Comments on "Trust, but verify: lessons from a large scale test of economics' research reproducibility"

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...

Other 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 (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. 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.

2. In 2004, Ben Bernanke, then editor-in-chief of the AER defines the journal's 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 AER's 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 AER's 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.

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.

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, which was not the case in Gleditsch and Metelits (2003).

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?

Minor comments:

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

References

[1] Bernanke, B. S. (2004). Editorial statement. The American Economic Review, 94 (1), 404–404.

[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.