Organizational Research Methods
14(4) 735-740

DOI: 10.1177/1094428109355485
http://orm.sagepub.com



# R is for Revolution: A Cutting-Edge, Free, Open Source Statistical Package

**Steven Andrew Culpepper[1] and Herman Aguinis[2]**

**Abstract**

The authors review the open source statistical package R. R allows researchers to implement statistical techniques including linear modeling, linear and nonlinear multilevel modeling, factor and principal component analysis, structural equation modeling, item and reliability analysis, time series modeling, and meta-analysis, among others. R presents several advantages over other statistical packages because it is updated on an ongoing basis, is free, is capable of creating high-quality graphics that are difficult to create with other packages, and includes important simulation capabilities. Some limitations of R include the need to learn a new programming language, difficulties handling missing data for new users, and relatively limited support and documentation. R is not yet popular in the organizational sciences but, given its ongoing improvement and many positive features, we predict that it will soon be.

**Keywords**

quantitative research, statistical computing, psychometrics, computer package, Monte Carlo simulation, free software

A steady change is occurring within industry and academe: researchers are adopting an open source statistical package named R. In fact, R received widespread publicity in a *New York Times* article (Vance, 2009), which noted that organizations such as Google, Merck, Pfizer, and Bank of America are using it. R is available at no charge for a variety of operating systems (Maindonald & Braun, 2003). Researchers from diverse fields ranging from psychology, genomics, and economics are continuously developing new functions that perform novel and cutting-edge statistical analyses. Although R is not yet popular in the organizational sciences, we predict that it will soon be. Given the prominence and visibility of *Organizational Research Methods*, we also hope that this review will serve as a catalyst for the use of R by organizational scientists.

[1] Department of Mathematical & Statistical Sciences, University of Colorado Denver, Denver, CO, USA
[2] Department of Management and Entrepreneurship, Kelley School of Business, Indiana University, Bloomington, IN, USA

**Corresponding Author:**
Herman Aguinis, Department of Management and Entrepreneurship, Kelley School of Business, Indiana University, 1309 E. 10th Street, Bloomington, IN 47405, USA
Email: haguinis@indiana.edu

**Table 1.** Selected R functions

| Procedure | R Function | R Package | Base Package[a] |
|---|---|---|---|
| General data handling/utility functions | | | |
| Graphical user interface | Rcmdr | Rcmdr | No |
| Read SPSS or STATA data file | read.spss, read.dta | foreign | No |
| Graphical method | | | |
| 3D perspective plot | persp | Graphics | Yes |
| 3D scatter plot | scatterplot3d | scatterplot3d | No |
| Histogram, box plot, scatter plot | hist, boxplot, plot | graphics | Yes |
| Statistical method | | | |
| Bootstrap resampling | boot | boot | No |
| Factor analysis | fa | psych | No |
| Generalized estimating equations | geeglm | geepack | No |
| Item and reliability analysis | reliability | CTT | No |
| Linear and nonlinear mixed models | lmer, nlmer | lme4 | No |
| Linear models, analysis of variance | lm, anova | stats | Yes |
| Logistic regression | lrm | Design | No |
| Meta-analysis | rma.uni | metaphor | No |
| Multidimensional scaling | isoMDS | MASS | Yes |
| Multivariate analysis of variance | manova | stats | Yes |
| Principal components | princomp | stats | Yes |
| Structural equation modeling | sem | sem | No |
| Text mining | tm_filter | tm | No |
| Times series models | arima | stats | Yes |

Note: A complete list of R packages and corresponding functions can be found at http://cran.r-project.org/web/packages/. SPSS = statistical package for the social sciences.
[a]Packages that are not installed in the base R software can be easily downloaded for free.

## Overview of R Features

R can be downloaded for free at the following Web site: http://cran.r-project.org/. Table 1 provides an overview of R functions and packages relevant to organizational science researchers. In R, functions are specific procedures that perform a specific analysis (e.g., functions are to R as procedures are to statistical analysis system [SAS]) and R packages are independently downloadable syntax files that consist of many functions. The first section of Table 1 presents general utility procedures. For instance, R is capable of reading a wide array of file types. The default data entry functions allow users to read files with quantitative and/or qualitative variables in the .txt or .csv formats (e.g., read.table and read.csv). Table 1 presents several functions for reading statistical package for the social sciences (SPSS) and STATA data files (i.e., read.spss and read.dta, respectively).

R also allows researchers to implement commonly employed statistical techniques, such as linear models, linear and nonlinear multilevel models, factor analysis and/or principal components, structural equation modeling, and item and reliability analysis. Table 1 provides a list of commonly used data-analytic approaches in organizational research, which can be conducted using R. Additionally, at least two packages (specifically, the ''psych'' and ''QuantPsyc'') include a variety of functions specific to the field of psychometrics. For example, the ''psych'' package includes a function to perform parallel analysis, to simulate item-level data, and to compute polychoric correlation coefficients and the ''QuantPsyc'' package includes functions to create *z* scores, approximate power for *F* tests (note that the ''pwr'' packages offer researchers ability to compute statistical power estimates for other tests), and test for multivariate normality. Other packages allow researchers to assess the number of latent factors (''nFactors''), to conduct dichotomous and polytomous item response

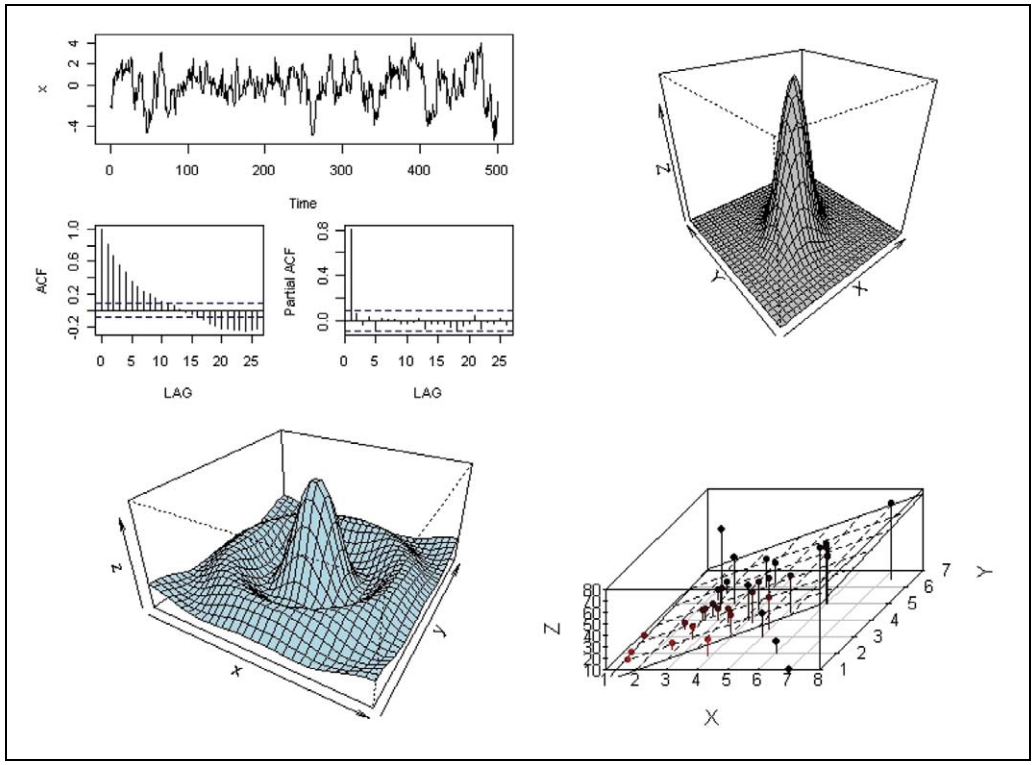


**Figure 1.** Examples of graphs created using R. ACF = Autocorrelation function.

theory analysis (e.g., the "ltm" package described by Rizopoulos, 2006), and to perform meta-analysis ("metafor"). As shown in Table 1, R includes functions and packages to perform many of the statistical methods that are most popular, as well as those that are becoming popular, in organizational research (Aguinis, Pierce, Bosco, & Muslin, 2009).

## Benefits of R

R has additional advantages over other (commercially available) statistical packages. First, because it is an open source package, it is updated on an ongoing basis. Thus, there is no need to wait for the next "big release" for R to incorporate important improvements. The latest and most novel statistical approaches and refinements are incorporated much faster than is the case with other more traditional commercially available packages.

Second, R is particularly suited for pedagogical purposes because it is free. Universities can save money by installing R in computer labs and using R in introductory and advanced statistics and measurement courses. Certainly, one downside is that R requires students to learn the R programming language (as discussed below). It is important to note that some graphical user interfaces (GUIs) are currently available (e.g., the Rcmdr function, as noted in Table 1), which allow users a more familiar point-and-click interface found in programs such as SPSS. However, the current R GUIs are not as refined as commercial software. As noted in Table 1, "Rcmdr" is a GUI that allows users to employ a variety of statistical methods; in addition, a statistical teaching plug-in tool ("RcmdrPlugin. TeachingDemos") is available to use with "Rcmdr" to demonstrate the logic of the central limit theorem, correlation, and confidence intervals.

Third, R is also particularly advantageous for creating publishable graphics. Producing meaningful graphics is often important to convey substantive findings. Some commercially available packages are less effective for producing publishable graphics. R includes a variety of functions and packages for creating advanced graphics. For instance, Table 1 includes a few functions for common graphics (such as histograms, box plots, and scatter plots) as well as functions for creating three-dimensional graphics. Figure 1 includes four panels with plots researchers can create with R, which may be difficult to create with other statistical packages. Specifically, the first panel includes figures useful for time-series analysis, the second presents the bivariate normal, the third includes a more general three-dimensional graph, and the last includes a three-dimensional scatter plot with a fitted regression plane.

Fourth, R is particularly useful for implementing Monte Carlo simulations. Specifically, R includes functions for matrix operations (e.g., "%*%" can be used for matrix products and "solve" to compute matrix inverses). R also includes functions to generate pseudorandom numbers from common (e.g., normal distribution, multivariate, normal distribution, and other exponential families) and more complex (e.g., generalized gamma and Wishart) probability distributions. Additionally, R allows users to create functions for generating data for a given simulation design and R includes several helpful functions to replicate functions (specifically, the "replicate" function is excellent for conducting a function thousands of times). For example, the following code creates a function to generate *t*-values for a two-sample *t* test given group sample sizes ("n1" and "n2"), a noncentrality parameter ("ncp"), and ratio of group standard deviations ("s2_s1"):

```
typeII <- function(n1,n2,ncp,s2_s1) {
g1 <- rnorm(n1)
g2 <- s2_s1*rnorm(n2)+ncp
sp <- sqrt(((n1-1)*var(g1)+(n2-1)*var(g2))/(n1+n2-2))
t <- (mean(g1)-mean(g2))/(sp*sqrt(1/n1+1/n2))
}
```

The following code replicates the function 5,000 times for group sample sizes of 100, a noncentrality parameter equal to 1, and the ratio of group standard deviations of 2 and stores it in an object named "output":

```
output<- replicate(5000,typeII(100,100,1,2))
```

Note that the results stored in "output" could be used to compute empirical Type I or Type II error rates. Researchers can also batch submit R scripts (i.e., syntax files) to improve computational time. Using such procedure, Aguinis, Culpepper, and Pierce (2010) conducted a Monte Carlo simulation that generated 15 billion 925 million individual samples including 8 trillion 662 billion 500 million individual scores. It is also important to note that several packages (namely, "multicore" and "taskPR") are available, which allow users to use parallel processing.

Finally, as noted above, R allows researchers to use many statistical methods. In fact, as of 2009, more than 2,000 R functions containing statistical methods and/or data sets were available to download. The availability of novel statistical methods provides organizational researchers the opportunity to employ novel methods to extend theory in ways currently used methodology does not allow. For example, R includes a text mining package ("tm") that could provide researchers with opportunities to pose and test new hypotheses, as well as building a bridge between quantitative and qualitative research paradigms. Additionally, numerous methodologists have created R packages to extend normal theory models. For example, R provides users the ability to use modern approaches as described by Wilcox (2001).

## Limitations of R

R is not without some limitations and/or drawbacks. First, and perhaps most important, R requires knowledge and competency in a new programming language. Arguably, R's programming language is designed for and by statisticians, so the language often offers users benefits and efficiencies over GUIs. A variety of texts exist, which effectively teach and guide readers to use the R programming language for the most commonly used statistical methods (Dalgaard, 2002; Maindonald & Braun, 2003; Muenchen, 2009), as well as books for more specific statistical topics such as time series or statistical computing (Cowpertwait & Metcalfe, 2009; Rizzo, 2008). Several articles also provide a general overview of R for new users (Cribari-Neto & Zarkos, 1999; Racine & Hyndman, 2002).

Second, missing data are common in organizational research and some R functions can be frustrating to use in the presence of missing responses. For example, the more common statistical methods (such as linear models) use listwise deletion by default and it is sometimes difficult to perform other functions (e.g., storing residuals and/or predicted values in the original data file) when data are missing. However, R does include some functions to help analysts curtail problems of missing data (e.g., the model.frame function returns cases with complete data for a given linear model). Additionally, the challenges associated with missing data are likely outweighed by novel multiple imputation packages for missing data, such as the "mi" package that offers various imputation methods.

Third, as noted above, R contains more than 2000 packages and each function has a corresponding user manual or help page. Despite the extensive online documentation help, pages are sometimes incomplete for the needs of new R users. Accordingly, users who need additional support for a given function will likely find online searchable blogs and archives quite helpful through the R homepage.

Finally, R can sometimes be difficult to use with a PC or laptop when data sets exceed a million observations. Researchers in the organizational sciences are less likely to study such large data sets, but organizations and businesses can have large data sets. R stores active data sets in short-term memory, so R will not function if the size of the data set exceeds working memory. Certainly, researchers can simply update their computing resources (i.e., increase working memory or use larger machines) or employ some R functions designed to handle large data sets (e.g., functions within the "bigmemory" package).

In conclusion, it is not surprising that R is growing in popularity within industry and academe. R offers organizational researchers access to cutting-edge, advanced statistical methods, in addition to commonly used methods in the social sciences. R is beneficial for introductory and advanced teaching statistics courses and R produces high-quality graphics to convey substantive findings. Additionally, given the wide selection of pseudorandom number generators and unique functions, R is a useful language for writing and implementing simulation studies. As R continues to develop, one area for additional innovation is improving existing GUIs to provide more user-friendly interfaces that will allow researchers to take advantage of all the benefits that R has to offer.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

### Funding

The author(s) received no financial support for the research and/or authorship of this article.

### References

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95*, 648-680.

Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. (2009). First decade of Organizational Research Methods: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods*, *12*, 69-112.

Cowpertwait, P. S. P., & Metcalfe, A. V. (2009). *Introductory time series with R.* New York: Springer.

Cribari-Neto, F., & Zarkos, S. G. (1999). R: Yet another econometric programming environment. *Journal of Applied Econometrics*, *14*, 314-329.

Dalgaard, P. (2002). *Introductory statistics with R.* New York: Springer.

Maindonald, J., & Braun, J. (2003). *Data analysis and graphics using R.* Cambridge: Cambridge University Press.

Muenchen, R. A. (2009). *R for SAS and SPSS Users.* New York: Springer.

Racine, J., & Hyndman, R. (2002). Using R to teach econometrics. *Journal of Applied Econometrics*, *17*, 175-189.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software*, *17*, 1-25.

Rizzo, M. L. (2008). *Statistical computing with R.* Boca Raton, FL: Chapman & Hall/CRC.

Vance, A. (2009, January 6). Data analysts captivated by R's power. *The New York Times*. Retrieved on November 20, 2009, from http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?_r=1

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy.* New York: Springer.

## Bios

**Steven Andrew Culpepper** (http://math.ucdenver.edu/~sculpeppe/) is an assistant professor of Mathematical and Statistical Sciences at the University of Colorado Denver. He received his PhD in educational psychology and quantitative methods in 2006 from the University of Minnesota, Twin Cities. His research interests include statistics, research methods, and psychometrics.

**Herman Aguinis** (http://mypage.iu.edu/~haguinis/) is the Dean's research professor and a professor of Organizational Behavior and Human Resources at Indiana University's Kelley School of Business. He has published five books, about 70 journal articles, and 20 book chapters and monographs on a variety of human resource management, organizational behavior, and research methods and analysis topics.