“Beyond the PDF” in Empirical Economic Research

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

Presented at the ASSA Annual Meetings
San Francisco, January 5, 2016

This work is supported by a grant from the Alfred P. Sloan Foundation.

Many thanks to Jan Hoeffler for the incredibly creative and effective work he has done to make this session, as well as the post-AEA replication workshop, possible.

This talk is about two things:

A vision of a fundamental change in the ways that economists communicate the results of their empirical research

Some practical steps that could be taken to move professional norms and practices toward that vision

The vision might be a bit grandiose and Utopian.

But the steps I propose for moving toward it are modest and realistic.

The vision:

It becomes standard practice for authors of empirical papers to assemble comprehensive documentation that makes it possible for interested readers to replicate (computationally reproduce) their work.

The vision (continued):

Authors would take the construction of such documentation as seriously as writing the paper (the text plus tables, figures, appendixes)

Readers would study the documentation as carefully as they study the paper itself.

The vision (continued):

No one would take seriously a paper for which such documentation was not available

No one would think that she had seriously studied or engaged with the paper unless she had taken the documentation, sat down at a computer, and played with the data and code.

“Playing with the data and code” can mean several things—and it is in those things that the value of replication documentation lies.

Verification

Examination, exploration, and extension

Transparency/understanding/communcation

Cumulative research

If the point of replication documentation is that readers should be able to use it for these purposes, what must be true about the documentation? What information must it contain, and how should that information be communicated?

Documentation should make possible ***complete replication***

Documentation should be ***usable*** or ***realistic***

What could be done to move us in the direction of this imagined vision?

I.e., what steps could be taken to make it more common for authors to construct ***complete*** and ***realistic*** replication documentation for their empirical papers, and for readers to make constructive use of it by “playing around” with the data and code?

A first critical step:

Writing a set of instructions, guidelines or standards—let’s call it a ***convention***—that tells authors what they should include in the replication documentation.

Seriously??

Haven’t a variety of conventions/standards/guidelines already been written (e.g., the AER data availability policy, TOPS, DA-RT)?

And anyway, how would simply writing a convention induce anyone to adopt it?

The *AER* data policy states (in part; emphasis added):

> … 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.

In formulating a convention for documenting empirical research, one encounters some tensions:

It needs to be concrete enough to provide meaningful guidance, but it must be flexible enough to accommodate the diverse kinds of empirical papers economists write.

Again to provide meaningful guidance, it needs to be somewhat detailed, but it must not be so long and complex that understanding and following its specifications is not overly burdensome to authors.

Here is a brief summary of a convention designed to balance the competing demands:

> Include command files (code written in the syntax of whatever software you used for the project) that are sufficient for a complete replication of all the results reported in the paper (including all data processing that generated the final data set, as well as the analysis performed on the final data set that generated the results).

Brief summary of a convention (continued):

Don't include any data in the documentation.

Include a ReadMe file that explains what everything included in the documentation is, and gives step-by-step instructions for how one can use the documentation to replicate the paper.

The rationale for this convention:

The requirement concerning the command files ensures ***complete*** replicability.

The ReadMe file is for the sake of ***realism***.

But why no data?

And do we need different specifications for cases in which data are public vs. confidential?

I have in fact written a draft of a convention based on these principles—I am calling it the DRESS convention (Documenting Research in the Empirical Social Sciences).

I have not posted it publicly, but will share it with anyone who asks me for it.

Is it realistic to think that simply defining a convention for research documentation will cause anyone to change her behavior or move us to a world like the vision with which I began this talk?

Do I have time to do a quick advertisement for Project TIER?

PROJECT TIER

TEACHING INTEGRITY
IN EMPIRICAL RESEARCH

www.haverford.edu/TIER