Transparent and Reproducible Research with R

Daniel Anderson[1] & Joshua Rosenberg[2]

[1] University of Oregon
[2] University of Tennessee, Knoxville

Correspondence concerning this article should be addressed to Daniel Anderson, 5262 University of Oregon. E-mail: daniela@uoregon.edu

Proposal narrative: AERA Professional Development Proposal

The purpose of this proposal is to provide participants with an introduction to tools for open, transparent, and reproducible analysis workflows. Specifically, we discuss R Markdown for integrating text with code and *GitHub* for collaboration and documentation of the project history.

Prerequisite skills or knowledge needed for course participation

All participants should have an interest in conducting open and transparent analyses. This training will be most useful to those with experience planning, carrying out, and writing up the results of a research project. We will specifically discuss moving from existing frameworks to reproducible frameworks using R and R Markdown.

All participants should also have at least a basic familiarity with R and be comfortable with the idea of working in a coding environment. Note that the presenters have partnered with Datacamp (https://www.datacamp.com) to help provide a platform for less experienced users to get "up to speed" prior to the training. Datacamp is an online learning platform for R (and related data science technologies) that includes direct instruction and opportunities to practice and apply the learned skills, all within an online platform. All participants will have access to the full suite of Datacamp modules for one month prior to, and one month following, our in-person training.

Users with less experience with R will be asked to complete the Introduction to R, Working with the RStudio IDE: Part 1, and Introduction to the *tidyverse* modules prior to the training. Following the training, we recommend all participants re-visit the skills they have learned by completing the Reporting with R Markdown module.

Target course participants

Target audience includes graduate students, emerging or early-career researchers, and continuing researchers interested in their own or their students'/trainees' work becoming more open, transparent, and reproducible. Because both an overview of ideas related to reproducibility as well as a number of quantitative and computational tools and approaches will be described, participants with less experience can benefit from developing a more conceptual understanding of open science; they can turn to the tools later on. This training is aimed at beginners who have little to no experience using R Markdown or version control with *git*/*GitHub*, but who feel comfortable learning code and at least a basic understanding of R.

Rationale

The basic premise of reproducible research is that quantitative analyses should be conducted and documented with sufficient clarity that independent researchers could reproduce all the results, exactly. Initially, this may sound relatively straightforward—of course research findings should be reproducible—and we may like to think the most research in education adheres to these principles. Unfortunately, this is generally not the case. For example, in a large-scale review of growth models published from 2007-2012 across 47 education and psychology journals, Stevens, Nese, and Tindal (2013) found that documenting

even relatively routine procedures, such as how much missing data were involved and how the missing data were handled, was extraordinarily difficult. Indeed, in the vast majority of cases, this information was simply missing. These findings were likely the result of, at least in part, researchers attempting to summarize their study without getting bogged down in the details. The findings, however, are congruent with Buckheit and Donoho (1995), who argue "an article about computational science in a scientific publication is **not** the scholarship itself, it is merely the **advertising** of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures" (p. 5, emphasis in the original). This may seem a somewhat extreme view, but it clearly articulates that a journal article on its own is generally insufficient for the accumulation of scientific evidence. That is, it is difficult to build off the work of others, or verify study results, if you only have access to the published findings. It is important to note as well that reproducibility to not imply "correctness", but rather transparency in process. Indeed, part of the reason reproducible research is so important is that it allows other researchers a means to verify the process, analysis, and ultimately the validity of study findings.

Conducting reproducible research. Considerable recent attention has been paid to open and reproducible research in science generally (Bartling & Friesike, 2014; National Academies of Sciences & Medicine, 2018), but also in educational research (Cook, Lloyd, Mellor, Nosek, & Therrien, 2018; McBee, Makel, Peters, & Matthews, 2017; Zee & Reich, 2018). What is often lacking, however, is a clear description and tutorial on how to actually begin engaging in open and reproducible research. This training seeks to fill that gap by providing an initial introduction to tools to help make the process more tractable. In particular, we advocate for *conducting science publicly* through the publication of living code that is modified and updated as the project matures, along with *literate programming* a concept introduced by Knuth (1984) that weaves substantive text from the manuscript with analysis code. Although literate programming has its own learning curve, leading to an initial dip in productivity, it can eventually lead to massive gains in efficiency by all tables, figures, and in-text references to statistics (e.g., sample means) updated automatically each time the document is rendered. There is therefore no hand entering of model results into tables, which can be error prone, and any tweaks to the model or data (including new data be added to the research) results in the entire document being updated automatically. The manuscript is therefore *dynamic* relative to the analysis (Xie, 2016). This process may sound complicated to implement, and it was even a few short years ago. Yet, the toolkit for producing dynamic, reproducible documents is rapidly expanding and is now far more accessible for the applied researcher. When this process is paired with a version control system such as *git*, and made publicly available through platforms such as *GitHub*, the project and process is far more open and transparent. Importantly, however, our training also discusses how to make specific parts of the project (e.g., the raw data) *not* available publicly to ensure research participant privacy.

Learning objectives

List and clearly define the learning objectives and purpose(s) of the course.

- Understand why reproducibility is an increasingly important consideration for educa-

tional researchers
- Know about some of the efforts to make reproducible research more feasible and better supported within and outside of educational research
- Learn about tools, particularly those related to the R software environment, for carrying out and distributing reproducible research
- Build a network of others interested in reproducible research through involvement in a mailing list (not sure this is remotely a good idea :))

Course content

Describe the topics and issues that the course examines. This should include a description of the course structure (i.e., lecture, small group interactions, hands-on demonstrations), overview of the course, discussion of the course focus, and an overview of the planned activities.

References

Bartling, S., & Friesike, S. (2014). *Opening science: The evolving guide on how the internet is changing research, collaboration and scholarly publishing*. Springer.

Buckheit, J. B., & Donoho, D. L. (1995). Wavelab and reproducible research. In *Wavelets and statistics* (pp. 55–81). Springer.

Cook, B. G., Lloyd, J. W., Mellor, D., Nosek, B. A., & Therrien, W. (2018). Promoting open science to increase the trustworthiness of evidence in special education.

Knuth, D. E. (1984). Literate programming. *The Computer Journal*, *27*(2), 97–111.

McBee, M., Makel, M., Peters, S. J., & Matthews, M. S. (2017). A manifesto for open science in giftedness research.

National Academies of Sciences, Engineering, & Medicine. (2018). *Open science by design: Realizing a vision for 21st century research*. Washington, DC: The National Academies Press. doi:10.17226/25116

Xie, Y. (2016). *Dynamic documents with r and knitr*. Chapman; Hall/CRC.

Zee, T. van der, & Reich, J. (2018). Open education science. *AERA Open*, *4*(3), 2332858418787466.