McCullough, B. D.

**Article**

# 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

**Provided in Cooperation with:**
Kiel Institute for the World Economy – Leibniz Center for Research on Global Economic Challenges

*Suggested Citation:* 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, ISSN 1864-6042, Kiel Institute for the World Economy (IfW), Kiel, Vol. 12, Iss. 2018-52, pp. 1-13, https://doi.org/10.5018/economics-ejournal.ja.2018-52

This Version is available at:
https://hdl.handle.net/10419/181451

Economics
The Open-Access, Open-Assessment E-Journal
Vol. 12, 2018-52 | August 15, 2018 | http://dx.doi.org/10.5018/economics-ejournal.ja.2018-52

# Quis custodiet ipsos custodes?: Despite evidence to the contrary, the American Economic Review concluded that all was well with its archive

*B. D. McCullough*

**Abstract**
In 2011, the Annual Report of the Editor of the American Economic Review reported that the journal's data-code archive was functioning well, and no changes were made to the archive rules. This was based on an audit of the archive that the Editor had commissioned. In point of fact, all was not well with the archive: the archive did not support the publication of reproducible research. The rules for the archive should have been changed and were not; thus the American Economic Review continued to publish articles that were not reproducible. The cause of reproducible research was set back many years. Currently it appears that the AER intends to reproduce articles prior to publication; this would be a horrible mistake.

(Published in Special Issue The practice of replication)

**JEL** B40
**Keywords** Replication; reproducible research

**Authors**
*B. D. McCullough*, Drexel University, Philadelphia, PA, USA, bdm25@drexel.edu

**Citation** B. D. McCullough (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 (2018-52): 1–13. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-52

**Received** September 2, 2017 **Published as Economics Discussion Paper** September 28, 2017
**Revised** July 17, 2018 **Accepted** August 3, 2018 **Published** August 15, 2018

## 1 A general discussion of principles about how one should do a replication

Before we can discuss replication, we need to define it. The word is used in many different and sometimes conflicting ways, both within and across disciplines. The title of a recent news article from *Nature* (Baker 2016) describes this problem accurately: "Muddled meanings hamper efforts to fix reproducibility crisis". This confusion harms research and retards progress. Clemens (2015) performed a yeoman's job in classifying forty one (!) different uses of the word "replication" within economics. Clearly there is a need for a standard taxonomy. For present purposes, the concept of "narrow replication" (a.k.a "reproducibility") suffices: The data and code in the archive reproduce the published results. (The actual purpose of an archive is broader than this, a point we shall make shortly.)

As an example of this need for clear thinking and precise definitions when talking about replication, consider the recent paper by Chang and Li (2018) that has received much attention. In their abstract they write:

> "We successfully replicate the key qualitative result of 22 of 67 papers (33%) without contacting the authors. Excluding the 6 papers that use confidential data and the 2 papers that use software we do not possess, we replicate 29 of 59 papers (49%) with assistance from the authors."

When they write that they cannot attempt papers that use software they don't possess, it implies that they do have the relevant software for the papers that they do attempt to replicate. This is an important point; they are not porting the code to other software, which would be a *prima facie* reason to encounter numerical differences. On the other hand, it is unclear whether they used the same software on different platforms, *e.g*, RATS 9.2 on Windows vs. RATS 9.2 on Linux, which could also induce numerical differences. Notice also the phrase "key qualitative results". One might think that what Chang and Li actually did is *confirm* key qualitative results, *not* reproduce them. Consider the following quote from their paper (p. 7):

> "We define a successful replication as when... for example, if the paper estimates a fiscal multiplier for GDP of 2.0, then any multiplier greater than 1.0 would produce the same qualitative result (i.e., there is a positive multiplier effect that government spending is not merely a transfer or crowding out private investment)."

On its face, Chang and Li's criterion for replicability does not make sense. Think about it: Using the same data, same code, and same software, the original author gets 2.0 while Chang and Li get 1.0 and they think this is a successful "replication". The number "1.0" most certainly does not replicate or reproduce the number "2.0" when using the same data and code! Chang and Li (p. 2) write, "Using the author-provided data and code replication files, we are able to replicate 22 of 67 papers (33%) independently of the authors by following the instructions in the author-provided readme files." If Chang and Li used the same data, same code, and the same software to get 1.0 when the original paper shows 2.0, then Chang and Li *prove* that the paper is not reproducible; they have shown the exact opposite of what they intended to show! If Chang and Li used a different version of the same software, then the best they could claim in such a situation is that at least one

of the two software packages in numerically inaccurate, not that anything has been reproduced. We can be quite confident that Chang and Li did not actually reproduce the results of 22 papers, and the actual number is probably much lower than 22. The question isn't whether the replication "supports the conclusions" of the original paper, the question is whether there exists data and code that reproduces the published results.

A clear distinction between reproducible and replicable is important. The recent article by Camerer et al (2016) clearly involved replication: they ran the same experiments on different subjects. In its earliest use in the physical sciences, to "replicate" an experiment meant to perform a second experiment in conditions similar to a first experiment, with the intent of confirming or disproving the result of the first experiment. With the advent of widespread computing, in about 1990 the geophysicist Claerbout coined the term "reproducible research" to refer to reproducing published results, typically using the same data and code but also allowing the coding to be done in a different language, as long as the published results are reproduced.

This hair may be further split between reproducible and repeatable. Imagine taking someone's data and code, running it on a different computer, and getting a different answer. One would say that the results might be repeatable, but they are not reproducible (Easterbrook, 2014). This is what Chang and Li find when the paper's result is 2.0 and they get 1.0 by running the same data and code: The published result is repeatable, but *not* reproducible.

It is also important to note that reproducibility does not imply correctness. For example, Donohue and Levitt's (2001) article on abortion may have been reproducible, but it was not correct. In the course of reproducing the article using the author's own data and code, Foote and Goetz (2008) discovered a coding error that invalidated the article's results.

The purpose of a journal's data/code archive is to ensure that the journal's published results are reproducible. This is a miminal standard that is easy to understand: either the results of an article can be reproduced or they cannot: it is a binary decision. To argue that "some of the results are replicable" or "the important results are replicable" is to admit that the article's results are not reproducible. We can quibble over how many significant digits matter as far as the conclusions of the paper are concerned, but as far as reproducibility is concerned, using the same data, code and software should reproduce the exact same published result. If the published result is 12.345, any numerical result that rounds to 12.345 reproduces the published result.

If the code is ported to another package, or the same package on a different platform (*e.g.*, Windows vs. Linux, or AMD vs. Intel), or a different version of the same program (*e.g.*, v1.0 vs. v1.01), then rounding differences can and should be excused. For linear procedures with moderately-sized datasets, there should be ten digit agreement, for nonlinear procedures there may be as few as four or five digits of agreement. See McCullough and Vinod (1999) for details. The audit team can make note of these differences and still classify the article as reproducible.

Now that we have definitions established, we can discuss procedure. For computational research, it is very easy. Put the data and code in the same folder, and run the code. Barring minor accommodations such as different operating systems (*e..g*, the author uses Stata in Windows while the replicator uses Stata in Linux), if it fails to execute (the code doesn't run), the person who prepared the data and code has failed to provide evidence that the article's results are reproducible, and the article should be labeled as such. It is *not* the duty of the would-be replicator to spend

valuable time trying to make the data and code work. To require this is to permit the original author to engage in cost-shifting; he spends less time preparing his replication files, and the replicator spends her time trying to make sense of data and code that doesn't work. If the data and code run but do not reproduce all the published results, she does not spend her valuable time trying to fix the data and code so that they do reproduce the published results. Even if she succeeds in this effort, it remains the case that the data and code *that are in the archive* do not reproduce the published results. She should inform the editor that the article has failed to replicate, how it has failed to replicate, and let the editor notify the original author. If he does not swiftly respond with data and code that reproduce all the published results, the article should be flagged as not reproducible. In general, no explanation of the extent of the non-reproducibility should be given, for this invites sloppy research. (Of course, if he used version 1.0 of the software and she used version 1.1, this is not a failure to reproduce, since the same algorithm was not applied to the data.)

If the article is not computational in nature and perhaps requires human judgment for classification, then the article should enumerate protocols so that another person would arrive at the same classification. This was a part of the Hoxby/Rothstein debate. Hoxby created her controversial variable on the number of streams by looking at a map and counting "all streams that were at least 3.5 miles long and of a certain width on the map" (Hoxby 2000, p. 1222), but she provided no further details. What was this "certain width"? Was it 1mm or 5mm in width? While this information probably should not have been in the actual article, it should have been in the archive. This lack of detail all but insured that no one else would be able to reproduce her work. As Rothstein wrote (Rothstein, 2007, pp. 2033-34): "Where Hoxby reports five larger streams in Fort Lauderdale, I counted 12, and a research assistant - working independently - counted 15". The bottom line is that other researchers, working independently, could not get the same result she did. Her paper was not reproducible.

Hoxby could easily have ensured that her work was reproducible. She could have included in the archive the specific map she used, and the specific rules that she used for counting. It is not the duty of the journal to ensure that these details are in the archive, it is the duty of the author to ensure that others can reproduce her work. This case makes clear that simply placing "data" in the archive, without a complete data dictionary that describes the provenance of all the data, is an ineffective means of ensuring reproducibility.

Both a coding error and a failure to document the provenance of a constructed variable plagued a paper by Levitt (1997). He had intended to give more weight to crimes that are less variable, but a coding error gave more weight to crimes that are more variable. Correcting this coding error reversed the conclusions of the article. Moreover, Levitt's key instrumental variable on the timing of elections could not be reproduced by McCrary (2002). When Levitt (2002) attempted to reconstruct this variable, he could not do it! If anything demonstrates the need for archive rules that cover primary data, this is it.

Based on their work covering complete analysis of several archives, McCullough et al. (2008) provide "Recommendations for an effective archive".

## 2 An explanation of why the "candidate" paper was selected for replication

Dewald et al. (1986) called into question the reproducibility of published economic research. They considered possible solutions to the problem, in particular they dismissed the idea of a "replication policy" that requires authors to supply data and code to would-be replicators after publication (primarily due to agency problems – once the article's published, the author has no incentive to spend time organizing replication files). They concluded that only a mandatory data/code archive might solve the problem, provided that the data and code were deposited *before* publication. The above notwitstanding, then-editor of the *American Economic Review* Orley Ashenfelter instituted a "replication policy". McCullough and Vinod (2003) confirmed that, as predicted, the *AER* replication policy did not produce reproducible articles. In response, then-editor Bernanke (2004) adopted a data-code archive. Before he could implement policies to ensure that the archive would result in the *AER* publishing reproducible research, Bernanke left academia and resigned his post as *AER* editor. In his first Annual Report, Bernanke's successor introduced the following boilerplate (Moffit, 2011) that is found almost verbatim in the Editor's Reports through the end of his term (even through that of his successor):

> In 2004, the Review began to require that authors of accepted papers who employ data in econometric exercises, simulation models, or experiments agree to post their data and programs on the journal Web site unless an exemption for proprietary data is requested and granted. The policy was strengthened in 2005 with more systematic enforcement and with greater attention to searching for alternative means of data access for papers requesting exemptions. Table 8 shows the number of papers in each of the 2009 issues containing data analysis, the number of exemptions granted, the number of authors who complied on the first round (defined as supplying data after receiving the acceptance letter detailing the requirement), and the number of authors who complied after a later reminder. Full compliance was achieved for all issues.

This boilerplate is the only time the Editor mentions the archive, and no one reading the "full compliance" sentence would have any reason to think that the data and code in the archive was doing anything other than reproducing the published results. There is no other mention of the archive until his final report in 2011, in which he wrote (pp. 686–687):

> In the summer of 2008, the *AER* conducted an exercise to check the submitted files of *a random set of papers* [emphasis added] to check for compliance with the policy, which requires the submission of both programs and data and an explanation of how to use them. A report prepared by Philip Glandon, included as an Appendix to this report, describes the project and the results. The vast majority of authors complied with the intent of the policy but a small fraction submitted materials that were either incomplete or that would have made replication difficult....Mr. Glandon's report contains additional details on the project and recommendations for strengthening the *AER*'s data posting policy.

Naturally, if there had been anything seriously wrong with the archive, the Editor would have taken steps to address the problem. To the casual reader, the Editor's remark suggests that all

was well with the *AER* archive, save perhaps the occasional glitch. The casual reader may well wonder whether the recommendations to strengthen the archive were even necessary. After all, if the archive policy needed to be strengthened, surely the editor would do it. As shown above, many authors followed this (misleading) storyline and reported that all was well with the *AER* archive. Here are some examples:

1. "The *AER* conducted a self-review and found relatively good, though still incomplete, compliance with its data sharing policy (Glandon 2010)." (Christensen and Miguel 2018)

2. "Roughly 80% 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). The replicated results generally agreed with the published results." (Breure and Hoogerwerf 2011)

3. "For instance, one in five articles examined from the 2006–2008 period in *AER* did not fully satisfy the requirement that results be reproducible from submitted data and code, leading the journal to require review by contracted grad students (Glandon 2010)." (Nylan 2015)

4. "The project on which Glandon reports covered replication of 39 articles published between 2006 and 2008 in the AER; about 80% of the submissions satisfied the spirit of the data availability policy." (Karolyi 2011)

5. "...Glandon (2010), who replicates a selected sample of nine papers only from the *AER*." (Chang and Li, 2018)

Yet, the Appendix supports none of the above characterizations. In fact, the Audit did not have a single successful replication! Economists familiar with the replication literature were more circumspect in considering the Editor's equivocations. Dewald and Anderson (2014, p. 208) wrote, "In 2004, *AER* editor Ben Bernanke adopted a mandatory data and program code archive. Compliance has been excellent, *at least according to the annual reports of the editor*." [emphasis added]

On the other hand, a critical reader might focus on the use the word "intent". He might then consider the difference between: (1) The vast majority of authors complied with the intent of the policy and (2) the vast majority of authors complied with the policy. A reader who merely glanced at said Appendix might not have reason to question the assertion. Someone who read the appendix carefully, especially someone who knows something about data-code archives, would realize that the two sentences are orthogonal.

The Appendix in question is called "Report on the *American Economic Review* Data Availability Compliance Project" (hereafter referred to as "the Appendix") and it offers as its primary piece of "evidence" its Table 1, reproduced below as our Table 1.

The reader's attention is directed to the end of the last line: 79%. This is the only number that might bear on the editor's claim that the "vast majority" of articles are compliant. Yet, even if the number is correct, it still admits that one out of five articles is not replicable, which is far from satisfactory. Curiously missing from Table 1 is the number of articles that were successfully replicated. After all, if the archive is functioning properly, then many articles should be replicated.

**Table 1:** Data and Code Submissions by Year of Publication

| | 2006 | 2007 | Mar-08 | Total |
|---|---|---|---|---|
| Articles published | 98 | 100 | 22 | 220 |
| Articles subject to data policy | 61 | 63 | 11 | 135 |
| Articles investigated | 13 | 24 | 2 | 39 |
| With "readme" file | 12 | 23 | 1 | 36 |
| | (90%) | (96%) | (50%) | (92%) |
| With complete submission | 7 | 12 | 1 | 20 |
| | (54%) | (50%) | (50%) | (51%) |
| With proprietary data instructions | 1 | 10 | 0 | 11 |
| | (8%) | (24%) | (0%) | (28%) |
| Articles investigated believed replicable | 8 | 22 | 1 | 31 |
| without contacting the author(s) | (62%) | (92%) | (50%) | (79%) |

However, the 79% number is not correct, as is apparent from even a causal perusal of the table. The reader's attention is now directed to the beginning of the last line; note the words: "*believed* replicable"; not "replicable", which is the ostensible purpose for auditing a data-code archive, but "believed replicable". The 79% figured is arrived at, not by dividing the number of articles investigated (39) by the number of articles *actually* replicated, but by the number of articles *believed* to be replicable, which is 31. It is not unreasonable to suggest that the difference between an article actually replicated and an article believed to be replicable is the basis for the Editor's use of the word "intent" in his report, which would make the word "intent" a "weasel word", as Hayek would call it. Of course, the Editor's assertion about the "intent" of the policy makes sense only if the "intent" of the policy is only to have the authors submit something to the archive. If the intent is for the article to be reproducible, then obviously the policy is a failure.

The reader's attention is now directed to the second line of the excerpt from the 2011 Report, in particular the phrase "a random set of papers". If the audit had been performed on a random set of papers, then perhaps its conclusions to be extrapolated to the population from which it was drawn – the entire archive. However, anyone who gives a cursory read to the Appendix can see that the sample is anything but random. It is, in fact, a convenience sample. Therefore, conclusions from the sample cannot be extrapolated to the population.

So far we've just looked at the Appendix's Table 1. Actually reading the Appendix reveals that the state of the archive was much, much worse than the Editor would have had us believe. Of the nine articles that actually were subjected to a reproduction attempt, each was scored on a scale given in Table 2. The Appendix does not define "perfect". Does it mean that a numerical result is reproduced exactly, or only to within rounding error? What is a "minor discrepancy"? No one knows. The author of the Appendix (p. 696) fell prey to the confusion that surrounds the concept of reproducibility, and stumpled into the same trap that ensnared Chang and Li: "The replicated results generally agreed with the published results." This statement is specious. A replicated result *necessarily agrees* with the published result. If a computed result only *generally agrees* with the published result then the published result is *not* reproducible. Since the same data and code did not

give the published results, the Appendix proved that the nine papers examined are *not* reproducible. This is captured in Table 3, which accurately depicts the reproducibility status of the nine articles.

**Table 2:** Appendix's scoring system for replicability

| score | Appendix's classification | articles |
|---|---|---|
| 5 | perfect | 0 |
| 4 | practically perfect | 5 |
| 3 | minor discrepancies | 4 |
| 2 | potentially serious discrepancies | 0 |
| 1 | serious discrepancies | 0 |

**Table 3:** The True State of the Appendix's Nine Replicated Articles

| score | "narrow replication" | articles |
|---|---|---|
| 1 | reproducible | 0 |
| 0 | not reproducible | 9 |

## 3 What the AER should have done

The fourth and fifth sections of the this paper are supposed to be "A replication plan that applies these principles to the "candidate" article" and "A discussion of how to interpret the results of the replication", but in the present case both of these sections have been mooted. A draft of this paper described how to doublecheck the Appendix in an attempt to reproduce its Table 1, but a referee correctly described this as pointless. Even if Table 1 could be reproduced it tells us nothing: knowing precisely how nine non-randomly selected articles were classified tells us nothing.

The important question is this: How should the audit of the *AER* archive been conducted?

My co-authors and I have used archives from several journals to attempt to reproduce hundreds of articles. According to Table 1, in 2006 the *AER* published 61 articles subject to the archive policy and 63 the following year. *Prima facie*, there is no good reason that the Editor should have been satisfied with anything less than attempting to reproduce all the articles published in an entire year, nor should he have drawn any conclusions based on a less-ambitious sampling scheme. Additionally, the criterion should be binary: either an article is reproducible in its entirety or it is not; there should be no degrees of reproducibility admitting that nonreproducible results are acceptable.

Consider auditing the archive for a specific year. For each article, obtain the requisite software, and see if the data and code in the archive will reproduce all the published results. Period. If not, article should be flagged in the archive as not reproducible. Interested researchers can determine whether that extent of irreproducibility affects the article's conclusions. It is not the job of the journal to adjudicate whether a particular discrepancy or set of discrepancies affects an article's conclusion. If someone thinks that the lack of reproducibility matters, that person will write an article and try to get it published.

At this stage, *every* number in the article must be reproduced. If the article says, "1.234" and the replicator finds "1.235", that is a failure to reproduce (subject to the previously-mentioned exceptions for minor rounding error due to operating systems, etc.). This can be noted in a supplementary file that the editors can append to the archive entry for that article. Whether this difference really matters can be determined by interested researchers; Let the interested researchers write an article if the difference in the third decimal really matters. Persons who conduct Monte Carlo studies should run their programs several times to make sure their results are stable to the number of reported digits.

As to documentating primary sources (*e.g.*, the "Hoxby streams problem"), that can be left to interested researchers. Again, if an interested researcher finds that he cannot get from primary resources to the archived dataset, he can write an article and try to get it published. As to whether the software matters, an interested researcher can port the code to another package and see if results change. If they change appreciably, he can write an article and try to get it published.

After the audit is completed, the editor can assess the results and determine whether performance is satisfactory and, if not, implement changes to the rules that are likely to lead to better performance. A follow-up audit can reveal whether the rules changes worked. The point is not so much to insist on 100% reproducibility – we are all human and no system is perfect – but we should know how well the system is working and, if it's not working well enough, take steps to improve its performance. At present, no one knows how well the archive is working, and that it the purpose of an audit.

What of articles that use proprietary data? According to 2106 report of the editor (Goldberg, 2016; Table 7), some 30% of articles published in the *AER* in 2015 used proprietary data and were granted exemptions from the archiving policy. If there exists some mechanism whereby other researchers can gain access to the data and verify the published results, then this can be done by interested researchers, not the audit team. On the other hand, if there is no mechanism whereby the published results can be verified by third parties, then this "research" falls outside the scope of the scientific method, and should not be published in a leading journal. Let it be published in a lesser journal, and reserve publication in the leading journal for someone who can answer the same question in a way that can be reproduced by others!

Some journals, especially in the discipline of political science, engage in "prereplication", *i.e.*, the journal itself reproduces all the results in the article before publishing it. The recent advertisement by the *AER* for a data editor suggests that the *AER* is considering this model; the advertisement states that the duties of the data editor will be:

- Design, in collaboration with the AEA journal editors and the AEA Executive Committee, a comprehensive strategy for archiving and promoting the curation of data and code that guarantees to the extent possible reproducibility of research and addresses the challenges above.
- Determine the staff and computing resources necessary to implement the strategy.
- Oversee the hiring of staff and implementation of the new policy.

It would be a horrible mistake for the *AER* to engage in prereplication. There are two primary reasons that this would be a mistake.

First, placing the onus on the journal relieves the researcher of the responsibility to produce data and code that reproduce the published results, and is an incentive to sloppiness; our culture should not encourage this.

Second, and much more importantly, only a few journals have the resources to do this. It would be wrong for the flagship journal to do as a first resort that which most journals cannot do at all. The *AER* should be showing the not-so-well-endowed journals that it is possible to publish reproducible research without a large capital investment in software and people to do the checking. In particular, simply by being willing to publish failed reproduction attempts, a journal can enforce the discipline necessary to ensure that authors submit data and code that reproduce the published results. To date, no journal in the profession does this. Only if the *AER* attempts this and it does not work should the *AER* resort to doing the replications itself.

## 4 Conclusions

I confess that I only got into replication by accident. By the turn of the century, Vinod and I had published articles showing that software was inaccurate, but we had great trouble getting people to believe that the inaccuracies (if they existed!) really mattered. Our great idea was to take articles from the *AER*, get the data and code from the authors, and port the code to other packages. We expected that different packages would give different answers. We did not expect that (1) authors wouldn't honor the replication policy or (2) that something published in the *AER* would not be reproducible. Sure, we had read Dewald, Thursby and Anderson, but that was 15 years prior, surely people are doing replicable research now plus, that was the *JMCB* and this is the *AER*. Were we surprised!

There is still too much naivete about replication in the economics profession. True, there are more journals than ever with archives. But let's not kid ourselves that the mere existence of archives ensures the reproducibility of the published results. Much more needs to be done, and the editors of the journals need to take the lead. Each journal with an archive should conduct a serious audit to determine whether it is publishing reproducible research and, if not, effect changes.

The problem with the "audit" commissioned by the AER is that it did not use the correct standard for reproducibility. It is not sufficient that, using the code and data provided by the authors, that one get something that is merely consistent "with the spirit" of the original article. If the authors provide the data and code, there is no reason why one should not be able to exactly reproduce the published results. Results that are not identical to the published article constitute a "failed reproduction". Had the Appendix used this standard, and/or had the AER employed this standard in interpreting the results from the Glandon report, the true, sorry state of the AER archive would have been revealed earlier. This highlights the importance of having a correct standard of replication.

The Appendix and its endorsement by the *AER* Editor set back the cause of reproducibility in economics by several years. Kudos to the current editors of the AER, Duflo and Hoynes, for publicly admitting that the archive isn't working – this implies that they are serious about publishing reproducible research in the pages of the AER, and this bodes well for the future of economic research in all journals, as long as the *AER* does not decide to engage in "prereplication".

Just as many journals followed the lead of the AER in creating an archive, so will these journals follow the lead of the AER in ensuring that its published results are reproducible.

**Acknowledgements** I thank the editor and the referees for their comments, which improved the paper.

## References


Bernanke, B. S. (2004). Editorial statement. *American Economic Review*, 94(1): 404. http://www.climateaudit.info/pdf/aereditorial.pdf

Baker, M. (2016). Muddled meanings hamper efforts to fix reproducibility crisis. *Nature News*. https://www.nature.com/news/muddled-meanings-hamper-efforts-to-fix-reproducibility-crisis-1.20076

Breure, L. and Hoogerwerf, M. (2011). Data availability policies: ideal and practice. Working paper. Department of Informatics, University of Utrecht. https://xposre.nl/cliodap/DAP_Ideal+Practice_1-1.pdf

Camerer, C. F., Anna, D., Eskil, F., Ho, T. H., Jürgen, H., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T.,Razen, M. and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*. http://science.sciencemag.org/content/early/2016/03/02/science.aaf0918

Chang, A. C. and Li, P. (2018). Is economics research replicable? Sixty published papers from thirteen journals say "often not". *Critical Finance Review*, 7. https://www.nowpublishers.com/article/Details/CFR-0053

Christensen, G. S. and Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*. Forthcoming.

Clemens, M. A.(2015). The meaning of failed replications: a review and proposal. *Journal of Economic Surveys*, 31(1): 326–342. https://doi.org/10.1111/joes.12139

Dewald, W. G., Thursby, J. G. and Anderson, R.G. (1986). Replication in empirical economics: the Journal of Money, Credit and Banking Project. *The American Economic Review*, 76(4): 587–603. https://www.jstor.org/stable/1806061?seq=1#page_scan_tab_contents

Dewald, W. G., Anderson, R. G. (2014). Replication and reflection: A decade at the *Jouranl of Money, Credit and Banking*. In Szenberg, M. and Ramrattan, L. *Secrets of economics editors*. The MIT Press.

Donohue, J. and Levitt, S. (2001). The impact of legalized abortion on crime. *Quarterly Journal of Economics*, 116: 379–420. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=174508.

Easterbrook, S. M. (2014). Open code for open science? *Nature Geoscience*, 7: 779–781. https://www.nature.com/articles/ngeo2283

Foote, C. L. and Goetz, C. F. (2008). The impact of legalized abortion on crime: comment. *The Quarterly Journal of Economics*, 123(1): 407–423. https://www.jstor.org/stable/25098902?seq=1#page_scan_tab_contents

Glandon, P. (2010). Appendix to the report of the editor: report on the American Economic Review data availability compliance project. Vanderbilt University. https://digital.kenyon.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1011&context=economics_publications

Goldberg, P. K. (2016). Report of the Editor: American Economic Review. *The American Economic Review*, 106(5): 700–712. https://www.aeaweb.org/articles?id=10.1257/aer.106.5.700

Hoxby, C. M. (2000). Does competition among public schools benefit students and taxpayers?. *American Economic Review*, 90(5): 1209–1238. https://www.jstor.org/stable/2677848?origin=JSTOR-pdf&seq=1#page_scan_tab_contents

Karolyi, G. A. (2011). The ultimate irrelevance proposition in finance? *The Financial Review*, 46(4): 485–512. https://doi.org/10.1111/j.1540-6288.2011.00309.x

Levitt, S. D. (1997). Using electoral cycles in police hiring to estimate the effect of police on crime. *The American Economic Review*, 87(3): 270–290. https://www.jstor.org/stable/2951346?seq=1#page_scan_tab_contents

Levitt, S. D. (2002). Using electoral cycles in police hiring to estimate the effect of police on crime: reply. *The American Economic Review*, 92(4): 1244–1250. http://pricetheory.uchicago.edu/levitt/Papers/LevittUsingElectoralCycles2002.pdf

McCrary, J. (2002). Using electoral cycles in police hiring to estimate the effect of police on crime: comment. *The American Economic Review*, 92(4): 1236–1243. https://www.jstor.org/stable/3083311?seq=1#page_scan_tab_contents

McCullough, B. D. and Vinod, H. D. (1999). The numerical reliability of econometric software. *Journal of Economic Literature*, 37(2): 633–665. https://ideas.repec.org/a/aea/jeclit/v37y1999i2p633-665.html

McCullough, B. D. and Vinod, H. D. (2003). Verifying the solution from a nonlinear solver: A case study. *American Economic Review*, 93(3): 873–892. https://www.aeaweb.org/articles?id=10.1257/000282803322157133

McCullough, B. D., McGeary, K. A. and Harrison, T. D. (2008). Do economics journal archives promote replicable research? *The Canadian Journal of Economics*, 41(4): 1406–1420. https://www.jstor.org/stable/25478330?seq=1#page_scan_tab_contents

Moffitt, R. A. (2011). Report of the editor: American Economic Review (with Appendix by Philip J. Glandon). *American Economic Review*, 101(3): 684–693. https://www.aeaweb.org/articles?id=10.1257/aer.101.3.684

Nylan, B. (2015). Increasing the credibility of political sciences research: A proposal for journal reforms. *Political Science & Politics*, 48(S1): 78–83. https://doi.org/10.1017/S1049096515000463

Rothstein, J. (2007). Does competition among public schools benefit students and taxpayers? A comment on Hoxby (2000). *The American Economic Review*, 97(5):2026–2037. https://www.jstor.org/stable/30034599?seq=1#page_scan_tab_contents