Reproducibility of Empirical Research:

Classroom Instruction and Professional Practice

Richard Ball	Norm Medeiros
Associate Professor of Economics	Associate Librarian
Haverford College	Haverford College
rball@haverford.edu	nmedeiro@haverford.edu

Presented at the ASSA Annual Meetings
Boston, January 5, 2015

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

This talk has two parts:

A description of how we teach our undergraduate students to conduct empirical research in a transparent way

A discussion of some lessons we have learned from this experience that have implications for transparency in professional research

The practices we teach our students are focused on a particular (narrow?) aspect of transparency:

computational reproducibility

Making empirical results computationally reproducible is all about

documentation

A real example of documentation of a student research paper:

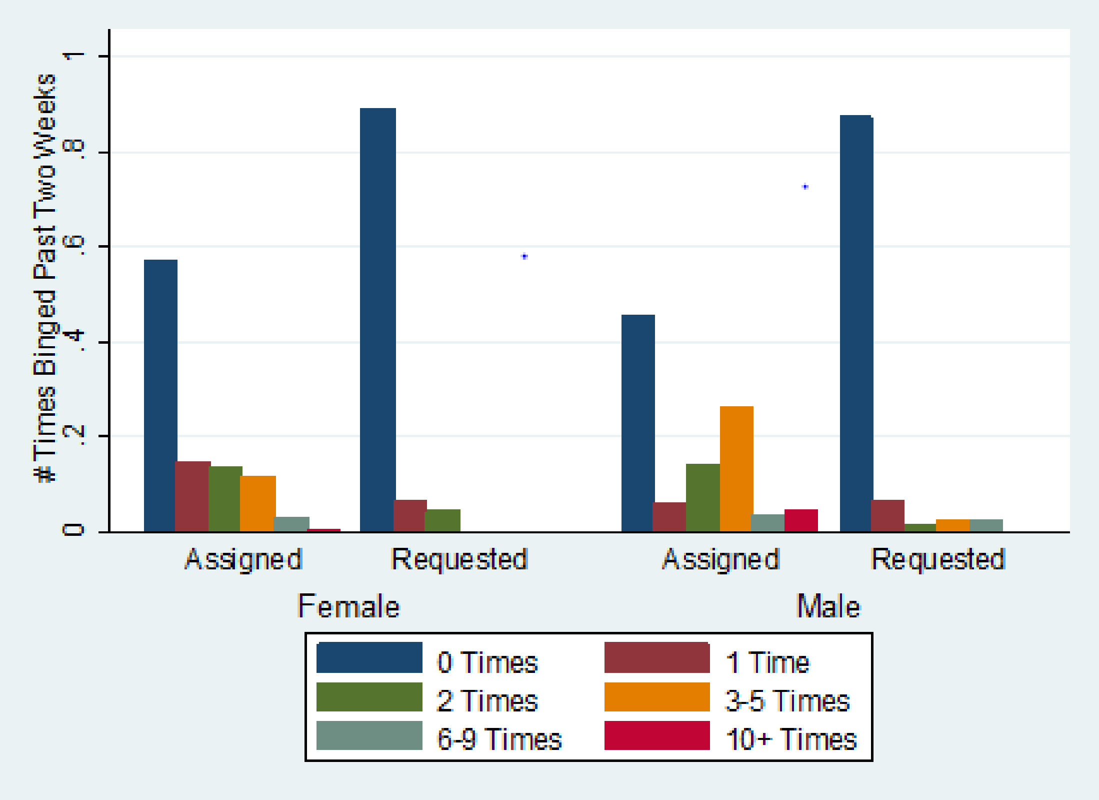
The Relationship Between Alcohol-Free Housing and Binge Drinking

Jonathan DeWitt, Dylan O’Connell and Ben Hart

Economics 204, December 6, 2013

One of the figures:

Figure 4
Drinking If Requested versus Assigned Alcohol-Free Housing



One of the tables:

```
Table 6: Difference in Proportion who Drink among Requested vs Assigned among males


Two-sample test of proportions                  No: Number of obs =      114
                                               Yes: Number of obs =       79
-----------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf.Interval]
-------------+---------------------------------------------------------------
          No |   .4561404   .0466488                         .3647104    .5475703
         Yes |   .8734177   .0374097                         .8000961    .9467393
-------------+---------------------------------------------------------------
        diff |  -.4172774   .0597963                      -.5344759   -.3000789
             |  under Ho:   .0707969    -5.89   0.000
-----------------------------------------------------------------------------
        diff = prop(No) - prop(Yes)                               z =  -5.8940
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0000         Pr(|Z| < |z|) = 0.0000          Pr(Z > z) = 1.0000
```

The documentation:

Read_Me.pdf

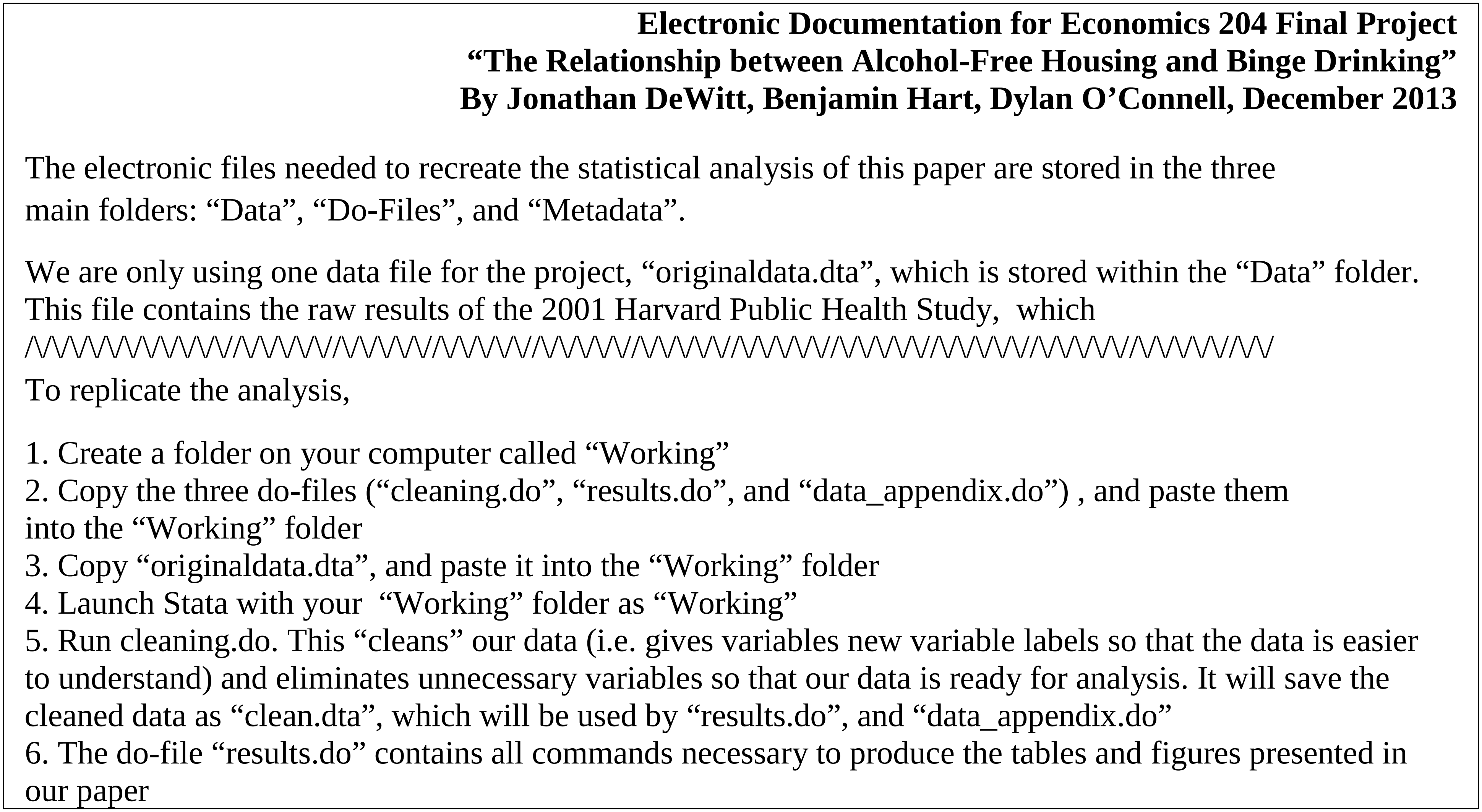
Original Data

originaldata.dta

Do-files

cleaning.do
results.do

The readme file:



Electronic Documentation for Economics 204 Final Project
“The Relationship between Alcohol-Free Housing and Binge Drinking”
By Jonathan DeWitt, Benjamin Hart, Dylan O’Connell, December 2013

The electronic files needed to recreate the statistical analysis of this paper are stored in the three main folders: “Data”, “Do-Files”, and “Metadata”.

We are only using one data file for the project, “originaldata.dta”, which is stored within the “Data” folder. This file contains the raw results of the 2001 Harvard Public Health Study, which

To replicate the analysis,

1. Create a folder on your computer called “Working”
2. Copy the three do-files (“cleaning.do”, “results.do”, and “data_appendix.do”) , and paste them into the “Working” folder
3. Copy “originaldata.dta”, and paste it into the “Working” folder
4. Launch Stata with your “Working” folder as “Working”
5. Run cleaning.do. This “cleans” our data (i.e. gives variables new variable labels so that the data is easier to understand) and eliminates unnecessary variables so that our data is ready for analysis. It will save the cleaned data as “clean.dta”, which will be used by “results.do”, and “data_appendix.do”
6. The do-file “results.do” contains all commands necessary to produce the tables and figures presented in our paper

Do-files: cleaning.do

```
**CLEANING.DO**
clear
use originaldata.dta

/*The variables we want to keep are

A1: How old are you?

A2: Are you male or female?

B8: Some universities have housing that is specially designated
As ‘alcohol-free.’ Do you live in this type of housing during
the current school year?
 /\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\//\/\/\/\/\/\/\/\/\//\/\/\/\/\/\/\/\/\//\/\/\/\/\/\/\/\/\//\/\/\/\/\/\/\/\/\//\/\/\/\/
G11: During your last year in high school, on how many occasions
did you have 5 or more drinks in a row?*/

keep A1 A2 B8 B9 B10 B11 C1 C2 C4 E2D G11

/*These variables already have labels, but we will now
relabel some with more precise and informative labels*/

label var A1 "AGE"
label var A2 "GENDER"
label var B8 "ALCFREE"
 /\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\//\/\/\/\/\/\/\/\/\//\/\/\/\/\/\/\/\/\//\/\/\/\/\/\/\/\/\//\/\/\/\/\/\/\/\/\//\/\/\/\/
 label var G11 "HSNUM5+"
 save clean.dta, replace
```

Do-files: results.do

```
**RESULTS.DO**
clear
use clean.dta
*Generate dummy variables for categories of drinking frequency
quietly tabulate C1, gen(C1_)
/\/\//\/\/\/\/\//\/\/\/\/\//\/\/\/\/\//\/\/\/\/\//\/\/\/\/\//\/\/\/\/\//\/\/\/\/\//\/\/\/\/\//\/\/\/\/\//\/\/\/\/\//\/\/
/* We now compare the drinking patterns of those in alcohol-free housing
who requested it and of those in alcohol-free housing who were assigned
to it.  We create the dummy variable B9_reqvsassign, where a 1 denotes that they
requested alcohol free housing, and 2 denotes that they were assigned to
alcohol-free housing. We then graph the drinking habits of both populations.*/
gen B9_reqvsassign = 1 if B9==1
replace B9_reqvsassign = 0 if B9==2
label value B9_reqvsassign yn
**Figure 4: Drinking if Requested vs Assigned Alcohol-Free Housing**
graph bar C1_*, over(B9_reqvsassign) over(A2) legend(label(1 "0 Times") label(2 "1 Time")  label(3 "2
Times")  label(4 "3-5 Times")  label(5 "6-9 Times") label(6 "10+ Times") ) ytitle("# Times Binged Past
Two Weeks")
** Table 6: Difference in Proportion who Drink among Requested vs Assigned among Males **
prtest C1_1 if A2==1, by(B9_reqvsassign)
```

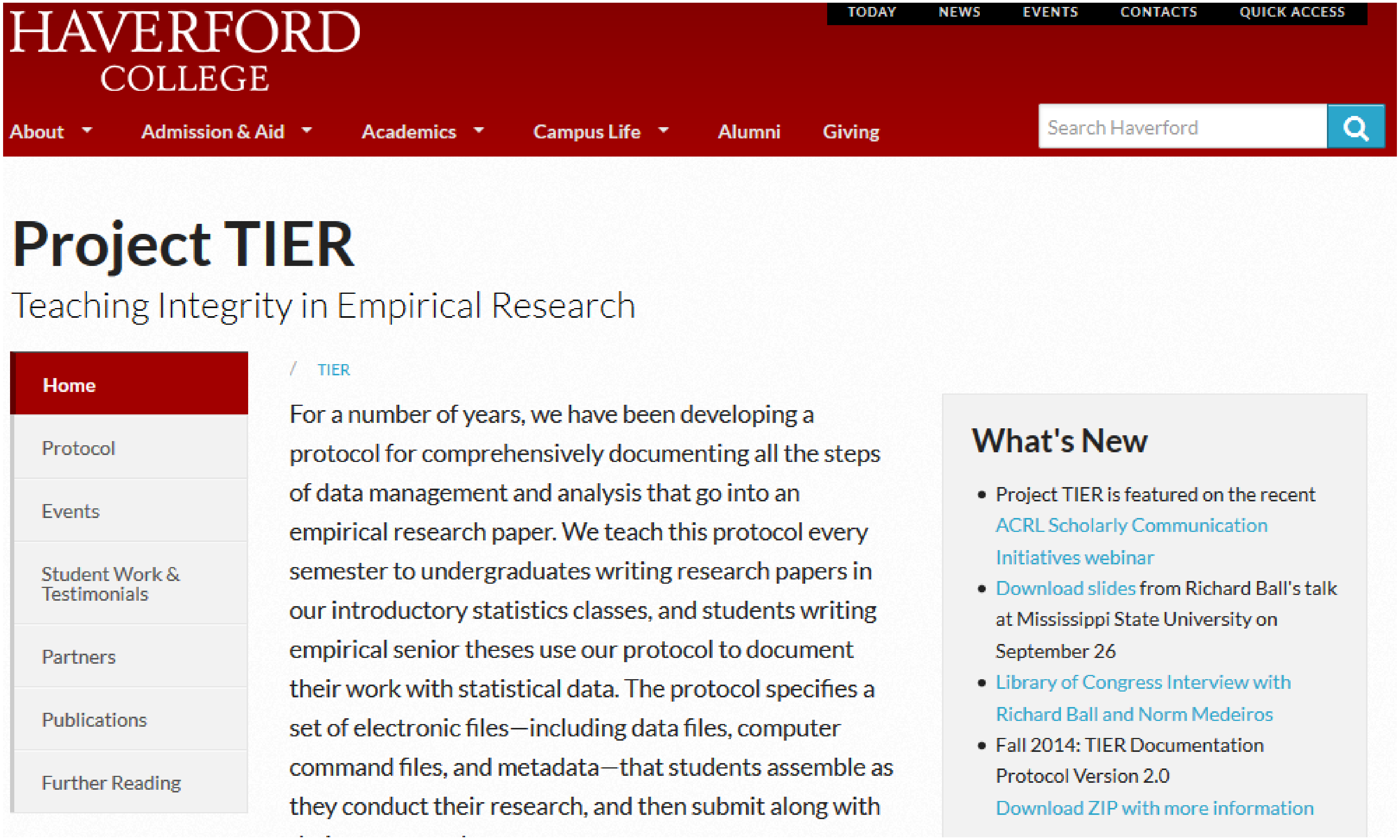
Students get it an appreciate it:

This semester I became acquainted with Stata, as well as with data documentation and analysis. As I had never utilized this tool before, I inevitably experienced occasional problems. However, by recording my data in a "do-file" for results, rather than plugging in codes into Stata, I was able to identify the error exactly… ***Rather than becoming lost in the tool and spending considerable time searching for errors, I was able to focus on the actual research and data analysis.***

…using Open Science Framework (OSF) as a platform aided the organizational structure of my research project. My entire team, including my professor and the research librarian, were able to access my team folders. OSF is organized in a way such that we had a team folder, as well as subset folders, including folders to hold our raw data, written works, imported data, data analysis and do-files. We were meticulous about dating our do-files in order to avoid confusion and so that we could readily refer back to changes that we made over time.

Overall, the combination of using correct data documenting techniques and OSF allowed me to better understand Stata and, at the same time, avoid the hassle of becoming lost in my own work… the do-files have the capability of including comments, which allows for the ready availability of my research by future scholars. ***Of course, the inability to replicate my data would cause my research to be useless and further scholarly work could not build on or add to it…***

More information: www.haverford.edu/TIER



Project TIER

Teaching Integrity in Empirical Research

For a number of years, we have been developing a protocol for comprehensively documenting all the steps of data management and analysis that go into an empirical research paper. We teach this protocol every semester to undergraduates writing research papers in our introductory statistics classes, and students writing empirical senior theses use our protocol to document their work with statistical data. The protocol specifies a set of electronic files—including data files, computer command files, and metadata—that students assemble as they conduct their research, and then submit along with

What's New

- Project TIER is featured on the recent ACRL Scholarly Communication Initiatives webinar
- Download slides from Richard Ball's talk at Mississippi State University on September 26
- Library of Congress Interview with Richard Ball and Norm Medeiros
- Fall 2014: TIER Documentation Protocol Version 2.0 Download ZIP with more information

Introducing very simple standards for documentation has fundamentally transformed the way we interact with our students as they conduct empirical research, and how we evaluate and respond to their work.

Can we draw any lessons from this experience that might be useful for the professional research community?

UPGRADING THE MEDIUM OF COMMUNICATION OF EMPIRICAL RESEARCH

A printed (or pdf) paper is inadequate

Comprehensive documentation that makes it possible for a reader to reproduce reported results is essential

--for understanding what an author did with the data

--to allow further exploration of the data and checks on the robustness of the analysis

--to facilitate cumulative progress in research

To achieve these purposes, documentation must include:

code for data processing as well as generating final results

citations or descriptions of original data that are detailed enough to allow a user to figure out how to actually get her hands on the exact data the author started with (barring confidentiality issues or other restrictions on access)

If we are given enough information to find the original data, and have the code that does all the processing and analysis, what need is there to include any processed data files (like the “final datasets”) in the documentation?

A shift in norms and expectations:

readers should use this documentation actively while reading a paper

authors should expect readers will do this, and prepare the documentation in such a way that this is feasible

Our view is that strategies for promoting those norms should focus on coordination and voluntary participation rather than sanctions and enforcement.