

Making Replication Documentation Useful

To You and Others: Purposes, Principles and Practices

Richard Ball

Professor of Economics, Haverford College
Director, Project TIER

Tomas Dvorak

Professor of Economics, Union College
2015-16 TIER Faculty Fellow

Cornell University Department of Applied Economics

September 14-15, 2017

Project TIER is supported by the Alfred P. Sloan Foundation.

DIMENSIONS OF THE RESEARCH TRANSPARENCY MOVEMENT IN THE SOCIAL SCIENCES

Computational reproducibility

Experimental replicability

Project registration and pre-analysis plans

P-hacking

Publication Bias

Resources for learning more:

Ted Miguel's spring 2015 graduate course on research transparency—syllabus and videos of 14 lectures

<http://www.bitss.org/education/economics-270d/>

Miguel and Christensen, forthcoming in JEL

http://emiguel.econ.berkeley.edu/assets/miguel_research/78/Transparency-JEL-2016-12-20.pdf

BITSS MOOC

<https://www.bitss.org/events/mooc-transparent-and-open-social-science/>

Key initiatives:

Berkeley Initiative for Transparency in the Social Sciences

www.bitss.org

Center for Open Science

<https://cos.io>

COMPUTATIONAL REPRODUCIBILITY OF SOCIAL SCIENCE RESEARCH: HISTORICAL CONTEXT

Serious problems recognized decades ago, and despite some progress, they persist

Concern about the reproducibility of published economic research was sparked by a 1986 study known as the “*Journal of Money, Credit and Banking (JMCB)* Project.”

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

The *JMCB* Project

Editors of the *JMCB* attempted to reproduce the statistical results reported in a large sample of the empirical papers published in that journal in the preceding five years.

Requests for replication data and code were sent to authors of 154 papers.

In 37 cases (24%), the authors did not reply to the request.

In 24 cases (16%), the authors replied, but either refused to send data and code, or said they would but never did.

In 3 cases (2%), the authors said they could not provide the data because it was proprietary or confidential.

In the remaining 90 cases (58%), the authors sent *some information* in response to the request.

The *JMCB* Project (continued)

Out of the 90 submissions received, the first 54 were investigated for completeness and accuracy.

Out of the 54 submissions that were investigated, the documentation provided by the authors of the papers successfully replicated the results of their papers in only 8 (15%) of the cases.

The remaining 46 (85%) of the papers could not be replicated because the information the authors submitted was insufficiently complete or precise.

Conclusions of the *JMCB* Project

The authors of the *JMCB* study concluded:

“Our findings suggest that inadvertent errors in published empirical articles are a commonplace rather than a rare occurrence.”

and

“...we recommend that journals require the submission of programs and data at the time empirical papers are submitted. The description of sources, data transformations, and econometric estimators should be so exact that another researcher could replicate the study and, it goes without saying, obtain the same results.”

Subsequent studies show problems persist. A few examples:

McCullough, Bruce D., Kerry Anne McGeary, and Teresa D. Harrison (2006). "Lessons from the *JMCB* Archive," *Journal of Money, Credit and Banking* 38(4): 1093-1107.

McCullough, Bruce D., Kerry Anne McGeary, and Teresa D. Harrison (2008). "Do Economics Journal Archives Promote Replicable Research?" *Canadian Journal of Economics* 41(4): 1406-1420.

Hoeffler, Jan (2014). "Teaching Replication in Quantitative Empirical Economics." Presented at the Meetings of the European Economic Association and the Econometric Society, Toulouse, France, August 28. <http://www.eea-esem.com/eea-esem/2014/prog/viewpaper.asp?pid=3108>.

Chang, Andrew C., and Phillip Li (2015). "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not.'" Finance and Economics Discussion Series 2015-083. Washington: Board of Governors of the Federal Reserve System, <http://dx.doi.org/10.17016/FEDS.2015.083>.

Fixing reproducibility problems means fixing **replication documentation**

Better guidelines and standards need to be **formulated**

And then somehow researchers need to be induced to **adopt** them

But haven't a lot of standards and guidelines for replication documentation been formulated already?

Journals have policies for replication archives (e.g., AEA journals <https://www.aeaweb.org/journals/policies/data-availability-policy>)

DA-RT: <https://www.dartstatement.org/>

TOPS: <https://cos.io/our-services/top-guidelines/>

BITSS manual: <http://www.bitss.org/resources/manual-of-best-practices/>

ALSO:

TIER Protocol: <http://www.projecttier.org/tier-protocol/>

DRESS Protocol: <http://www.projecttier.org/tier-protocol/dress-protocol/>

PURPOSES OF REPLICATION DOCUMENTATION

Not catching mistakes

Rather:

Exploration

Experimentation

Extension

PRINCIPLES

Complete—“soup-to-nuts”

Portable

The “seriously, folks” principle

PRACTICES

Establish a fixed folder structure

Pay attention to the working directory

Use relative directory paths

Let's see some examples:

A toy demo: The midlife crisis paper

A real research paper: Joseph Price & Justin Wolfers, 2010. "Racial Discrimination Among NBA Referees," *The Quarterly Journal of Economics*, MIT Press, vol. 125(4), pages 1859-1887, November.

Both examples use a Stata/Word cut-and-past approach.

Folder Structure

Figure out what works for you, but generally:

- one main project folder

 - pdf of paper

 - subfolder for data

 - subfolder for code

 - subfolder for supporting information (like citations of sources and codebooks for original data)

 - read-me file

That whole packet is the medium of communication

The idea is that while someone is working with your rep doc, they install the whole packet onto their computer—keep the folder structure and file organization intact while they work with your stuff

In Data folder: assuming data are public—need original data files—before you have processed them at all, in whatever format they were in when you first got them

(or else use “netuse” if there is a stable site your software can grab the files from)

---What about intermediate data files?

---What about analysis data files?

In code folder: soup to nuts: commands that read the data from the original data files all the way to command that generate the figures, tables and other results you report in your paper—and all processing in between

all one long script?

separate for separate stages of analysis (import, process, analyze)?

different scripts for different data sources?

--Put tons of comments in code

-----iterate programming??

Pay attention to the working directory:

--for each command file, choose a folder that should be designated as the wd when the user runs the command file, and put a comment at the top of the do file indicating which folder that is

--suggested conventions:

---- always designate the main project folder that contains all the rep doc as the working directory

-----avoid using change directory commands

----instead, use relative directory paths