Peer review of Stage 1 Registered report titled “How Does Model Specification Impact Statistical Power and Type I Error Rate in Moderated Mediation Analysis? A Registered Report”

# Review template

For this review, I am using a modified version of a review template that was originally developed by Dr Warrick Roseboom with the help of other colleagues in the School of Informatics and the School of Psychology at the University of Sussex (Dr Maxine Sherman, Dr Peter Lush, Prof Zoltan Dienes).

# Reviewer details

**Name:** Dr Reny Baykova

**Affiliation:** School of Informatics and School of Psychology, University of Sussex

**Relationship to the authors:** None

**Expertise in the topic/field:** I have experience with Stage 1 registered reports. I have prepared a Stage 1 registered report in the past and I have also reviewed a Stage 1 registered report before. I have no experience running moderated mediation analyses myself, but I have theoretical knowledge of the subject. I have practical experience with simulating data, and implementing different types of regression models, including mixed effects models.

# Summary of paper

|  |
| --- |
| Fossum, Montoya, and Anderson a proposing to conduct a pre-registered simulation study investigating how model misspecification (overspecification, underspecification, and complete misspecification) affects statistical power and type I error rates in moderated mediation analysis. The overall aim of the paper is to provide researchers with guidelines for sample size planning for studies that involved moderated mediation. |

# Overall comments

|  |
| --- |
| I think the authors have set out to achieve an admirable goal – to help other researchers better plan moderated mediation studies. However, there are some aspects of the manuscript that I am hoping the authors can clarify.  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.  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.  I do believe this paper could be a very useful resource for other researchers, and I am looking forward to the authors’ responses before I make a recommendation. |

# Transparency

|  |  |
| --- | --- |
| 1. Will the data and code to reproduce the results be available on a public repository? (Make sure this is noted in an “availability statement” or similar) | yes/no/NA/**comment** |
| I couldn’t find a data availability statement. Can the authors clarify what resources will be made available and add an availability statement? | |
| 2. Have you checked the code/data? | yes/no/**NA**/comment |
| 3. Does the code run without requiring unreasonable amendments? | yes/no/**NA**/comment |
| 4. Will the proposed analysis use inferential statistics?  (E.g. any statistical test/model/method against a criterion.) | **yes**/no/NA/comment |
| 5. Is there a justification provided for the sample size in relation to this test(s)? If yes, is the provided justification reasonable? | yes/no/**NA**/comment |
| Not applicable; sample size is a predictor varied as part of the simulations. | |
| 6. Have the authors minimized all discussion of post hoc exploratory analyses, apart from those that must be explained to justify specific design features? Maintaining this clear distinction at Stage 1 can prevent exploratory analyses at Stage 2 from being inadvertently presented as pre-planned. | yes/**no**/NA/comment |
| 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. | |
| 7. Does the stage-1 registered report robustly reduce researcher degrees of freedom? (Ideally, for pre-registered tests there should be no discernible researcher degrees of freedom.) | yes/**no**/NA/comment |
| 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. | |
| 8. Have the authors clearly distinguished work that has already been done (e.g. preliminary studies and data analyses) from work yet to be done? | yes/**no**/NA/comment |
| 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”. | |

# Content: Introduction

|  |  |
| --- | --- |
| 1. Are the hypotheses of the study clearly expressed or otherwise able to be found in the text (copy or paraphrase below if possible)? | yes/no/NA/**comment** |
| 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.   1. 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.   1. 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.  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”.  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” | |
| 2. Are the presented hypotheses well-motivated by the presented literature? | yes/**no**/NA/comment |
| 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.  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. | |
| 3. Are the presented hypotheses based on theories or assumptions that are reasonable or supported by the literature?[[1]](#footnote-1) | yes/no/NA/**comment** |
| 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. | |
| 4. Is the presented literature review a fair reflection of the existing literature?[[2]](#footnote-2) | yes/no/NA/**comment** |
| I am not too familiar with the literature around moderated mediation so I cannot provide a detailed review on this aspect of the paper.  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?  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?  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? | |
| 5. Is previous work cited appropriately?[[3]](#footnote-3) | yes/no/NA/**comment** |
| I am not too familiar with the literature around moderated mediation so I cannot provide a detailed review on this aspect of the paper.  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. | |

# Content: Design, analysis, and results

|  |  |
| --- | --- |
| 1. Is the proposed paradigm appropriate for testing the hypotheses?[[4]](#footnote-4) | **yes**/no/NA/comment |
| I think the paradigm is overall appropriate for testing the hypotheses, but I have some questions about the choice of levels of some of the predictors.  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?  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.  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? | |
| 2. Are there any clear methodological confounds/errors in the proposed analyses? (provide a full explanation of why, citing references where necessary). | yes/no/NA/**comment** |
| 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.  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.  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.  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.  \*It is also not clear what the random intercepts would model – would they be random by-sample intercepts?  \*\* 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? | |
| 3. For proposals that test hypotheses, have the authors explained precisely which outcomes will confirm or disconfirm their predictions? | yes/no/NA/**comment** |
| 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.    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.  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?  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? | |
| 4. Have the authors avoided the common pitfall of relying on conventional null hypothesis significance testing to conclude evidence of absence from null results? Where the authors intend to interpret a negative result as evidence that an effect is absent, have authors proposed an inferential method that is capable of drawing such a conclusion, such as [Bayesian hypothesis testing](https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00781/full) or [frequentist equivalence testing](https://journals.sagepub.com/doi/full/10.1177/2515245918770963)? | yes/**no**/NA/comment |
| 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. | |
| 5. Are there any analyses that you can think of that have not been considered and would potentially reflect alternative interpretations? | **yes**/no/NA/comment |
| 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? | |
| 6. Do figures make sense and have clear labelling and explanation? (give suggestions of how the figures could be made more intuitive for you). | **yes**/no/NA/comment |
| 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. | |
| 7. Have the analysis methods been applied correctly (as specified) as far as you can tell?[[5]](#footnote-5) | yes/no/**NA**/comment |
| 8. Have you checked the presented statistics for errors? (at least using “statcheck”) | yes/no/**NA**/comment |
| 9. Are there any clear methodological confounds/errors in results? (provide a full explanation of why, citing references where necessary). | yes/no/**NA**/comment |

# Content: Discussion

This section of the review template is not applicable as this is a stage 1 registered report, but I have left the questions here for reference.

|  |  |
| --- | --- |
| 1. Does the Discussion provide a fair natural language account of what happened in the experiments in relation to stated hypotheses? | yes/no/**NA**/comment |
| 2. Does the Discussion provide a fair account of the links between the obtained results and motivating results outlined in the Introduction? | yes/no/**NA**/comment |
| 3. Are there any other relevant relationships between the results observed in this study and other literature that have been overlooked by the authors?[[6]](#footnote-6) | yes/no/**NA**/comment |

# Wildcard

|  |  |
| --- | --- |
| 1. Does the paper develop/deploy a method in a new and/or interesting way? | yes/**no**/NA/comment |
| Using data simulation to inform study design is not a new idea, but it is an immensely powerful approach and I hope that as more papers show people how to do it, it will become more widespread, and help researchers design better studies. | |
| 2. Does the paper provide a link between known results in an interesting way? | yes/no/**NA**/comment |
| Not applicable as this is a stage 1 registered report. | |
| 3. Is there generally something about this paper that makes you go “Wow, I wish I had thought of that”? | **yes**/no/NA/comment |
| I would be very keen to be involved in a project like this which aims to help other researchers solve a particular problem and do better research. | |

# Other comments

Please identify basic mistakes that you notice like typos and grammatical errors (suggested edits added using tracked changes), or sections of text that could otherwise be expressed more clearly. Provide suggested alternative phrasing if you wish.

|  |
| --- |
| I have done a deep dive on typos only up to page 13.   1. Page 2: “Mediation analysis provides a way of examining whether a proposed mediator variable (e.g., peer victimization) serves as a mechanism by which one variable affects another (e.g., discrimination affects internalizing).” 2. Page 2: “In this paper, we focus on model specification (where moderation is allowed to occur in the mediation model) and its implications for sample size planning and power.” 3. Page 3: “Because statistical power depends on sample size, the goal of sample size planning is to find the optimal balance between maximizing power and minimizing wasted resources (Maxwell & Kelley, 2011)”. 4. 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).” 5. Page 3: “Götz, O’Boyle, Gonzalez-Mulé, Banks, and Bollmann (2021) conducted a large-scale review of mediation analyses in psychology journals…”. 6. Page 3: “To our knowledge, no prior studies have examined whether current moderated   mediation analyses are well-powered…”.   1. Page 4: “Prior research in moderation analysis suggests that detecting more   interactions and higher-order interactions requires larger sample sizes (McClelland & Judd, 1993). However, this issue has not been explored in the context of moderated mediation models.”   1. Page 4: “Next, we conducted a simulation study which examines how different model specification decisions impact type I error and power.” 2. Page 6: “While some of the notation in the following equations is the same, the values will not necessarily be equal.” 3. Page 6: “…where a0 is the intercept, a is the effect of X on M, and εMi is the residual.” 4. Page 9: “The equations for the conditional indirect effects and the index of moderated   mediation are unique to the model being estimated...”   1. 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…”. 2. Page 11: “To understand how a data analysis would perform if the model is misspecified, it is helpful to distinguish the data-generating process (DGP) from the model used for the data analysis.” 3. Page 12: “It fits the criteria for under-specification because at least one path involved in the indirect effect…”. 4. Page 13: “The index of moderated mediation from Model 14 is ab3, which should be 0 based on the DGP.” 5. 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”? 6. Page 14: “Power was only assessed for over-specified models because we hypothesized including additional interactions could reduce statistical power.” 7. 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).” 8. 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 hypothesize that…” 9. Page 15 (I am not sure what was meant by this sentence, so my suggested edits might be wrong): “Related to RQ1, we examined what factors affect statistical power for the index of moderated mediation for under-specified models.” 10. 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).” 11. 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. 12. 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”. 13. 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? 14. 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? 15. 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 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.” 16. 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. 17. 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. 18. Page 21: “The hypothesis 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).” 19. 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. 20. 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…”. 21. 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? |

1. For example, hypotheses that are based on homuncular accounts of cognition (e.g. “we track time using an internal stopwatch, therefore if X then Y”) are unreasonable. It is also common to misinterpret measures as theories, e.g. intentional binding is an implicit measure of sense of agency, therefore when I measure temporal binding I am measuring implicit agency. Are any caveats/limitations like this acknowledged, either in the introduction or the discussion? [↑](#footnote-ref-1)
2. Note, this doesn’t mean comprehensive. Not all papers need to be review papers. Make sure to provide full references to papers you believe to be relevant but missing. [↑](#footnote-ref-2)
3. Cited work not relevant, too general, misinterpreted, retracted, unreliable or debunked, citing secondary sources, overuse of review papers. If no, provide explanation. [↑](#footnote-ref-3)
4. e.g. when making claims about how participants’ perception of a stimulus differs between conditions, does the task measure performance (sensitivity) or just propensity to report (bias); if you don’t know what this means refer to Kingdom and Prins, Psychophysics: A Practical Introduction, Chapter 2. Another common problem is the experimental manipulation not being precise enough, e.g. assuming that TMS can selectively target particular cells; TMS to the neck would not be an appropriate way of testing the causal role of baroreceptors in interoception, because TMS would stimulate the whole neck, not just baroreceptors. [↑](#footnote-ref-4)
5. e.g. no optional stopping in frequentist test; no impossible degrees of freedom/t values/p values. See Appendix A in doi:10.1177/2515245918806489 for comprehensive list of things to look out for. [↑](#footnote-ref-5)
6. Note, this is not meant for you to request references to your own work unless actually related. [↑](#footnote-ref-6)