- Kenneth Hung
- Response to reviews
- Reference: Statistical Methods for Replicability Assessment

We thank both referees for their thoughtful questions and helpful suggestions. Please find a point-by-point response below.

1 Response to Reviewer 1

- *The current paper provides some previously lacking statistical rigor in the analysis of an impactful replication study, and establishes that publication bias due to selection of significant effects is consistent with the observed findings. This is important because it provides one possible explanation or mechanism for this widely discussed outcome. I believe this aspect of the paper is important enough that it would make sense to include it as a focal point in the abstract of the paper.*

  We agree and have added this to the abstract.

- *I found the statement of Assumption 1 a bit confusing to parse. It's a run-on sentence. The word "literally" can be omitted in the preceding paragraph as well.*

  Noted and rewritten.

- *The plots on the right sides of Figures 2–4 are excellent for providing intuition. Is there a way to make this intuition clear as well for the estimators of FDP? I suspect it should be obvious that the estimates of FDP will be lower than* 64% *when applied to the RP:P data, perhaps there is a figure that would show this.*

  We believe that Figure 6 is the best at delivering the intuition for where this conclusion came from, and we have added more detailed descriptions there now.

2 Response to Reviewer 2

- *Should these both be* 35%*? Do you mean that all of the* 35% *that declined did so by at least* 20%*, or that* 35% *of the* 35% *that declined did so by at least* 20%*?*

  For this data set the estimates for how many declined, and how many declined by 20%, happen to coincide, but we understand why this created confusion.

  We mean that all of the 35% that declined did so by at least 20% (or at least, that our estimates for how many effects declined, and how many

effects declined by 20%, happen to coincide). We understand why this created confusion, however, and we have replaced 20% with 25%.

- *You mention "in contrast to the earlier internal comparison method" (pg 12) but it's not very clear what it refers to and the phrase doesn't appear again in the paper. I think you mean the the estimator at the top of the page?*

  Yes, that is what we meant. We have rewritten to clarify what we are referring to.

- *In general, I found it quite easy to get lost in the notation. You might find a notation table useful for clarity.*

  Thank you for this suggestion, we have added a table.

- *You comment on the sizing of replication studies to achieve fixed power under the original effect size, and how this negatively impacts replication chances due to inflated original effect sizes. It seems like there could be a role for post-selection inference here to more appropriately size those replication studies. If so, it might be an interesting point for the discussion.*

  Yes, we agree this is interesting and we have added to "Future work" section.

- *As an aside, I would actually question Assumption 1. I think strongly significant results are often significantly easier to publish than studies with p-values near* 0.05*. This would be in the opposite direction of the pile-up phenomenon that you mention, though certainly both could happen together in different regions of p-value. I suppose that if recent suggestions for a much lower significance threshold were accepted, Assumption 1 would likely hold more strongly.*

  We agree that it would be better to have more detailed information about this, especially for meta-analyses of studies in the recent past or future. With all the attention paid to replicability recently, it is likely that scientists' application of the 5% significance threshold has become less automatic since 2008, when the RP:P studies were published. A more sophisticated assumption can replace Assumption 1 if we have a model for the propensity for publication, e.g. Andrews and Kasy (2018). We hope future work can refine this. With the additional data from preprints and registered reports, it might be possible to try assessing how behavior is changing.

References

Isaiah Andrews and Maximilian Kasy. Identification of and correction for publication bias. *GitHub*, pages 1–85, May 2018.