I enjoyed reading this Stage 1 report! It’s clearly written and laid out, and I noted no major flaws in the logic or plan. I have a concern about the under-specified condition that I describe below, a number of clarifying questions about the design of the simulation and the analysis plan, and then a bunch of tiny typo-level comments.

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.

**Simulation Method**:

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.

“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.

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 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.

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.

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.

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.

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.

**Figures/Tables:**

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.

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.

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.

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.

**Typos and other super minor things:**

p. 2 “whether a proposed mediator” 🡪 “whereby a proposed mediator”

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 publicaations overall?

p. 2 “it’s” 🡪 its

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

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.

p. 4 “methods” 🡪 “method’s”

p. 4 “types I”

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

p. 9 “model being estimates” 🡪 estimated

Table 1 note: “assume dichotomous moderated is with” 🡪 moderator has?

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

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

p. 11 “researcher[s] may need”

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

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

p. 12 “can introduce excessive collinearity, especially with the interactions” 🡪 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?

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.

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?)

p. 13 “base don”

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.” 🡪 this can also be a consequence of some kinds of underspecification, right?

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

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

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

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

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.

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

p. 21 “will be supported would be supported”