would produce another phantom effect (i.e., the whack-a-mole effect whereby leaving something out causes that omitted effect to go somewhere else) but I think you just meant that by under-specification you include situations in which a parameter is under-specified but it may also be over-specified w.r.t. other parameters.

    **We have revised the description and organization of the different types of specification used. You are correct that we meant that a parameter may be under-specified and other parameters can be over-specified, and this would still be considered underspecification. This was a difficult case to consider, so if the reviewer has other suggestions for how to describe these cases, we would be open to alternatives. We would like to note that we plan to, in exploratory analyses, compare purely underspecified models (models which are only underspecified) to mixed models (models with both over- and under-specification). However, we have not mentioned this in the Stage 1 submission, as exploratory analyses should be omitted to make clear what is expected a priori vs. a posteriori.**

31. p. 12 “where none of the paths included in the indirect effect are correctly moderated” is unclear – does “included in the indirect effect” refer to the DGP or the analysis model? does “correctly moderated” mean that they are correctly allowed to be moderated? The end of this sentence “should be 0 according to the DGP” also totally confused me for a while – I eventually figured out what you mean by complete misspecification but this description is not clear. It might be helpful to describe that for “complete misspecification” to occur, (a) the DGM contains exactly one moderated path in the indirect effect, that is left out of the analysis model, and (b) the DGM contains exactly one unmoderated path that is included in the analysis model. (Is that right?)

    **We have revised this section to make this point more clear. We now describe complete misspecification as "where the DGP includes moderation on a path that is moderated in the data analysis model, and the data analysis model includes**

**moderation of a path that is not moderated in the DGP" (page 10). The specific example you gave is (conditions a and b) are correct given this context; however, we wanted to provide a definition that is generalizable to more complex models if needed (e.g., serial moderated mediation). Additionally, we provide an example of specific models to make our definitions more clear.**

32. p. 13 “base don”

    **Addressed**

33. p. 13 “By incorrectly specifying where the moderation occurs in the model, researchers may get biased estimates of the paths, coming to incorrect conclusions about which paths are moderated.” à this can also be a consequence of some kinds of underspecification, right?

    **That is correct. We now make very clear how parameter bias may be problematic for both under-specified models and completely misspecified models. By incorporating parameter bias into our study hypotheses, this is much clearer in the current version of the manuscript. Page 10 is now revised to include "Under-specification omits important elements of the DGP, which could bias parameters and lead to incorrect conclusions about which paths are moderated (Yzerbyt et al., 2018). This is a risk of the minimalist approach to model misspecification."**

34. p. 14, “This study examined the effect of model specification (over-, under-, or correctly specified)” – also completely misspecified.

    **Addressed**

35. p. 14, “continuous moderators and X focal predictor variables” – is the X meant to be there?

    **Addressed**

36. p. 14, “Power was only assessed for over-specified models because … and under-specified models because…” the “only” threw me off here. I first read it as “only for over-specified models” but then you added under-specified models. Then I read it as “only because”, as in “we only assessed power for this reason”. I think it may be clearer to say that only power was assessed for over- and under-specified models, and only type-I error was assessed for completely misspecified models

**Thank you for this suggestion. Those sentences are revised and substantially updated for clarity in the Performance Metrics section of the manuscript in the Method section.**

37. p. 15 “type I error rate will increase as the number of incorrectly moderated paths increases” – clarify incorrectly moderated?

    **We have revised the sentence to clarify. "and type I error rate will be highest in cases when the data analysis model moderates both the direct effect and the wrong path for the indirect effect compared to the DGP that only moderates the indirect effect (H2b)."**

38. p. 19 “a path with an interaction term from the DGP must be included in the data analysis model”: clarify that “must” follows from the way you defined over/under-specification.

    **The Performance Metrics section has been revised and no longer includes this sentence.**

39. p. 19 “we use the criteria from Bradley (1978)” clarify what these were (.025 /.075?)

    **Yes, thank you. We've clarified that Bradley's liberal criterion is [.025, .075].**

40. p. 21 “will be supported would be supported”

    **Addressed**

**Reviewer 2: Pier-Olivier Caron**

1. 5000 replication is probably enough to answer the hypotheses. Even though, with my own experience with moderation analysis, 50000 would be preferable to estimate parameters and their variability (which is not an objective of the project).

   We believe that 5,000 replications is sufficient, especially for the context of mediation as each sample needs to be bootstrapped 1,000 times which can make run times much longer than in moderation contexts where bootstrapping is not typically used. The largest possible Monte Carlo error for 5,000 replications occurs when power is exactly 0.5 (the probability at which the variance of a binomial distribution is maximized). At this value the Monte Carlo error is .007, meaning that we can determine differences in power reliably in the 1-2% range which we determined to be sufficient for this project.

2. I am unsure what will be the outcome, more specifically, the guideline that will emerge from this study. I am worried the message might misleading for applied researchers: to emphasize on rejection rates rather than on correctly specifying a model can send the inappropriate message that “model specification is less important than mere statistical power”. I see this future article to be cited specifically for this purpose.

   **Thank you for raising this issue. We have revised the manuscript to emphasize that correct model specification should always be prioritized above mere statistical power. We believe the manuscript still provides an important contribution, as researchers are often unsure whether to lean toward over-specification or under-specification. For example, in many consulting meetings researchers hypothesize a specific path is moderated in the indirect effect but do not have clear hypotheses about the indirect effect. Often in these cases, researchers ask statistical consultants what they should do. Indeed these experiences are largely the inspiration for the work presented here. We have tried to emphasize more throughout the manuscript that correct specification is the goal, and we are trying to provide guidance in cases where one is unsure about including or excluding moderation on a specific path.**

   **Please also see the response to Reviewer 1's (Mijke Rhemtulla) first comment as it is closely related, though somewhat distinct from this comment.**

3. The study did not vary mediated paths, and many paths are kept null. Mediated paths (X to M, M to Y) are all fixed to 7% (.26 when standardized), the direct path X to Y is null. They are known to have differential effects on rejection rates. Keeping them fixed may be desirable at this stage, but it greatly reduces the scope of the project.

   **We purposefully did not vary the mediated path coefficients for two reasons: 1) this would have made an already complex simulation design more complex, and 2) we do not aim to provide specific sample size guidelines based on the results of the simulation. By limiting the specific mediation effect sizes we hope to avoid the temptation to use this paper as a reference for generating specific sample size recommendations. We have removed reference to this goal, and have clarified more that the purpose of the paper is to help researchers understand the consequences of over- and under-specification, as well as understand the relative sample size needs depending on model specification. We do intend to revisit the issue of sample size planning by pointing to existing tools which researchers can use to plan sample size for their specific models. If fact, in our preparation of this manuscript we identified that the WebPower tool has recently expanded the set of available models for sample size planning in moderated mediation; however, they do not calculate power for the Index of Moderated Mediation. We have contacted the tool developers and they have expressed a willingness to update the tool to allow for this functionality. These changes would greatly expand the availability of power analysis for moderated mediation, and we would be excited to recommend**

**this approach once it is implemented. We expect these updates to be implemented by the time we submit a Stage 2 version of the paper. If not, our research team may adapt our simulation code to create a Shiny App which could be used for power analysis in moderated mediation contexts.**

**Reviewer 3: Reny Baykova**

1. First, several sub-sections of the methods section have been written in past tense which suggests that substantial parts of the analysis may already have been completed. The way the manuscript is written suggests that the data has already been simulated, the data analysis models have been fit to the simulated data, power/ type I error rates have been calculated. It appears that the only part of the analysis that has not been completed already is the fitting of the logistic regressions. If this is the case, can the authors please state which parts of the analysis have been completed and what countermeasures they have taken to ensure rigour and bias control? If I have misunderstood and the analysis steps I listed have not been completed already, the tense of the "Method" section needs to be changed. A statement specifying which parts of the analysis have been completed already and which parts are yet to be completed would also be beneficial.

   **Thank you for suggesting clarification as to which parts of this study have already been completed. While we wrote the method section in past tense to avoid having to change tenses at the time of the possible Stage 2 submission, we have not yet completed the work described in past tense yet. To clarify, we have included the suggested statement at the end of the manuscript after the Data Availability Statement, reading "Stage 1 Registered Report: As the time of submission as a Stage 1 registered report, pilot data have been generated and analyzed as part of the first author's dissertation study. However, data for this study have not yet been generated and no analyses have been completed. Simulation code has already been written to generate data, and the script for data analysis has also already been written. Both are available on the OSF page for the study: https://osf.io/vgkdt/."**

2. Moving on, pre-registered and exploratory hypotheses as well as pre-registered and exploratory analyses are not differentiated sufficiently clearly. In addition, most of the listed hypotheses and analyses are described as exploratory. For example, out of 6 listed hypotheses, 4 are described as exploratory. It is reasonable to include some mention of plans for exploratory analyses if these inform the design of the study. However, if a manuscript is to be considered as a registered report, I would expect that most of the predictions and analyses would be pre-registered rather than exploratory. In its current form, I am not sure how suitable this manuscript is for a stage 1 registered report.

**For this Stage 1 manuscript, we have removed all exploratory hypotheses and analyses. We have reorganized our hypotheses and analyses, and either specified what we expect to see for each previously exploratory hypothesis, or removed it completely. We plan to include a separate, clearly labeled exploratory analyses section in the Stage 2 submission.**

3. I couldn't find a data availability statement. Can the authors clarify what resources will be made available and add an availability statement?

   **Thank you for suggesting a Data Availability statement. This has been added at the beginning of the Method section (where it is required for the journal AMPPS), stating "All data will be made available on the OSF page for this study. The GAUSS simulation code to generate the data, a .csv file of the simulation results, and the R analysis script will all be posted at https://osf.io/vgkdt/."Our analysis script is already posted.**

4. Post hoc exploratory analysis plans make up most of the paper and they are not clearly discernible from pre-registered analysis plans. Four out of the 6 hypotheses presented in the paper are described as exploratory. The analysis section presents exploratory analyses mixed in with pre-registered analyses which makes it difficult to differentiate what is pre-registered and what is not. Pre-registered and exploratory analyses should be presented in different sections, using sub-headings for clear signposting.

   **We have removed or revised all exploratory analyses to be fully specified.**

5. Because the paper presents a lot of exploratory analyses and predictions, and these are not clearly separated, I would say that the manuscript does not sufficiently reduce researcher degrees of freedom.

   **We hope that the revision to remove or revise all previously exploratory analyses addresses this concern about researcher degrees of freedom.**

6. Most of the methods section is written in the past tense, suggesting a lot of work has already been done but the authors have not included a statement directly stating what work has been completed and what work has not. There is also a sentence in the introduction, page 4, which suggests that the whole study has already been completed: "we conducted a simulation study which examines how different model specification decisions impact type I error and power".

   **While it is written in past tense to avoid having to change tenses prior to publication, we have not yet completed this study. We appreciate the above suggestion to include a statement about what work has and has not been completed at the time of the Stage 1 submission, and that is included as the last section in the main manuscript.**

7. The manuscript lists 6 hypotheses, 4 of which are described as exploratory. Therefore, I consider the 2 remaining hypotheses as pre-registered. I would say that the hypotheses are overall stated clearly, but I am not sure of the difference between hypothesis 1a and 1b (more details below). In addition, the exploratory hypotheses are not described as such until quite late in the introduction, and until this point, readers may misunderstand that all goals of the manuscript would be pre-registered. I would suggest that the authors discern the pre-registered from exploratory predictions of the papers from the very start.

   Pre-registered hypotheses:
   Hypothesis 1a (page 14): “We hypothesize that [the] statistical power of the index of moderated
   mediation will be higher for correctly specified models compared to over-specified models.” – I think this hypothesis is clearly stated.

   Hypothesis 1b (page 14): “We also hypothesize that power will be higher for models with fewer moderated paths” – Here it is not immediately clear if this hypothesis refers only to models which are over-specified or refers to both correctly specified and over-specified models. I understand the hypothesis to refer to the number of paths across both over- and correctly specified models, but after reading further, it seems my interpretation was wrong. Therefore, it would be good to be clearer. It might also be good to include a bit of a discussion on how hypothesis 1b differs from hypothesis 1a. Overspecified models have more moderated paths than correctly specified models, so hypothesis 1a follows directly from hypothesis 1b. Therefore, it is not clear how or why the effect of overspecified vs correctly specified models would be different from the effect of the number of moderated paths.

   **We appreciate the opportunity to clarify and we have made the distinction between these hypotheses clearer in the manuscript. Specifically, H1a focuses on comparing power between correct and over-specified models. H1b, however, is a sub-group analysis comparing power across the number of paths in over-specified models only. We now more clearly differentiate the goal of H1b in the current study section, and we report the distinction in the multilevel logistic regression models we use to test these hypotheses in the analysis plan section.**

8. **See also response to Reviewer 1 Comment #7**

9. Exploratory hypotheses:
   Hypotheses 1.1a and 1.1b (page 15): “We will examine the effect of under-specification compared to correct specification (H1.1a), the number of moderated paths in the model (H1.1b).” – I think some words are missing in the second part of the sentence (potential suggested edits listed below under “Other Comments”). The “Stage 1 Snapshot” lists a hypothesis that “model under-specification will lead to elevated type 1 error rates for the

index of moderated mediation", however in the Stage 1 registered report, it says that for under-specified models the outcome variable of interest will be power, not type I error.

**We appreciate the thorough attention to detail regarding these hypotheses and the Stage 1 Snapshot. Between submitting the Stage 1 Snapshot and submitting the Stage 1 manuscript, we had thought further about our definition of model under-specification, and we had broken down the original under-specification into complete misspecification and under-specification. Complete misspecification still aligns with type 1 error rate because the index of moderated mediation from the data analysis model should be 0 according to the DGP, while under-specification is a special case where a moderated mediation still exists, but the model data analysis does not quite match the DGP, aligning with the outcome of statistical power instead of type 1 error rate. We have reorganized the hypotheses to better clarify our distinct goals and we describe these in the current study and analysis plan sections. Specifically, Hypotheses 1a-1c focus on over-specified models, where the interest is only on power and parameter bias. Hypotheses 2a-2b focus on underspecified models, where the interest is only on power and parameter bias. Hypotheses 3a-3b focus on completely misspecified models, where the focus is only on Type I error rate and parameter bias.**

10. Hypotheses 2a and 2b (page 15): "We treated these models as exploratory, though we hypothesize that type I error rate would be too high in completely misspecified models (H2a) and type I error rate will increase as the number of incorrectly moderated paths increases as well (H2b)." – Here it is not immediately clear what would be considered as "too high".

    **We will be using the liberal criteria from Bradley et al., (1978) to see if type 1 error rate is "too high" or "too low", and we have amended this sentence to include the criteria.**

11. Finally, one of the hypotheses listed in the "Stage 1 Snapshot" is not discussed in the Stage 1 registered report: "larger sample sizes will be needed for smaller effect sizes and over-specified models"

    **We appreciate the detailed reading of the Stage 1 snapshot. We have determined that we will be examining the role of sample size in an exploratory manner based on the results of the primary simulations. As exploratory analyses are not typically mentioned in Stage 1 manuscripts, we removed these plans from the Stage 1 document.**

12. The literature review provides evidence that estimating power for moderated mediation analysis is a problem which justifies why conducting this study is important. However, as far as I can tell, the introduction does not present previous evidence on how model

misspecification or the number of paths in a moderated mediation model affect power and type I error rate.

**There is not previous literature on this topic. Model misspecification has been largely unexplored in the context of moderated mediation. This is exactly the purpose of this study: to investigate the role of model misspecification in power and type I error for these models. We have incorporated previous literature from moderation that suggests omitting moderated paths leads to bias (Yzerbyt et al., 2018); however, even in this context the literature is quite sparse.**

13. I think the introduction focuses a lot on the goal of the paper to provide researchers with guidelines on study design and doesn't build enough support for the hypotheses.

    **We have revised and reorganized the hypotheses such that we feel that this comment has been addressed.**

14. I would say that it is reasonable to predict that model over-specification and increasing the number of moderated paths will result in lower statistical power, but I think it would be good for the manuscript to include a more in-depth justification to support the pre-registered hypotheses.

    **Thank you, we have added in additional support in the introduction to justify this pre-registered hypothesis. When describing the maximalist perspective, which we connect to over-specification.**

15. To me, the introduction provided a clear overview of moderated mediation and the challenges of designing moderated mediation studies. I have a few questions. First, on page 4, the authors discuss 2 potential approaches to model specification – maximalism and minimalism. Are there any previous papers which discuss this distinction?

    **Yes, we are drawing on these terms used in the structural equation modeling and multilevel literature. We have now included a few citations in the text when introducing the approaches, including Barr et al., 2013 arguing for the use of maximal models for confirmatory factor analysis in psycholinguistics, and We have additionally integrated the terms throughout the introduction.**

16. Second, on page 9 it says, "bootstrapping is the recommended method for conducting inference [on the index of moderated mediation] because it is commonly used in mediation analysis already…". This doesn't sound like a very strong argument. Wouldn't a better argument be that the index of moderated mediation is not normally distributed?

    **Thank you for this suggestion. We have updated our rationale accordingly, and removed reference to it being commonly used in mediation analysis already. The sentence now reads, "Bootstrapping is the recommended method for conducting inference because the index of moderated mediation is not normally distributed,**

**and bootstrapping has been shown to perform better than other methods in simulation studies while not inflating the type I error rate (..." at the end of the Index of Moderated Mediation section of the Introduction.**

17. Third, on page 19 it says, “Because there is no comparison group for type I error, and previous simulations in mediation analysis have found that type I error rates are often differ from 0.05 for correctly specified model, we use the criteria from Bradley (1978) and Serlin (2000) to classify type I error rates as overly conservative or liberal”. Can the authors provide some references to previous work to justify that previous simulations have found type I error rates often differ from 0.05? Also what exactly are the overly conservative and liberal criteria?

    **We have cited two previous simulation studies that found bootstrapping can be overly conservative for testing the index of moderated mediation and conditional indirect effects: , Yzerbyt et al., (2018) and Coutts (2023). We have also included Bradley's criteria for overly conservative (below .025) and liberal (above .075) in the manuscript on page 16.**

18. I would suggest including a table or list of all papers that were included in the systematic review. The database is great, it includes additional papers which makes it a bit hard to pinpoint the exact papers that were used in the review.

    **Thank you for exploring the database. We've uploaded a spreadsheet of just the articles included in the systematic review to OSF, and we've put the link in the Systematic Review section of the main manuscript (https://osf.io/pf2gh). This also includes coded information from the articles, such as model used and sample size.**

19. The manuscript states that selected levels of sample size used in the simulations are based on previous studies, but it is not immediately clear how the levels were selected. For example, looking at the database, the smallest sample size in a previous study was 29, while the smallest sample size in the simulations is 50. Could the authors elaborate on how the levels for sample size were selected? In addition, could the authors include a histogram of sample sizes across the studies?

    **We used deciles to select the sample sizes used in this simulation, with the innerdecile range (10th percentile and 90th percentile) as the minimum and maximum in order to eliminate outliers. We have clarified this in the manuscript by saying "Based on deciles (with rounding) from the systematic review, using the 10th and 90th percentiles (interdecile range) as the maximum and minimum, we used sample sizes of 100, 150, 200, 250, 300, 400, 500, 750, and 1000" as the second sentence of the Simulation Procedure section. We have also included box plots of sample sizes found, separated out by model number.**

20. The “Systematic Review” sub-section in the introduction justifies the selection of the six models that are examined in the paper. However, I have some questions about the model-selection process. On page 13 the manuscript states that “… we chose the six most commonly used moderated mediation models…, accounting for 86% of published models from the systematic review”. Was the goal to cover precisely 86% of the published models (and if so, why 86%?) or was the goal to select precisely 6 models (and if so, why 6?)? Also, looking at the database, model 15 was used in 12 studies, but model 9 was used in 16 and model 21 was used in 13 but these models were not included in the study. Why was model 15 chosen instead of models 9 or 21? Can you also include a table of the exact papers that were used for the review? The database is great, it includes additional papers which makes it a bit hard to pinpoint the exact papers that were used in the review. In addition, as the models are first presented in the sub-section “Introduction to Moderated Mediation” which comes before “Systematic Review”, this organization of the paper makes the justification for selecting these models appear seemingly post-hoc at first reading.

    **We have revised the Systematic Review section to clarify and address these concerns, and have opted to move it to an Appendix because our goal journal has a word limit and we are trying to stay within that guideline for a smooth transition of this registered report. Models were indeed selected after looking at the results of the systematic review, since one of the goals of the systematic review was to figure out which moderated mediation models were used most commonly. We mention this in the introduction to set up how these models were chosen: “We conducted a systematic review of 411 articles to understand which models are most commonly used in practice, and six models emerged (Models 7, 8, 14, 15, 58, and 59; see Appendix A for more details on the systematic review).”. In the appendix, we clarify that the goal was not to cover precisely 86% of published models, but rather to pick a subset of models that were used commonly so as to make the results as useful as possible to researchers. We also did not initially set out to choose precisely 6 models either, but after looking at the breakdown of models used in the systematic review, these six models were chosen based on their use frequency and the fact that they only include one moderator variable. Five of the models were much more commonly used (accounting for 6-31% of published models from the systematic review), then Model 15 was also included, accounting for 2.9% because it complimented the other models used. Importantly, it also only had one moderator, instead of two different moderators like Model 9. From the systematic review to the current database, Model 21 has grown in popularity, since originally it only accounted for less than 2.5% of published models. Extending this research to include multiple moderator models, especially with the growth seen in the use of these models, will be mentioned as an important future direction.**

21. Finally, can the authors comment on why they are making a distinction between dichotomous and continuous moderators and predictors, and what is the reason for

using an incomplete design? What is the reason for not using models 58 and 59 when the moderator was continuous?

**The note in Figure 1 now addresses this concern, along with an in text description in the Index of Moderated Mediation section. The index of moderated mediation, which is how we are testing for significant moderated mediation using these models, is not defined for Models 58 and 59 when the moderator was continuous, making a complete design impossible.**

22. There are some aspects of the logistic regression analysis that I am confused about and might benefit from a bit more explanation. First, it is not clear how many logistic regressions will be fitted to the data.

    **Thank you for this need for clarification. The hypotheses section now more clearly outlines each logistic regression analysis that will be fitted to the data.**

23. On page 19, “Analysis plan” starts with “To test our hypotheses about model specification on power and type I error rate, we will use multilevel logistic regression with random intercepts only to predict rejection”*. This suggests that the whole analysis will consist of only 1 logistic regression, and reading further, this logistic regression will include a main effect of each factor and all possible interactions between all factors (“all possible two-way through six-way interactions”)**. However, multiple different logistic regressions are presented afterwards and they also don’t follow this structure - none of them include a random intercept and only one of them includes a two-way interaction between model specification and the number of moderated paths (there are no other interactions detailed anywhere). So, I do not understand what the logistic regression described on page 19 will be used for, or what it refers to.

    **We appreciate the opportunity to revise our analysis plan section to better clarify how and when the multilevel logistic regression models will be used. Rather than describing these upfront, we now describe the appropriate model for each hypothesis separately. We also clarify which hypotheses are tested using these multilevel logistic models (H1a, H1b, and H2a) and which hypotheses are tested using descriptive patterns instead (H1c, H2b, H3a, and H3b). Moreover, when a multilevel logistic regression is run, two will be fit (one for continuous moderators and one for dichotomous moderators).**

24. Next, further discussion is needed to justify the need for fitting multiple logistic regressions, and the exact combination of predictors included in each logistic regression. Can the authors explain why they have decided to use two separate logistic regressions to test hypotheses 1a and 1b? Also, why does the logistic regression to test hypothesis 1b includes an interaction between model specification and the number of moderated paths, but the logistic regression to test hypothesis 1a doesn’t? Can the authors elaborate on their decision to fit separated logistic regressions are fitted when the

moderator is dichotomous and continuous? These are very specific questions, but the comment applies to the whole analysis plan.

**We have now clarified all of these points in the manuscript and we appreciate that you have pointed out these areas of confusion. Upon further reflection of our hypotheses and based on other reviewer comments, we now no longer include the test of the interaction in the models. Specifically, we test H1b as a subgroup analysis because the hypothesis is only relevant for over-specified models. In the current version of the manuscript, we more carefully describe the model for H1b as distinct from H1a and the rationale for this. Specifically, H1b does not include the predictor of model-type because the analysis is limited to over-specified models, unlike H1a, where the focal interest is in the model-type effect and in comparing across over-specified and correctly-specified models. Finally, we have also clarified the rationale for fitting separate logistic regressions for dichotomous and continuous W. This was necessary because our simulation design is not fully crossed, and some moderated-mediation models allow for a dichotomous moderator, and others do not (see explanation on page 7 in the Introduction to Moderated Mediation section of the introduction).**

25. I am also confused about what the outcome variables in the study are. On page 18, the manuscript says that there are two outcome variables of interest – power and type I error rate. Both are calculated as the proportion of samples that have an index of moderated mediation with a confidence interval that does not include 0 (described on page 9). However, the outcome variables in the logistic regressions described from page 19 onwards are not power/type I error, but rejection rate – which is defined as whether the confidence interval of the index of moderated mediation includes 0 or not. So, the outcome variable of interest is not power/type I error rate, but rejection rate. I am also not sure of the difference between "rejection", introduced but not defined on page 18, and "rejection rate", defined on page 19.

    **Thank you for suggesting this clarification. Rejection rate refers to the proportion of times within a condition where the confidence interval for the index of moderated mediation does not included zero. When the true index of moderated mediation is zero in the population, the rejection rate refers specifically to the type I error rate. When the true index of moderated mediation is non-zero in the population, the rejection rate refers specifically to power. Thus, the general quantity of rejection rate in statistics takes on more specific definitions based on the population parameter. We have made this point more visible in the performance metrics and analysis plan sections. We have also clarified how these definitions of power vs type I error rate map onto our hypotheses given that the population parameter is nonzero for over and under-specified models, but zero for completely misspecified models.**

26. *It is also not clear what the random intercepts would model – would they be random by-sample intercepts?

    **The random intercepts are for the data analysis model, since all six models are used to analyze each generated sample of data (data analysis model is a within subjects factor). We have revised the description of the random intercepts to clarify this, "...with random intercepts for the data analysis model (within-subjects factor since each generated sample of data is analyzed using all six data analysis models) to predict rejection." in the analysis plan on page 16.**

27. ** Another question related to the first paragraph under “Analysis plan” (page 19). The authors say that they will use “an approach aligned with type II sums of squares”. Does this mean that type II sums of squares will be used, or some other approach?

    **We have removed this sentence from the manuscript and instead elected to provide our code for maximum clarity in terms of our approach.**

28. Here I will focus only on the two pre-registered hypotheses. I think the authors explain how they will interpret the results of the logistic regressions in relation to hypothesis 1a relatively clearly. The hypothesis is associated with 2 pre-registered logistic regressions (one when the moderator is continuous and one when the moderator is dichotomous) and whether it is supported or not depends on whether the main effect of model specification is significant. However, the discussion of how the results will be interpreted in relation to the hypothesis (page 20) is preceded by a description of 3 or 4 models (depending on the reader’s interpretation). It would be better if exploratory analyses which have no bearing on the pre-registered hypotheses are described separately to avoid confusion.

    **We have removed all exploratory analyses from the Stage 1 manuscript.**

29. The conditions for rejecting or failing to reject the null hypothesis of hypothesis 1b are less clear (page 21). The logistic regression fit for this analysis contains 6 predictors, and hypothesis 1b would be supported “by a significant effect of number of moderated paths”. Then the manuscript continues with “we hypothesize that more moderated paths would lead to lower power in overspecified model[s], and so we hypothesize that all coefficients will be negative”. Here is it not clear whether “all coefficients” refers to all the coefficients in the logistic regression and whether this will be used to shape conclusions regarding hypothesis 1b. Then, the paragraph continues by saying that two logistic regressions will be fitted to the data – one when the moderator is continuous and one when the moderator is dichotomous. It is not clear to me how all this fits together. As a general comment which also applies to the presentation of the analysis for hypothesis 1a, starting the paragraph saying that 1 logistic regression will be fitted to the data, but further down changing that to two (and then including additional exploratory models) is quite confusing.

**We have clarified the language of "all coefficients" so that we now state exactly what the coefficients need to convey for full or partial support of the hypothesis. Specifically, in the Analysis Plan section of the manuscript, the coefficients refer to the two sequential comparisons between 1 and 2 paths, and 2 and 3 paths, for over-specified models only. If both of these coefficients are significant in the predicted (negative) direction, H1b is fully supported. If only one is significant in the predicted direction, H1b is partially supported. We have also better clarified the multilevel logistic regression model that we use for H1b and how this is distinct from H1a in the analysis plan. Regarding models for dichotomous and continuous W, we now clarify that for each logistic regression model, the model is fit separately for each type of W. e have also removed all mention of exploratory analyses and instead incorporated those into the full analysis plan.**

30. In addition, in several places, the authors say that if the results go in the opposite direction of what was hypothesized, “we will interpret these results appropriately” (e.g. page 20). Can the authors elaborate on what this means?

    **Thank you for bring this (along with the exploratory analyses) to our attention. We have revised or removed all of our exploratory analyses, and included specifics of how we would interpret results for each of our hypotheses. "We will interpret these results appropriately" no longer appears in the manuscript.**

31. Finally, on the top of page 20, it states that “only effects with an odds ratio greater than 1.68 will be considered meaningful”. Does this mean that conclusions regarding the hypotheses will be reached based on both odd ratios and p-values?

    **We have adjusted our criteria to no longer look for “meaningful effects” and focus on significance tests  at a .001 level.**

32. Non-significant p-values will be used to conclude that an effect is absent. For example, on page 20: “Hypothesis 1a will be supported if we find a significant coefficient for model specification (over vs correct) such that power is lower when models are over-specified. We would expect to see this in both models with continuous W and dichotomous W. However, if this is non-significant in both models, we would conclude that over-specification does not negatively impact power”. The accurate interpretation of a non-significant p-value is that you can’t reject the null hypothesis. If the authors want to draw conclusions about the absence of an effect, equivalence testing or a Bayesian approach would be required.

    **We appreciate you raising this important point. In conjunction with responses from other reviewers and the editor, we have chosen to focus simply on *p* values for determining statistical significance in Hypotheses 1a, 1b, and 2a and we will not interpret a non-significant *p* value as the absence of an effect. We also chose to focus on descriptive statistics for our other hypotheses.**

33. As stated above, I would suggest including an analysis that would allow drawing conclusions in favour of the null hypothesis. In addition, I am not sure how the suggested analyses related to the paper's ultimate goal of providing researchers with recommendations for study design. For example, the "Stage 1 Snapshot" states that one of the core questions the manuscript seeks to answer is "what sample sizes are sufficient to detect moderated mediation?", but this is not discussed in the registered report. Can the authors elaborate on that?

    **Thank you for the suggestion to include an analysis that allows for drawing conclusions in favor of the null hypothesis. We have clarified our inferential methods, and will be careful to not draw conclusions in favor of the null hypothesis. We are also cautious to not have our recommendations be specific sample sizes to use for different models, but rather a discussion of factors that are important to consider (anticipated interaction effect size, model specification, sample size) in designing studies using moderated mediation analyses, and instead recommend methods for sample size determination for individual studies.**

34. Figure 1 is great for visualising the different models, but the small indices on the right column are not legible. I had to zoom to 350% to be able to read them.

    **We have revised Figure 1 in accordance with comment 10. . We hope that the increased font size is sufficient to be read without requiring zooming in to that magnitude.**

35. I have done a deep dive on typos only up to page 13.

    **We thank you for taking the time to do this deep dive on typos. We have made all track change suggestions as suggested, and for the comments requiring further clarification, it is made in this response and in the main manuscript.**

Page 2: "Mediation analysis provides a way of examining effects whether a proposed mediator variable (e.g., peer victimization) serves as a mechanism by which one variable effectsaffects another (e.g., discrimination affects internalizing)."

a. Page 2: "In this paper, we focus on model specification (where moderation is allowed to occur in the mediation model) and it'sits implications for sample size planning and power."
b. Page 3: "Because statistical power depends on sample size, the goal in of sample size planning is to find the optimal balance between maximizing power and minimizing wasted resources (Maxwell & Kelley, 2011)".
c. Page 3: "Low power has been cited as a common source of problems in the scientific literature (Ioannidis, 2005), particularly with respect to the replicability crisis (Anderson & Maxwell, 2017; Earp & Trafimow, 2015)."
d. Page 3: "Götz, O'Boyle, Gonzalez-Mulé, Banks, and Bollmann (2021) conducted a large scalelarge-scale review of mediation analyses in psychology journals…".

e. Page 3: “To our knowledge, no prior studies have examined whether current moderated
f. mediation analyses are well well-powered…”.
g. Page 4: “Prior research in moderation analysis suggests that detecting more interactions and higher higher-order interactions requires larger sample sizes (McClelland & Judd, 1993). However, this issue has not been explored in the context of moderated mediation models.”
h. Page 4: “Next, we conducted a simulation study which examines how different model specification decisions impact type I error and power.”
i. Page 6: “While some of the notation in the following equations is the same, the values will not necessarily be the equal.”
j. Page 6: “…where a0 is the intercept, and a is the effect of X on M, and εMi is the residual.”
k. Page 9: “The equations for the conditional indirect effects and the index of moderated mediation are unique to the model being estimatesestimated...”
l. Page 11: “Currently, for models other than 7 and 14, there are no tools available to assist with this process, meaning that researchers may need to create their own Monte Carlo simulation…”.
m. Page 11: “To understand how a data analysis would perform if the model is misspecified, it is helpful to distinguish the data data-generating process (DGP) from the model used for the data analysis.”
n. Page 12: “It fits the criteria for a under-specification because at least one path involved in the indirect effect…”.
o. Page 13: “The index of moderated mediation from Model 14 is ab3, which should be 0
based don the DGP.”
p. Page 14: “These effects were examined across a variety of realistic conditions: sample sizes, effect size of the interaction, and both dichotomous and continuous moderators and X focal predictor variables.” – Can you specify which interaction you are referring to here and what you mean by “X focal predictor variables”?

**The term "focal predictor" in reference to the X variable has been removed, and interaction specified to be the interaction term(s) included in the model. We have also included Table 1 and reference it in place of that sentence to clarify the simulation conditions..**

q. Page 14: “Power was only assessed for over-specified models because we hypothesized including in additional interactions could reduce statistical power.”
r. Page 14: “We hypothesize that the statistical power of the index of moderated mediation will be higher for correctly specified models compared to over-specified models
(H1a).”
s. Page 14: “If the DGP is Model 7, for example, Models 8, 58, and 59 are over-specified.
Models 8 and 58 have two moderated paths but Model 59 has three. we We hypothesize that…”
t. Page 15 (I am not sure what was meant by this sentence, so my suggested edits might be wrong): “Related to RQ1, we examined these what factors affecting

affect statistical power for the index of moderated mediation for under-specified models."

**Thank you for pointing out the need for clarification. The sentence now reads "Research Question 1 examines the consequences of the maximalist approach: specifically, how over-specification impacts statistical power of the index of moderated mediation and parameter bias" on page 11.**

u. Page 15 (I am not sure what was meant by this sentence, so my suggested edits might be wrong): "We will examine the effect of under-specification compared to correct specification (H1.1a), and the effect of the number of moderated paths in the model (H1.1b)."

**We have removed Hypothesis 1.1 parts a and b, and revised this section to not include this sentence.**

v. Page 16: "We only used Models 58 and 59 for generation and analysis when the moderator was dichotomous." – This sentence is a bit too easy to misunderstand. At first, I thought that this means that when the moderator was dichotomous, you used only models 58 and 59. Reading further, it became clear that when the moderator was dichotomous, you used all 6 models, when it was continuous – you didn't use 58 and 59.

**Thank you for letting us know how the sentence can be misunderstood. It is revised to read "Models 58 and 59 were not used for generation and analysis when the moderator was continuous." on page 12.**

w. Page 17: "Effect size on the interaction term and sample size were varied." – just for clarity can you specify what interaction you are referring to (for example, Equation (6) contains two interaction terms)? It would also be good to have all levels of all predictor variables described in the same section. Currently, these are spread across "Simulation Conditions" and "Simulation Procedure".

**We have moved all of the related information from the two sections into the "Simulation Procedure" section. In this section, we specify that the effect size on the interaction term was varied as a condition, but if there are multiple interaction terms included in the same model, they remain consistent (page 14).**

x. Page 18: "Data analysis models were then fit to each sample of generated data. Models were analyzed using the percentile bootstrap confidence interval set at 95% with 1000 bootstraps (Efron & Tibshirani, 1994)." – just to clarify, was this procedure repeated for each of the 5000 samples of simulated data?

**Correct, this procedure was repeated for each of the 5000 samples of simulated data. We have included that clarification in the manuscript on page 15, reading "Each of the 5000 samples were analyzed with all four (continuous *W*) or six (dichotomous *W*) analysis models. Inference for the**

**index of moderated mediation was conducted using the percentile bootstrap confidence interval set at 95\% with 1000 bootstraps (Efron & Tibshirani, 1994)."**

y. Page 18: ”We calculated rejection rate for the index of moderated mediation for…” – rejection rate has not been defined yet. Is this the same as “rejection” which is defined towards the bottom of page 19?

**Correct, the rejection rate is the same in all cases. The definition from page 19 has been moved up to the first time rejection rate is mentioned.**

z. Page 19: “Because there is no comparison group for type I error, and previous simulations in mediation analysis have found that type I error rates are often differ from 0.05 for correctly specified models, we use the criteria from Bradley (1978) and Serlin (2000) to classify type I error rates as overly conservative or liberal.”

aa. Page 19: “To test our hypotheses about model specification on power and type I error rate, we will use multilevel logistic regression with random intercepts only to predict rejection.” – It is not clear if “only” refers to the model having only random intercepts but not random slopes, or to rejection being the only outcome variable.

**Thank you for pointing this out. The "only" refers to the model only having random intercepts but not random slopes. We have removed the word "only" from this sentence.**

bb. Page 20: “To test Hypothesis 1…” – previously, the authors have defined Hypothesis 1a, Hypothesis 1b, Hypothesis 1.1a, and Hypothesis 1.1b, but I can’t find a definition of Hypothesis 1, so I am not sure what hypothesis is being discussed here.

**We have reorganized our hypotheses, and removed any mention of a more general hypothesis (like Hypothesis 1) without specifying if it refers to part a, part b, or both. We have also changed hypotheses having to do with under-specified models to be Hypothesis 3a and 3b for clarity.**

cc. Page 21: “The hypothesis will be supported would be supported by a significant effect of the number of moderated paths in the analysis model (tested using a Wald test for the set of coefficients).”

dd. Page 21: “To test Hypothesis 2…” – previously, the authors have defined Hypothesis 2a and Hypothesis 2b, but I can’t find a definition of Hypothesis 2.

**We have reorganized our hypotheses, and removed any mention of a more general hypothesis (like Hypothesis 2) without specifying if it refers to part a, part b, or both.**

ee. Page 22: “In addition, we will explore what factors predict type I error rate using a multilevel logistic regression with 5 main effects and all possible interactions…”.

ff. Page 28 onwards: The pages over which the appendix is spread out are numerated as 1-6. Also, the text in the table is a bit difficult to read because the text in one column spreads across multiple pages. Would it be possible to summarize each point a bit more succinctly so that the contents of each cell are confined to one page only?

**The page numbering for the appendix is updated to be a continuation of the numbering system from the manuscript. Additionally, it is revised and summarized more succinctly so each column is limited to one page.**