General notes

We have revised the paper for grammar and language throughout.

We have revised the definition of reproducibility, replicability to be clearer, and with reference to the 2019 NASEM report, which is often considered the authoritative definition. We first reference this for clarity in the introduction, but define it more clearly in the "Methodology" section. This was previously in the introduction. Note that we explicitly address the issue that in our experimental setup we used the term "replication" which now should be interpreted as "reproduction", see Footnote 6.

We have renamed Section 2 "Description of Reproduction Procedure" to "Methodology" and Section 3 "Are data policies enough to ensure full reproducibility?" to "Results".

We have also manually reviewed the write-in responses provided by the students that described reasons for failure to reproduce, which has lead to a reclassification of several outcomes from "Other" to "Confidential data". This affects any numbers that were computed as "conditional on data being available", but does not affect the absolute numbers that were fully and partially reproduced (see new Section 4.2).

Referee 1

Summary: I greatly appreciate this paper for many reasons. First, it is latest large scale reproduction exercise of a journal that introduced a data availability policy (DAP) since the seminal work done by Dewald et al. (1986), and later by McCullough, McGeary and Harrison (2006), who analyzed the resources put online by the Journal of Money, Credit and Banking (JMCB) for the years 1996-2003. Here, the authors consider the journals of the AEA, namely the AEJ:AE. This is particularly interesting due to (1) the leading position of these journals in our research field and (2) since the AER was pioneer as concerned the reproducibility policy (with the JMCB and the Journal of Applied Econometrics). Indeed, the AER was the first top-5 economics journal to introduce a DAP back in 2004 (Bernanke, 2004) followed by the Journal of Political Economy in 2006, whereas the Quarterly Journal of Economics was a late adopter in 2016. Second, the authors introduce a second interesting discussion about the impact of reproducibility about citations and more generally on the incentives to do a reproducible research. They obtain two major results. First, they found a moderate reproducibility rate of 37.78 % (41.98 % conditional on non-confidential data). Second, they show that replicability of papers did not provide a citation bonus. The paper is well written and the framework of the reproducibility evaluation is precisely detailed. I bet this article that shows that reproducibility does not guarantee citations will be widely cited.

Comments:

1. This paper can be viewed as an assessment of the DAP, with a focus on the

DAP of the AER. First, I think that it could be useful (i) to define what is a DAP in general and

We have expanded the discussion of what the specific DAP is.

(ii) to discuss the DAPs in the landscape of academic journals in economics. Maybe the reader should be surprised to learn that although significant made in recent years, economic journals with a DAP are still in the minority. Out of a sample of 343 economic journals, Höffler (2017) identifies 158 journals (46%) without any DAP and 49 journals (14%) that merely recommend that their authors provide the codes and data on request. Since 2016 all the top-five journals have a DAP. However, even when they exist, these policies remain generally not very coercive and sometimes fall within the realm of communication alone.

We define what we understand by "light" enforcement, and describe it for this journal, as well as providing a brief overview of the literature.

Höffler (2017) mentions that only 28 journals have a DAP in which the deposit of codes and data is mandatory for publication, while the others only recommend the deposit of these resources. Why is this important to note? It may imply that the reproducibility rate obtained in this study could be considered as a kind of upper limit of the reproducibility rate we would obtain on other reviews.

We have included references to the literature on the effectiveness of DAPs (both Höffler and Vlaeminck come to mind), and alluded to the upper limit in the conclusion. We also discuss how for this particular journal, actual reproducibility, if undertaken by a more skilled replicator, might actually be higher.

2. In 2004, Ben Bernanke, then editor-in-chief of the AER defines the journals first replication policy. This DAP requires authors to deposit codes and data in an open archive, except when confidential data are used. This paper is not the first assessment of the AERs DAP.? In 2008, the AER launched an audit to assess the quality of the data and code contained in its online data archive. To do so, six economics Ph.D. students selected a sample of 39 empirical articles, out of the 135 published articles subject to the data policy between 2006 and 2008. The conclusions of this internal audit were published by Glandon (2011) who states that roughly 80 percent of the submissions satisfied the spirit of the AERs data availability policy, which is to make replication and robustness studies possible independently of the author(s). Indeed, out of the selected 39 articles, 11 are based on proprietary data and 20 have the appropriate code and data posted on the journal repository. Considering that these 31 papers are believed to be replicable, the report concludes that around 80% (31 out of 39) of the papers comply with the data policy. The rather optimistic conclusion drawn by Glandon is vigorously criticized by McCullough (2018). He notes that nothing guarantees that the shared code and data allow duplicating the results of the 20 sample papers. Furthermore,

the reproducible study itself only concerns a sample of nine papers. For five of these papers, the results have been fully reproduced, whereas for the other four, the results were only partially reproduced. Thus, for McCullough (2018), the American Economic Review archive did not support the publication of reproducible research and, consequently, the rules of the archive should have been drastically amended. The current study will close this debate about the effectiveness of a standard DAP operated in a top-5 journal, that is only based on data and code archive without any control of the resources shared by the authors.

A reproducibility rate of 37.78 % (or 41.98 %) obtained from a rigorous reproducibility study for the top journal in economics, argues for an end to DAP policies only based on archive. This is the ultimate argument to generalize the reproducibility assessment of the submitted papers and the pre-publication verification of materials provided by authors, as it is done by the data editor. Here, the authors should mention the change in reproducibility policy of the AEA with the appointment of one of the authors as data editor for all the journals operated by the Association. Since then, similar positions have been created at Review of Economic Studies, Economic Journal, Management Science, etc.

We now refer to the changes that happened after this exercise was concluded (and in part because of the conclusions supported by earlier versions of this paper).

3. The authors consider a sample of 303 articles. They should explain how they have been selected from the population of published articles over the period.

We have, in line with similar comments from Referee 2, streamlined the explanation of the sample selection.

4. The authors show that the reproducibility does not improve the potential of citations of the published articles. I am personally convinced by the argument. But, the authors should mention the opposite results obtained in the literature on this point. For instance, Gleditsch and Metelits (2003) show that an article for which the original data are made available has twice as many citations as an article for which the data are not available. Similarly, Höffler (2017) shows that the availability of numerical resources increases the visibility of academic journals. I think that the authors should insist on the fact that they control for the journal quality by comparing papers which have been published in the same [paper]=[journal], which was not the case in Gleditsch and Metelits (2003).

Thank you for this point. We have expanded the discussion around various discussions of citation benefits, which can go in multiple directions, depending on the baseline in the literature (across or within journal comparisons). The within-journal comparison was particularly salient for us.

5. I am convinced by the fact that the question to know if the reproducibility increases (or not) the academic outcome (citations, future publications, etc.) is important, but the authors should also consider the reverse effect. What are the determinants of reproducibility? Why some authors (those of the 37%) spend effort and time to make their work reproducible?

It could be interesting to propose a binary regression model to explain the probability of being reproducible with some individual factors. The database considered here could allow to answer to some interesting questions. Is there an effect of the seniority of the authors measured by the h-index? Is there an impact of the tenure? Is there a generational effect? Is there an impact of working in a top-academic institution?

Thank you for the excellent suggestion. We have added a section on this, using data from OpenAlex that proxies for much of this (we did not measure generational effects, but do believe based on our own observations that there is something to it – "publishing experience" must be sufficient here)

Minor comments:

1. Introduction, page 4: the authors should mention the appointment of data editors in other top-5 or top-field journals.

Done.

References

[1] Bernanke, B. S. (2004). Editorial statement. The American Economic Review, 94 (1), 404404.

[2] Glandon, P.J. (2011). Appendix to the report of the Editor: report on the American Economic Review data availability compliance project, American Economic Review, 101(3), 696-699.

[3] Gleditsch N.P., C. Metelits (2003) Posting your data: Will you be scooped or will you be famous? International Studies Perspectives 4, 89-97.

[4] McCullough B.D. (2018) Quis custodiet ipsos custodes? Despite evidence to the contrary, the American Economic Review concluded that all was well with its archive. Economics: The Open-Access, Open-Assessment E-Journal, 12 (52), 1-13.

Referee 2

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.

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).

We appreciate the referee pointing this out. We have cleaned this up (see also the general comments). Part of the confusion comes from the fact that the exercise reported here was conducted mostly prior to the "consolidation" (incomplete, as it is) of the terminology used. We try to make that clearer in the setup of the Methodology section.

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 13, 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.

We link to the policy as it was posted in 2019, and as far as we know, it was unchanged except in scope of application since 2005 (after Bernanke's editorial announcement in the AER). The policy changed in 2019 through the efforts of one of the authors of this paper. Information about application of the policy is current as of 2018, and presumably several years prior to that date, but we have no information on application of the policy since 2005. This is all in the new Background section.

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.

We agree with the referee that (A) is more complex, and warrants a deeper investigation. On the other hand, there are no such investigations that we are aware of, so providing some information, even if cursory, is a contribution. Finally, it is important to understand at least some rough correlates of non-provision of data to assess whether "reproduced" in this exercise is a lower bound, and could be improved. We point to cascad for an example of how "confidential" data

can be assessed when the right conditions are met, and Table 3, while reporting results collected in this exercise, also sets up that part of the conclusion. We have thus chosen to leave Table 3 in, but appropriately caveat it in line with the referee's comments. [TODO]

> *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?*

We now explain the non-systematic (but non-informative) selection and assignment of papers in the methodology section, in a new segment called "Article selection". The appendix table explaining the sample has been fixed as well.

> *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.*

This (intentional) upward bias relative to the unfiltered is now acknowledged (it was present in earlier versions, apologies for it having been lost in edits). We do note that this bias is primarily to alleviate lack of skills in the replicator group, and probably still leads to an underestimate of the real reproducibility of these articles.

> *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.*

Again, apologies for this edit error on our end. The appendix with details is now included, and we have overall simplified the exposition of the sample. We constructed a smaller dataset, and have two main data set: one of assessed papers for which we had complete questionnaires, and one of reproduced papers, both used consistently throughout the 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.

We have split these results into what is now Section 4.1 and 4.2. .

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.

We have expanded the discussion around why reproducibility might generate a citation bonus. We have expanded the analysis, with a larger sample covering all papers in our analysis and the full sample period (albeit with the caveat that we cannot, with the new data source, measure the impact for the whole sample of characteristics at the time of publication). We have looked at the effect on citations 4 years later, but also at the effect over time. However, with this extended sample, the increased presence of data editors across many econ journals, but in particular at the journal being studied, might lead to a change in citation practices. We therefore also included a dummy for the post-2019 period – which shows no effect.

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.

Thank you for pointing this out. Perignon et al (2022) came out after we originally wrote this article. We have added a discussion about how our paper is positioned in the broader realm of reproducibility verifications of the published literature, and mention in particular this paper as a clean measure of author skills in a "laboratory" setting.

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.

We agree that there are many degrees of freedom here, but we disagree that it is not a useful measure. In all three cases, we cannot exclude replicator error (even in the – reported – fully reproducible case). The partially reproducible cases may still be reproducible by more skilled replicators. We do note that the very artificial scenario in Perignon et al lends itself to a more precise definition on practical grounds – there are only six numbers for each team to verify. The typical paper has a dozen tables with several hundred numbers, which is much harder to quantify as cleanly.

It is also important to report partial reproducibility, because even older texts sometimes play loose with the definitions. While we are uncertain whether the glass is half empty or half full, it is definitely not empty, and that, we find, is an important finding.

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.

We leave this for future research.

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.

Done. We have changed it to “Reproduce to Validate: a Comprehensive Study on the Reproducibility of Economics Research”.

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

Done, together with the older JAE.

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?

The process of enforcement is now described in the Background section, with a short explanation in the introduction.

> *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".*

Thank you for pointing that out. We clarified our contribution in the first paragraph of section 2, background, writing more directly that we assess whether the availability of replication packages, in compliance with a posted data availability policy, leads to reproducibility

> *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.*

Recruitment is described as part of the methodology section, and is not part of any class (these are paid replicators who happen to be undergraduates): "*Over the course of five summers, we recruited undergraduate students (typically but not always rising seniors) for the reproducibility exercise as part of summer research, to serve as assessors and replicators.*"

We disagree with the referee in multiple points. It is not clear why undergraduates should not be able to reproduce the articles in the scientific literature. They are required to read them, and learn the basic tools. In fact, as we demonstrate in this exercise and in Vilhuber et al (2022), undergraduates are mostly perfectly capable of reproducing these articles.

We also dispute the referee's assertion that a replicator contacts the authors in the case of problems or question. This was first emphasized by King (1995) and reiterated by multiple authors in this field afterwards (e.g. McCullough et al 2006, Glandon 2010): Replication materials should allow a replicator to reproduce the analysis without information or interaction with the author. The need to contact the author is already, in our mind and those of others in the

literature, a failure. In part, that is unlikely to be successful, based on studies that have looked into requesting replication materials. Furthermore, that is likely to only work in a brief period after publication, assuming the authors are still in academia. The same way that the article has to stand on its own, and not rely on subsequent correspondence to explain this paragraph or that expression, the replication materials can and should stand on their own, because that is the most likely outcome.

This is different, of course, in pre-publication verification, or during the development phase of a paper, where interactivity is key to the scientific progress.

While we disagree on the specifics, the referee's questioning of this point has led us to explain a bit more the motivation for this particular setup in the article. **We have moved the mention of "not contacting the authors" to a more prominent position within the Methodology section.**

> *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.*

We have actually removed this from the current version of the paper, but we did clarify that the replicators had access to most software. We identify the one case where software was not availble (Postgresql).

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

We agree about Figure 1, but we very much liked Footnote 15. We are sorry to see it go.

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

The question is from the "Entry" questionnaire, which asked specifically about data. As we note in the text, "*all articles had some supplementary materials*".

> *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.*

Thanks for pointing this out. We have removed this.

> *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.*

We have merged with the discussion of reproduction success in section 4.1.4.

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).

We kept the main specification in the main text and moved other tables in the appendix. Thank you for pointing that out, we added captions to every tables explaining the sample, as well as the covariate definitions.

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.

Correct. Done.

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.

We have corrected this to the correct "five", thank you for pointing this out.

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

This is no longer relevant, and has been removed.

Space in the definition of t after equation 1.

Done.

Page 13: similar

Done

Page 13: replicability ratio

Thanks. Done.

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

Done, though we have removed the footnote.