Project TIER:

Teaching Transparency in Empirical Research

Richard Ball
Associate Professor of Economics
Haverford College
rball@haverford.edu

www.haverford.edu/TIER

@Project_TIER

Presented at the INET/YSI Workshop on
Replication and Transparency in Economic Research
San Francisco
Day 2: January 7, 2016

Project TIER is supported by a grant from the Alfred P. Sloan Foundation.

Documentation for Computational Reproducibility of Quantitative Social Science Research: Purposes

Replication documentation can be used for several purposes, all of which are important:

Confirmation: verifying that there were no errors in the computations conducted for the study.

Robustness checking: assessing the robustness of the results reported in the paper.

Extension: initiating new research that extends or builds on the paper.

Communication: recording concisely and unambiguously the ways in which the data were manipulated in preparation for analysis, and the procedures that generated the reported results.

Documentation for Computational Reproducibility of Quantitative Social Science Research: Principles

To serve the purposes we have identified, replication documentation should satisfy the following principles:

Complete replicability: The documentation should make it possible to conduct a replication of the study that begins with raw data identical to those with which the author started the research, processes them as necessary to prepare them for analysis, and finally executes the commands that generate the results reported in the paper.

Independent replicability: All the information necessary to replicate the study should be included in the documentation. In particular, it should not be necessary to request any additional information from the author.

Realism: The documentation should be clearly enough organized and presented that it is realistic to expect a reasonably competent researcher to be able to conduct a complete and independent replication of the study without undue difficulty.

Compare these principles to the *AER* data availability policy:

> … Authors of accepted papers… must provide to the *Review*…the data, programs, and other details of the computations sufficient to permit replication. These will be posted on the *AER* Web site.
>
> …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…*** 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 PROCESS OF CONSTRTING DOCUMENTATOIN THROUGHOUT THE COURSE OF CONDUCTING A RESEARCH PROJECT

A prologue to the process: readings and exercises to get students oriented for the research and documentation exercise they are about to begin

Readings:

Articles with examples or surveys of professional publications that fail to replicate

Articles about why documentation and replicability are important

Exercises:

Have students work through the Project TIER demo project

Have them use documentation prepared by students who did a project in a past semester to replicate that previous project

Have them replicate a real published paper, in the Hoeffleresque tradition

For undergrads, it is probably a good idea to break the process down into "***installments***," each of which involves a written deliverable (with electronic files, when the work has progressed far enough).

For example:

--Ideas about possible topics and data sources

--Proposal—more fleshed out topic and sources—and narrowed down to one firm choice

--Literature review

--Data Report, with Data Appendix appended—and replication data and code

--Draft of the paper—with replication data and code

--Final paper—with complete replication data and code

Implementing the TIER Documentation Protocol: Key Steps in the Process of Conducting the Whole Research Project

At the very beginning, set up the (empty) folder structure; then collect and create the various documents that go in the various folders as you go along.

That set of folders will be where you keep all your work for the project.

The first time you get your hands on a new original data file:

Put an unmodified version in the Original Data folder.

And immediately write the section of the MetaData Guide for this original data file; if the original data files needs any supplementary metadata documents, put copies of those documents in the Supplements folder.

Put an importable version of the original data file in the Importable Data folder.

Processing the data: transforming the original data files into your (final, clean) analysis data files:

> All the steps of data processing—from importing or reading the original data files, to cleaning, modifying and combining them, to saving the analysis data file(s)—should be preserved in the form of do-files with commands that execute all the steps.
>
> You will often spend weeks incrementally building up the do-files that do your data processing.
>
> > Keeping your do-files clean and up-to-date throughout this process is critical.
> >
> > Remove "detritus" from your do-files at the end of every work session so that you are left only with clean, well-organized commands that implement part of the processing you want to do.

Put lots of comments in the command files—not only to help others understand what you did, but to remind yourself in the future of what you did.

Do not keep multiple copies of slightly different versions (with slightly different names) of a single-do file. To keep track of the most recent version, put the date at the end of the name of each do-file.

> (If you are nervous about trashing earlier versions, create a subfolder called "history," and put obsolete versions there instead of trashing them. When you have finished the project and are cleaning up your final documentation, just delete that whole history folder.)

Choose one folder to which your software's working directory should be set at all times when the do-files are running, and do not use "change directory" commands in the do-files.

> Instead, use ***relative directory paths*** to open or save files in other folders. (Look up "relative directories" in the index to Scott Long's *Workflow of Data Analysis Using Stata*.)

There is no need to include copies of "intermediate" data sets in your documentation.

> (In fact, it is usually not necessary to include copies of "analysis" data sets.)

The Data Appendix

Like a codebook for your cleaned/processed/final data set

Construct a complete data appendix ***before you begin your analysis***

When students work this way, it fundamentally transforms the way I can communicate with them and advise them

(Using a cloud platform like OSF helps, but it is not necessary)

And it makes it possible for me to evaluate their papers much more critically and constructively

But how much class time do I have to devote to this? What gets dropped from the curriculum to make room for it? How much more time does it take me to grade papers when I have to consider the documentation as well as the paper?

Please keep in touch!

Ask us for help—in your own research or in teaching this stuff to students or research advisees.

Please let us know if you are using any resources or ideas you have taken from Project TIER.

Please let us know of new ideas you have or resources you develop.

We would like to post stuff from lots of people on our website.