Master Tutorial

TITLE

Big Data Systems with R

SHORTENED TITLE

Big Data Systems with R

ABSTRACT

Beginners to R are often intimidated by various ways to analyze data. Three popular workflows include base R, data.table, and the tidyverse, each with different syntax. This tutorial session will explain how each system works and the corresponding benefits/drawbacks when analyzing big data. Bring your laptop for this interactive session (download session materials here: <https://bit.ly/32ag86B>).

PRESS PARAGRAPH

R can be intimidating when transitioning from statistical software such as SPSS or SAS. Unlike SPSS or SAS, R is a programming language and has many ways to approach the same problem. Three common frameworks for analyzing and cleaning data include base R, “data.table”, and the “tidyverse”. Beginners to R cannot know which paradigm to adopt for a given problem and, worse, how to read and understand code across different packages. This session will act as a Rosetta stone, showing new R users how to read code across the various frameworks and which one might be the most appropriate for a given problem, especially when analyzing big data where differences between methods can yield significant performance improvements.

WORD COUNT: 1571

**Big Data Systems with R**

R (R Core Team, 2019) is an open-source programming language that is designed for statistical computing (Hornik, 2013). R can perform anything from standard data analysis (e.g. Multiple Regression, Hierarchical Linear Modeling, or Structural Equation Modeling) to machine learning and natural language processing to highly specialized computations that may be unique to a scientific field. R is a programming *language* and not just a statistical analysis package. By some measures, R has become one of the ten most popular programming languages (Cass, 2018). R’s popularity may be partly due to the large ecosystem of support pages, books, blogs, tutorials, and R specific conferences.

Many data scientists and practitioners can contribute to R by writing new and unique software, called “packages” in R. At the time of this writing, the Comprehensive R Archive Network (CRAN) contains 14,750 available packages, including packages to read data in varying formats (e.g., readr, open.xlsx, haven, rjson, officer, vroom), access databases (e.g., DBI, odbc, RSQLite), clean data (e.g., dplyr, tidyr, stringr, reshape2), perform data analyses and machine learning (e.g., infer, caret, xgboost, randomForest, survival, DALEX), visualize results (e.g., shiny, ggplot2), and interface with other programming languages (e.g., Rcpp, reticulate, RJava). These packages, just like R itself, are free of charge.

R has recently been ranked as the highest domain-specific programming language and one of the most popular tools among data scientists (Cass, 2018; O’Grady, 2019). Although several Graphical User Interfaces (GUI) exist for R (e.g., Deducer; Fellows, 2012; R Commander; Fox and Bouchet-Valat, 2019), sophisticated analyses and production ready workflows require intimate knowledge of the R programming language. As with many programming languages, one can perform the same analysis in many different ways.

This Master Tutorial will teach attendees how to develop skills in using various R ecosystems, including for reading data, combining and cleaning data, and analyzing data to produce output. Emphasis will be made on the performance tradeoffs and constraints for each of the systems when analyzing large data sets and extracting data from databases, which can magnify differences in execution time. Special attention will be given to transitioning from exploratory analyses, where the focus is typically on the speed of writing code, to producing code that can easily be transferred and run on various systems, where the purpose is typically replication on different computers and/or large and repeated analyzes across an entire organization. Moreover, understanding how different R paradigms work can help with comprehending various solutions to different questions (commonly posted on the numerous R online forums) and applying/adapting those solutions to individual problems.

*Proposed Session*

R has different frameworks, each with different syntax and performance differences for various problems. Many of the common frameworks (such as the tidyverse, Wickham, 2017 and data.table, Dowle and Srinivasan, 2019) are nearly mutually unintelligible, so understanding how to do an analysis using a given system does not necessarily translate to another system. Moreover, both data.table and the tidyverse were developed with tradeoffs between the convenience of a quick analysis with the robustness and stability needed to perform analyses at scale across tens or hundreds of millions of records. The tidyverse and data.table typically have analogous functions or methods in base R (i.e., the set of packages that are initially installed with R and are developed by the R Core Team). Knowing how each of the three systems works allows researchers to choose the system most appropriate for a given data analysis problem. Commonly cited frustrations with R (such as the speed of analyses for large problems) can usually be ameliorated by choosing the system best designed to address that problem.

The proposed tutorial is an attempt at a Rosetta stone for the most commonly used R programming frameworks, lowering the barrier to entry for scientists and practitioners and helping researchers choose the best methodology for a given analysis. This tutorial will be interactive. Audience members are strongly encouraged to bring laptops and to have downloaded the materials ahead of time. For those who wish to follow along, we will make available all of the materials and R scripts at <https://bit.ly/32ag86B>. We request 80 minutes for the tutorial, with the approximate time for each topic as well as additional information provided below. Note that none of the authors are affiliated with the producers of any of the packages described and that there are no material gains (financial or otherwise) for them.

**Topic #1: Introduction to Different R Programming Frameworks (35 minutes)**

The most commonly used R programming frameworks include base R, “data.table”, and the “tidyverse”. Each framework was developed with a different rationale and for different types of analyses. For instance, base R is typically developed with an eye on backward compatibility, so that changes to base R rarely affect the ability of old code to result in the same output. Because of this stability, developers often rely on base R when code is highly distributed across an organization and where changes to the user interface can have high impact on organizational decisions. However, the emphasis on stability comes at a cost. Many of the functions in base R are inconsistent and unlikely to change due to the conservative philosophy of the R Core Team. Some known issues with base R are still part of R due to legacy code requiring this idiosyncratic behavior.

Conversely, the tidyverse was developed with the goal of a consistent programming philosophy, so that all packages within the tidyverse “share an underlying design philosophy, grammar, and data structures” (Tidyverse, n.d.). Moreover, many of the tidyverse packages emphasize non-standard evaluation (e.g., Wickham, 2019b), which makes it easy to perform simple exploratory analyses (as described in Grolemund and Wickham, 2017). For these reasons, tidyverse packages and functions (e.g., dplyr, tidyr, and ggplot2) are often users’ first introduction to the R language and make up 6 out of the top 10 most downloaded R packages in 2019 (Kopf, 2019). In fact, Matloff (2019a) lamented that “R is rapidly devolving into two mutually unintelligible dialects” and cited an R user who wrote that “one can code in the Tidyverse while knowing very little R.”

Unlike the “tidyverse”, “data.table” is less a set of packages developed around a programming philosophy and more a single package designed to speed up R computations. Where base R and the tidyverse uses copy on modify semantics (so that applying a function to an object does not change the original object), data.table uses reference semantics and tries to avoid copying objects. Modifying objects in place can lead to drastic improvements in performance, especially with respect to large data sets (e.g., Matloff, 2019b). However, the additional performance gains come at a cost, as data.table has a less intuitive syntax than other R frameworks and requires a basic understanding of programming concepts, such as deep vs shallow copies of data.

**Topic #2: Translating Analyses across Frameworks (20 minutes)**

Once each of the systems are explained, we will present simple data analytic examples and show how to approach the problem in base R, data.table, and the tidyverse. Special attention will be paid to understanding how code written in one system can easily be broken down to simple elements and reconstructed to work in a different system. For instance, if exploring a small sample of data using the tidyverse, one can often quickly rewrite code using data.table to gain performance benefits. The dtplyr package (Wickham, 2019a) can translate dplyr code to data.table syntax, but this translation is often slow and inexact. Understanding how data.table syntax works can ease the transition from the tidyverse to data.table and ensure that data.table code performs the same analysis and finds the same result.

**Topic #3: Performance Tradeoffs when Analyzing Big Data (15-20 minutes)**

All three systems have a majority of functions written in C or C++. However, the implementation details of each system can result in performance impacts (i.e., slow code). These impacts are magnified when analyzing tens or hundreds of millions of records, as is the typical case when pulling data from the internet. As unstructured online text data provides important insight into job attributes in different organizations, regions, or countries (including employee engagement, compensation, management issues, etc.), efficiently parsing and analyzing this data is crucial for testing out a variety of hypotheses and coming up with principled conclusions. During this section, we will provide real-world examples to show how each system performs under a variety of conditions and the cost-benefit tradeoffs with performance when analyzing data for a particular research problem.

**Topic #4: Wrap-up (5-10 minutes)**

Finally, the presenters will answer audience questions and help with technical problems encountered during previous sessions. The presenters will also provide materials for self-study and exploration. Participants should leave feeling confident about how to approach a given data analysis question in R and comfortable parsing code and analyses that others have written in any of the commonly used systems.

**Learning Objectives**

By the end of this workshop, you will be able to:

1. Understand how to read in data, clean data, and perform basic descriptive statistics with each of the commonly used R frameworks, including “base R”, “data.table”, and the “tidyverse”.
2. Know enough to be able to ask for help about a particular problem on online R forums and be able to translate code from one system to another system.
3. Comprehend the benefits and tradeoffs of each system when analyzing large datasets and determine whether a given R system is appropriate given a particular research problem.

**Presenter Information**

Jeff Jones

Assessment Researcher

Stripe

510 Townsend Street

San Francisco, CA 94103

Email: jeff.jones@kornferry.com

Membership Status: Member

Steven Nydick

Data Scientist Programmer

Korn Ferry

33 South Sixth Street

Suite 4900

Minneapolis, MN 55402

Phone: 612-373-3548

Email: [steven.nydick@kornferry.com](mailto:steven.nydick@kornferry.com)

Membership Status: Member

Benjamin Wiseman

Data Scientist Programmer

Korn Ferry

33 South Sixth Street

Suite 4900

Minneapolis, MN 55402

Phone: 612-373-3445

Email: benjamin.wiseman@kornferry.com

**Presenter Bios**

Jeff Jones is the People Analytics Assessment Researcher at Stripe, where he designs and implements analytics infrastructures, performs analytics, and is a subject matter expert for psychometrics and statistical methodology. He has published articles in theoretical and methodological journals such as *Psychometrika* and *Psychological Methods*, and is a coauthor on several CRAN and internal R packages. Jeff received his Ph.D. at the University of Minnesota in Psychometrics and Quantitative Psychology where he focused on creating new statistical methodology, asymptotic statistics, and higher-order geometry of statistical methodology.

Steven Nydick is a Data Scientist Developer at the Korn Ferry Institute, where he designs R-based tools and scoring algorithms. He is the lead author and maintainer of the catIrt R package as well as several internal R packages helping with everything from plotting to powerpoint generation to interfacing with servers. He has contributed to developing psychometric models and corresponding estimation algorithms that have been published in *Applied Psychological Methods* and the *Journal of Educational and Behavioral Statistics*. Steven received his Ph.D at the University of Minnesota in Psychometrics and Quantitative Psychology, where he primarily studied IRT-based adaptive tests for selection and classification. He also has an M.S. in Statistics from the University of Minnesota.

Ben Wiseman is a Data Science Developer at the Korn Ferry Institute responsible for maintaining and developing R-based automation tools, models, reports, and user interfaces. He has publications in entomology, ecology, and molecular evolution and has worked with and trained numerous clients in the military, public, and private sectors on a wide range of applications. Ben received his MSc from Lincoln University (New Zealand) in applied statistical modelling where he developed a user-facing geospatial AI platform for DOCs predator monitoring and control systems.

**References**

Cass, S. (2018). The 2018 top programming languages. Retrieved August 29, 2018, from <https://spectrum.ieee.org/at-work/innovation/the-2018-top-programming-languages>

Dowle, M., & Srinivasan, A. (2019). data.table: Extension of `data.frame`. R package version 1.12.2.

Fellows, I. (2012). Deducer: A data analysis GUI for R. *Journal of Statistical Software*, *49*, 1-15. Retrieved August 29, 2019, from <http://www.jstatsoft.org/v498/i08>.

Fox, J., & Bouchet-Valat, M. (2019). Rcmdr: R commander. R package version 2.5-3.

Grolemund, G., & Wickham, H. (2017). *R for Data Science: Import, Tidy, Visualize, and Model Data*. Sebastopol, CA: O’Reilly Media, Inc.

Hornik, K. (2017). The R FAQ. Retrieved from http://CRAN.R-project.org/doc/FAQ/R-FAQ.html

Kopf, D. (2019). The 10 most downloaded R packages in 2019. Retrieved from <https://www.theatlas.com/charts/NQ9tEMM58>

Matloff, N. (2019a). R vs. Python for Data Science. Retrieved from <https://github.com/matloff/R-vs.-Python-for-Data-Science>

Matloff, N. (2019b). TidyverseSkeptic. Retrieved from <https://github.com/matloff/TidyverseSkeptic>

O’Grady, S. (2019). The RedMonk programming language rankings: June 2019. Retrieved August 29, 2019, from <https://redmonk.com/sogrady/2019/07/18/language-rankings-6-19/>

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Tidyverse (n.d.). R packages for data science. Retrieved from tidyverse.org.

Wickham, H. (2017). Tidyverse: Easily install and load the “Tidyverse”. R package version 1.2.1.

Wickham, H. (2019a). dtplyr: Data table back-end for ‘dplyr’. R package version 0.0.3.

Wickham, H. (2019b). Programming with dplyr. Retrieved from

<https://dplyr.tidyverse.org/articles/programming.html>

**Appendix**

CV Jeff Jones

CV Steven Nydick

CV Benjamin Wiseman

Jeff Jones

Email: [Jeff.Jones@KornFerry.com](mailto:Jeff.Jones@KornFerry.com)

**Education:**

Doctorate in Quantitative Psychology and Psychometrics, University of Minnesota, October, 2013. Advisor: Dr. Niels Waller.

Bachelor of Science, Psychology, University of California, Davis, June 2006.

Bachelor of Arts, Japanese Language and Literature, University of California, Davis, June 2006.

Minor in Mathematics, University of California, Davis, June 2006.

**Employment History:**

Assessment Researcher, Stripe, 2020 – Present.

Director, Talent Analytics and Data Systems, Korn Ferry, 2017 – 2020.

Adjunct Professor, University of Minnesota, 2017 – Present.

Senior Manager of Analytics, Korn Ferry, 2015 – 2017.

Manager of Research and Analytics, Korn Ferry, 2013 – 2015.

Adjunct Professor, Hamline University, Fall 2013.

Graduate Instructor/Section Leader, University of Minnesota, 2006 – 2013.

**Awards:**

Korn Ferry Founder’s Award for Innovation, 2015.

Eva O. Miller Fellowship, 2012.

Graduate Summer Research Fellowship, 2009.

Graduate Research Partnership Program Fellowship, 2007.

**Publications:**

Jones, J. A. & Waller, N. G. (2016). Fungible weights in logistic regression. *Psychological Methods, 21,*

241-260*.*

Jones, J. A. & Waller, N. G. (2015). The normal-theory and asymptotic distribution-free covariance matrix of standardized regression coefficients: Theoretical extensions and finite sample behavior. *Psychometrika, 80,* 365-378.

Jones, J. A. & Waller, N. G. (2013). Computing confidence intervals for standardized regression coefficients. *Psychological Methods, 18,* 435-453.

Jones, J. A. & Waller, N. G. (2013). Abstract: The normal-theory and asymptotic distribution-free covariance matrix of standardized regression coefficients. *Multivariate Behavioral Research, 48,* 161.

Waller, N. G. & Jones, J. A. (2011). Investigating the performance of alternate regression weights by studying all possible criteria in regression models with a fixed set of predictors. *Psychometrika, 76,* 410-439.

Jones, J. A. (2010). GenCorr: An R routine to generate correlation matrices from a user-defined eigenvalue structure. *Applied Psychological Measurement, 34,* 68-69.

Waller, N. G. & Jones, J. A. (2010). Correlation weights in multiple regression.  *Psychometrika, 75,* 58-69.

Waller, N. G. & Jones, J. A. (2009). Locating the extrema of fungible regression weights. *Psychometrika,*

*74,* 589-602.

**Software:**

Wiseman, B., Nydick, S. W., & Jones, J. A. (2018). roperators: Additional operators to

help you write cleaner R code. R package version 1.0.1.

<https://CRAN.R-project.org/package=roperators>

Goebl, A. P., Jones, J. A., Dahlke, J., & Beatty, A. S. (2016). iopsych: Methods for

industrial/organizational psychology. R package version 0.90.

Waller, N. G. & Jones, J. A. (2015). fungible: Fungible coefficients and Monte Carlo

functions. R package version 1.3.

**Presentations and Workshops:**

Jones, J. A., Nydick, S. W., & Wiseman, B. (2019, April). *Web scraping with R.* Master Tutorial at the

annual meeting of the Society of the Industrial and Organizational Psychology, National Harbor, MD.

Jones, J. A., Nydick, S. W., & Wiseman, B. (2019, April). *Effective data wrangling and visualization with R.*

Master Tutorial at the annual meeting of the Society of Industrial and Organizational Psychology, National Harbor, MD.

Goebl, A. P., Jones, J. A., & Semmel, S. G. (2018, April). *Machine learning in R: A*

*tutorial and jam session.* Master Tutorial at the annual meeting of the Society of

Industrial and Organizational Psychology, Chicago, IL.

Thompson, I. B., Song, Q. C., Goebl, A. P., Hall, S., Meade, A. W., Newman, D. A.,

Wee, S., & Jones, J. A. (2018, April). *Machine learning techniques for multiple*

*criteria optimization.* Alternative Session at the annual meeting of the Society of

Industrial and Organizational Psychology, Chicago, IL.

Wendt, H., Goff, M., Jones, J. A., & Hezlett, S. A. (2017, May). *Examining relationships*

*between the Korn Ferry personality inventory and job engagement across*

*countries.* In S. Dilchert and D. Ones (Chairs), *An IRT based approach to*

*personality measurement: Some cross cultural examinations.* Paper presented at

the annual meeting of the European Association of Work and Organizational

Psychology, Dublin, Ireland.

Hezlett, S. A., Jones, J. A., Lewis, J., Goff, M., & Stirling, E. (2017, April). *What*

*motivates may alienate: Linking motivational factors to derailment risks.* In S.

Hezlett (Chair), *Maladaptation: Building the nomological net of derailing traits and behaviors.* Paper presented at the annual meeting of the Society of Industrial and Organizational Psychology, Orlando, FL.

Jones, J. A., Goebl, A. P., & Semmel, S. G. (2017, April). *Modern methods for I-O*

*psychologists: An interactive tutorial in R.* Master Tutorial at the annual meeting of the Society of Industrial and Organizational Psychology, Orlando, FL.

Schwall, A., Beatty, A., & Jones, J. A. (2017, April). *Data visualization with R.* Master

Tutorial at the annual meeting of the Society of Industrial and Organizational

Psychology, Orlando, FL.

Semmel, S. G., Jones, J. A., & Goebl, A. P. (2017, April). *What is machine learning?*

*Foundations and introductions to useful methods.* Master Tutorial at the annual meeting of the Society of Industrial and Organizational Psychology, Orlando, FL.

Jones, J. A., Goebl, A. P., & Semmel, S. G. (2016, April). *Handling big(gish) data in R:*

*An introductory and interactive tutorial.* Master Tutorial at the annual meeting of the Society of Industrial and Organizational Psychology, Anaheim, CA.

Goebl, A. P. & Jones, J. A. (2016, April). *An R package for I-O psychology simulation*

*building: iopsych.* Poster presented at the annual meeting of the Society of

Industrial and Organizational Psychology, Anaheim, CA.

Goebl, A. P. & Jones, J. A. (2016, April). *Creative performance is a viable criterion for*

*personnel selection.* Poster presented at the annual meeting of the Society of

Industrial and Organizational Psychology, Anaheim, CA.

Blazek, S. & Jones, J. A. (2016, April). *A year on the job: Simulations-based*

*assessments’ versatile utility.* In D. Guangrong (Chair), *The art and science of executive assessment: Research and practice.* Paper presented at the annual meeting of the Society of Industrial and Organizational Psychology, Anaheim, CA.

Blazek, S. & Jones, J. A. (2016, February). Streamlining your voice of the customer

program: Automating survey follow-ups, panel updates, and reports. Presentation given at Qualtrics Insight Summit, Salt Lake City, UT.

Jones, J. A. & Beatty, A. (2015, December). *R Workshop*. Workshop given at the

monthly meeting of Minnesota Professionals for Psychology Applied to Work, Minneapolis, MN.

Schwall, A., Beatty, A., & Jones, J. A. (2015, May). *Getting started with R: An*

*Interactive and Hands on Tutorial.* Master Tutorial at the annual meeting of the

Society of Industrial and Organizational Psychology, Philadelphia, PA.

D’Mello, S. & Jones J. A. (2015, May). *Demographic Differences in the Motivational*

*Drivers of Leaders*. Poster presented at the annual meeting of the Society of

Industrial and Organizational Psychology, Philadelphia, PA.

Goebl, A., Jones J. A., & Sharpe, E. (2015, May). *Relative Criteria Importance Depends*

*on Predictor Choices*. Poster presented at the annual meeting of the Society of

Industrial and Organizational Psychology, Philadelphia, PA.

Schwall, A., Lustenberger, D., Beatty, A., & Jones, J. A. (2014, May). *Getting started*

*with R: Examples and Lessons Learned.* Panel discussion at the annual meeting

of the Society of Industrial and Organizational Psychology, Honolulu, HI.

Huber, C. & Jones, J. A. (2014, May). *Principal components and factor scores in*

*multiple regression: A simulation.* Poster presented at the annual meeting of the

Society of Industrial and Organizational Psychology, Honolulu, HI.

Jones, J. A. & Waller, N. G. (2012). *The normal-theory and asymptotic distribution-free covariance matrix of standardized regression coefficients.* Annual Meeting, Society for Multivariate Experimental Psychology Conference, Vancouver, British Columbia.

Morris P. E. & Jones, J. A. (2005). *Using formal inference-based recursive*

*modeling to detect plausible interactions for multiple regression.* Annual Meeting, American Psychological Society Conference, Los Angeles, CA.

**Technical Reports and White Papers:**

Blazek, E. S., Jones, J. A., Lewis, J. L, & Orr, J. E. (2018). Develop and select the best

CEOs. Korn Ferry Institute, Korn Ferry.

Blazek, E. S., Jones, J. A., Lewis, J. L, & Orr, J. E. (2017). CEO staying power. Korn

Ferry Institute, Korn Ferry.

Lewis, J., Goff, M., Hezlett S., Jones, J. A., Li, T., Dai, G., & Deege, A. (2017). Korn

Ferry four dimensional enterprise assessment: Research guide and technical

manual. Version 17.1a—11/2017, Korn Ferry.

<http://www.kornferry.com/technical-manuals>

Blazek, E. S., Jones, J. A., Lewis, J. L, & Orr, J. E. (2016). Predicting financial gains.

Korn Ferry Institute, Korn Ferry.

Blazek, E. S., Jones, J. A., Lewis, J. L, & Orr, J. E. (2016). Leading indicators. Korn

Ferry Institute, Korn Ferry.

Blazek, E. S., Jones, J. A., Lewis, J. L, & Orr, J. E. (2016). Korn Ferry simulations-based

assessments predict CEO success: CEO outcomes research technical paper. Korn

Ferry Institute, Korn Ferry.

Dai, G., Davies, S., Goff, M., Jones J. A., D’Mello, S., Orr, J. E., Storfer, P., & Tang, K.

Y. (2014). Korn Ferry Leadership Architect: Research guide and technical manual. Version 14.2a—01/2016. <http://www.kornferry.com/technical-manuals>

Jones. J. A. & Waller, N. G. (2013). The normal-theory and asymptotic distribution-free

(ADF) covariance matrix of standardized regression coefficients: Theoretical

extensions and finite sample behavior. Technical Report 052513. University of

Minnesota, Twin Cities.

http://www.psych.umn.edu/faculty/waller/downloads/techreports/TR052913.pdf

Lewis, J. & Jones, J. A. (2016). Fit matters. Korn Ferry Institute, Korn Ferry.

<http://www.kornferry.com/institute/fit-matters>

Lewis, J., Goff, M., Jones, J. A., Hezlett S., Tang, K. Y., Dai, G., D’mello, S., Henry, L.,

Zes, D., Fetzer, J., Xie, C., & Scheer, P. (2015). Korn Ferry four dimensional executive assessment: Research guide and technical manual. Version 15.1a—01/2016, Korn Ferry. <http://www.kornferry.com/technical-manuals>

**Courses Taught:**

Introduction to Data Analysis/Statistics for Undergraduates

Analysis of Psychological Data for Graduate Students

Steven Nydick

Email: Steven.Nydick@KornFerry.com

**Education:**

PhD, University of Minnesota, Psychometrics/Quantitative Psychology, 2013.

Advisor: Niels Waller

MA, University of Minnesota, Psychometrics/Quantitative Psychology, 2012.

Advisor: Niels Waller

MS, University of Minnesota, Statistics, 2011.

Advisor: Sanford Weisberg

BS, Syracuse University, Mathematics and Psychology, 2006.

**Professional Experience:**

Data Scientist Developer, Korn Ferry, 2018 – Present.

Senior Psychometrician, Pearson VUE, 2016 – 2018.

Psychometrician, Pearson VUE, 2013 – 2016.

Research Assistant, University of Minnesota, 2013 – Present.

Intern in Psychometrics, ARRT, 2012 – 2013.

Intern in Psychometrics, ACT, 2011.

Graduate Instructor/Section Leader, University of Minnesota, 2007 – 2013.

**Awards:**

Doctoral Dissertation Fellowship, 2013

Graduate Research Partnership Program, 2010

Archimedes Prize in Mathematics, 2006

**Manuscripts Published and In Press:**

Wang, C. & Nydick, S. W. (2015). Comparing two algorithms for calibrating the restricted non-

compensatory multidimensional IRT model. *Applied Psychological Measurement*, *39*, 119-134.

Nydick, S. W. (2014). The sequential probability ratio test and binary item response models. *Journal of*

*Educational and Behavioral Statistics*, *39*, 203-230.

**Software:**

Wiseman, B., Nydick, S. W., & Jones, J. A. (2018). roperators: Additional operators to

help you write cleaner R code. R package version 1.0.1.

<https://CRAN.R-project.org/package=roperators>

Nydick, S. W. (2014). catIrt: An R package for simulating computerized adaptive tests. R package version

0.5-0).

**Presentations and Workshops:**

Jones, J. A., Nydick, S. W., & Wiseman, B. (2019, April). Web scraping with R. Master Tutorial at the

annual meeting of the Society of the Industrial and Organizational Psychology, National Harbor, MD.

Jones, J. A., Nydick, S. W., & Wiseman, B. (2019, April). Effective data wrangling and visualization with R.

Master Tutorial at the annual meeting of the Society of Industrial and Organizational Psychology, National Harbor, MD.

Nydick, S. W. (2016, April). The expected likelihood in computerized classification testing. Paper

presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Nydick, S. W. (2014, April). Multidimensional mastery testing with CAT. Paper presented at the annual

meeting of the National Council on Measurement in Education, Philadelphia, PA.

Nydick, S. W., Wang, C., & Xiong, X. (2014, April). Measuring multidimensional growth—a higher-order

IRT perspective. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Nydick, S. W., Nozawa, Y., & Zhu, R. (2012, April). Accuracy and efficiency in classifying examinees using

computerized adaptive tests: An application to a large scale test. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, BC.

Nydick, S. W., & Weiss, D. J. (2010, June). Accepting the null: No change in change CAT. Paper presented

at the IACAT conference on CAT, Arnhem, NL.

Nydick, S. W., & Weiss, D. J. (2009). A hybrid simulation procedure, evaluated for the development of

CATs. In D. J. Weiss (Ed.) *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.*

**Unpublished Manuscripts:**

Nydick, S. W. (2013). *Intro to R for Psychologists.* Minneapolis, MN: Author.

**Courses Taught:**

Introduction to Data Analysis/Statistics for Undergraduates

Honors Introduction to Data Analysis/Statistics for Undergraduates

Analysis of Psychological Data for Graduate Students

Benjamin Wiseman

Email: Benjamin.Wiseman@KornFerry.com

**Education:**

MS, Lincoln University, Applied Statistics, 2015.

BS, Lincoln University, Biostatistics, 2013.

**Professional Experience:**

Data Scientist Developer, Korn Ferry, 2018 – Present.

Owner, Wiseman Analytics, 2016 – 2018.

Information Services, DHS, 2015 – 2016.

Instructor, Lincoln University, 2013 – 2014.

Research Assistant, Lincoln University, 2011 – 2015.

Research Assistant, Seoul National University, 2011.

**Awards:**

Freemasons university scholarship

Forest and Bird research award

AGLS research scholarship

**Manuscripts Published and In Press:**

Wiseman, BH., Fountain, ED., Bowie, MH. He, S., Cruickshank, RH. 2016. Vivid molecular divergence over volcanic remnants: the phylogeography of Megadromus guerinii on Banks Peninsula, New Zealand. New Zealand Journal of Zoology

Fountain, ED., Pugh, AR., Wiseman, BH., Smith, VR., Cruickshank, RH., and Paterson, AM. 2015. On the captive rearing of Hadramphus tuberculatus (Pascoe 1877) (Coleoptera: Curculionidae: Molytinae):is ex-situ conservation the lesser of two weevils? New Zealand Entomologist.

Gillespie, M., Cruickshank, RH., Wiseman, BH., Wratten, S. 2013. Incongruence between morphological and molecular markers in the butterfly genus Zizina (Lepidoptera: Lycaenidae) in New Zealand.Systematic Entomology 38:151-163.

Fountain, ED., Wiseman, BH., Cruickshank, RH., and Paterson, AM. 2013. The ecology and conservation of Hadramphus tuberculatus (Pascoe 1877) (Coleoptera: Curculionidae: Molytinae). Journal of Insect Conservation 17:737-745.

**Software:**

Wiseman, B. W., Nydick, S.W., Jones, J (2018) roperators: Additional Operators to Help you Write Cleaner R Code. R package version 1.0-1).

Wiseman, B. W. (2015) Neurofriendly: Artificial Neural Networks Made Simple

Wiseman, B. W. (2015) Geofriendly: Easy Spatial Application of Artificial Neural Networks

**Presentations and Workshops:**

Jones, J. A., Nydick, S. W., & Wiseman, B. (2019, April). Web scraping with R. Master Tutorial at the

annual meeting of the Society of the Industrial and Organizational Psychology, National Harbor, MD.

Jones, J. A., Nydick, S. W., & Wiseman, B. (2019, April). Effective data wrangling and visualization with R.

Master Tutorial at the annual meeting of the Society of Industrial and Organizational Psychology, National Harbor, MD.

Wiseman, B. H. 2017 Data Science with Python. ESRI Developer Summit, Palm Springs, CA.

Wiseman, B. H. 2013 Messy data, messy models and applied statistics. Presented for Bio-Protection seminar, Lincoln University, New Zealand.

Marris, J. and Wiseman, B. H. 2012. Islands in the snow: Ecology, systematics and biogeography of the New Zealand beetle genus Protodendrophagus (Coleoptera:Silvanidae:Brotini). Presented at the New Zealand Ecological Society conference.

Cripps, M., McNeil, M., Patrick, H., Wiseman, B., Nobilly, F., Edwards, G. 2012. Invertebrate abundance and diversity in intensively managed dairy pastures.New Zealand Plant Protection Society Conference.

Wiseman, B. H., Cruickshank, R. H., Bowie, M. H., Fountain, E. D. 2011. Unexpected genetic variation in an endemic ground beetle: The molecular mystery of Megadromus guerinii (Coleoptera: Carabidae). 3rdAnnual Combined Australian and New Zealand Entomological Societies Conference

Wiseman, B. H. (2011). The curious case of Megadromus guerinii: phylogeographic oddities on Bank’s Peninsula. Presented to the Canterbury branch of the New Zealand Entomological Society.

**Courses Taught:**

Research and Analytical Skills

Geospatial Information Systems with Arc GIS

Business Statistics

Intermediate Statistics for Commerce