Master Tutorial

TITLE

Text Analytics and NLP with R

SHORTENED TITLE

Text Analytics with R

ABSTRACT

Text is messy! Extracting information from text data is not as simple as analyzing quantitative questionnaires. This tutorial session will walk you through how to clean, describe, summarize, display, and predict outcomes from text using the powerful statistical language R. Bring your laptop for this interactive session (download session materials here: <https://bit.ly/2KKXlHQ>).

PRESS PARAGRAPH

Data that is useful is not always easily amenable to quantitative analysis. Much of that data is in the form of unstructured text, whether in interviews, open-ended questionnaire responses, essays, or online comments. The statistical computer language R provides tools for cleaning and standardizing text for the purpose of summaries, displays, models, and prediction. Text mining tools can illuminate how employees feel about their job as well as determining areas of research missed by a questionnaire. This session will give a hands-on and step-by-step tutorial on how to clean and analyze text-based data.

WORD COUNT: 2155

**Text Analytics and NLP 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 to machine learning and natural language processing to highly specialized computations that may be unique to a scientific field. R is not just a statistical analysis package, but a fully-fledged programming *language*. R has even managed to become one of the ten most popular programming languages (Cass, 2018) partly due to the large ecosystem of support pages, books, blogs, tutorials, and R specific conferences. Several R packages, including dplyr (Wickham, François, Henry, and Müller, 2019), ggplot2 (Wickham, 2016), and data.table (Dowle and Srinivasan, 2019) have over 12,000 mentions on Stack Overflow (Robinson, 2017).

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.

The Master Tutorial will teach attendees how to leverage R and several packages to transform text data into statistical displays and predictive models. We will show users how to extract data from common text sources, including surveys, interviews, and online reviews/comments, standardize the text into a useful form, and predict quantitative and qualitative outcome measures. One can think of this Master Tutorial as a direct continuation of the R Master Tutorial at the 2019 SIOP conference on web scraping: after extracting data from the web, what do you do with that data? Attendees should be familiar with R and have both R and RStudio installed prior to the workshop. We will walk through and explain each line of code in detail, but we will have little time to review the basics of R itself.

*Proposed Session*

Many studies require parsing and analyzing existing data. Much of this data is in the form of unstructured text. Anyone who has ever read a book knows that unstructured text contains substantial information. Humans are specially attuned to spotting patterns in these types of stimuli. However, manually parsing millions of reviews, thousands of books, or hundreds of articles would take extensive time and effort. Computers can quickly parse information but lack the nuance to spot novel patterns contained therein. Cleaning and parsing unstructured text data for analyses requires special strategies. R has many tools designed to clean, describe, display, model, and predict text data (e.g., Munzert, Rubba, Meißner, and Nyhuis, 2015; Silge and Robinson, 2017). Unfortunately, I-O psychologists often lack the specialized training required to successfully parse unstructured data without tedious, manual coding methods. Building on several R-based master tutorials over the past few years (e.g., Schwall, Lustenberger, Beatty, and Jones, 2014; Schwall, Beatty, & Jones, 2015; Goebl, Jones, & Semmell, 2016; Goebl, Jones, & Semmell, 2018; Jones, Nydick, & Wiseman, 2019a; Jones, Nydick, & Wiseman, 2019b), this tutorial aims to break down useful R methods for I/O psychologists. Specifically, this tutorial can be seen as extension of Web Scraping and APIs with R (Jones, Nydick, & Wiseman, 2019b) that describes what to do with the data being extracted from the web. We will explain text mining as implemented by R in three parts.

The first part of the tutorial will introduce packages and methods for importing, cleaning, filtering, and structuring text data. We will discuss tokens, n-grams, stop words, and stemming, so that common patterns of text can easily be identified and isolated. Much of this processing employ several powerful R packages, including tm (Feinerer and Hornik, 2018), corpus (Perry, 2017), and tidytext (Silge and Robinson, 2016), each of which have tools to translate, organize, and simplify text for follow-up analysis.

The second part of the tutorial will take the structured text and show how to summarize and display useful aspects of the data that might suggest future areas of investigation. Specifically, we can pull comments about specific organizations from job boards and compare the sentiment of words in comments with the frequency of those words appearing across all comments within an organization (e.g., Silge and Robinson, 2017). These results can easily be displayed in, for example, a wordcloud (e.g., Fellows, 2018; Lang and Chien, 2018) or a sentiment barplot using ggplot2 (Wickham, 2016).

Finally, we will show how to create a predictive model with processed text inputs. For instance, we can use various machine learning models (such as Random Forests, e.g., Liaw and Wiener, 2002; or xgboost, Chen et al., 2018) to predict ratings from review text and n-grams in individual comments. These models can be useful when trying to predict employee engagement and voluntary turnover rates (which can have obvious negative financial effects for individual companies). Although our example relates to review comments, any unstructured text corpus can be processed in similar steps.

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 materials and R scripts at <https://bit.ly/2KKXlHQ>. 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. All packages are free-of-charge and complete.

**Topic #1: Text Parsing and Cleaning (25 minutes)**

In principle, text data can be treated like any other unit of data. One can draw bar plots, summarize the count of each piece of text, and relate individual pieces of text to an outcome. However, part of the difficulty with text analysis is that the unit of interest is typically contained within the text statement rather than being the statement itself. For example, one rarely finds two essay responses to have exactly the same words and structure, so a bar plot for an entire textual statement would typically be a uniform height of one across all responses. To make sense of text, one typically would have to find smaller units contained in the text that could be repeated within and across textual statements. These units are often called “tokens” (e.g., Silge and Robinson, 2017). Tokens can be anything from single words to n-grams (multiple words strung together) to complete sentences. These tokens are typically stripped of superfluous information, such as punctuation, that would cause two different tokens to be different even if their content were identical.  
 Once text statements are tokenized, they can be inserted as variables in predictive models. Unfortunately, many tokens might not be useful in summaries and displays. For example, “the” is the most common word in English (Oxford Corpus, 2011). If the word “the” appears as the largest word in a wordcloud of Glassdoor comments, one would not learn much about individual feelings toward a company. In fact, keeping the word “the” as a token would risk overshadowing the importance of other, more useful, words, and adding extraneous variance in predictive models. Words that are very common and are typically ignored when inferring meaning from a statement, such as “the”, are called “stop words”. Various languages have different sets of stop words, and R packages such as stopwords (Benoit, Muhr, and Watanabe, 2019) or tm (Feinerer and Hornik, 2018) have functions to remove these stop words. Certain professional disciplines also have very common words that add little value for analyses. These discipline-specific stop words should also be removed as part of the data cleaning step.

Stop words are not the only token feature adding additional variance to text. Language is filled with inbuilt operators to take text and change the appearance of words without changing the meaning. For instance, many languages have verb conjugations, so that “running”, “run”, “ran”, and “runs” look different but should be grouped together. Removing superfluous features of individual words is called “stemming” and typically implemented by a variant of the Snowball algorithm (Porter, 2001; and which is built into the corpus, Perry, 2017, and SnowballC, Bouchet-Valat, 2019, R packages). Similar text sometimes shows up as different tokens for other reasons, such as spelling variants, spelling mistakes, and synonyms. Although important, correcting these token variations is difficult, time consuming, project or outcome dependent, and beyond the scope of this tutorial.

**Topic #2: Summarizing Text Data (20-25 minutes)**

If various tokens appear with enough regularity, analyses could proceed by looking at how often certain tokens appear in the data using statistics such as “term frequency” and “inverse term frequency”. Term frequency refers to the number of times that a token appears in a collection of tokens and can be calculated by simply adding up the number of times a term appears (where one can count multiple occurrences of a term in a particular piece of text as a single appearance or separate appearances). The inverse term frequency (ITF) is simply the natural logarithm of the number of overall statements divided by the number of statements in which a particular term appears. ITF plots generally appear as positively skewed, with very common tokens near the mode of the distribution and very rare tokens in the tail. One common use of term frequencies is as the input to a wordcloud (e.g., Fellows, 2018) with terms sized and colored according to their term frequency value.

One could also perform simple token classifications, such as sentiment analysis (Jurafsky, n.d.). Sentiment scores can be as simple as assigning simple polarity to tokens, such as “positive”, “neutral”, or “negative”, with optional sentiment strength. As a common example, one could use these simple sentiment scores to predict global events from online comments or social media posts (e.g., predicting stock market prices from Twitter posts). Several R packages such as tidytext (Silge and Robinson, 2016), SentimentAnalysis (Feuerriegel and Proellochs, 2019), and syuzhet (Jockers, 2015) include functions to perform basic sentiment coding and analysis. The tidytext package has a simple dictionary of how individual words can be classified according to sentiment, whereas the SentimentAnalysis has tools to classify and display tokens, sentences, or entire documents according to sentiment score. The syuzhet package can classify tokens according to polarity and strength but also has tools to classify tokens according to specific emotions, such as joy, fear, disgust, anger, and surprise, rather than simple polarity. Sentiment classifications can be combined with wordclouds for dramatic illustration of how certain sentiments appear in a set of text statements.

**Topic #3: Predictive Models with Text Data (20-25 minutes)**

Once text statements are cleaned and dummy-coded, one can include those tokens in a statistical model or for prediction. As these tokens are categorical-type variables, they can be included as-is in any machine learning prediction algorithm, such as random forests (Breiman, 2001) or gradient boosted trees (Hastie, Tibshirani, and Friedman, 2009). These models can be used to predict outcomes, such as online ratings or engagement scores on a questionnaire. One could then estimate the engagement ratings of people with only text statements and, consequently, obtain an aggregate view of the engagement of an organization.

Alternatively, specialized methods and R packages exist that can build models explicitly around the structure of text, such as keyword analysis (using RKEA, Feinerer and Hornik, 2015) or latent semantic analysis (using lsa, Wild, 2015). Keyword analysis can predict keywords of future text by building a model given author-assigned keywords as well as text in which those keywords may appear. Latent semantic analysis is designed to extract latent variables from text sources where pairs or sets of words appear with various degrees of frequency. One could think of a latent semantic analysis as a textual equivalent to a principle components analysis where the variables are the terms or tokens and the dimensions are formed from a combination of similar terms. This analysis can serve as an approximation to a term document matrix and capture similarities across words that might be missed by simple stemming procedures.

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

Finally, the presenters will answer audience questions and help with technical problems encountered during previous sections. The presenters will also provide materials for participants to read for self-study and include links to useful materials for solving text analytics problems.

**Learning Objectives**

This workshop is designed to help you:

1. Explain steps that need to be taken to transform unstructured text data to a structured corpus, including tokenizing, stemming, and removing stop words.
2. Create basic visual displays and quantitative summaries of text data from R, including wordclouds and basic sentiment analysis.
3. Understand how to include text data in basic predictive modeling and how to run basic predictive text analyses in R.

**Presenter Information**

Jeff Jones

Director, Talent Analytics and Data Systems

Korn Ferry

33 South Sixth Street

Suite 4900

Minneapolis, MN 55402

Phone: 612-373-3418

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 Director of Talent Analytics and Data Systems at Korn Ferry where he oversees the organization’s production scoring services, designs scoring algorithms, 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**

Benoit, K., Muhr, D., & Watanabe, K. (2019). stopwords: Multilingual stopword lists. R package version 1.0.

Bouchet-Valat, M. (2019). SnowballC: Snowball stemmers based on the C ‘libstemmer’ UTF-8 library. R package version 0.6.0.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32.

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>

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., & Li, Y. (2018). xgboost: Extreme gradient boosting. R package version 0.71.2.

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

Feinerer, I., & Hornik, K. (2018). Tm: Text mining package. R package version 0.7-6.

Feinerer, I., & Hornik, K. (2015). RKEY: R/KEA interface. R package version 0.0-6.

Fellows, I. (2018). wordcloud: Word clouds. R package version 2.6.

Feuerriegel, S., & Proellochs, N. (2019). SentimentAnalysis: Dictionary-based sentiment analysis. R package version 1.3-3.

Goebl, A. P., Jones, J. A., & 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., & 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.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). Boosting and additive trees. In *The Elements of Statistical Learning (2nd ed.).* New York, NY: Springer.

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

Jockers, M. L. (2015). Syuzhet: Extract sentiment and plot arcs from text. R package version 1.0.4.

Jones, J. A., Nydick, S. W., & Wiseman, B. (2019a, 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.

Jones, J. A., Nydick, S. W., & Wiseman, B. (2019b, April). Web scraping with R. Master Tutorial at the annual meeting of the Society of the Industrial and Organizational Psychology, National Harbor, MD.

Jurafsky, D. (n.d.). Sentiment analysis: What is sentiment analysis. Retrieved August 14, 2019, from <https://web.stanford.edu/class/cs124/lec/sentiment.pdf>

Lang, D., & Chien, G. (2018). Wordcloud2: Create Word Cloud by “htmlwidget”. R package version 0.2.1.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18-22.

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2015). *Automated Data Collection with R*. New York, NY: Wiley.

Oxford Corpus (2011). The OEC: Facts about the language. Retrieved August 14, 2019, from [https://web.archive.org/web/20111226085859/http://oxforddictionaries.com/words/the-oec-facts-about-the-language](https://web.archive.org/web/20111226085859/http:/oxforddictionaries.com/words/the-oec-facts-about-the-language)

Perry, Patrick O. (2017). Corpus: Text corpus analysis. R package version 0.10.0.

Porter, M. (2001). Snowball. Retrieved August 14, 2019 from <https://snowballstem.org/>.

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/.

Robinson, D. (2017). The impressive growth of R. Retrieved August 29, 2018, from

https://stackoverflow.blog/2017/10/10/impressive-growth-r/

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.

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.

Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R.

*The Journal of Open Source Software*, *1*. Retried August 13, 2019, from <https://joss.theoj.org/papers/10.21105/joss.00037>.

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. Sebastopol, CA: O’Reilly Media,

Inc. Retrieved August 13, 2019, from https://www.tidytextmining.com/.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer.

Wickham, H., François, R., Henry, L., & Müller, K. (2018). dplyr: A grammar of data manipulation. R

package version 0.8.3.

Wild, F. (2015). Lsa: Latent semantic analysis. R package version 0.73.1.

**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:**

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

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>

**Patents:**

U. S. Patent 10,346,804: "[Determining job applicant fit score](https://patents.justia.com/patent/10346804)", July 9, 2019.

**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