Evolution of Statistical Software and Quantitative Methods in Educational Research

Brandon LeBeau[1] & Ariel Aloe[1]

[1] University of Iowa

Author Note

Correspondence concerning this article should be addressed to Brandon LeBeau, Psychological and Quantitative Foundations, University of Iowa, Iowa City, IA 52245. E-mail: brandon-lebeau@uiowa.edu

Abstract

Statistical software is the enabling tool of quantitative research studies and the availability and use of the software can greatly shape which methods are used by researchers. Software that is more accessible is likely to have more users and the methods implemented within the software limits the methods accessible to researchers. Open source software, (e.g. R), has reduced these barriers by making cutting edge statistical methods available to researchers through add-on packages. In addition, the idea of reproducible analyses has grown significantly within the statistics and medicine disciplines. This paper aims to explore the evolution of statistical software within educational research using a research synthesis to establish the state of affairs.

Keywords: Research Synthesis, Statistical Software, Quantitative Methods

Evolution of Statistical Software and Quantitative Methods in Educational Research

Objectives

The purpose of this paper is to explore the evolution (or lack thereof) of statistical software usage over time in education. As this usage is likely tied closely to the methods they are employed, the interaction between software usage and quantitative research methods will also be explored. Research synthesis methods will be used to explore these trends over time in published educational research journals.

Theoretical Framework

Statistical software is the enabling tool to performing applied data analysis. Statistical methods that are implemented within software will increase their usage (particularly if the software is also user-friendly) by applied analysts and are likely taught more frequently in methodology courses at universities. However, the user-friendly nature of software (i.e. point and click graphical interfaces, ability to manipulate data by hand) also can severely limit the ability for research to be reproducible; a recent topic of intense discussion in biostatistics, medicine, and pyschology (Asendorpf et al., 2013; Ioannidis, 2014; Iqbal, Wallach, Khoury, Schully, & Ioannidis, 2016; Roger D. Peng, 2009; Roger D Peng, 2011; Stodden, 2012). The reproducibility crisis has pointed the finger at statistical software more directly with a strong push in some disciplines for analyses to be script based and posted with the published journal article, often described as the gold standard.

This idea of reproducibility has not seemed to fully enter the educational research domain. SPSS is likely the most common statistical software program used in educational research. Although SPSS has many common and advanced statistical techniques and it is possible to have a reproducible analysis, the default behavior within this program is often not script based and can create bad habits (i.e. editing raw data directly in the gui interface, running analyses without saving a script, etc.). Statistical software that is primarily command line, programs such as R, Python, or SAS, offer easier reproducible frameworks as

all data manipulations or analyses are saved in scripts that can be re-ran in the future. A data script can be thought of as a cockpit flight recorder in which every single step that was done to the original data going from data collection to final tables and figures was script based. Under a reproducible framework, the raw data are never altered directly, they should always be altered programmatically through a script. This keeps a log of the data manipulations that happened in the data analysis cycle. For example, R has packages that aid in the ability to create living data documents that contain text and analysis code within a single document (see Allaire et al., 2017; Xie, 2015, 2017).

The reproducible analysis framework has many advantages, including a transparent analysis process that could be validated by others or even simply the ability to investigate the data analysis completed months or years previously. Unfortunately, the current academic research framework has many barriers that limit the reproducibility. First, applied researchers may not be users of primarily command line or script based statistical software. This limits the ability to create a reproducible framework from the start. Secondly, researchers are not incentivized to conduct an analysis in a reproducible framework. Namely, the publish or perish aspect of academic research limits the sharing of statistical code partly due to the increased chance of criticism upon evaluation of the code used for the analysis. Finally, many journals and even the American Psychological Association (APA) publication manual (American Psychological Association, 2010) states that common software or programming languages need not be cited. This could even be interpretted by some as not needing to mention. Unfortunately, if the software used for a data analysis is not reported, the ability to recreate the analysis drops even more due to differences in estimation, handling of missing data, or other software specific settings.

This paper aims to explore the state of affairs in statistical software usage in education. Particular attention will be made to which software is currently being used in published educational research as well as how this has changed over the last twenty years. Secondly, this paper also aims to explore how frequently open-source software tools are used and to

explore evidence of reproducible analysis framework being implemented. These aims will be explored using research synthesis methods.

Research Questions

1. To what extent has the statistical software usage shifted over time in published analyses?

   - If there is evidence of a shift, is there evidence this shift differs based on quantitative method or journal?

2. To what extent are published analyses citing statistical software?

   - Has this changed over time and across journals?

3. To what extent are open-source software tools used?

   - Is there evidence of reproducible analyses being employed?

Methods

Research synthesis methods (Cooper, 2016) will be used to explore the evolution of statistical software and quantitative methods in educational research. More specifically, the statistical software used for the analysis will be coded in additional to the specific quantitative methods (i.e. linear regression, hierarchical linear model, etc.). Additional meta data will also be coded including, journal, article title, author information, article keywords, year published, and any mention of supplementary materials. This information will be used to explore descriptive trends in the data over time, by journals, and methods.

The research synthesis will gather data from a handful of education journals that primarily publish empirical data analysis. The search will not include journals that the primary focus is methodological, the use of software in these journals would likely be a different population than those that are data analytic in nature. Therefore the following journals were selected to be searched from 1995 onward:

- American Educational Research Journal (AERJ)

- Educational Researcher (ER)

- Educational Evaluation and Policy Analysis (EEPA)

- Higher Education (HE)

- Journal of Educational Psychology (JEP)

- Journal of Experimental Education (JEE)

- Journal of Teacher Education (JTE)

- Journal for Research in Mathematics Education (JRME)

- Sociology of Education (SE)

Data and Software

All journal articles published between 1995 through 2016 will be organized into EndNote. Within EndNote, the find pdf feature will be used to gather the published documents from each journal. This pdf database will then be searched using the *pdfsearch* R package (LeBeau, 2016, R Core Team (2017)). This package allows for keyword searching directly within pdf documents. This will be the primary data collection method. The software keywords searched for can be seen in Table 1. Table 1 also shows keywords to be used to search for statistical models and estimation methods.

Table 1

Search keywords used in search of published journal documents

Search	Keywords
Software	"SPSS Statistics", "SPSS Modeler", "SPSS", "R", "R-project", "R project", "SAS", "JMP", "STATA", "MATLAB", "Statistica", "Statsoft", "Java", "Hadoop", "Python" , "Minitab", "Systat", "Tableau", "Scala", "Julia", "Azure Machine Learning", "Mplus", "LISREL", "AMOS", "HLM[0-9]", "HLM [0-9]"
Statistical Models	Regression, Hierarchical Linear Model (HLM), Linear Mixed Model (LMM), Multilevel Model, Analysis of Variance (ANOVA), Structural Equation Modeling (SEM)
Estimation	Maximum Likelihood, bayesian,

A handful of articles will be randomly selected to be coded manually by reading the document to evaluate the accuracy of coding from the *pdfsearch* package.

Initial Results

Currently, published journal articles between the year 2000 and 2014 for AERJ, EEPA, and JEE have been obtained and the software keyword search has been performed with the *pdfsearch* package. Table 2 reports the number of unique published articles for each journal

that contained at least one software match. The table shows that many published journal articles do not list the software used for the analysis. Some of this may be due to the APA publication manual stating that "reference entries are not necessary for standrad software and programming languages… and even SAS and SPSS" (American Psychological Association, 2010, p. 210). If the goal of analyses are to be transparent and reproducible, this is bad practice. The larger percentage of articles returning a match for JEE may be due to this journal serving a dual role, part methodological part applied analysis. Methodological articles may be more likely to state which statistical program was used to perform the simulation or which language a new algorithm was implemented in.

Table 2

Number of citations with mention of software by journal

Journal	Count	Number of Citations Obtained	Percentage
AERJ	158	443	35.67
EEPA	77	188	40.96
JEE	112	208	53.85

Figure 1 shows the statistical software matches across all articles in the current sample. This figure shows that R had the most matches followed by SAS and SPSS. Other more specialized software follows these general statistical languages.

When exploring the differences across the three journals in the current sample, some notable differences appear (see Figure 2). SAS, SPSS, LISREL, and Mplus are more common in JEE compared to other software programs. EEPA is more likely to use Python, Julia, or STATA and AERJ has a more consistent distribution across the sofware programs.

Prior to AERA, more search terms, more journals, and gathering more article meta data will be included in the research synthesis. These additional data sources will allow a better understanding of the current state of affairs in software usage, reproducible analysis, and software citation/mention within educational journals.
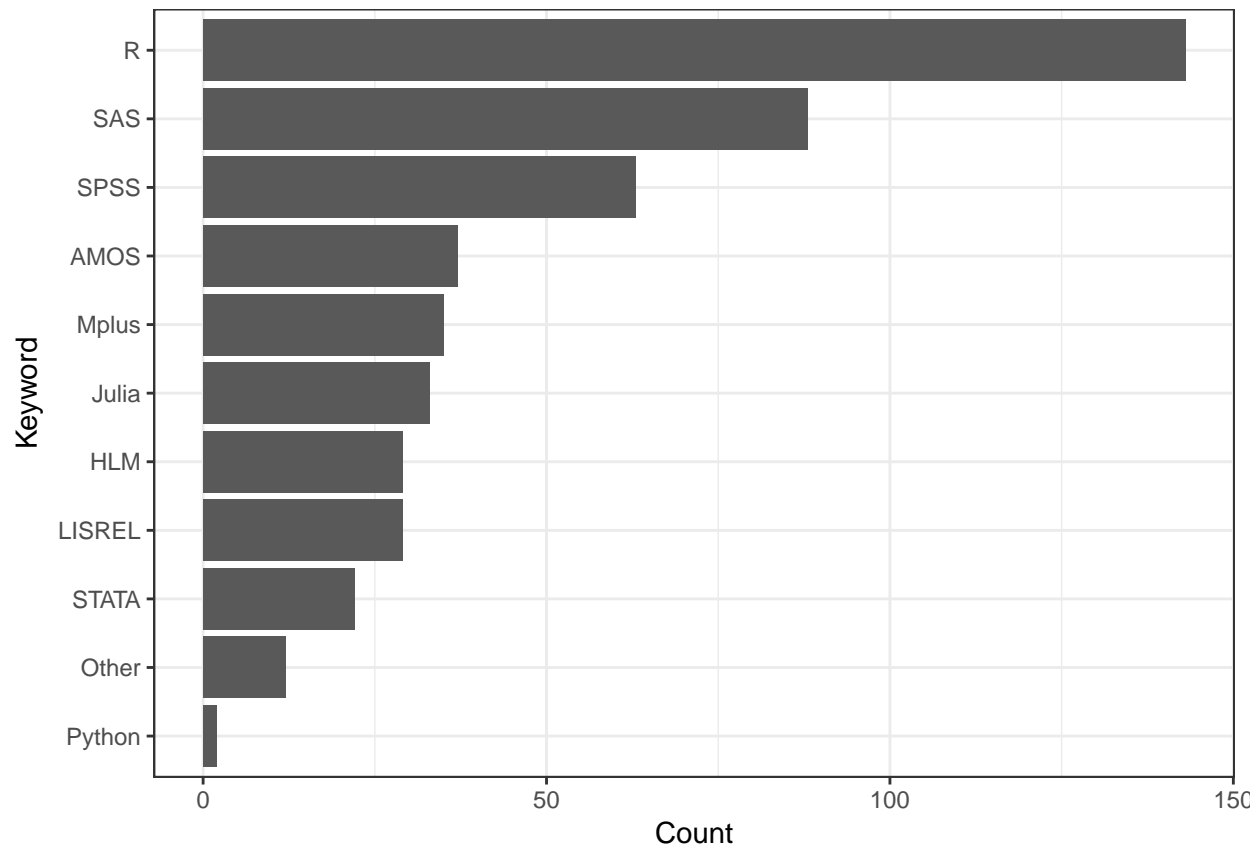
Figure 1. Software count overall

Scholarly Significance

Understanding the state of affairs in software usage in education is important to help disentangle and highlight the importance of citing statistical software and a reproducible research framework. Much more attention has been given in other content areas such as biostatistics, medicine, and psychology and the same standards should apply in education. Recommendations will be given as to which information regarding statistical software should accompany journal articles. This study also aims to go a step further than others (see Muenchen, 2017) in that the software usage will be explored with respect to journals and also quantitative methodology. This will allow for a deeper exploration of potential interactions between software usage and citations across different journals and methodologies over time.
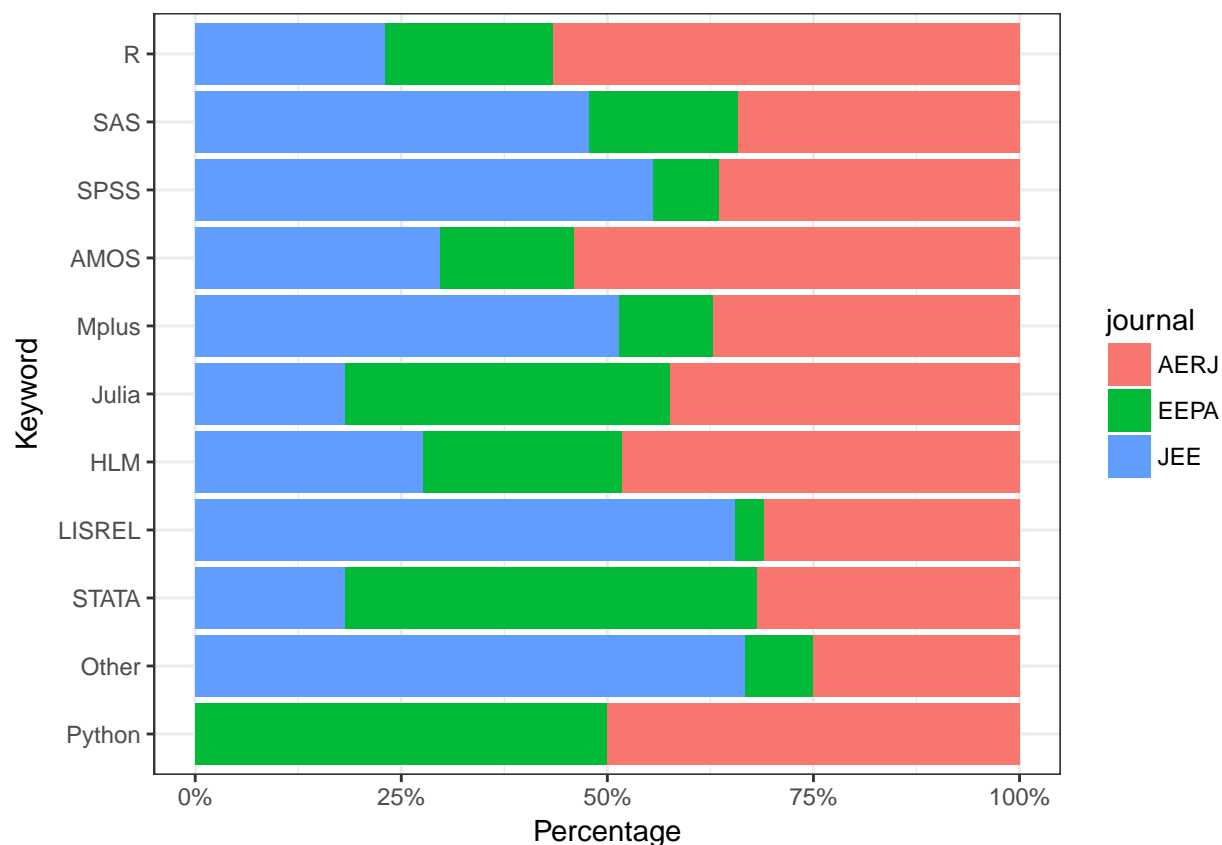
Figure 2. Software counts by journal

References

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., . . . Arslan, R. (2017).
Rmarkdown: Dynamic documents for r. Retrieved from
https://CRAN.R-project.org/package=rmarkdown

American Psychological Association. (2010). *Publication manual of the american
psychological association*. Washington, D.C.: American Psychological Association.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., . . .
others. (2013). Recommendations for increasing replicability in psychology. *European
Journal of Personality*, *27*(2), 108–119.

Cooper, H. (2016). *Research synthesis and meta-analysis: A step-by-step approach* (Vol. 2).

Sage publications.

Ioannidis, J. P. (2014). How to make more published research true. *PLoS Medicine*, *11*(10), e1001747.

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., & Ioannidis, J. P. (2016). Reproducible research practices and transparency across the biomedical literature. *PLoS Biology*, *14*(1), e1002333.

LeBeau, B. (2016). *pdfsearch: Search Tools for PDF Files*. Retrieved from https://github.com/lebebr01/pdfsearch

Muenchen, R. A. (2017). The popularity of data science software. Retrieved from http://r4stats.com/articles/popularity/

Peng, R. D. (2009). Reproducible research and biostatistics. *Biostatistics*, *10*(3), 405–408. doi:10.1093/biostatistics/kxp014

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Stodden, V. (2012). Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science & Engineering*, *14*(4), 13–17.

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from http://yihui.name/knitr/

Xie, Y. (2017). *Knitr: A general-purpose package for dynamic report generation in r*. Retrieved from http://yihui.name/knitr/