

Report on the American Economic Review Data Availability Compliance Project

Philip Glandon*

Vanderbilt University

November 2010

*I thank John Siegfried and Robert Moffitt for helpful comments on previous drafts of this report, and Noel Whitehurst of Vanderbilt University for outstanding research assistance. The findings of this report were presented at the 2009 Joint Statistical Meetings in Washington, D.C.

Replicating empirical results is an underemphasized, yet essential element of scientific progress. The rewards for replicating are often low (relative to original contributions) while the costs can be substantial. A researcher who sets off to do a serious replication study is likely to find that the task is tedious, more difficult than anticipated, and prone to souring relationships with colleagues. Add these deterrents to the fact that faculty don't get promoted by replicating others' results, and it is no surprise that published replication studies are rare.

Replication and robustness studies have been difficult to conduct because they usually require cooperation from the author(s). Researchers frequently fail to keep documented, well organized, and complete records of data and data processing programs underlying published articles, and are less than enthusiastic when asked to help replicate their work. A highly visible instance of this type of behavior became headline news in November 2009 when the so-called "climategate" controversy broke. Stolen email records from the University of East Anglia revealed that researchers had resisted transparency and gone out of their way to prevent skeptics from analyzing their data. While no evidence has indicated that these researchers were covering up errors, their deliberate efforts to prevent others from independently verifying their results has reflected poorly on an important body of research, and science in general.

Dewald, Thursby, and Anderson's (1986) "Journal of Money, Credit and Banking (JMCB) Project" initiated interest in replication in economics. With financial support from the National Science Foundation (NSF), the *JMCB Data Storage and Evaluation Project* began requesting data and programs used in articles published in the *JMCB* during or after 1980. After attempting to replicate results using these data, Dewald et al. concluded that the economics profession needed to improve the replicability of empirical results by persuading journal editors to secure a copy of data and computer code used by authors of empirical research *prior* to publication, when their leverage is most effective. By depositing

this material in a public data archive, the cost of replicating results declines, thereby increasing the incentive for authors to carefully check data and its manipulation prior to publication. Subsequent work by McCullough et al. (2006) emphasized the importance of including the computer code in data archives. The initial adoption of archiving by journal editors was slow. Anderson and Dewald (1994) reported that the pioneering *JMCB* discontinued requesting data in 1993 and that only two economics journals were routinely requesting data in 1993, and neither secured programming code¹. Since 2000, several economics journals, including *The American Economic Review (AER)*, have adopted systematic data availability policies.² The *AER* began encouraging authors of empirical papers to submit electronic copies of their data to the online archive as early as September 2003. The *Review's* data availability policy was strengthened and became mandatory in March 2005.³

In summer 2008, the *American Economic Review* conducted a Project to evaluate the quality of the data and processing code contained in its online data archive. The objectives of the project were: (1) to assess the extent to which authors complied with the *AER's* data submission policy; (2) to evaluate how easily results could be replicated; and (3) when the materials supplied were complete, to attempt detailed replications without contacting the author(s). To accomplish this task, six economics PhD students from four universities each selected several empirical articles published in the *AER* between March 2006 and March 2008⁴. 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. The students

¹ The *JMCB* apparently resumed collecting data in 1996.

² Others include: *Econometrica*, *The Journal of Applied Econometrics*, *The Journal of Money Credit and Banking*, *The Journal of Political Economy*, *The Review of Economics and Statistics*, and *The Review of Economic Studies*.

³ Some authors of articles published as early as December 2002 voluntarily submitted their data and code to the archive.

⁴ Jose Azur, Princeton University; Yonatan Ben-Shalom, Johns Hopkins University; Laura Gee, University of California San Diego; Philip Glandon, Vanderbilt University; Benjamin Horne, University of California San Diego; Jon Samuels, Johns Hopkins University.

downloaded the data and code from the data archive, evaluated both for completeness, and, when possible, attempted to replicate published results using only the materials found in the archive.

All authors submitted *something* to the data archive. 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). The replicated results generally agreed with the published results. There remains, however, room for improvement both in terms of compliance with the policy and the quality of the materials that authors submit.

I. Background

There are two general types of replication studies. Each serves a unique purpose. The *Journal of Applied Econometrics* refers to the two types of replication studies as “narrow” and “wide”. Hamermesh (2007) calls them pure replication and scientific replication. 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 narrow replication. Wide, or scientific, replications investigate whether results hold under different analytical techniques, other data sources, or small perturbations to the data used. For example, Vinod (2009) provides an algorithm for checking the perturbation sensitivity of estimated coefficients.

Dewald, et al. (1986) provided the impetus for replicating empirical results published in economics journals. They found that “inadvertent errors in published empirical articles are commonplace rather than a rare occurrence” and that replicability improves when authors are required to submit their data and computer code prior to publication. Data submission policies improve accuracy in two ways. First, the process of compiling data, instructions, and code into a user friendly form causes researchers to find and fix more of their own mistakes prior to publication. Second, reducing the cost of replication studies

increases the incentive to minimize errors and discourages fraud. Additionally, submitted data files are an excellent resource for economics students to learn cutting edge empirical and computational techniques.

Simply requiring authors to submit their data prior to publication may not be sufficient to improve accuracy. McCullough et al. (2006) reported that as of 2003, fewer than 10 percent of the submissions to the *JMCB* data archive could be used to replicate the published results. The broken link in the replication process usually lies in the procedures used to transform raw data into estimation data and to perform the statistical analysis, rather than in the data themselves.

Because accuracy of research results is a public good, an otherwise unencumbered market will deliver less accuracy than is socially optimal. Journal editors can do two things to narrow the gap between the market outcome and the social optimum. As suggested in Dewald, et al. (1986), editors can require authors to provide electronic copies of their data and analytical procedures *before* articles are published. They can also encourage authors to undertake serious replication studies by publishing them⁵. For example, *The Journal of Applied Econometrics (JAE)* and *The Indian Journal of Economics and Business (IJEB)* express a specific interest in receiving manuscripts that attempt to replicate previously published work. *The JAE* occasionally includes a special replication section which reports the results of both “narrow” and “wide” replication attempts of articles published in top economics journals. *The IJEB* has published several successful and unsuccessful replication attempts.

The evolution of the *AER*'s data availability policy illustrates the important role that journal editors play in improving replicability. In a 2003 *AER* article, McCullough and Vinod demonstrated the importance of verifying solutions to non-linear estimation obtained by various software packages. The authors wished to apply their analysis to all of the empirical articles from the June 1999 issue of the *AER*.

⁵ McCullough (2006) provides several recommendations for effectively implementing data archives.

Unfortunately they had to abandon this part of the project because only half of the authors would comply with the *AER*'s policy that "Details of computations sufficient to permit replication must be provided." Subsequently, the *AER* announced in the March 2004 issue that henceforth, authors of accepted papers would be required to submit all data and code prior to publication. Beginning with the December 2004 issue, the data availability policy was printed in the front matter of every issue (just below the table of contents) and the editor instructed *AER* staff to begin enforcing the policy as of the March 2005 issue.

II. Brief Project Description

The *AER*'s data availability policy is:

...to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Details of the computations sufficient to permit replication must be provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the requirements above cannot be met.

Author(s) of accepted papers are only required to submit estimation data and not the raw unprocessed data. With this in mind, the 2008-09 *AER* Project asked the following questions of each article selected:

- 1) Was something submitted to the data archive?
- 2) Did a "readme" document describe the contents of the file and provide replication instructions?
- 3) Did the submission include all of the necessary data files or instructions for obtaining them?

Of the articles satisfying item (3) that did not use proprietary data, nine were selected for detailed replication attempts.

Replicators downloaded the zip file associated with each article investigated. The zip files typically contain a “readme” document and several files containing data, computer code, or other documents needed for replicating results. The replicators obtained the files in the same manner that anyone could and did not contact the author(s) for additional information.

The investigators read each article, paying careful attention to the details of the empirical analysis. Next, they reviewed the contents of the archive files, relying on the “readme” document (when available) to help understand what each file contained and how it was used in the original empirical analysis. The replicators then analyzed the submitted files to determine whether the authors complied with the *AER*’s data submission policy. Compliance with the policy requires the author(s) of accepted articles to provide all data and computer code used to generate the empirical results unless the data are proprietary, in which case detailed instructions for obtaining the data should be included.

Finally, if the archive contained everything required to replicate the study, four of the investigators attempted to execute full replications⁶. They compared results and assigned a score of 1 to 5, indicating how closely the replicated results matched the published results. A score of 5 indicates that all figures and tables could be replicated perfectly and a 1 indicates significant and irreconcilable discrepancies between the replication attempt and the published results. The appendix describes the rating system in further detail.

III. Results

A. Policy compliance results

All of the 39 articles selected for the Project had something available in the *AER* data archive, and all but three included a readme file. One of the three that did not have a readme file did contain

⁶ Jose Azur, Laura Gee, Philip Glandon, Jon Samuels

programs that were detailed enough to make replication possible. The other two were incomplete submissions.

The data archives for twenty of the articles examined were complete. This means that after reviewing the contents of the data archive entry, the replicator believed that a full replication was possible using the data and programs provided. An additional 11 (28 percent) of the articles used proprietary data and complied with the data availability policy by including specific instructions for obtaining the data in lieu of the data themselves. Therefore, roughly four out of five papers met the spirit of the *AER* data availability policy, which is to make replication and robustness studies feasible without contacting authors. It was the opinion of the investigators that all but two of the articles (95 percent) could be replicated with little or no help from the author(s). Table I summarizes the results by year.

B. Replication attempt results

Of the 39 articles sampled, 20 contained data archives sufficient to attempt a detailed replication. The investigators attempted to replicate nine of these. A score on a scale of one to five was assigned to summarize how closely the replicated results matched the published results. Five of the articles received a score of four, and four received a score of three. Articles that received a four were replicated almost exactly. The usual reason for not receiving a perfect score (five) was that some details were missing from the instructions for producing a table or figure. The investigators believed that with substantial effort, they could have replicated the results perfectly. A score of 3 meant that there were several small discrepancies between the replicated and the published results. 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.

Table I: Data and code submission results by year of publication

| | 2006 | 2007 | Mar-08 | Total |
|--|-------------|-------------|---------------|--------------|
| Articles Published ⁷ | 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 |
| | (92%) | (96%) | (50%) | (92%) |
| With complete submission ⁸ | 7 | 12 | 1 | 20 |
| | (54%) | (50%) | (50%) | (51%) |
| With proprietary data instructions | 1 | 10 | 0 | 11 |
| | (8%) | (42%) | (0%) | (28%) |
| Articles Investigated believed replicable without contacting the author(s) | 8 | 22 | 1 | 31 |
| | (62%) | (92%) | (50%) | (79%) |

⁷ These totals exclude articles published in the (May Issue) Papers and Proceedings of the AEA.

⁸ A complete submission contained all of the data and code needed to replicate every single table and figure presented in the article. A submission that was incomplete but in compliance with the policy contained instructions for obtaining the missing material (either in the data appendix or in the data archive submission).

IV. Discussion and Conclusion

The AER data submission policy has largely been successful. Four out of five authors complied with the intent of the policy and no serious errors were detected in the nine replications attempted in the Project. Empirical studies published in the *AER* after 2005 are much easier to replicate than those published prior to the *AER*'s mandatory data availability policy. Some of the submitted data packages were thorough enough to serve as examples for graduate students wishing to learn how to perform cutting edge empirical analysis.

There remains some opportunity for improvement in the quality of the data and programming packages submitted to the archive. One in five did not fully satisfy the intent of the policy (to enable independent replication) and many more could have eased replication. In order to improve the quality of the submitted files, the publication office of the American Economic Association in Fall 2009 hired economics PhD students to examine the data files submitted before publication. If the files appear to be incomplete, authors are now contacted to update their data submissions prior to publication.

The next step for improving replicability in economics is to develop generally accepted replication principles (the acronym GARP would be appropriate if it were not used elsewhere in economics). Koenker and Zeileis (2009) provide an overview of how recent developments in software can help researchers improve replicability. They discuss specific examples of data management software, programming environments, and version tracking systems that reduce the burden of maintaining accurate and useable records of empirical analysis.

A potential shortcoming of the *AER* data submission policy is that only final estimation data must be submitted. This means that the code containing all of the procedures applied to the raw data such as variable creation, data "cleaning", and sample selection, are not technically required to be

submitted (although many authors voluntarily include them). This highlights a specific weakness that a set of replication principles might address. The most commonly used software packages were Stata, Matlab, and Gauss. Authors using these applications (and many others) will typically write several programs for transforming raw data (potentially from several sources) into estimation data. Additional programs will then use the estimation data as an input for the econometric analysis that results in the tables and figures found in an article. Since difficulties in replicating work arise in both steps, a more complete policy would require authors to submit all of the programs used to transform the raw data files into the tables and figures found in the paper. This should be the ultimate goal for data archives because it leaves no ambiguity about what procedures the authors conducted to perform their analysis (provided that one can become familiar with the application used).

References

- Anderson, Richard G., and William G. Dewald.** 1994. "Replication and Scientific Standards in Applied Economics a Decade after the Journal of Money, Credit and Banking Project." *Federal Reserve Bank of St. Louis Review*, 76(6): 79-83.
- 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.
- Hamermesh, Daniel.** 2007. "Viewpoint: Replication in Economics." *Canadian Journal of Economics*, 40(3): 715-733.
- Koenker, R., and Achim Zeileis.** 2009. "On Reproducible Economics Research." *Journal of Applied Econometrics*, 24(5): 833-847.
- McCullough, B.D.** 2007. "Got Replicability? The Journal of Money, Credit and Banking Archive." *Econ Journal Watch*, 4(3): 326-337.
- McCullough, B.D., Kerry Anne McGeary, and Teresa D. Harrison.** 2006. "Lessons from The Journal of Money, Credit and Banking Archive." *The Journal of Money Credit and Banking*, 38(4): 1093-1107.
- McCullough, B.D. , and Hrishikesh D. Vinod.** 2003. "Verifying the Solution from a Nonlinear Solver: A Case Study." *American Economic Review*, 93(3): 873-892.
- Vinod, Hrishikesh D.** 2009. "Stress testing of econometric results using archived code for replication." *The Journal of Economic and Social Measurement*, 34(2/3): 205-217.

Appendix

Replication Accuracy Rating System

| | |
|----------|---|
| 5 | Perfect <ul style="list-style-type: none">• All tables and figures, including those in the appendix can be replicated precisely. |
| 4 | Practically Perfect <ul style="list-style-type: none">• Whenever complete instructions are available, replicated and published results are equivalent• Some figures or tables may be time consuming or difficult to reproduce due to missing instructions. However, these exhibits should be approximated and though results may differ slightly, there is no reason to doubt their accuracy. |
| 3 | Minor Discrepancies <ul style="list-style-type: none">• Minor discrepancies exist between the replication and published replicated results that would not affect the conclusions of the paper (if they were truly errors).• e.g. Estimated coefficients match but standard errors are slightly different and do not substantially change the results of important hypothesis tests.• Difference between 3 and 4 could be the result of computer program versions and operating systems |
| 2 | Potentially Serious Discrepancies <ul style="list-style-type: none">• Differences exist between published and replicated results that may indicate problems with the empirical analysis. |
| 1 | Serious Discrepancies <ul style="list-style-type: none">• Substantial differences exist between important empirical results that can't be reconciled.• These discrepancies indicate that an error in the analysis has probably lead to incorrect conclusions. |