Introduction

William H. Knapp III

March 25, 2016

1 What Is Reproducibility?

Science is facing a crisis. Public trust in science and scientists is abysmally low. Basic science has been politicized to such an extent that governmental leaders are writing laws to keep scientists from advising governmental organizations on issues directly related to their areas of expertise. Coupled with economic downturns, it's no surprise that low levels of trust in the scientific enterprise has resulted in funding cuts that undermine national security.

Are there reasons to be skeptical of science? Absolutely. Science is a skeptical endeavor. Good scientists are skeptical of one another and of their own thought processes, theories, and hypotheses. But they're not blind skeptics. They understand the scientific method, the tentative nature of hypotheses and theories, and the value of evidence.

Unfortunately, scientists are humans faced with pressures to publish (i.e. publish or perish). I'm a fan of these pressures as they ensure that valuable resources are allocated to those who have a history of making contributions to the gradual progression of science. However, these pressures don't exist in a vacuum. When coupled with publication biases to publish interesting and statistically significant results, this pressure could create a perfect storm in which various forms of academic fraud become more attractive.

The Dutch social psychologist Diederik Staple published dozens of articles using fraudulent research practices. Although the dangers fraud poses to scientific progress and partial solutions have been known for years, the furor resulting from the huge scale of of Stapel'sfraud, has helped psychologists and other scientists to reexamine the causes of and partial solutions to academic fraud.

As Roediger's article indicates, replication is an important part of the puzzle. However as replications aren't as interesting to publishers, especially when the replication doesn't work, it's only part of the puzzle.

Another part of the puzzle is pre-registration of journal articles, in which an article is accepted based on the merits of its methods and potential contributions to the literature, as opposed to the significance of the results. The "pre" in pre-registration means that the articles are basically accepted before publication before any data are collected. This takes off pressures to massage, p-hack, or fabricate data.

The part of the puzzle we tackle here is related to reproducibility. Although reproducibility is used within many sciences to refer to all parts of the research process from data collection and to post-analytic interpretation of results, I'll use reproducibility in the sense it's used in data science (i.e. taking the same raw data, running the same, analyses, and obtaining the same results.) I'll refer to the efforts to duplicate the methods that are undertaken to determine whether previously identified effects replicate as replication.

Although I distinguish between reproduction and replication, many of the concepts we'll discuss related to reproduction directly apply to replication too.

As reproduction focuses on using the same analyses on the same data to obtain the same results, a reproducible result requires others to have access to the original data and analyses. Many journals are now requiring authors to make their data available to others during reviews of their submissions for publication and / or post-publication. These actions are part of and the open-data initiatives.

2 Opening Science

2.1 Open Formats

2.1.1 .csv files

For data to be truly open, it should be in a format that people can open and analyze without proprietary software. It should also be in a format that makes data analysis as easy as possible. Thus, we'll be using comma separated values (.csv) files to contain our data. .csv files are simple text files that can be opened in Word, Excel, LibreOffice, and text editors. As the name might imply, different values are separated by commas. Typically a csv file will start with a row of headers (i.e. titles), separated by...duh, duh, Duhhhh...headers, that indicate what data follow. Each subsequent row represents a single observation of all the relevant variables. The raw csv file for a data set might look something like the following.

subject,sex,score

1,male,74

2,female,87

3,male,82

4,female,91

5,female,78

Opening the file in a spreadsheet program might look like the following.

subject	sex	score		
1	male	74		
2	female	87		
3	male	82		
4	female	91		
5	female	78		

To create a csv file is as easy as opening up a plain text file, entering comma-separated headers on the first row, entering comma-separated values for each observation in the second row, and saving the file as a YOUPICKTHENAME.csv file.

Making things even easier. If you're used to using spreadsheets, you can enter the data directly into most spreadsheets and then ”Save As” a YOUPICKTHENAME.csv file.

2.1.2 .md and .rmd files

In addition to using .csv files for holding your data, you'll use .md files to explain the different components of your project and .rmd files to conduct your analyses. Both .md files and .rmd files are simple text files. The md in both of these formats refers to Mark Down. Programs that can process .md and .rmd files can produce formatted html or pdf from the text contained in those file. In addition to the text in these files, there are some simple commands that we'll go over in the next assignment that will allow you to use .md and .rmd files to their fullest. By sharing your data in .csv files and by sharing your analyses using .rmd files, anyone who has access to those files can reproduce your results.

2.2 Open/Free Software

2.2.1 R

R is at the core of the analyses we'll be conducting and the figures we'll be creating. R is an open-source, free statistical software package that can be extended with different packages that we'll use to create publication worthy graphs and analyze our data.

2.2.2 RStudio

Although R is the engine of our analytic car, RStudio is the body. RStudio is an opensource environment for working with R. You'll also use RStudio to edit/create the .md and .rmd files that will comprise your homework. RStudio makes working with R much easier as it contains separate sections to edit documents, work directly with R, and other functions that will make our lives easier.

2.2.3 Git

Git is also an open-source software project that is used for tracking changes to the files you're creating and editing. With Git, you'll track your changes and upload your work to an open-access online service Github.

2.2.4 Github

Git hub is an website that allows you to take projects you're tracking with Git on your own computer and share your work with others. We'll be using this service to complete your homework and also to share your senior project including, the data, the analyses, and the final products of your work: an APA manuscript and a poster.

As we'll be using this open-access online service, you and your partner will need to make a decision. Do you wish to share your identity with others? If you're planning on doing your best, it might be wise to identify your selves so you have a tangible product you can show future prospective employers or graduate school admissions committees. If you're not planning on doing your top work, you might want to use an alias that won't identify you. If you're especially concerned about your privacy, you can delete your project and even your profile from Github after the end of this term.

2.2.5 Google Docs

To make things easier for collaborating with your partner on data entry. I recommed using a Google Docs spreadsheet that you share with your partner for entering the data you collect. After you're finished entering data, you can click "File," then "Download As," and finally "Comma Separated Values."

You'll also use Google Docs to complete editing your APA style manuscript. When you're finished editing the final draft, you can click "File," then "Download As," and finally "PDF document."

2.2.6 Dropbox

To collaboratively work on your power point presentations, I recommend that you and your partner download and install Dropbox, so you can share your presentation and work collaboratively to create and edit it.

2.3 Maintaining Confidentiality

Not all data should be open. Research participants have rights that researchers need to protect. Part of these protections involve maintaining confidentiality for non-anonymous research participation. Without participants, behavioral research would come to a standstill. To provide for a pool of potential participants that is plentifully packed—sorry, I've always appreciated alliteration—it's both practical and prudent to take precautionary protections to prevent personally identifying information from being published. In other words, don't release information that would release your participants' identities.

For the atypical research project in which the predictor variables could be used to identify individual participants, keeping the data or a good portion of it closed would be necessary. For other projects, personally identifying information should be kept out of the data files. In both types of projects, forms, documents, or other media containing personally identifying information need to be protected.

Should you collect consent forms your participants sign, you should securely store them (e.g. in a locked cabinet). Many of you will receive emails containing the data you wish to analyze. You should save the emails in a password protected folder and delete the originals from the server.