PROJECT TIER

TEACHING INTEGRITY IN
EMPIRICAL RESEARCH

www.haverford.edu/TIER

@Project_TIER

Documenting Statistical Research:
What We Can Learn About it By Teaching About It

Richard Ball
Co-Director, Project TIER
Associate Professor of Economics, Haverford College

Presented at the Dataverse Community Meeting
Harvard Medical School
July 11-12, 2016

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

The mission of Project TIER (Teaching Integrity in Empirical Research) is to make principles and practices of transparency and reproducibility a standard and integral component in the research training of students in the social sciences.

We ***develop practical methods and tools*** for reproducible research that suit the needs of students at early stages in their education, and prepare them to continue conducting research transparently throughout their professional careers.

Through ***a variety of outreach activities,*** we share these tools and methods with faculty from across the social sciences who are interested in introducing them to students in their quantitative methods classes and their thesis or dissertation advisees.

Project TIER was officially launched in 2013, but it developed gradually and organically out of our experience teaching introductory statistics classes for economics majors and advising senior theses at Haverford College over many years.

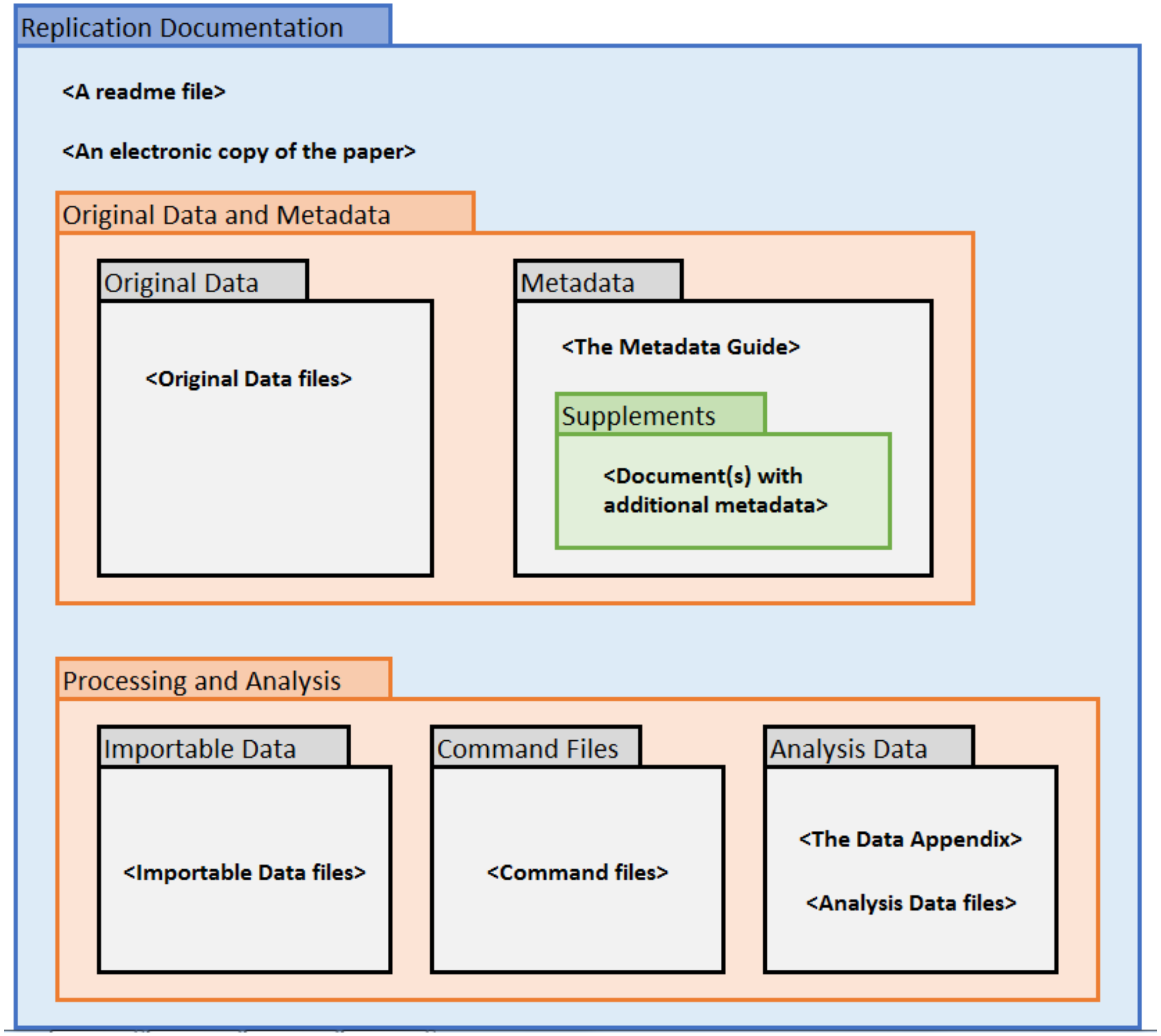
Our efforts to date have focused mainly on promoting the “TIER Protocol” for documenting an empirical research project.

The TIER Protocol consists of two main parts:

Specifications for a set of electronic files, including data, code and explanatory information, that serve as comprehensive replication documentation for a study.

Guidance about the ***process*** of conducting a study in such a way that the construction of this documentation is an integral part of the research that is carried out incrementally over the entire course of a project, rather than an afterthought tacked on at the end.

The ***specifications*** can be summarized by this figure:



The guidance about the ***process*** gives students tips about things they can do at each stage of the research so that when they have finished the project, the replication documentation is also complete.

A crucial element of this process is doing all your work in command files.

> When students finish work for a day, the command files are the tangible product of their labor and are what they should save.
>
> They need not, and probably should get out of the habit of, saving edited data files.

For the last seven or eight years, students writing papers for our introductory statistics class, as well as advisees writing empirical senior theses, have been following the TIER Protocol as they conduct their research and assemble replication documentation for their projects.

They have been doing so with high rates of success, and without undue difficulty or effort.

> On the contrary, the Protocol provides structure that tremendously increases the efficiency with which students work and reduces the difficulty of the exercise.

Adhering to the Protocol also enhances ***communication:***

--among students collaborating on a group project

--between students and the instructor/advisor

This enhanced communication depends critically on the use of some kind of web platform for sharing and archiving research documents—that is what led us to begin using Dataverse.

Project TIER's ***outreach activities to date*** have consisted mostly of

--Faculty Development Workshops held twice a year at Haverford College

Dates and details for the fall 2016 Faculty Development Workshop will be posted on our website and tweeted by the end of this month.

--Annual TIER Faculty Fellowships (for which the second cohort has just been selected) for faculty from around the country to collaborate with us in developing and disseminating methods and curriculum for reproducible research.

--Speaking at events like this

--Launching a rudimentary web page

Activities that have been initiated recently or will be initiated soon include:

--Faculty Development Workshops held at locations other than Haverford College

--Workshops for grad students in individual graduate programs

--Launching a new, professionally-designed web site to present Project TIER resources more effectively

And we are beginning to develop new resources to complement the TIER Protocol:

--Curriculum for using markup languages to make statistical work reproducible (using R Markdown as well as a similar package that works with Stata)

--Developing a structured set of extended exercises to teach students some of the data management, visualization and analysis tools they need to conduct original research

--Further integration of research management platforms—like Dataverse—into Project TIER

Implications for professional standards for replication documentation

The value of replication documentation for an empirical paper is not exclusively—indeed, probably not primarily—to make it possible to detect errors.

Instead, replication documentation can be used for more constructive purposes, namely:

> --**communicating** to the reader all the steps taken and decisions embodied in the processing and analysis of the data
>
> --making it possible for an interested reader to **experiment** with or **extend** the empirical work done for a paper

What qualities does replication documentation need to have if it is to serve those purposes?

> The documentation should include command files containing code that takes the original data for the project, processes them as necessary to prepare them for analysis, and then executes the procedures that generate the results reported in the paper.
>
> It should be well-enough organized and include sufficient explanatory material (including a read-me file) to enable a reasonably competent researcher to independently replicate the study and reproduce all the reported results—exactly, and without undue difficulty.

In broad terms, all of that is well-understood (if not always practiced).

But our experiences teaching students to document their work to a high standard have highlighted for us the importance of a few issues that are typically given insufficient consideration in discussions about professional standards:

> --The importance of a fixed directory structure (a fixed hierarchy of folders and subfolders in which the replication documentation for a study is stored).
>
> --The importance of fine-grained citations of data in their "native environments."
>
> --The ***unimportance*** of developing special policies for what replication documentation should consist of when data are confidential.