A Framework for Really Reproducible Research

Scott Jackson

January 28, 2013

Contents

Part I

Conceptual framework

Chapter 1

Introduction

1.1 Motivation

We live in an exciting but perilous time for science. The rise of the internet into a mature infrastructure, the continuing advances in personal and large-scale (e.g., cloud) computing, great strides in terms of data collection and analysis of types we could barely contemplate 10 or 20 years ago, and so on, provide an exciting new global realm of scientific discovery and collaboration. On the other hand, because of various domains of economic and social change, science is also under attack. Basic scientific education (including "hard" and "soft" sciences) is in jeopardy in many spheres of public education. Funding for basic science is becoming harder to come by, and more competitive. Several severe cases of academic fraud have received a lot of exposure in the popular press. So while the potential for advances in sciences, including social sciences, has never been greater, issues of risk, accountability, and demonstrating value to society loom heavy over the academic landscape.

An absolutely critical piece for both sides of this picture is the concept of *reproducible research*. On the one hand, as possibilities for new data sources and analyses and collaborations explode, reproducibility is key, in order to maintain trust and order within the scientific community. For example, some recent advances in statistical methods that have recently become much more accessible (e.g., mixed-effects models, Bayesian analysis) are still not fully integrated into the field. This means both that some researchers are employing methods they do not (yet) fully understand, and that some journal reviewers are resistant to new methods, even if they do not have good reason to, simply because they are unfamiliar. More transparent, reproducible methods of employing these and other more novel analyses would greatly facilitate the ability to share, evaluate,

and critique these methods.

Reproducible methods promote transparency and trust. When people outside the academic scientific community can pick up and replicate analyses and results, it can help break down the "ivory tower" metaphor. It increases accountability and decreases the possibilities for fraud and scandal. If anyone with a computer can re-run and inspect for themselves some important analysis of (e.g.) climate change, voter fraud, economic disparity, health issues, etc., then there is far more opportunity to bring discourse of such topics into the realm of facts and better decisions, and out of the realm of heresay and fact-twisting partisanship. There is a growing practice of people circulating various plots and graphs, showing things like debt growth under Democrats vs. Republicans, etc. etc. But without an ability to replicate the methods (and directly inspect the data) that went into creating such graphs, there is no real reason to trust any of them. A bar graph can lie just as easily as anything else, especially if you can't see how it was made.

Within academia, there has been a growing recognition and dissatifaction of the problem of replication. Essentially, studies are rewarded (by publication and dissemination) for showing effects, while failed replications of such studies may have an extremely difficult time getting published, even though many well-done failed replications should cast significant doubt on the initial published effects. To make matters worse, replication is more difficult and resource-consuming if the original study is not very thoroughly described. By making replication easier, we can save costs and time by reducing the amount of resources wasted on failed replication attempts. There are some interesting current projects trying to address the so-called "file drawer" problem of unreported failed replications, but increased reproducibility is a critical piece of making such efforts successful.

Currently, reproducible research is a fairly hot topic. The rise in popularity of the open-source statistical software R has generated a fair amount of interest, because this software encourages and enables reproducible research in a way that the more popular commercial software packages (SAS, SPSS, etc.) do not. The ideas of *literate programming* (Knuth 1984) have spilled over from the programming world into other academic disciplines, and many of the tools developed for programmers to work with code are turning out to be useful for reproducible research in other fields. There are websites dedicated to the topic, countless blog posts on the virtues of R and Sweave for reproducible research, and in a new column in the journal *Chance* dedicated to ethics in statistics, prominent statistician Andrew Gelman started off by examining a case of non-reproducible results (in the practical sense, that the authors refused to share the data that would allow for reproduction). A boycott movement against big for-profit publishers like Elsevier and proposed legislation like the Research Works Act have

re-invigorated the dialogue about open access to publication of scientific results. So with all this current interest and influx of new tools, what's the purpose of this thing you're reading? And conversely, why should *you*, gentle reader, bother to care? As the first part of the answer, I will discuss who this book is intended for.

1.2   Who is this book for?

Full disclosure: this book is primarily a selfish effort. I am putting this system together to improve my own methods and productivity, and having it all written down in an organized way is a helpful way for me to put my thoughts together into a system that's documented and clear. But this point also gets to the heart of it: I believe that adopting more reproducible methods will lead not only to Better Science, as alluded to by the intro above, but a more efficient and productive workflow. In other words, I believe that Really Reproducible Research (more on what I mean by that in section 2) is not just The Right Thing for Science, but The Way to Get More Done and Published.

Because of this last point, I intend my primary audience (after myself) to be academic scientists. It also follows from my self-centered goals that the specifics will be geared towards academics working in linguistics, and psycholinguistics. However, I expect that people from other fields could find the discussion and implementation helpful, and easily adapted. To frame it another way, ask yourself the following questions:

- Have you ever picked up an old paper of yours and wished you could remember some detail of how the data was collected/analyzed?

- Have you ever needed to update a figure/table/statistical analysis after a change in the data or analytic procedure?

- Have you ever read someone else's paper and wished you could see exactly how they did their analysis?

- Have you ever been frustrated in how much time you spend chasing down and re-typing/re-formatting the same set of references across multiple papers?

- Have you ever had a request for your data/analysis/other details from a paper and shuddered at the effort needed to share it in an accessible way?

- Have you ever had a problem with inconsistency in a paper, where the stats are from one data set, but the figures (or summary tables, or stats

in another section) are taken from a different data set (e.g., after some additional data, or some additional data-cleaning, or something)?

- Have you ever lost track of what kinds of data-cleaning (outlier trimming, transformations, missing data, etc.) have been performed on a data set, and which ones were applied to results in a given paper or presentation?

- Have you ever gone through some laborious data-organization or analysis process (e.g., sorting/labeling/tweaking/cleaning things in Excel by hand), only to have to do it over and over when you discover mistakes or when the data changes in some way?

- Have you ever taken hours to carefully construct some kind of complex figure or diagram by hand (e.g., graph, flowchart, theoretical model, syntactic tree), only to have to re-format it for a journal submission, or a talk handout, or a PowerPoint presentation, or some other formatting issue?

If you are still in the early stages of your career, and are unsure about whether anything like this may happen to you, just do a quick poll of your advisor, more advanced students, etc. If none of these things apply to you, you are likely either (a) not an academic, (b) an academic in a non-scientific field, or (c) already doing a fantastic job doing reproducible research. But if any of these things apply, and you like the idea of doing something about it, then my hope is that this book will help.

Finally, this book does *not* assume you already have facility with programming, etc. Many of the implementation tools I'll discuss in Part II involve some level of savvy in programming, using command-line tools, and other things normally associated with steep learning curves. My intention is to present arguments for why these tools are worth the effort to learn and use, but I will start out assuming that the reader is a user of commercial products like the Windows or Mac operating systems, programs like Microsoft Word and maybe a little of Excel, and a graphical stats package like SPSS or JMP, if anything. My hope is that this book could be picked up by people early in their academic careers and applied as they go. My greater hope is that the ideas will be appealing enough and the implementation easy and effective enough that even experienced, established academics could find some utility in improving some of their habits and/or tools. People tend to get entrenched, though, so I'm not holding my breath on the latter group. But one can hope...

1.3 Goals

So what exactly do I hope to accomplish with this book? What exactly should you, the reader, expect to be able to get out of it? To return to my motivations and audience, I would like to enable linguists and psycholinguists (and others, perhaps) to produce Really Reproducible Research, from soup to nuts. Currently, there are bits and pieces of resources and ideas spread around multiple fields and websites and repositories. My purpose here is to collect what I think are the best of the best, and assemble them into a system of principles, tools, and methods that will work well together for a "complete" system of Really Reproducible Research. Additionally, many resources on reproducble research (including the website of that name) are geared primarily towards computational or statistical work, and their principles can be summed up as "share your data, include your code, and make your code legible to others." These principles are certainly relevant, but they don't capture the whole messy system of producing scientific research that is truly reproducible.

Therefore, on the one hand I aim to present a more general discussion and system for carrying out reproducible scientific research beyond "include your code," and on the other hand, I aim to provide a very specific configuration of tools geared towards carrying out reproducible research in linguistics and psycholinguistics. The book is organized with these goals in mind. In the rest of this first part of the book, the discussion will remain tool agnostic in general, although the principles discussed will end up favoring some kinds of tools over others. The goal of this part of the book is to lay out the principles and concepts for what it means to carry out Really Reproducible Research, and what the benefits and drawbacks might be.

The second part of the book makes this more concrete by spelling out a particular implementation. The implementation is partly a set of software recommendations and partly a set of workflows, procedures, and methods for doing typical research tasks in a way to support Really Reproducible Research. There will be plenty of room for customization, because I don't expect that any two researchers will want to do things in exactly the same way, but the goal is to be as specific and concrete as possible, so that you are not left wondering about how to connect the dots. Some suggestions for alternatives will be included, but I will focus on tools that I use and that I think are best for the job, and I will not go through an exhaustive review of tools I'm less familiar with.

Finally, the third part of the book is a set of tutorials designed to enable you to use the tools in the implementation. For example, I discuss Emacs and org-mode as major tools in the implementation. Most people are not Emacs or org-mode users. Both Emacs and org-mode have extensive documentation,

including books, tutorials, and tons of articles spread across the web. However, existing documentation is both more and less than what you would need to implement the system I describe in Part II. They are *more* in the sense that there are *tons* of functions in both Emacs and org-mode that may be great features and very useful, but not relevant or necessary for the system I outline here. Existing tutorials and documentation also provides *less* than what I do here, in the sense that using these tools in the specific way I describe in Part II may not be obvious, even if you worked your way through the general manuals or tutorials already available. In other words, my goal is not to teach you Emacs for all general purposes, but rather to teach you how to use Emacs in the system of Really Reproducible Research. Even if you know Emacs, there may be something useful for you in my tutorials, but there will also be lots more to learn about Emacs after you've mastered my tutorials. And as I mentioned in the previous section, I will assume that you have experience with Word, and that's about it, so you should approach the tutorials with minimal anxiety.

With these three parts, my ultimate goals are to (1) describe what I think are the critical elements of reproducible research and convince you that these are worthy and useful goals, (2) describe a concrete system for achieving reproducible research in the real world of working academia, and (3) enable readers with no knowledge of the tools I describe to learn and apply these tools in their own personal approach to reproducible research. This way, I hope that the end result of this book is not just a series of suggestions, but the actual means to implement and improve upon my idea in your own work.

Chapter 2

What is "reproducible"?

Thus far, I have only hinted vaguely at what *reproducible* really means. I have frequently used the phrase "Really Reproducible," implying that some values of "reproducible" may be less than desired. In this chapter, I will tackle the definition of "reproducible" in a more systematic way. I argue that "reproducible" is a continuum, and even more so, a two-dimensional continuum. With this understanding, we are in a better position to zero in on appropriate principles and standards in defining a target for what Really Reproducible means.

The first approximation of reproducible comes from the general idea in the scientific method that results should be able to be replicated. That is, I can present some data and an analysis, and in order for it to qualify as "good science," it should be possible for someone else to also collect similar data and perform a similar analysis and get (generally) the same result. Put another way, if no other scientist/lab in the world can get the same results you can, that's a big problem.

But when you start thinking about this seriously, it's apparent that this raises two questions. Reproducible by who? How "similar" must the data, analysis, and results be for it to qualify as "reproducible"? These are what I call the *domain* and *range* of reproducibility.

2.1   Domain: the audience

The first dimension of reproducibility is the *domain* or the *audience*. In short, *who* do you expect to be able to reproduce your work? On one end of the continuum is yourself. If you cannot reproduce your own work, how could you expect anyone else to? On the other end of the continuum is virtually anyone in the world, which is probably almost always impossible. Figure 2.1 illustrates a

few important values for the domain.

2.2 Range: the precision

2.3 The minimum

2.4 Publication standards

2.5 Wide dissemination

Chapter 3

The scientific loop

Part II

Implementation

Chapter 4

Guiding principles

Chapter 5

Summary of tools

5.1 Emacs

5.2 Org-mode

5.3 Git

5.4 LaTeX

5.5 Python

5.6 R

5.7 (Emacs) Lisp

Chapter 6

Scholarship

Chapter 7

Data collection

Chapter 8

Data analysis

Chapter 9

Sharing

Chapter 10

Collaboration

Chapter 11

Putting it all together

Part III

Tutorials

Chapter 12

Walkthrough: traditional linguistics paper

Chapter 13

Walkthrough: psycholinguistics paper

Bibliography

Knuth, D.E. (1984). "Literate programming". In: *The Computer Journal* 27.2, pp. 97–111.