

Reply to Review

Obels, Lakens, Coles, Gottfried, Green

23-11-2019

20-Nov-2019

Dear Dr. Lakens:

Thank you for submitting your General Article (AMPPS-19-0139) entitled “Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology” to *Advances in Methods and Practices in Psychological Science* (AMPPS). The manuscript has now been reviewed, and the reviewer comments appear at the end of this letter.

Both reviewers provided helpful suggestions for clarifying the methodology by providing additional detail. In the manuscript, it says “we considered an article reproducible when we could get the same main results as represented in the paper with at best minor changes to the analysis scripts” (p. 3). The reviewers noted that it could be helpful to specify how the main results were identified, and also what were considered to be minor changes. Reviewer 1 also asked for an operational definition of “reproducible” (e.g., was a result considered reproducible if there were only minor deviations between the reported and reproduced result?). The reviewers also requested further details about the reasons for low inter-rater reliability, the way you calculated average time to reproduce analyses, and the issues that were encountered when attempting to reproduce R scripts. I think these added clarifications could give readers an even fuller picture of how to maximize the reproducibility of their own work.

On a related note, Reviewer 2 mentioned that they would be interested to see your suggestions regarding developing standards and educating researchers about best practices. I agree with this reviewer that this is potentially an important contribution of the current paper, but I also see the recommendations you make on pp. 5-6 as an effort to provide such guidance. I'll leave it up to you as to whether you think there is more to be said about standards and educating researchers at this point.

If you choose to submit a revision, please include include a letter detailing your point-by-point responses to each reviewer comment and indicating how you changed the manuscript to address them. The revised manuscript may undergo further peer review. We ask that you submit your revision within three months. Please let us know if you will not be able to meet this deadline.

To submit your revision, log into manuscript central and enter your Author Center, where you will find your manuscript title listed under “Manuscripts with Decisions.” Under “Actions,” click on “Create a Revision.” Your manuscript number has been appended to denote a revision.

IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant or outdated files before completing the submission.

Once again, thank you for submitting your manuscript to AMPPS and I look forward to receiving your revision.

Sincerely,

Alexa Tullett Associate Editor, *Advances in Methods and Practices in Psychological Science*

Reviewer(s)' Comments to Author: Reviewer: 1

Comments to the Author Manuscript ID: AMPPS-19-0139 Manuscript title: Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology

Summary

This article reports an observational study designed to examine data sharing practices and computational reproducibility in the context of Registered Reports published in the field of psychology. Computational

reproducibility is a minimal standard we should expect from all scientific manuscripts that involve data analysis and this is therefore an important topic to address.

The study is generally well designed, the reporting is reasonably clear and concise, conclusions appear well calibrated to the obtained evidence, limitations have been appropriately considered, and, generally, transparent research practices have been adopted (data and analysis scripts are publicly available and shared in a computational environment which enables reproducibility of the article). I am therefore satisfied that the article makes a quality contribution to the literature on this topic.

I have a few comments and suggestions which I hope might be a helpful. Major Comments

1. It is reported that “After evaluating our initial coding round, we considered an article reproducible when we could get the same main results as represented in the paper with at best minor changes to the analysis scripts.” – could a more precise operationalization of these variables be provided? It is not always clear what the main results of a paper are for example, particularly if the paper reports multiple outcomes and/or experiments.
2. “Reproducibility” is not actually defined operationally. Presumably, if values matched exactly, they were considered reproducible, but was any deviation from exactly identical considered to be non-reproducible? It is also not entirely clear which values were examined. For example, were all numerical values related to the ‘main results’ examined, including say, sample sizes, means, standard deviations, degrees of freedom, test statistics, p-values? Or just some subset of these? More detail would be welcome.
3. The findings are reported in terms of how many “articles that could be computationally reproduced” (from the abstract) but it is also stated that only the “main findings” were checked.
4. According to statements in the discussion, it seems that the reproducibility assessments conducted have not been documented in a step-by-step manner or themselves made reproducible. This is explained on the grounds that it would have taken a lot of time - but it does seem a bit unfortunate given the topic and recommendations of the paper. Without this level of detail, it is not clear to me how an independent researcher could systematically repeat and verify the methods employed. Minor Comments
5. The description of Hardwicke et al. (2018) in the introduction states that “Eleven articles could be reproduced...”. Actually Hardwicke et al. report that they only examined a subset of the results reported in each article they examined.
6. In the introduction it states that “The novelty of Registered Reports may attract early adopters who are also exploring other novel developments aimed at improving research practices in psychology, such as data and code sharing. We hypothesized that researchers who publish Registered Reports would also be likely to share data and code in public repositories”. It may also be worth mentioning that journals that adopt Registered Reports are probably more likely to be adopting policies that encourage or mandate data sharing.
7. Later in the manuscript it states that “We examined how many authors voluntarily shared data and code” – but it is not clear how it was determined that the sharing was voluntarily (rather than a response to a journal mandate or request from a reviewer/editor).
8. Several comparisons are made to a study by Hardwicke et al. (2018) and so it might be worth noting that in that study the researchers were assessing analytic reproducibility whereas in the present study the focus was on computational reproducibility. The two terms are very similar and sometimes used interchangeably – however in the former case, the researchers wrote their own code from scratch in an attempt to repeat the original analysis as it was reported in the original studies (most original analysis scripts were not available). By contrast, in the present study, the focus was on re-running the scripts provided by the original authors. This difference may be particularly pertinent to the comparison of time taken to complete the reproducibility checks (for example).
9. The structure of the paper is a little erratic, with both methodological information and discussion blending into the results section and some results appearing in the introduction.

10. The low inter-rater reliability between coders trying to execute scripts was surprising because this seems (naively perhaps!) like a fairly straightforward mechanical task to perform – could more details and examples could be provided to illustrate why there were problems with this.
11. The study does not appear to have been pre-registered (there is no mention of it). Although some benefit may have been gained by pre-registering, this omission is not overly concerning given that (a) the study is being reported; (b) the study is fairly clearly exploratory and descriptive in nature. Nevertheless, I would recommend stating the absence of a pre-registration explicitly and providing a disclosure statement which states whether all data collected, all measured variables, and all analyses conducted have been reported in the manuscript.

Reviewer: 2

Comments to the Author This manuscript discusses issues related to computational reproducibility within Registered Reports in psychology. Computational reproducibility is an aspect of (open) science that is, in my view, often overlooked or regarded as being of lesser importance (compared to, for example, preregistration or other forms of reproducibility).

Overall, I think this manuscript does a good job of conveying not only the issue that there are still many findings out there that lack the data and code to reproduce results, but also that achieving reproducibility entails much more than merely sharing that data and code. I think it is particularly worrisome that of the 35 RRs in your sample that linked to data and code, only 20 contained actual reproducible results. This means that even those researchers who appear to have the best intentions, still often don't get it right. Especially taking into account that these are RRs, which are up there in the *league of good research practices*.

Naturally, I checked out the OSF page associated with this manuscript, and tried to reproduce it all. Even though I had some issues with the Markdown script (I'm running R 3.6.1 which does not run papaja), your README file provided me with enough information to get me the results I wanted to see. I also appreciate the use of a comprehensive codebook and the multiple output files (.tex, .Rmd etc.). The low inter-rater reliability just shows how hard it is to make something completely reproducible and accessible for as many people as possible, on as many different systems as possible.

Generally, I am positive about this paper. I think it is a timely contribution to the literature, and relevant to any researcher working with (shareable) data and code. I very much like that you've included novel technological solutions such as Code Ocean and Whole Tale, as well as references to examples of best practices. Overall, I would accept this article for publication, and the only reason I ticked "revise and resubmit" is because of the following small remarks and clarification questions, that are easy to solve. In order of the manuscript (not importance):

Page 2, method section: you state you found 81 articles that were published in psychology, which I think should be 79.

Page 2: inconsistent number of decimal points (e.g., in percentages).

Page 3, there is a random "+" after the Gertler & Bullock reference.

Page 3, sentence starting with "We used a dichotomous" is worded a bit weird.

Page 3: I did not fully get the following part: "even if variables could only be found with reasonable effort". Do you mean "data" that could be found?

Page 3 - You state you would consider an analysis reproducible, even if you had to make "minor changes". A bit earlier on the page you reference the possibility of rewriting SPSS .sps syntax to R syntax. Does this fall under the minor changes you mention? (I would call that a major change). In general, could you clarify the minor changes a bit more?

Page 4 – It wasn't immediately clear to me if the average time to reproduce analyses includes everything, from reading the manuscript for the first time to matching the reported and reproduced results? If so, I commend you on your speed (even though I can assume that it must have felt very tedious at times)!

Page 5 - In the discussion section, you state issues you encountered with SPSS scripts. However, the R analyses took longer to reproduce, and the issues with R scripts were not clear to me. In the paragraph before the SPSS script one, you state researchers should be clear on the order of R scripts- was that the biggest issue with the R scripts?

Page 5 - You introduce in the “Nevertheless,” paragraph 4 points researchers should focus on (ABCD). The next paragraph (“First,”) has A, the paragraph after that (“Second,”) has B and D, and the paragraph after (“Based”) has C. It is minor but the change in order distracted me while reading it for the first time.

Page 7 - You state developing standards and educating researchers about best practices could help progress. Are there any examples that you could add here? If this is such an important point, I think some clarification could help.

Datafile, codebook tab: I don’t fully understand the included_language variable on line 20. To me it reads as if the programming language should be included in the final dataset and I do not understand why.

That’s all! Please let me know if I can provide any clarification on some points.