Referee report for manuscript CJE-2022-128

This paper investigates the reproducibility level of most papers published in the *American Economic Journal: Applied Economics* between 2009 and 2018. In this literature, a given paper is said to be *reproducible* if another researcher can regenerate all the results displayed in the paper by re-running the original computer code on the original data. The authors report a 38% reproducibility rate, which is higher than the levels found in previous research in economics. In addition, they find that reproducible papers do not attract more citations than less-reproducible or non-reproducible papers.

I very much appreciate the topic of the paper, which I think is extremely important for our profession. The *AEJ-Applied* experiment conducted by the authors is large and very interesting. The insights produced could be useful to draft reproducible policies in economics academic journals. However, my overall feeling is that the submitted manuscript is still preliminary. To make an impactful contribution, the paper still requires significant work. Below my comments, which I hope will prove useful to the authors.

Main comments:

1. This paper is about *reproducibility*, and it gives a very clear definition of it on page 3 citing Bollen et al. (2015); this is the ability to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator. This paper is NOT about *replicability*. The latter concept means that the results remain valid when one changes the methodology and/or the data (see footnote 8 for another excellent definition). Both terms are very important and remain poorly understood by most researchers in economics. Yet the current paper uses **both terms interchangeably**: e.g. our replication success rate is relatively moderate; Page 4: our replication exercise was conducted; Page 5: the expected level of replicability of an article; Title Table 1; Page 12: one replicator was able to replicate the results; etc. As a result, someone not familiar with this terminology would conclude that both terms are synonyms (which is wrong) and someone familiar with this terminology would wonder why the authors keep talking about replicability when they only check reproducibility (which is problematic).

2. The paper analyzes a given policy but what is the policy exactly? What are the authors supposed to provide? All computer code and data, except if they are confidential, along with a readme file? On page

3, the paper says the data must be “properly documented and […] readily available to any researcher for purposed of replication”. Does that mean that the data must be stored in a public repository, or can the authors commit to readily provide them to other interested researchers when politely asked to do so? What does the policy say exactly about confidential data? Did the policy change between 2009 and 2018? It would be useful to clearly copy and paste the exact policy and provide any relevant additional information allowing the readers to fully understand its scope.

There are several ways to assess this policy:

A. One could check whether the authors strictly comply with it and measure the percentage of paper sharing their code and data, as requested by the policy. In addition, one would check whether the authors not sharing all their data have a good reason for not doing so (e.g. they use confidential data or non-sharable commercial data).
B. One could measure if this policy is sufficient to guarantee the reproducibility of the results published by the journals. To do so, one could download the code/data from a repository and try to run the code and compare the regenerated and published results.

While both analyses would be interesting per se, the current paper focuses on B (see title of Section 3) but also tries to do a little bit of A when contrasting in Table 3 papers (i) using confidential data and not sharing, (ii) not using confidential data and not sharing, and (iii) sharing data. Using confidential data does not imply that no data can be shared. Often, papers combined several datasets, some of which could not be shared while others could. As a result, a given paper could in theory be in both rows: “Confidential Data” and “Data was provided”. The point I would like to make is that to do B, one just needs to tell us whether all data/code are available, some code/data are missing, or no code/data are available. This is very much in the spirit of what is done in Table 6. You can do that without investigating the reason(s) for any missing resources. For me, studying the reasons is another interesting question, but a complex one, which requires a much rigorous approach than what it currently done.

3. The empirical analysis is not easy to follow. First, it is never mentioned in the paper that AEJ-Applied published a total of 390 papers between 2009 and 2018. However, the considered sample only includes 303 papers. How were these 303 papers selected from the entire population? For instance, in 2013, 38 papers were published but only 10 of them are in the sample. Same for the year 2015, 24 papers in the sample out of 36 published papers this year. Initially I thought it was because the missing papers were not using STATA or Matlab but Table 4 suggests this is not reason. An alternative reason is that

theoretical papers got excluded but my understanding is *all* papers in AEJ-Applied have some empirical results of some sorts. So, it cannot be the reason either. Please tell us what were the decision rules to generate this sample?

Out of these 303 papers, 342 assessments are conducted (see footnote 16) and for each paper, the authors select the most successful approach. This is biasing the results upwards and needs to be acknowledged.

Then we see that, out of the 303 papers, 209 have their data available. However, when turning to the documentation-clarity analysis, there are only 180 papers left. Footnote 17 indicates that the attrition is due to some missing assessment questionnaires, but the main text indicates that some questionnaires were filled in ex post… Honestly, this is very hard to follow. Eventually, I was happy to see that an Appendix was supposed to describe the various steps, but the appendix was missing in the submitted manuscript I received and appeared with a ?? sign in the pdf. Overall, I suggest the authors to construct a cleaner and smaller dataset and keep using the same one during the entire analysis.

5. There are two main empirical results in the paper. The first (and main) one is about the level of reproducibility at AEJ-Applied and the second one is about the link between reproducibility and future citations. While both results are currently in the same section (3), I suggest putting them in separate sections (3 and 4) and significantly improving the section on citations. Two potential extensions could be: improve the theoretical motivation of the effect of reproducibility on citations and enrich the empirical analysis (currently there are only 77 papers as the analysis is only on the first part of the sample period). In particular, I do not see why the authors cannot include all papers in the analysis: measuring the h-index of all authors on the publication year and measuring the number of citations 3 (or 5) years later.

6. The authors need to compare their set-up and results with those in “Reproducibility of Empirical Results: Evidence from 1,000 Tests in Finance” by Perignon et al. (2022, WP SSRN). Unlike the current manuscript, Perignon et al. (2022) does not rely on academic papers published in a leading economic journal but on a large-scale controlled experiment in finance.

7. Full reproducibility and no reproducibility are both clear outcomes. The former means all results have been perfectly regenerated while the latter means no results have been regenerated (typically the code

does not run or generates very different results). However, unlike in Perignon et al. (2022), there is no former definition for partial reproducibility in the current manuscript. Consequently, it remains a vague concept in the current paper. Say a given paper has 15 tables and the student is able to regenerate 14 of them perfectly. That's partial reproducibility. Similarly, it is also partial if the 15 tables are reproduced but the second decimals of all results are different. Alternatively, the reproducibility is partial if the student only reproduces the first table with the summary stats but nothing else. As there are way too many degrees of freedom, very little weight can or should be put on the partial reproducibility results.

8. It would be particularly interesting to compare the level of reproducibility of the papers published by the *AEF_Applied* before the enforcement of the systematic pre-publication reproducibility check conducted by a data editor (post 2018). The current sample period would become the control sample period and the post 2018 would be the treatment sample.

Minor comments:

9. I very much like the title of the paper, but it has already been used in a similar context by Willis and Stodden (2020): Trust but verify: How to leverage policies, workflows, and infrastructure to ensure computational reproducibility in publication. Harvard Data Science Review, 2 (4). Consequently, the title needs to be changed.

10. On page 2, when the authors talk about academic journals committing to publish replication papers, they could mention the newly introduced *JPE Micro*.

11. On page 3, what does that mean that the paper studies “a particular journal's data availability policy, combined with light enforcement”? Do the authors imply that AEJ-Applied enforced its DAP lightly? Is it an assumption or a result from the current study?

12. On page 4, the authors say their first contribution is to “provide an estimate of reproducibility standards of a journal that imposed, from its creation, a data availability policy”. I do not understand. It would be better to write something more direct, such as “provide an estimate of the reproducibility level

of the papers published by the AEJ-Applied, which is a journal having since its inception a strict data availability policy”.

13. The papers use undergrad students as replicators. I understand why this is the case as this is part of a course at Cornell University. However, the reason why some people talk about a potential reproducibility crisis in science is because some published papers by researchers cannot be reproduced by their *peers*. That is really problematic indeed. However, publications in peer-reviewed academic journals are not supposed to be read, and reproduced, by undergrads. A related point is that, according to the protocol, the replicators do not contact the authors of the published paper. This should not be put forward as an advantage. In reality, when a given researcher aims to reproduce the results of a given paper, he or she would definitely contact the original authors in case of problem or question. Both features (undergrads and no contact) bias the reproducibility success rates downwards.

14. I was a bit lost about the software used. On page 3, the replicators are supposed to only have knowledge in Stata and Matlab but in Table 4, some papers use SAS, R, SPSS, and Excel. On page 7, when discussing the versions of the software, the authors mention Stata, SAS, and SPSS but not Matlab. This needs to be clarified.

15. Figure 1 and footnote 15 do not bring much. They both should be removed.

16. In Table 3, data means data + code, not just data. Is my interpretation correct?

17. On page 5, the authors talk about machine-readability and machine reproducibility. They mention that they will return to it in the conclusion, but this is nowhere to be seen on page 21.

18. There is no need for 3.2.2 to be a section. It is just a correlation coefficient. It could be merged with 3.2.1 and basically just have all these results in a Section 3.2.

19. In Tables 13-15, please add a caption allowing one to understand the table without going back to the main text and write the variable name more explicitly (e.g. what is avghindex:confidential_data?). Table 15 is just the log version of Table 14 and should go to the appendix (or keep T15 in the main text and put T14 in the appendix).

20. In the conclusion, it is not really "a journal that introduced a data availability policy". It is a paper that was created with a data availability policy since day 1.

21. In the third paragraph of the conclusion, the point about complex changes should come first as it is related to the reproducibility rate. And the point about the potential citation bonus should come afterwards.

Typos:

Abstract: since 2005 whereas the journal was created in 2009.

Page 4: four summers from 2014 to 2018. Then it becomes five summers on page 9.

Footnote 12: Since September 2018, the team has used…

Space in the definition of t after equation 1.

Page 13: similar

Page 13: replicability ratio

Footnote 18: No capital letter at the beginning of the sentence and no dot at the end.

References:

Perignon C., O. Akmansoy, C. Hurlin, A. Menkveld, A. Dreber, F. Holzmeister, J. Huber, M. Johannesson, M. Kirchler, M. Razen, U. Weitzel (2022), Reproducibility of Empirical Results: Evidence from 1,000 Tests in Finance. SSRN working paper.

Willis C. and V. Stodden (2020): Trust but verify: How to leverage policies, workflows, and infrastructure to ensure computational reproducibility in publication. *Harvard Data Science Review*, 2 (4).