Not So Fast My Friend: The Rush to R and the Need for Rigorous Evaluation of Data Analysis and Software in Education

Michael Harwell
University of Minnesota

Commercial data analysis software has been a fixture of quantitative analyses in education for more than three decades. Despite its apparent widespread use there is no formal evidence cataloging what software is used in educational research and educational statistics classes, by whom and for what purpose, and whether some programs should be recommended over others. This paper argues that the rise of the R *data analysis software has intensified the need for rigorous evaluations of these programs to identify their strengths and weaknesses in ways that provide educators with guidance in choosing programs. Examples of research activities to produce a literature to guide these choices are described.*

Commercial data analysis software such as SPSS (SPSS Inc., 2011), SAS (SAS Institute, 2012), and Stata (2013) appears to have played an important role in quantitative analyses in educational research for more than three decades. This software has also played an important role in training students in quantitatively-oriented education classes such as those in statistics and measurement (Curtis & Harwell, 1998).

The last 15 years have seen the rise of an important competitor to these commercial programs known as R. R is "… an integrated suite of software facilities for data manipulation, calculation and graphical display" (R Core Team, 2013, p. 2). Essentially, R is a repository of programs (called packages on the R website) that perform specific analyses such as ANOVA, multilevel modeling, and latent class analysis, support data manipulation, and produce data plots. These packages are typically written by statisticians, submitted to the R Core Team and vetted, and then made

available for use on the R website. Individuals writing the packages do not receive financial compensation for their work.

R became available publicly in 1995 as open source software, meaning that users can modify the software and distribute it without restrictions as long as they comply with licensing requirements. R appears to be widely used in academic settings but less so in business settings where commercial software still seems to be dominant (Muenchen, 2014). Wegman and Solka (2005) provide a short history of the origins of several data analysis programs including R.

The central role data analysis software plays in educational research and in many education classes, and the rise of R as a competitor to commercial software, suggests it is important to examine literature documenting the use and impact of these programs in education. The rationale for doing so is that these programs vary in their capabilities, technical support available to users, user friendliness, and in some cases accuracy of results (e.g., Keeling & Pavur, 2007), and these differences likely affect the extent to which software meets the needs of educators and their students. Thus there is a need to identify which programs are being used, by whom and for what purpose, and whether some programs should be recommended over others.

What Data Analysis Software is Most Frequently Used in Education?

Remarkably, there appears to be no evidence cataloging the use of data analysis software in education in the past 20 years, and only modest evidence of its use in non-educational fields during this time (e.g., Dembel, Partridge, & Geist, 2011). Perhaps the best known current source for tracking the popularity of data analysis software is Muenchen (2014), who uses benchmarks like sales and downloads and internet discussion groups such as Stack Overflow which is a

"..website for professional and enthusiast programmers" (http://stackoverflow.com/). An examination of Muenchen's results, which are updated several times per year, provides evidence that the popularity of R has grown dramatically in the past several years, whereas the popularity of commercial software like SPSS, SAS, and Stata has remained steady or declined. For example, Muenchen (2014) reported that as of May 2013 the number of internet blogs devoted to these programs was eight (Stata), 40 (SAS), 0-3 (Others which presumably includes SPSS), and 452 (R). Still, which data analysis software is most widely used in education is unclear.

Relatedly, there are no rigorous evaluations of the capabilities, technical support, or user-friendliness of these programs. The few existing comparisons of data analysis software are informal and focus on commercial software (e.g., Prvan, Reid, & Petocz, 2002). On the other hand, there is plenty of informal evidence that many educators and students have switched to R within the past 10 years and that this movement has intensified in the past five years. Perusals of educational conference papers and published research articles, the appearance of R in educational statistics course syllabi posted online, and the number of online R resources with a focus on educational researchers such as tutorials (e.g., http://faculty.smu.edu/kyler/training/sera_r_2012/R_intro_SERA_2012_4up.pdf) and how-to documents (e.g., http://www3.nd.edu/~kkelley/publications/chapters/Kelley_Lai_Wu_Using_R_2008.pdf), are consistent with evidence of its dramatic growth provided by Muenchen (2013). This evidence is more qualitative than quantitative but suggests that there is a rush to R by many educators.

Why Has R Grown in Popularity?

The apparent growth in R's popularity can be attributed to four complementary factors. First, and probably most important, is that R is free, in contrast to its commercial

competitors. This feature of R plays out in multiple ways that likely enhance its popularity. For example, many academic research staff, faculty, and students need data analysis software in their research and classes but cannot afford to purchase commercial software. The fact that R can be downloaded for free onto a PC and is available at the user's convenience rather than having to use school computing facilities that provide commercial software, is a very attractive feature.

Second, R offers packages that perform advanced analyses not found in most commercial data analysis software. An examination of the 5,166 packages available in R as of this writing (see http://cran.cnr.berkeley.edu/) provides compelling evidence of the depth and breadth of available analyses. For example, R offers a package that performs multilevel modeling of binary repeated measures data that is unavailable in commercial programs like the HLM multilevel modeling software (Bryk, Raudenbush, & Congdon, 2011), and a package to synthesize correlation matrices in a meta-analysis that is similarly not available in commercial software.

Third, many R programs provide output that is quite detailed and as such may enhance data-analytic activities such as model checking. Fourth, R graphing programs have a well-deserved reputation of producing professional-looking graphs or figures many of which are virtually camera-ready for publication. By comparison, the graphs produced by many commercial data analysis programs are coarse and certainly not camera-ready. Collectively these features offer researchers a powerful set of data-analytic tools and help to explain the apparent growth in popularity of R in education.

Should Educators Abandon Commercial Data Analysis Software and Embrace R?

The advantages of R and its growing popularity raise the question of whether educators should abandon commercial

data analysis software and embrace R. In considering this question it is important to examine two factors.

The lack of evidence documenting how well data analysis software meets the needs of educators.

As noted above there are no rigorous evaluations documenting how well or how poorly R or any data analysis software meet the data analysis needs of educators. This is important because differences in the capabilities, technical support, and user-friendliness of these programs are likely relevant to the needs of educators and their students. However, the absence of rigorous evaluations of these programs means there is no formal basis for recommending one program over another. This extraordinary gap in the literature has certainly not deterred the R community from passionately arguing in various settings such as blogs and discussion groups that R is superior to commercial software and should be embraced (Similar passion for commercial data analysis software does not seem to exist). But advocacy and evidence are not the same.

The ability of R and the R community to support the learning and work of educators and their students.

A second factor in considering whether educators should abandon commercial data analysis software and embrace R is the ability of R and the R community to support their learning and work and that of their students. This factor speaks to the likelihood that many in the R community and many educators have different views of data analysis software that will impair the ability of the former to support the latter.

Specifically, many in the R community seem to view data analysis software as more than a tool; it is a partner of sorts, one that facilitates interactions between user and software in ways that lead to a deeper understanding of important data patterns than is typically possible with

commercial software. The comprehensive and detailed data manipulation, data analysis, and data plotting capabilities of R arguably facilitate such interactions. The focus of the R community on ever more complex statistical methods and the almost daily inclusion of new R packages may also support such interactions. In this view problems with finding adequate technical support or concerns over the lack of user-friendliness of R are trumped by the promise of deep understanding.

On the other hand, many educators appear to view data analysis software as a tool not a partner, one that should be relatively comprehensive in its data manipulation, data analysis, and data plotting capabilities, provide adequate technical support, and be at least moderately user-friendly. Educators with this view may find problems with available technical support for R and concerns over its user-friendliness are not trumped by the promise of a deeper understanding of data patterns. It's likely that educators and students who are familiar with commercial software such as SPSS are likely to find that R is comparatively much (much) more difficult to learn and use and the technical support for doing so is much (much) less useful. The primary reason is that R resembles a computer programming language such as Fortran (1995), and educators and students familiar with software such as SPSS or SAS are likely to find the complex and terse nature of R code and supporting documentation for learning and using R less than friendly.

How does R fare if only simple analyses are needed? Below are R commands to perform descriptive analyses for an outcome variable y as well as a two sample t-test using the independent variable sex after reading in an SPSS datafile:

```
install.packages("e1071")
library(e1071)
install.packages(pkgs="foreign",dependencies=TRUE)
```

```
library(foreign)
Rdataex <- read.spss("Rdataex.sav",to.data.frame = TRUE)
plot(density(Rdataex$y))
summary(Rdataex$y)
t.test(y~sex, data=Rdataex, var.equal = TRUE)
```

The first five lines are commands to read in R programs and an SPSS datafile called Rdataex. These commands resemble computer code and must be typed because R has a limited graphical user interface (GUI), in contrast to commercial software like SPSS, SAS, and Stata.

Educators unfamiliar with R might argue that the above commands do not seem that unfriendly, to which a pithy response would be "Try to use them with your data and see what happens. And then try to find technical support that enables you to fix the problems you encountered." More concrete evidence of the disagreeable nature of R can be found at the Holy Grail of unfriendly R code: The document http://cran.r-project.org/doc/contrib/Short-refcard.pdf on the R website which provides a reference card (i.e., cheat sheet) for novices.

Fortunately there are resources to help users decipher the hieroglyphic nature of R, such as ***simpleR*** (Verzani, 2002), *R for dummies* (Meys & de Vries, 2012), *Discovering statistics using R* (Field, Miles, & Field, 2012), the RKWard website (http://sourceforge.net/apps/mediawiki/rkward/index.php?title=Main_Page), a growing number of GUIs for R (http://cran.r-project.org/web/packages/Rcmdr/index.html), and free online tutorials (http://www.r-bloggers.com/list-of-free-online-r-tutorials/). An examination of these resources suggests that their helpfulness depends heavily on a user's statistical background and to a lesser extent their computer skills; educators and students who have strong backgrounds in these areas will likely find

these materials helpful after investing some time and effort, whereas those with more modest statistics backgrounds may see little return on even substantial investments of time and effort to learn to use these resources. The reason is simple: Technical support for R seems to be largely focused on users with strong statistics backgrounds.

The explanation for the uneven technical support for learning and using R probably lies in R's origins, the complexity of R reflected in the breadth of available data analysis packages, its non-profit status, and intended user audience. R packages and technical support are written to support the work of an audience of statisticians not a broad user audience with variable statistics skills. It seems self-evident that a company would probably not make a profit selling R because of the uneven technical support and its demonstrable lack of user-friendliness. The nature of R's primary audience, coupled with the absence of a profit motive, also means that there is little pressure on the R community to improve the technical support and user-friendliness of R.

Mitchell (2007) perfectly captured the problematic nature of R:

> I regret to say that I have had enormous difficulties learning and using R. I know that R has a great fan base composed of skilled and excellent statisticians, and that includes many people from the UCLA statistics department. However, I feel like R is not so much of a statistical package as much as it is a statistical programming environment that has many new and cutting edge features. For me learning R has been very difficult and I have had a very hard time finding answers to many questions about using it. Since the R community tends to be composed of experts deeply enmeshed in R, I often felt that I was

> missing half of the pieces of the puzzle when reading information about the use of R (it often feels like there is an assumption that readers are also experts in R). I often found the documentation for R quite sparse and many essential terms or constructs were used but not defined or cross-referenced. (pp. 24-25)

An advocate of R might offer the diplomatic response that learning to use R is consistent with knowing what you are doing statistically, and if you do not know what you are doing learning any data analysis software is problematic. This is probably true but what arguably sets R apart from commercial software like SPSS, SAS, and Stata is the limited technical support for users who need it the most, a problem that is exacerbated by R's lack of user-friendliness. What could account for this state of affairs? Follow R-oriented blogs and discussion groups and an explanation emerges, one that raises additional concerns about an embrace of R by educators. Burns (2007), in responding to Mitchell's critique, captured this perspective:

> Though statistics is vast, I'll simplify it to two extremes. There is statistics in the lesser sense: "I need to find a plausible sounding hypothesis test that gives me a p-value less than 5% so I can publish my work." If this is as far as you are going, then R is not for you. Your search will be much more efficient in a traditional statistics package. Alternatively, there is statistics in the large sense: "I want to know what my data have to say. …. If your goal is to find what is in your data, then sooner or later R is likely to provide you functionality which can't be found elsewhere. (p. 2)

Burns' comments can be summarized as follows: R is for those who want to learn something large from their data in which case good statistical skills are important; those who are content to learn something small and want to use R, are, more than anything else, a nuisance and should instead use software like SPSS, SAS, or Stata. How widely Burns' perspective is shared within the R community is unclear but it certainly has a presence. More importantly, the uneven technical support is prima facie evidence that users with more modest statistics backgrounds are likely to struggle to find the technical support they need. This in turn raises concerns about the willingness of the R community to support the learning and work of educators and their students who have varying views of data analysis software and varying statistics backgrounds.

What Should Happen Next?

Moving forward at least two complementary lines of research are needed to (1) catalogue the data analysis software being used in education, by whom, and for what purpose, and (2) construct a literature based on rigorous evaluations of data analysis software that provides guidance to educators choosing among programs.

Cataloguing which data analysis software is used in education is an important activity that should inform subsequent evaluations of software. The information collected would generally include the data analysis software currently used, by whom and for what purpose, why a particular program (or programs) were chosen, cost, what capabilities of the software are needed, adequacy of the technical support, and user-friendliness. Information about a respondent's statistics background and computer skills would also be important to obtain.

This information could be collected by surveying educators working in different settings, for example,

academic research staff and faculty in universities, colleges, and non-profits, and/or individuals belonging to professional associations like the American Educational Research Association and its many special interest groups, as well as students. Requests to provide information about data analysis software are likely to be well received as educators and their students are likely to have strong opinions on this topic. Moreover, summaries of survey results should represent publishable work as this information is likely of considerable interest to educators.

A second line of research is needed to provide rigorous evaluations of data analysis software using benchmarks drawn from the technology literature that has identified desirable features of software. For example, general benchmarks would likely include comprehensiveness, adequacy of technical support, and user-friendliness, whereas more specific benchmarks might focus on the ability of programs to efficiently handle large datasets, data manipulation capabilities such as ease of transposing datafiles, and the advantages of GUIs vs. code. The goal of this work would be to generate a literature that provides educators and their students with sound guidance in selecting the data analysis software that best fits their needs. Studies providing in-depth comparisons of data analysis software would almost by definition be multidisciplinary through the involvement of educators with different kinds of expertise (statistics, statistics education, learning, educational technology), and individuals with expertise in computer science.

Rigorous evaluations of the extent to which these programs support learning in the classroom are also needed. Targeted groups would include undergraduate and graduate students in educational statistics classes and might also include high school students in AP Statistics. This research would presumably draw on the statistics education literature for good models for evaluating the ability of software to

support statistics learning (e.g., Fitzallen & Brown, 2006; Garfield & Ben-Zvi, 2009).

Naturally evaluations of data analysis software should be both qualitative and quantitative. For the latter, comparisons of software based on educators and students obtained from randomized cluster designs (What Works Clearinghouse [WWC], 2011) in which different software is randomly assigned to academic units within a university or college or to different sections of the same statistics class, would be especially welcome. However, the history of studies in the statistics education literature attempting to identify models and practices that enhance student learning suggests that such studies are unlikely. Still, quantitatively-oriented evaluations that employ quasi-experimental designs in which data analysis software has been self-selected that attend to the deficiencies of this design (e.g., selection bias) (WWC, 2011) should be welcome.

Summary

There is little doubt that the hegemony of commercial data analysis software in education over the past 30 years is ending due to the rise of the R software. While many educators appear to be switching to R it is striking that little formal evidence of the ability of commercial software or R to meet the needs of educators is available. In the absence of such evidence educators must rely on their experience to guide their choice of software which, while valuable, is insufficient for the field as a whole.

This paper argued that a two-fold approach is needed to provide educators with guidance for choosing data analysis software that meets their needs. First is to catalogue the software currently used in education, by whom, and for what purpose. Second is to develop a literature of rigorous evaluations of software that builds on the cataloguing of software in education and includes comparisons of their

comprehensiveness, technical support, and user-friendliness. Developing a literature of this nature is a substantial task but, given the importance of data analysis software in education, is consistent with an educational landscape that is teeming with evaluations of educational models, programs, and practices. Until a literature evaluating data analysis software is available the rush to R or any data analysis software is unwarranted.

References

Bryk, A. S., Raudenbush, S. W., & Congdon, R. (2011). *Hierarchical linear and nonlinear modeling* (version 7) [Computer software manual]. Lincolnwood, IL: Scientific Software International.

Burns, P. (2007). R relative to statistical packages: Comment 1 on technical report number 1 (Version 1.0) strategically using general purpose statistics packages: A Look at Stata, SAS and SPSS. Retrieved from http://www.burns-stat.com/pages/Tutor/R_relative_statpack.pdf

Curtis, D. C., & Harwell, M. R. (1998). Preparing graduate students in educational statistics: A national survey. *Journal of Statistics Education*, *6*.

Dembe, A. E., Partridge, J., & Geist, L. (2011). What are the most common statistical software applications used by health services researchers? *BMC Health Services Research, 11* (*252*), 1-6. doi: 10.1186/1472-6963-11-252

Fitzallen, N., & Brown, N. (2006). Evaluating data-analysis software: Exploring opportunities for developing statistical thinking and reasoning. In N. Anderson & C. Sherwood (Eds), IT's Up Here for Thinking. Proceedings of the Australian Computers in Education Conference.

Fortran 90 User's Guide. (1995). Mountain View, CA: Sun Microsystems, Inc.Keeling, K. B., & Pavur, R. J. (2007). A comparative study of the reliability of nine statistical software packages. *Computational Statistics & Data Analysis*, *51 (8*), 3811-3831.

Meys, J., & de Vries, A. (2012). *R for dummies*. West Sussex, England: Wiley.

Mitchell, M. N. (2005). Strategically using general purpose statistical packages: A look at Stata, SAS, and SPSS (Technical Report Series, Report Number 1, Version Number 1). Statistical Consulting Group: UCLA Academic Technology Services. Retrieved from http://www2.jura.uni-hamburg.de/instkrim/kriminologie/Mitarbeiter/Enzmann/Lehre/StatIIKrim/Mitchell_2007.pdf

Muenchen, R. (2014). The popularity of data analysis software. Retrieved from http://r4stats.com/articles/popularity/

Prvan, T., Reid, A., & Petocz, P. (2002). Statistical laboratories using Minitab, SPSS and Excel:A practical comparison. *Teaching Statistics, 24 (2*), 68-75.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from: http://www.R-project.org/(Version 3.0.1).

SAS Institute. (2004). *SAS/STAT 9.1 User's guide* [Computer software manual]. Cary, NC: SAS Institute.

SPSS Inc. (2011). *Command syntax reference* (Version 20.0) [Computer software manual]. Chicago, IL: SPSS Inc.

Stata Press. (2007). *Stata base reference manual* (Release 10). [Computer software manual]. College Station, TX: Stata Press.

Veranzi, J. (2005). *Using R for introductory statistics*. New York: Taylor & Francis.

Wegman, E. J., & Solka, J. L. (2005). Statistical software for today and tomorrow. Retrieved from http://binf.gmu.edu/~jsolka/PAPERS/ess2542_rev1.pdf