Teaching and Practicing

Integrity in Empirical Research

Documenting and Replicating Data Management and Analysis

Richard Ball
Associate Professor of Economics
Haverford College

www.haverford.edu/TIER

Mississippi State University
September 26, 2014

The broad context:

Concerns about transparency/robustness/credibility of empirical research in economics and other social sciences.

These concerns have many facets:

--pre-registration of research hypotheses and methods

--“p-hacking”, specification mining

--replicability of experimental results

--replicability of computational results

A brief history of computational replicability of empirical research in economics

The Big Bang:

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.

Recent Observations

Case studies:

McCullough, B. D., & McKitrick, R. 2009. Check the numbers: The case for due diligence in policy formation. Vancouver, British Columbia: Fraser Institute Studies in Risk & Regulation.

The Reinhart-Rogoff affair:

Reinhart, C. M., & Rogoff, K. S. (2010). Growth in a time of debt. *American Economic Review, 100*(2), 573-578.

Herndon, T., Ash, M. & Pollin, R. (2013). Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff. Working paper. Political Economy Research Institute, University of Massachusetts, Amherst.

The Picketty affair:

Piketty, Thomas (2014). *Capital in the Twenty-first Century*. Cambridge: Belknap Press of Harvard UP.

Giles, C. “Data problems with *Capital in the Twenty-First Century*. *Financial Times* blog. May 23, 2014.

Many journals now have “data policies” and maintain on-line “data archives.” But their success has been mixed:

Enforcement and quality-control are not consistent.

The policies are stated very broadly, and offer little specific guidance about what constitutes adequate documentation.

The standard the documentation must meet is very low:

Only the “final data set” and code that uses it to produce reported results must be submitted.

Preservation of original data files and documentation of construction of final data sets not required.

For example, see the data policy for the *American Economic Review* at www.aeaweb.org/aer/data.php.

The *American Economic Review* data policy (in part):

> For econometric and simulation papers, the minimum requirement should include the data set(s) and programs used to run the final models, plus a description of how previous intermediate data sets and programs were employed to create the final data set(s). Authors are invited to submit these intermediate data files and programs as an option; if they are not provided, authors must fully cooperate with investigators seeking to conduct a replication who request them. The data files and programs can be provided in any format using any statistical package or software. Authors must provide a Readme PDF file listing all included files and documenting the purpose and format of each file provided, as well as instructing a user on how replication can be conducted.”

The Evolution of Project TIER:

Project TIER developed organically in the context of an introductory statistics class for undergraduate economics majors.

> Experience gained over several years of assigning, advising and reading student research papers for the class led to the development of a set of instructions—a pdf document—for assembling comprehensive documentation for a research paper.
>
> We call these instructions a "protocol."
>
> The protocol specifies a set of electronic documents that students submit at the same time they turn in their completed research papers.

The principle underlying the protocol:

Documentation of an empirical research project should allow

easy

exact

soup-to-nuts

replication of all reported statistical results.

Our students now routinely produce documentation that meets this standard, both for papers for introductory statistics classes and for senior theses.

The files specified by the protocol include:

all **original data files** from which data for the project were extracted

metadata required to understand and interpret the contents of the original data files (e.g., pdfs of, or links to, codebooks or users' guides)

command files containing code that executes all the steps required to get from the original data files to the empirical results reported in the paper, including:

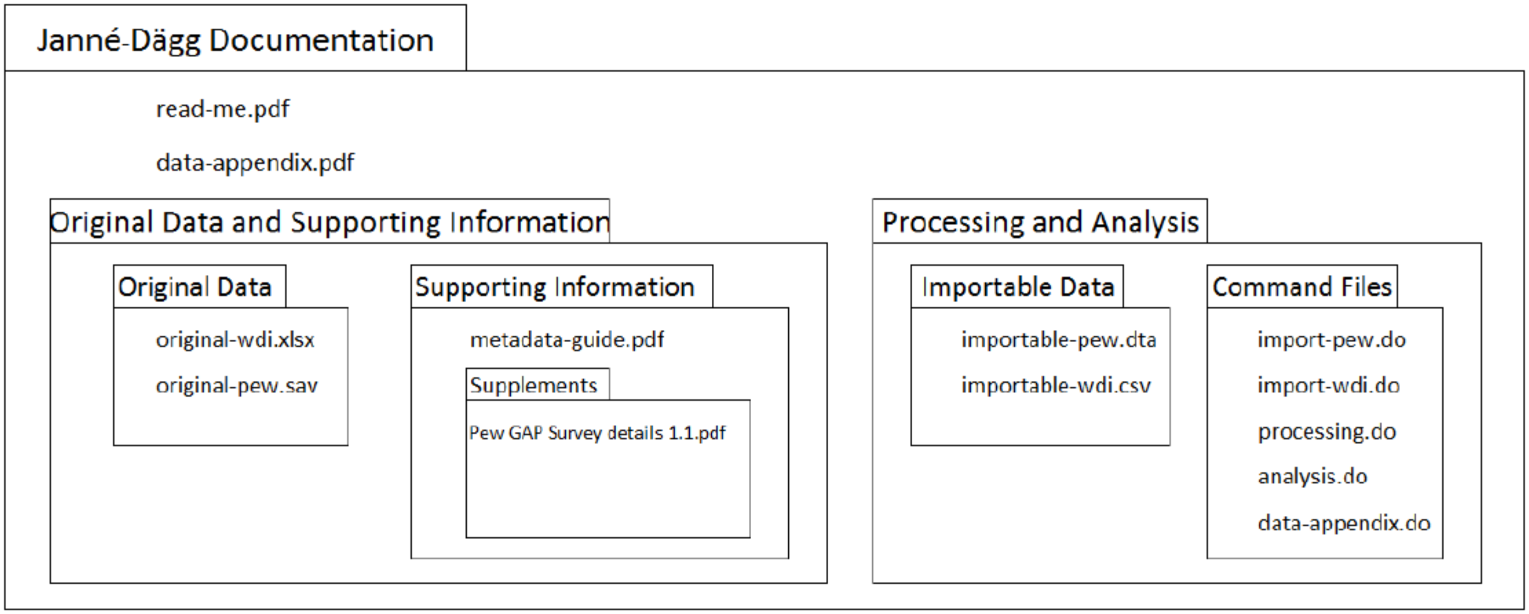
> importing original data files; cleaning, merging, and all other processing required to construct the "final data set(s)" used for analysis; generating all the statistical results, tables and figures presented in the paper

a **data appendix** for the “final data set,” providing definitions, sources, coding and descriptive statistics for all variables in the “final data set(s)”

a **readme file** that describes all the documents included in the electronic documentation

Illustration of TIER Protocol: Organization of Documentation Files

Using the Hypothetical Janné-Dägg Paper as an Example



Web platforms for organizing and documenting empirical research

Open Science Framework (https://osf.io/)

Especially good for managing the process/workflow of research; works well for collaborative projects

Dataverse (http://thedata.harvard.edu)

A little “heavier.” Good for archiving. As a workflow tool, maybe best for experienced data users and information specialists.

Details of the protocol can be found at the Project TIER website: www.haverford.edu/tier.

More details and discussion can be found in a paper we have written:

> Ball, R. and Medeiros, N. (2012). Teaching Integrity in Empirical Research: A Protocol for Data Management and Analysis. *Journal of Economic Education* 43(2): 182-189.

The protocol is a work in progress; we are continually refining it. There are many ways it could still be extended and improved—and we would love your help!

There are pedagogical benefits to teaching students to document their research that are independent of the broad issue of robustness of social science research:

> Students organize their work much better and understand what they are doing much more clearly when they have to do everything in command files.
>
> I can advise them much more effectively throughout the project when I can get my hands on their data and code and play around with them. Similarly, I can evaluate and comment on their final papers much more insightfully.
>
> When students keep track of and understand what they are doing with their data, they gain confidence that they can learn and discover real, meaningful things.
>
> The entire documentation exercise reinforces some fundamental educational and life principles: transparency; integrity; understanding how you arrive at the claims you make and knowing how to justify or substantiate them.

Two useful web platforms for managing, sharing and archiving research data, metadata and command files:

Dataverse: https://thedata.harvard.edu/dvn/

Open Science Framework: https://osf.io/

Project TIER has posted resources on both of these sites; see the following pages for more information.

Dataverse: Several examples of student research, documented according to the TIER protocol, are available at http://thedata.harvard.edu/dvn/dv/TIER.

> Some of the projects are senior theses of recent economics majors at Haverford College. For these, both a pdf of the thesis and all the accompanying documentation, are publicly available for download.
>
> Some of the projects were completed as an assignment for an introductory statistics class at Haverford College in the fall of 2013. For these, titles and some metadata for the paper and the supporting files are publicly accessible, but the files cannot be downloaded without permission.

Open Science Framework: We have posted a “hypothetical” research paper that was written to serve as a demonstration of the TIER documentation protocol at https://osf.io/he87a/.

When you go to that URL, the “Overview” button on the menu near the top of the page will be highlighted/shaded, and you will be able to open and read the files, but not download them. If you would like to download them, click on the “Files” button just to the right of the “Overview” button.

Begin by reading the document with the name “READ ME FOR PROTOCOL 2.0 DEMO.pdf.” That document describes the documents posted at this OSF site, and explains how to use them to learn about the protocol.