McCullough, Bruce D.

Working Paper

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

Economics Discussion Papers, No. 2017-78

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

Suggested Citation: McCullough, Bruce D. (2017) : Quis custodiet ipsos custodes? Despite evidence to the contrary, the American Economic Review concluded that all was well with its archive, Economics Discussion Papers, No. 2017-78, Kiel Institute for the World Economy (IfW), Kiel

This Version is available at:
https://hdl.handle.net/10419/169137

Economics
The Open-Access, Open-Assessment E-Journal

Discussion Paper
No. 2017-78 | September 28, 2017 | http://www.economics-ejournal.org/economics/discussionpapers/2017-78

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 made no changes in the archive rules. This was based on an audit of the archive that the editor has commissioned. The audit was performed by a graduate student who apparently had no experience with archives, and the audit concluded that all was largely well with the archive. 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.

(Published in Special Issue The practice of replication)

JEL B40
Keywords Replication; reproducible research

Authors
B. D. McCullough, ✉ Department of Decision Sciences & MIS, Drexel University, Philadelphia, PA, USA, bdmccullough@drexel.edu

Citation B. D. McCullough (2017). Quis custodiet ipsos custodes?: Despite evidence to the contrary, the American Economic Review concluded that all was well with its archive. Economics Discussion Papers, No 2017-78, Kiel Institute for the World Economy. http://www.economics-ejournal.org/economics/discussionpapers/2017-78

Received September 2, 2017 **Accepted as Economics Discussion Paper** September 12, 2017
Published September 28, 2017

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 purposes of maintaining an archive, the concept of "narrow replication" (a.k.a "reproducibility") suffices: The data and code in the archive reproduce the published results.

As an example of this need for clear thinking and precise definitions when talking about replication, consider the recent paper by Chang and Li (2017) that has received much attention. In their abstract they write (p. 2):

> "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."

Notice the phrase "key qualitative results". What Chang and Li think they are doing is *confirming* key qualitative results, *not* replicating them. To see this more clearly, consider the following quote from their paper (p. 7):

> "We define a successful replication as when... [f]or 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)."

However, they are not even confirming the published results because they are using the same data and code as the original author. On its face, Chang and Li's criterion for replicability is utter nonsense. Think about it: Using the same data and same code, 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 and same code to get 1.0 when the original paper shows 2.0, then Chang and Li *prove* that the paper is not reproducible because the authors could not provide data and code that reproduce the published result. 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.

This 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 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 consitute reproducibility, but in the end the decision is binary. (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.)

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 replicable. In general, no explanation of the extent of the non-replicability 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 2000:1222), but she provided no further details. What was this "certain width"? Was it 1mm or 5mm in width? This lack of detail all but insured that no one else would be able to reproduce her work. As Rothstein wrote (Rothstein 2007:2033-34): "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.

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

Recently the *American Economic Review* posted an advertisement for a "Data Editor". The advertisement cited problems with its archive, the most notable being: "Posted code often does not run or does not actually replicate the results." This contradicts the assertions made years ago by then-editor of the *AER*, Robert Moffitt. Moffitt had commissioned a graduate student, Philip Glandon, to conduct an audit of the *AER archive*. Glandon's (2011) report, which is the subject of the present paper, was the basis on which Moffitt (2011:687) assured readers of the *AER* that "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." In response to Glandon's report, Moffitt made no changes to the *AER* archive policy, implying that all was well with the archive.

The fact of the matter is that all was not well with the archive, and Glandon's report failed to make this clear; Moffitt did not make any changes to the *AER* archive policy and the *AER* continued to publish nonreproducible research. If Glandon had written his report competently, the *AER* would have been forced to take action to fix the archive years ago. Glandon's report and Moffitt's uncritical acceptance of it set the goal of reproducible economic research back by several years.

The editor's job is to ensure that the papers published in his journal can be relied upon. The editor *should have known* that the archive was failing. Many authors have cited Glandon's appendix as a basis for asserting that the *AER* archive is fulfilling its function of ensuring the published results are reproducible. Here are some examples:

1. "[Glandon] replicated a selected sample of nine papers only from the AER." (Chang and Li 2017)

2. "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)

3. "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)

4. "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)

5. "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)

Yet, Glandon's (2010) report supports none of the above characterizations. In fact, Glandon's report did not have a single successful replication!

Dewald, Thursby and Anderson (1986) called into question the replicability 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 replicable articles. In response, then-editor Bernanke (2003) 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, Robert Moffit, introduced the following boilerplate that is found almost verbatim in the Editor's Reports through the end of his term (even through that of his successor, Goldberg):

> 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 Moffitt 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. 686-7):

> In the summer of 2008, the AER conducted an exercise to check the submitted files of a random set of papers 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 AERs data posting policy.

Naturally, if there had been anything seriously wrong with the archive, the Editor would have taken steps to address the problem. This lack of action implies that no action was needed, that all was more or less well with the archive.

To the casual reader, Moffitt'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 Moffit's lead and reported that all was well with the *AER* archive.

Economists familiar with replication literature were more circumspect in considering Moffit's blandishments with respect to the archive. Dewald and Anderson (2014:208) 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 wonder why Moffit chose to 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 Moffit's assertion. Someone who read the appendix carefully, especially someone who knows something about data-code archives, would discover that the two sentences are orthogonal.

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

	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%)

Table 1: Glandon's Table I: Data and Code Submissions by Year of Publication

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.

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 Moffitt's use of the word "intent" in his report, which would make the word "intent" a "weasel word", as Hayek would call it.

(Even if a belief in an article's replicability, rather than an article's actual replicability, is the relevant criterion, any conclusions drawn about the sample cannot be extrapolated to the archive in general: the 39 articles were not randomly sampled but instead constitute a convenience sample.)

Pointedly 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.

So far we've just looked at Glandon's Table 1. Actually reading the Appendix reveals that the state of the archive was much, much worse than Moffitt would have had us believe.

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

To replicate Glandon's report, we need to know some details of how it was done. According to Glandon (p. 696),

> "Narrow, or pure, replication seeks to precisely reproduce the tables and charts using the procedures described in an empirical article. The purpose of narrow replication is to confirm the accuracy of published results given the data and analytical procedures that the authors claim to have used. The *AER* Project was aimed exclusively at at narrow replication."

Clearly, "narrow replication" seeks to reproduce *the* tables and charts, *all* of them, not *some* of them. The purpose is to confirm the accuracy of published *results*, *all* of them, not just *some* of the published results. If narrow replication seeks only to reproduce some of the results, Glandon would have said so, and he would have specified which results qualify as meriting reproduction.

On the other hand, Glandon did write (p. 696) that the purpose of the audit was, in part, "to assess the extent to which authors complied with the AER's data submission policy." If the policy only required submission – regardless of whether or not the submitted data and code would reproduce the published results – then the 79% figure that Glandon and Moffit referred to as proving the success of the archive is justified. To the contrary, if the policy requires submitting data and code that reproduce the published results, then the archive is a failure. Yet, the *AER Data Availability Policy* explicitly states:

> Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to publication, the data, programs, and other details of the computations sufficient to permit replication.

This raises an important question: When the policy clearly states that replication is the goal, why did Glandon (and Moffitt) accept mere submission of files in lieu of files that replicate?

Rather than simply make a binary decision, "Do the data and code reproduce the published results?", the students graded each of the nine articles on Glandon's scale. With respect to a score of 3, Glandon adds, "While these discrepancies could not be reconciled, they were immaterial to the conclusions of the paper and may have been the result of differences in software versions used." "Immaterial" is in the eye of the beholder, thus rendering Glandon's criteria purely subjective. As a consequence, Glandon's results may be repeatable but tehy will not reproducible, just as was the case for Hoxby's "streams" variable. How many times has a minor result from one article been cited in another, when that minor result was ancillary to the main point of the article? According to Glandon, those minor results need not be dependable. Science is a building process, and all its blocks must be credible. It is not enough to say that the non-replicable parts of an article are not germane to the main conclusion of the article, they must also be

not germane to any future article; and this, Glandon cannot guarantee. The results of Glandon's system applied to the nine articles is given in Table 2.

score	Glandon's criteria	articles
5	perfect	0
4	practically perfect	5
3	minor discrepancies	4
2	potentially serious discrepancies	
1	serious discrepancies	0

Table 2: Glandon's scoring system for replicability

Notice that the investigators awarded a score of five to no article. None of the articles had data and code that reproduced all the results. This is complete failure. The investigating team could not find a single article for which the data and code could reproduce all the results of the article, and from this, Glandon reached the conclusion that the archive was functioning well! Moffitt repeated this misbegotten conclusion in his annual report, and the entire profession still thinks the archive is functioning well when in fact all available evidence suggests that it is failing.

If Glandon had desired the reader to reach a reasonable conclusion, he would have said, "None of the articles was reproducible but some of them were partially reproducible." Instead, Glandon (p. 696) fell prey to the confusion that surrounds the concept of reproducibility, and fell into the same trap that ensnared Chang and Li: "The replicated results generally agreed with the published results." This is more nonsense. 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. In the history of replication, no one has ever defined a published paper to be reproducible if some portion of its results could not be reproduced. Since the same data and code did not give the published results, Glandon proved that the nine papers he examined are not reproducible. This is captured in Table 3, which accurately depicts the reproducibility status of the nine articles he examined.

If I wanted to replicate Table 1, I should ask Glandon for the following:

1. A list of the nine papers he analyzed, and the procedure by which the nine were selected.

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

Table 3: The True State of Glandon's Nine Articles

2. A precise description of the 1-5 rating system, so that two independent researchers would apply the same score to the same paper. In particular, Glandon needs to give rules for classifying a discrepancy between a reproduced result and a published result as "immaterial to the conclusions of the paper".

3. A precise description of how one can "believe" a paper to be reproducible merely by examining the archive, so that two independent researchers would characterize the same paper the same way.

4. A justification for considering a "belief" that a paper is reproducible to be more important than a paper actually being reproducible (why is the last line of Table 1 included?).

4 A discussion of how to interpret the results of the replication

If Glandon were able to provide the requested information (which should have been in his paper, to ensure that it was reproducible) it would be a straightforward job to analyze the nine papers and apply his definitions. However, the criteria for his rating system is necessarily so subjective that one could not obtain his results exactly. Nonetheless, were one able to obtain the same numbers that he did, then I could assert that his paper was reproducible, *but it would still not be correct!* Several important questions have been raised:

1. When the *AER Data Availiability Policy* explictly states that submitted files support replication, why does Glandon emphasize merely submitting files rather than submitting files that replicate?

2. Why did Glandon count articles that were "believed to be replicable" rather than articles that actually were replicable?

3. Why did Moffitt accept the above two actions by Glandon? Why did Moffitt simply not ask, "What proportion of papers are reproducible?"

4. Why did Moffitt appoint a graduate student with no replication experience to audit the *AER* archive?

My co-authors and I have used archives from several journals to attempt to reproduce hundreds of articles. According to Glandon's Table I, 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 Moffitt 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.

5 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 ago, surely people are doing replicable research now plus, that was the *JMCB* and this is the *AER*. I was once as naive as Moffitt. The difference, though, was that in 2008 there was much evidence that none of the archives in economics was functioning well. Moffitt should have been aware of this. Glandon was but a graduate student at the time he wrote the report. No one without a experience with archives should have been commissioned this task. Ultimately, the responsibility for this failed report lies with the person who commissioned the report, not the person who actually wrote it (and was too young to realize he had no business doing anything like this).

Had Moffitt actually commissioned a thorough audit of the archive, he would have discovered that it was not functioning and he would have had to fix it. As such, Glandon's report 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. 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.

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.

References

Anderson, Richard, William H. Greene, B. D. McCullough and H. D. Vinod (2008), "The Role of Data/Code Archives in the Future of Economic Research" *Journal of Economic Methodology* **15**(1), 99-119

Bernanke, Ben S. (2004), "Editorial Statement," *American Economic Review* **94**(1), p. 404

Baker, Monica (2016), "Muddled meanings hamper efforts to fix reproducibility crisis," *Nature News*, doi:10.1038/nature.2016.20076

Breure, Leen and Maarten Hoogerwerf (2011), "Data Availability Policies: Ideal and Practice," working paper Department of Informatics, University of Utrecht

Camerer,Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu (2016), "Evaluating replicability of laboratory experiments in economics," *Science* **351**(6820), 1433-1437

Chang, Andrew C. and Phillip Li (2017), "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not',"

Critical Finance Review (to appear)

Christensen, Garret and Edward Miguel (2018), "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, forthcoming

Clemens, Michael A. (2015), "The Meaning of Failed Replications: A Review and Proposal," *Journal of Economic Surveys* **31**(1), 326-342

Dewald, William G., Jerry G. Thursby and Richard G. Anderson (1986), "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project," *The American Economic Review* **76**(4), 587-603

Dewald, William G. and Richard G. Anderson (2014), "Replication and Reflection: A Decade at the *Jouranl of Money, Credit and Banking*," Chapter 13 in *Secrets of Economics Editors*, Michael Szenberg and Lall Ramrattan, eds., Cambridge, MA: MIT Press

Donohue, John and Steven Levitt (2001), "The Impact of Legalized Abortion on Crime," *Quarterly Journal of Economics* **116**, 379-420

Foote, Christopher and Christopher Goetz (2008), "The Impact of Legalized Abortion on Crime: Comment," *Quarterly Journal of Economics* **123**(1), 407-423

Easterbrook, Steve M. (2014), "Open Code for Open Science?" *Nature Geoscience* **7**, 779-781

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

Hoxby, Caroline (2000), "Does competition among public schools benefit students and taxpayers?" *American Economic Review* **90**(5), 1209-1238

Karolyi, G. Andrews (2011), "The Ultimate Irrelevance Proposition in Finance?" *The Financial Review* **46**(4), 485-512

McCullough, B. D. and H. D. Vinod (2003), "Verifying the Solution from a Nonlinear Solver: A Case Study," *American Economic Review* **93**(3), 873-892

McCullough, B. D., Kerry Anne McGeary, and Teresa D. Harrison (2006), "Lessons from The Journal of Money, Credit and Banking Archive," *Journal of Money Credit and Banking* **38**(4), 10931107

McCullough, B. D., and H. D. Vinod (1999), "The Numerical Reliability of Econometric Software," *Journal of Economic Literature* **37**(2), 633-665

McCullough, B. D., and H. D. Vinod (2003), "Verifying the Solution from a Nonlinear Solver: A Case Study," *American Economic Review* **93**(3), 873-892

Moffitt, Robert A. (2011), "Report of the Editor," *American Economic Review* **101**(3), 684-689

Nylan, Brendan (2015), "Increasing the Credibility of Political Sciences Research: A Proposal for Journal Reforms," *PS: Political Science & Politics* **48**(S1), 78-83