Comment on "Quis custodiet ipsos custodes?: Despite evidence to the contrary, the American Economic Review concluded that all was well with its archive," by B.D. McCullough, Drexel University

Comment by: Andrew C. Chang, Board of Governors of the Federal Reserve System, a.christopher.chang@gmail.com

The views and opinions expressed here are mine and are not necessarily those of the Board of Governors of the Federal Reserve System

Submitted papers for the special issue on "The practice of replication" for *Economics: The Open-Access, Open-Assessment E-Journal* need to meet four requirements, of which this article covers all four. I will briefly outline how McCullough meets the necessary requirements for the special issue and follow with a short discussion. The necessary requirements for the special issue are below in *italics*.

Necessary requirements:

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

McCullough defines the procedure to replicate as: "For computational research, it is very easy. Put the data and code in the same folder, and run the code. (pg. 4) 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. (pg. 4)" with a definition of "success" as: "linear procedures with moderately-sized datasets, there should be ten digit agreement [for successful replication], for nonlinear procedures there may be as few as four or five digits of agreement [for successful replication]", (pg. 3-4)

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

The candidate paper is Glandon (2011), which was an assessment of the American Economic Review's (AER) data availability policy. The rationale for selecting Glandon (2011) is (at least in my reading of McCullough): "Glandon's report and Moffitt's uncritical acceptance of it set the goal of reproducible economic research back by several years." (pg. 5), with a similar conclusion in the abstract.

3) A replication plan that applies these principles to the "candidate" article.

See Section 3, in particular asking Glandon for: (1) "A list of the nine papers he [Glandon] analyzed, and the procedure by which the nine were selected.", (2) "A precise description of the 1-5 rating system, so that two independent researchers would apply the same score to the same paper", (3) "A precise description of how one can "believe" a paper to be reproducible", and (4) "A justification for considering a "belief" that a paper is reproducible to be more important than a paper actually being reproducible" (pg. 11-12).

4) A discussion of how to interpret the results of the replication (e.g., how does one know when the replication study "replicates" the original study).

For the nine (unnamed) AER replication attempts Glandon (2011) doesn't provide actual estimates, but instead classifies attempted replications on a 1-5 subjective scale. If a would-be replicator were able to match all of the hypothetically provided estimates that underlie the 1-5 classification, then naturally Glandon's article would be successfully replicated. However, if a would-be replicator obtained different underlying estimates than Glandon (2011), presumably the would-be replicator could still come to the same assignments on the 1-5 scale as Glandon, but such assignments would be subjective and would most likely result in discrepancies between independent researchers, a point that McCullough notes: "the criteria for his [Glandon's] rating system is necessarily so subjective that one could not obtain his results exactly." (pg. 12), or at least it would not be possible without an exact match. Glandon's Table 1, which includes summary statistics about the AER's replication archive and no actual replication attempts, would also require either snapshots of the AER's archive at the time that Glandon did the study, or at least a list of papers that fall into each cell of the table.

Discussion:

I believe that the main points missing from McCullough's procedure to replicate Glandon (2011) (McCullough Section 3) is *how* you would acquire the necessary information from Glandon (list of papers, procedure through which they were selected, description of the rating system, description of "beliefs"), and to *what* lengths you should go about getting that information. As I discuss in my submission for this issue, Chang (2017), in principle you could spend an infinite amount of time and effort trying to obtain such necessary information, so it is necessary to use some kind of stopping rule. McCullough's last procedure point, "A justification for considering a "belief" that a paper is reproducible to be more important than a paper actually being reproducible", is unnecessary for replication.

Setting aside Glandon's summary statistics in Table 1, I think that attempting to replicate the nine attempted (unnamed) replications from Glandon's article would not lead to any new insights about data availability policies in general, or of the AER in particular, because the sample design of Glandon is ill equipped to do so. As Glandon (2011) notes (emphasis mine): "A total of 39 articles (29 percent of the relevant empirical study population of 135) were selected *primarily based on the students' interest in the topic*." (pg 696). The purpose of Glandon's article was to evaluate the AER's data availability policy; something that cannot reasonably be done by this sample design.

Any replication attempt of Glandon would, at most, only yield insights about the nine (unnamed) articles that Glandon attempted to replicate. Even if Glandon were to provide you with the article list and estimates from the nine attempted replications, and you were to confirm the estimates from the replications, the most you could say is that you had confirmed that a nonrandom subset of estimates from a subjectively selected sample of articles from the AER could be replicated, and that another nonrandom subset of estimates from those same articles could not be replicated.

For some applications, namely verifying work for the archival record, McCullough's binary classification scheme for "successful" replication seems appropriate. But for many other applications a binary classification is too crude, and even verifying work for the archival record only requires precision to the reported number of significant digits. At the individual estimate level, most variables are continuous (and even discrete outcomes can be thought of as the outcome of a latent continuous process). So obtaining an estimate of a continuous variable that is a few ε from the published number could be thought of as just statistical noise and should not immediately classify a replication as a "failure".

If you have the exact econometric software, operating system, hardware, code, and same vintage of data as the authors, then you should be able to run their programs by pressing the GREEN GO BUTTON and your output should be within machine precision of the original authors (i.e. computational confirmation given the same circumstances as the authors). But most of the time at least one of those elements is missing from your toolkit when attempting a replication. You are rarely, if ever, in the same circumstances as the authors. As such, deviations of your replication estimates from the authors' estimates could come from anywhere,[1] so most replication attempts that are at least somewhat constrained by replicator effort have to allow for some degree of deviation of replication estimates from published estimates and still qualify as a "success".

The question is how much tolerance is allowable for "success"? Outside of the case where you have the exact econometric software, operating system, hardware, code, and same vintage of data as the authors, I don't think there is a clear answer. Allowable tolerance depends on the context, which could include a researcher-specific loss function.

References:

Anderson, Richard G. (2017), "Should You Choose to Do So... A Replication Paradigm," Economics Discussion Papers, No. 2017-079, Kiel Institute for the World Economy.

Chang, Andrew C. (2017), "A Replication Recipe: List Your Ingredients Before You Start Cooking," Economics Discussion Papers, No. 2017-074, Kiel Institute for the World Economy.

Glandon, Philip 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

[1] A point also shared by Anderson (2017), another submission to this issue.