THE JOURNAL OF ECONOMIC EDUCATION, *43*(2), 182–189, 2012

ISSN: 0022-0485 print/2152-4068 online
DOI: 10.1080/00220485.2012.659647

Routledge
Taylor & Francis Group

ECONOMIC INSTRUCTION

Teaching Integrity in Empirical Research: A Protocol for Documenting Data Management and Analysis

Richard Ball and Norm Medeiros

This article describes a protocol the authors developed for teaching undergraduates to document their statistical analyses for empirical research projects so that their results are completely reproducible and verifiable. The protocol is guided by the principle that the documentation prepared to accompany an empirical research project should be sufficient to allow an independent researcher to replicate easily and exactly every step of the data management and analysis that generated the results reported in a study. The authors hope that requiring students to follow this protocol will not only teach them how to document their research appropriately, but also instill in them the belief that such documentation is an important professional responsibility.

Keywords *documentation, empirical research, replication*

JEL codes *A22, B4, C8*

This article describes a protocol for documenting statistical analysis that we developed for use by undergraduates conducting empirical research projects. The guiding principle of the protocol is that the documentation should allow an independent researcher to replicate every step of the data management and analysis and to generate the same results. To achieve this objective, students create and assemble a collection of electronic documents that they turn in with their printed papers. These documents are of several types, including raw data files, computer command files, and supporting information that we refer to as "metadata." Economics majors writing senior

Richard Ball is an associate professor of economics at Haverford College and the corresponding author (e-mail: rball@haverford.edu). Norm Medeiros is an associate librarian of the College and a coordinator for Bibliographic and Digital Services at Haverford College (e-mail: nmedeiro@haverford.edu).

The authors thank participants in the AEA Conference on Teaching Economics and Economic Research, Stanford University, June 1–3, 2011, for valuable discussion and comments, with particular thanks to Vera Brusentsev, the discussant for the paper, for especially insightful and constructive ideas. The article has also benefitted from the thoughtful reports of two referees.

This article is based on a paper that was presented at the National Conference on Teaching Economics held at Stanford University on June 1–3, 2011.

theses, as well as students writing research papers for an introductory statistics class, have used this protocol successfully. In the case of senior theses, we make this documentation publicly available, along with electronic copies of the theses themselves, in an online archive maintained by the Haverford College library.

Our primary motivation for teaching this protocol is that preserving documentation of statistical analysis is simply good research practice. According to standards of the American Statistical Association (1999), authors of research involving statistical analysis have an ethical responsibility to "[p]romote sharing of (nonproprietary) data and methods" and "[a]s appropriate, make suitably documented data available for replicate analyses, metadata studies, and other suitable research by qualified investigators." As we discuss below, the record of the economics profession with respect to this standard of replicability has been generally poor. We hope that requiring students to follow our protocol will not only teach them how to document their research appropriately, but also instill in them the belief that such documentation is an important professional responsibility.

In addition to serving the important objective of ensuring the replicability of results, introducing students to our protocol has generated a variety of ancillary pedagogical benefits. Requiring students to produce comprehensive documentation of their empirical work leads them to improve the organization and coherence of their data management and analysis throughout the entire course of their research. It also makes it possible for the instructor to provide better guidance to students as they work on their projects and to evaluate and comment on the completed projects more insightfully. Most broadly, by teaching students that they can and should take steps to ensure the replicability and verifiability of their statistical work, we hope to contribute to their appreciation of the principles of integrity and accountability. In all their work—whether empirical or nonempirical, and whether in economics or in any other field—they should understand how they have reached the conclusions that they state, and be prepared to substantiate or document their arguments and evidence.

The next section of this article enumerates the principles that underlie the standard of replicability that our protocol for documenting empirical research is intended to achieve and compares them to principles and standards of replication that have been defined and adopted elsewhere. The subsequent section reviews the practices that typically economists have followed with respect to documentation and replicability of empirical research. The story is a sad one and underscores the importance of teaching students to do better.

We then present the nuts and bolts of our protocol—the electronic documents we ask students to create and preserve, the information those documents should contain, and how they should be formatted and organized. Our main reason for discussing the nuts and bolts of our protocol is that explaining how certain standards of replicability can be achieved in practice serves to highlight the principles and objectives those standards are meant to achieve. Our purpose is not to propose the particulars of the protocol that we have developed as a uniquely ideal system for students—or anyone else—to use to document empirical research. We continue to revise and develop the protocol, and further refinements are certainly possible. We expect others will see ways to improve the protocol, and we hope they will communicate their ideas to us.

After presenting the nuts and bolts of the protocol, we discuss the online thesis archive maintained by the Haverford College Library that we use to make the documentation assembled by economics majors for their senior theses available to the public, so that anyone interested in replicating their results or exploring their data further is able to do so.

The final discussion and conclusion include a description of some ancillary pedagogical benefits—beyond the direct benefits related to the replicability of empirical results—that arise when we ask students to follow our protocol, as well as broader reflections on the contributions of this exercise to the education of undergraduates.

REPLICABILITY OF EMPIRICAL RESEARCH: STANDARDS AND PRINCIPLES

Soup-to-Nuts Replication

Our protocol for documenting empirical research is intended to achieve a standard of replicability that we characterize as "soup to nuts." To meet the soup-to-nuts standard, the documentation for a project must contain a number of elements.

First, copies of all data files used for the project must be preserved and included in the documentation, in exactly the form in which they were first obtained by the researcher, before they were modified in any way.

Second, every raw data file should be accompanied by one or more additional documents with whatever metadata—information about file structure and format, variable definitions and coding, sampling methods and weighting, etc.—a user would need to be able to understand and interpret the contents of the data file.

Third, the documentation should include files containing commands that instruct the statistical software used for the project to execute all the steps of data processing and analysis that were conducted for the project. The commands in these files should begin by importing the data contained in the raw data files, then carrying out all the operations—cleaning, organizing and combining data from different files, generating new variables, etc.—necessary to create the final versions of the data files that were used for the project, and finally implementing the procedures and analyses that generate all the reported statistical results—including tables, figures, and quantitative findings stated in the text of the paper.

Finally, to make the idea of replicability a reality, this documentation must be made publicly available, unless allowing public access would violate intellectual property rights or compromise the privacy of survey respondents or research subjects. An independent researcher with access to the appropriate statistical software could then replicate the empirical work conducted for the project in its entirety—i.e., from soup to nuts—simply by downloading the data and command files, and then running the command files.

Partial Replication

In some contexts, the standard of replication applied to empirical research is one that we would characterize as "partial." Partial replicability differs from soup-to-nuts replicability in a critical way: Partial replicability requires the preservation only of the processed data in the form used for the final analysis, and command files with instructions that implement the analyses on the processed data to generate the results of the study. In contrast to soup-to-nuts replicability, partial replicability does not require the preservation of raw data files or the commands that transform the raw data into the final, processed data files used in the analysis.

An important example of partial replication being adopted as the standard for documentation of statistical analysis can be found in the Data Availability Policy of the *American Economic Review* (*AER*) (American Economic Association, n.d.). The *AER* requires authors to submit "the data set(s) and programs used to run the final models," but does not require authors to submit the original, unprocessed data sets and code with the commands that process the data as necessary to create the data sets "used to run the final models." Authors are required only to submit "a description of how previous intermediate data sets and programs were employed to create the final data set(s)" (American Economic Association, n.d.). We agree with Glandon (2011, 699) that "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," because this "leaves no ambiguity about what procedures the authors conducted to perform their analysis . . . ."

REPLICABILITY OF EMPIRICAL RESEARCH: PRACTICE

Historically, economists have paid little attention to the replicability of their empirical research and have done a poor job of documenting it. Dewald, Thursby, and Anderson (1986) brought attention to this issue with a study of research published in the *Journal of Money, Credit and Banking* (*JMCB*). Beginning in 1982, the editors of the *JMCB* adopted a policy of asking authors of empirical papers published in the journal to submit data sets and computer code that could be used to replicate the empirical results reported in the papers. Dewald, Thursby, and Anderson (1986, 591) reviewed the first 54 papers published under this new policy, and found that the documentation was "complete enough to allow an attempt at replication" for only 8 of them, and 14 unambiguously failed to meet this standard. Problems like incorrect or imprecise citations of data sources, and failures to define variables and describe how they had been transformed, were found in the documentation of all of the remaining 32 papers. In addition, more thorough attempts were made to actually carry out the replication of the results of 9 selected papers for which the documentation appeared complete. In 2 cases these attempts were unambiguously successful, in 2 others they were unambiguously unsuccessful, and the 5 remaining cases lay somewhere between these extremes.

In the 25 years since the publication of the Dewald, Thursby, and Anderson (1986) study, many economic journals have adopted policies requiring authors of empirical papers to submit accompanying documentation. Links to the online archives containing this documentation for seven journals can be found at http://www.rfe.org/showCat.php?cat_id=9, and a number of other major journals maintain archives not linked from that site. But even when authors are required to submit this kind of documentation, and even when authors are aware that this documentation will be posted publicly on the Internet, the standards of transparency and replicability that are attained typically remain very low. McCullough, McGeary, and Harrison (2006), for example, found that, among more than 150 empirical articles published in the *JMCB* between 1996 and 2002, the documentation available in the *JMCB* archive could successfully reproduce the results of fewer than 15 papers. Similarly, Glandon (2011, 698) found that among 39 selected empirical articles published in the *American Economic Review*, the posted documentation was "sufficient to attempt a detailed replication" for only 20.[1]

This lacuna in the practice of academic economists has had real consequences. There have been numerous cases in which questions raised about the empirical results of influential papers

on critical policy issues, published in leading journals, have led to protracted debate and sometimes heated controversy. The Hoxby–Rothstein dispute over Hoxby's paper on the educational consequences of competition among schools is a well-known example (Hoxby 2000, 2007; Rothstein 2007); another is the Pitt and Khandker–Morduch and Roodman exchange about the empirical evidence on the effects of microcredit programs on social outcomes (Pitt and Khandker 1998; Khandker 2005; Roodman and Morduch 2009). Both of these debates encompassed methodological questions that went beyond the issue of replicability, but in both cases difficulties encountered in replicating results proved to be a significant obstacle in the exegesis and resolution of the methodological questions. McCullough and McKitrick (2009) describe ten additional instances of published, peer-reviewed, empirical research that had major influences on public opinion and public policy, but for which sufficient data and documentation to allow replication were not made available in a timely manner. In all of these cases, when the data and computer files were eventually made available, both the originally reported results and the policy implications had to be reversed or substantially revised—in some cases after policy decisions based on the original results had been taken.

Despite calls for better documentation, it is not at all clear what the best means for establishing norms or incentives for appropriate documentation of empirical research would be, or how likely it is that such efforts would succeed. Promoting broad changes in attitudes and practices throughout the economics profession looks like a tough nut to crack. As educators, however, we can at least set the norms and incentives that guide the work of our students. Efforts that instructors make to hold their students to a high standard of replicability and documentation of their research have the potential to create a "trickle-up" effect that could begin to exert some broader influence as our students join the ranks of professional economists.

THE NUTS AND BOLTS OF SOUP-TO-NUTS

In this section, we briefly outline the content and organization of the files that we require students to assemble to document their empirical research. For readers interested in more detail, we have created a Web site on which we have posted an expanded version of this article, a set of documentation files for a hypothetical senior thesis that we have created to illustrate how our protocol is implemented, and a copy of the complete set of instructions for documenting research papers and theses that we ask our students to follow. The URL for the Web site is www.haverford.edu/economics/faculty/rball/soup_to_nuts.php.[2]

The documentation we ask students to create includes electronic files of various types:

- Raw data files;
- Importable data files;
- Metadata;
- Command files, written in the syntax of the software used for the project;
- A readme file.

We use the term "raw data file" to refer to an electronic file containing statistical data that is in exactly the form in which the student first obtained it, before it was edited or modified in any way. When raw data files are not in the proprietary format of the statistical software that will be used for the project (e.g., the .dta format for Stata or the .sav format for SPSS), we ask students

to also submit "importable" versions that are saved in a format (typically some kind of delimited text) that can be read by the software. The metadata files contain all the information a user would need to understand the contents of the data files, such as variable definitions, coding schemes, and sampling methods.

To ensure soup-to-nuts replicability, the command files included in the documentation should contain instructions that (i) begin by directing the software to open or import the importable data files, then (ii) complete all the steps involved in processing the data (cleaning, merging, recoding, generating new variables, etc.) required to prepare them for analysis, and finally, (iii) execute commands that use the processed data to generate the results reported in the thesis. Finally, the readme file lists all the files included in the documentation, describes the content and purpose of each, and explains how the importable data files and the command files can be used to replicate the analysis that generated the statistical results reported in the paper.

MAKING THE DOCUMENTATION PUBLICLY ACCESSIBLE

Haverford College, like many academic institutions, maintains an online repository of scholarly work created at the college. Like some, but fewer, institutions, Haverford includes undergraduate senior theses among the materials archived in the repository. To date, more than 1,100 senior theses, from all the academic departments of the college, have been deposited in this online collection. The homepage for the thesis archive can be found at http://thesis.haverford.edu. In most cases, an electronic copy of the printed thesis is the only document posted in the archive. For economics majors who write empirical theses and assemble electronic documentation of their statistical work according to our guidelines, however, we use this archive to also make the accompanying documentation files available to the public.

Haverford's digital repository is maintained on DSpace, a widely deployed open source application jointly developed by MIT and Hewlett-Packard. Haverford's DSpace instance has been given the name "Triceratops." This platform accommodates not just text files, but also nontextual objects, such as image, music, and video. Moreover, it provides a robust authorization module that can negotiate copyright concerns and embargo requirements.[3] The files made available on Triceratops for each senior thesis include the full text of the thesis and all the documentation files assembled by the student.

Two recent theses provide good examples. The Triceratops record for the 2010 senior thesis of Laura Costanzo can be accessed at http://triceratops.brynmawr.edu/dspace/handle/10066/4899. Since none of the data were proprietary and Ms. Costanzo agreed to allow public access to her materials, the thesis, data files, command files, and the readme file are all publicly available. Using instructions included in the read_me.pdf file, a researcher from anywhere in the world can download these files and replicate the analysis.

The Triceratops record for the 2010 senior thesis of Siobhan Neitzel (http://triceratops.brynmawr.edu/dspace/handle/10066/4820) illustrates the capability of DSpace to restrict access to a selected subset of the files, while allowing public access to others. In this case, some of the data, having to do with league standings and statistics for soccer teams in the United Kingdom, were assembled by the students from public sources, and she gave her permission for unrestricted access. Some of the data, however, were obtained from the British Household Panel Survey (BHPS), which releases data by permission only. Accordingly, as indicated on the Triceratops

record, access to the BHPS data has been limited to archive staff, but no restrictions have been placed on access to the other files in the documentation.

DISCUSSION AND CONCLUSIONS

Our primary motivation for developing our documentation protocol is to teach students the importance of transparency and replicability of data management and analysis. In addition, learning the protocol helps students acquire the technical tools they need to execute and document their empirical research effectively. We hope that as some of our students go on to careers as professional economists, their heightened awareness of the importance of replicability may contribute to higher standards for documenting empirical research throughout the profession as a whole.

Teaching students to follow our protocol also has produced several ancillary pedagogical benefits. The process of constructing and saving commands in do-files not only generates command files that can be used for later replication, but also benefits the students at all stages of their work with the data. When students know they will have to turn in command files that reproduce everything they do, their data management and analysis tend to be much more organized and efficient, and their understanding of what they are doing tends to be much greater, than when they use their statistical software to execute commands interactively. Many students initially balk at the idea of working with command files instead of in an interactive mode, but they quickly come to realize how much easier it is to modify their work, experiment, and keep track of what they have done when they preserve all the instructions that are executed in the course of their management and analysis of the data.

In addition, when students are required to document their work carefully, the instructor's ability to provide useful comments and guidance on the work is greatly enhanced. When students ask for help while they are working on a project, the instructor can respond much more effectively when all the steps of the students' work are transparent and replicable up to the point at which the question arose. And when assessing the final paper, the instructor can use the electronic documentation to do "live" checks of the statistical work and explore the data further, making it possible to make much more detailed and insightful comments on the student's work.

Most fundamentally, the experience of assembling complete documentation for their research serves the students as an exercise in responsibility and integrity. It requires them to adhere to an important principle, namely that they should not turn in papers in which they make statements or claims that they cannot verify or substantiate. We believe that this principle should apply no less to statements about results obtained from statistical analyses than it does to any other statements. It is violated when students are allowed to turn in papers that report statistical results without preserving documentation that can be used to replicate them. Requiring students to ensure that their empirical results are replicable sends them the message that they must accept responsibility and accountability for all of their work—statistical and otherwise. For students who are planning to pursue a career in economics as well as those who are not, we believe that this is an important message.

NOTES

1. For additional discussion of current practice in economics with respect to the documentation and replicability of empirical research, see Anderson et al. (2008); McCullough, McGeary, and Harrison (2006, 2008); and Vinod (2001).

2. Several related books may interest instructors and thesis supervisors. Long's (2009) book is an excellent and detailed guide to organizing data management and analysis in ways that facilitate the construction and assembly of the documentation our protocol requires. (Although it is written expressly for Stata users, its main lessons are easily transferable to any other statistical package.) For an excellent guide to the entire process of writing research papers, aimed at undergraduate economics majors and including several chapters on working with statistical data, see Greenlaw (2005). And for pointers and principles of good writing, see McCloskey (1999).
3. When students turn in their senior theses, they are also asked to sign release forms granting the library permission to post their work in the thesis archive. These releases indicate the level of access to be granted, which the student can choose to be unrestricted open access, access limited to authenticated users of Haverford's network, or access allowed only for archive administrators and Haverford economics faculty. The student also can request to have the thesis embargoed until a specified release date.

REFERENCES

American Economic Association. n.d. The *American Economic Review:* Data availability policy. American Economic Association. http://www.aeaweb.org/aer/data.php (accessed June 15, 2011).

American Statistical Association, Committee on Professional Ethics. 1999. Ethical guidelines for statistical practice. American Statistical Association. http://www.amstat.org/about/ethicalguidelines.cfm (accessed November 1, 2010).

Anderson, R., W. 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.

Dewald, W., J. Thursby, and R. Anderson. 1986. Replication in empirical economics: The *Journal of Money, Credit and Banking* Project. *American Economic Review* 76(4):587–603.

Glandon, P. J. 2011. Appendix to the report of the editor: Report on the *American Economic Review* Data Availability Project. *American Economic Review Papers and Proceedings* 101(3):695–99.

Greenlaw, S. 2005. *Doing economics: A guide to understanding and carrying out economic research*. Mason, OH: South-Western College Publishing.

Hoxby, C. 2000. Does competition among public schools benefit students and taxpayers? *American Economic Review* 90(5):1209–38.

———. 2007. Does competition among public schools benefit students and taxpayers? Reply. *American Economic Review* 97(5):2038–55.

Khandker, S. R. 2005. Microfinance and poverty: Evidence using panel data from Bangladesh. *World Bank Economic Review* 19(2):263–86.

Long, S. 2009. *The workflow of data analysis using Stata*. College Station, TX: Stata Corporation.

McCloskey, D. 1999. *Economical writing*. 2nd ed. Long Grove, IL: Waveland Press Incorporated.

McCullough, B. D., K. A. McGeary, and T. D. Harrison. 2006. Lessons from the *JMCB* archive. *Journal of Money, Credit and Banking* 38(4):1093–1107.

———. 2008. Do economics journal archives promote replicable research? *Canadian Journal of Economics* 41(4):1406–20.

McCullough, B. D., and R. McKitrick. 2009. Check the numbers: The case for due diligence in policy formation. Vancouver, British Columbia: Fraser Institute Studies in Risk & Regulation. Accessed on September 7, 2011 at: http://www.pages.drexel.edu/∼bdm25/DueDiligence.pdf.

Pitt, M. M., and S. R. Khandker. 1998. The impact of group-based credit on poor households in Bangladesh: Does the gender of participants matter? *Journal of Political Economy* 106(5):958–96.

Roodman, D., and J. Morduch. 2009. The impact of microcredit on the poor in Bangladesh: Revisiting the evidence. Center for Global Development Working Paper No. 174. Washington, DC: Center for Global Development. Accessed on September, 7 2011 at: http://www.cgdev.org/content/publications/detail/1422302.

Rothstein, J. 2007. Does competition among public schools benefit students and taxpayers? Comment. *American Economic Review* 97(5):2026–37.

Vinod, H. D. 2001. Care and feeding of reproducible econometrics. *Journal of Econometrics* 100:87–88.