R you ready? Using the R programme for statistical analysis and graphics

Gregor Štiglic

email: gregor.stiglic@um.si

Austyn Snowden

email: A.Snowden@napier.ac.uk

Roger Watson

email: r.watson@hull.ac.uk

Leona Cilar

email: leona.cilar1@um.si

Abstract

Introduction: Research in nursing has a big potential to influence current and future nursing practice and nursing profession. For conducting research, nurses often use various statistical programmes. Less known and used in nursing is R.

Methods: In this study we used the dataset used in previous study to examine the construct validity of the TEIQue-SF. A step-by-step CFA using a laavan package was conducted.

Results: A total of 938 students of nursing, midwifery and computer science in two Scottish Universities participated in the study.

Discussion:

Conclusion: R is a free and easy available programming language for conducting many different statistical analyses in healthcare and nursing research. Nevertheless, R requires some

basic statistical and specific programming knowledge that nurses must have to use it. Thereby additional education may be required.

Key words: R, statistics, data analysis, graphics, nursing.

Introduction

Nursing researchers are becoming increasingly well-versed in statistical methods with copious evidence of the use of advanced methods embedded in sophisticated research designs (Zellner & Boerst, 2007). Nevertheless, it is evident that nursing researchers largely depend on commercial statistical packages - principally SPSS® (IBM, 2019) the package formerly known as the Statistical Package for the Social Sciences and developed by the IBM Corporation. Other commercial packages are available such as Stata® (STATA, 2019), developed by Stata Corp, and SAS (SAS, 2019), developed by the SAS® Institute for Advanced Analytics. Each of these packages has its advantages and proponents (de Smith, 2018). Adherents to particular packages tend to fall within particular subjects; for example, SAS® is often favoured in Medicine while SPSS® tends to be adopted by researchers in nursing and psychology. These packages offer excellent analytical facilities, and most have user-friendly interfaces compatible with both Windows and Mac computers. The exception is SAS®, which is not available for Mac computers. Nevertheless, they have limitations. They all offer a different range of analyses (for example, Stata® offers non-parametric item response theory analysis which SPSS® and SAS® do not and the bolt on package to SPSS® for structural equation modelling (AMOS®: Analysis of Moment Structures) is only available to run on Windows software (Acock, 2005). Graphical facilities are limited in SPSS®, despite the popularity of the package and creating publishable figures and tables is a multi-step process. While individual academics rarely pay for licences for statistical packages, which are covered by institutional agreements, the packages are

updated annually and require licenses and these can cost over $US100,000 (https://www.quora.com/How-much-does-SAS-cost; accessed 2 September 2019) (it should be noted that some SAS® software is available free for non-commercial use).

The alternative to these packages, which compete commercially for space in the research environment, would be a facility that is virtually unlimited in its analytical capacity, flexible and regularly updated, has excellent graphical facilities and is available open source at no cost. Such a package—or packages—is available within the R Project for Statistical Computing (R Development Core Team, 2011; Paura & Arhipova, 2012). However, while we are aware of some nurses who make us of R and these are becoming more common, its use in nursing research remains limited. While we acknowledge that there is no formal, fair and direct way of comparing statistical packages, we wish to propose that the use of R should become more widespread in nursing research and in this article we offer: an overview of the capabilities of R; some examples of its application in our own work; and one practical example with the concomitant R coding in order that others interested in this specific method may use that coding and use our example as an exercise. We hope that this encourages others to explore R themselves and become part of the international R community whereby they may find support for their own analyses and contribute—as we have—to the development of the statistical packages with which they become familiar.

What is R?

R may be described in many ways but, essentially it is programming language that is specifically applied to statistical computing and graphics. It is open source—free to use anywhere in the world—and contains packages that are created by statistical analysts around the world. The packages are published under a Creative Commons license whereby, while

retaining attribution, they make the packages and relevant coding available universally. R adherents are responsive to comments and suggestions regarding their packages and many are updated regularly as and when required (Beaujean, 2013; Venables, Smith, & R Core Team, 2019).

R is free to use because it is, essentially, crowd-funded; users and institutions are invited to make contributions to the upkeep of the system, but this is not compulsory. Packages that are available via R, are submitted by individuals or groups of programmers and then rigorously tested before being made available via CRAN (The Comprehensive R Archive Network). To use R the software needs to be downloaded and installed from The R Project for Statistical Computing webpage. R is available for Linux, Mac and Windows and the software is regularly updated (Ozgur, Kleckner, & Li, 2015). At the time of writing, version 3.5.3 (Great Truth) is available—each version is given an obscure and sometimes amusing name; version 3.5.1, for example, was call 'Feather Spray'.

The process of installing R on your own computer, therefore, involves visiting The R Project for Statistical Computing webpage, following the 'download R' link which takes you to a page of CRAN mirrors. These webpages are all identical in content, but you are advised to select a CRAN mirror geographically close to you. Currently there are nearly 50 CRAN mirrors across the world and some countries have several. Selecting a CRAN mirror takes you to the different R packages for the three platforms referred to above and, depending on your operating system, you select that version of R. You then install R from that link and the software is stored in a library on your computer's hard drive and an R logo is installed on your desktop from which the software can be accessed. There are some basic calculations that can be carried out with R installed but, to conduct more complex statistical analyses, you need to install the appropriate packages; there are over 10,000 packages to choose from and sometimes several packages need

to be installed to run simultaneously. Instructions are provided during the installation process. In the process of installing packages a CRAN mirror must be selected again.

R can be run by opening the software and then loading the relevant package for your analysis. The package performs the statistical calculations, but the user must enter the appropriate code to run the programme. This is one of the disadvantages of using R; the 'learning curve' for using R is very steep and for someone unfamiliar with computer programming languages, it can seem quite intimidating; it can also be very time-consuming to learn R and, once learned, to run analyses using it. This is the least user-friendly aspect of R. Nevertheless, copious guidance is available online, and most packages have manuals which are open source and easily available online. To ease the use of R most users copy and paste their usual codes into other documents and then simply paste them back into R each time they use it. Alternatively, R programmes can be written and uploaded to run analyses automatically. Most users install RStudio® which is an open source (for the basic package) IDE (integrated development environment) which facilitates the running of R and has the advantage of storing coding already used which can then simply be selected and re-entered into any analysis. Once R is installed—preferably to be run with RStudio®—and the relevant packages have been installed, you are ready to start using it.

Once installed and some basic mastery of R has been achieved, users have access to over 10,000 statistical packages (https://blog.revolutionanalytics.com/2017/01/cran-10000.html; accessed 2 September 2019). Users will also find that R affords them considerable flexibility in terms of the wide formats of data that can be imported (including from the popular commercial packages referred to above). Data can be manipulated very easily, and R has excellent graphics capabilities. However, '*caveat emptor*' applies as, while packages undergo considerable scrutiny prior to being included in R and they can be updated, they do not some with a guarantee and, unlike commercial packages, R does not come with customer care. There are often several

packages performing the same function and, in addition to mastering the programming language of R and the concomitant and regular errors that can be made, users have to contend with frequent and obscure error messages which they may neither understand not be able to circumvent.

Methods

Example Data

In this paper we use the dataset originally used in a study by Snowden et al. (2015) to examine the construct validity of the Trait Emotional Intelligence Questionnaire Short form (TEIQue-SF). Data were collected from 938 students of nursing, midwifery and computer science in two Scottish Universities. TEIQue-SF is a 30-item trait emotional intelligence measure introduced by Petrides (2009) who derived it from his larger 130-item based TEIQue questionnaire (Freudenthaler et al. 2008). Despite the reduction in the number of items it is still possible to identify four factors (well-being, sociability, self-control & emotionality) of emotional intelligence that can be measured by this questionnaire. The dataset contains information on gender of the participants, 30 items of the TEIQue-SF questionnaire with pre-calculated total TEIQue-SF score as well as scores for all four factors – i.e. well-being, sociability, self-control & emotionality.

The dataset is available in the comma separated values (CSV) format as a supplement of this paper. The R source code, which can also be downloaded from supplementary material contains an example of the R command needed to read the data into R programming environment. Additionally, the source code also provides the R commands to conduct all steps of an example provided to the readers of this paper.

Factor Analysis

To demonstrate the capabilities of R programming language on a simple example, we decided to supplement the paper with the dataset and source code of the R programming script with corresponding R commands to run the factor analysis. ~~The aim of factor analysis is to reduce "the dimensionality of the original space and to give an interpretation to the new space, spanned by a reduced number of new dimensions which are supposed to underlie the old ones" (Rietvel & van Hout, 1993, p. 254). It is mainly used in fields such as medicine and nursing, economics, behavioural and social sciences, and geography (Yong & Pearce, 2013).~~

~~Two types of factor analyses that are mostly used are Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). CFA is used to confirm hypotheses on existing latent variables and it uses path analysis diagrams to represent variables and factors, while EFA is used when there is a need for identifying the latent variables that are underlying a set of variables. Similarly, Principal Components Analysis (PCA) is used to reduce a set of variables to a smaller set of factors (Yong & Pearce, 2013). More specifically, PCA aims to find a new reduced set of variables, equal in number to the original set of variables, where these synthetic variables are uncorrelated (Rossiter, 2017).~~

In this article we demonstrate the use of R programming language to perform CFA in a reproducible way using a *laavan* package (Oberski, 2014). CFA is just one of the Structural Equation Models (SEM) special cases supported by *laavan* package and other packages within R have the capacity to run CFA. Usually it is used to test the fit of data to measurement models (Graham et al., 2003) and can be frequently met in studies where a measure is used in a new environment or language for the first time.

Many different approaches were proposed to assess the fit of the model to the data. Some of the more popular fit statistics include Comparative Fit Index (CFI), the Tucker Lewis Index (TLI) and Root Mean Square Error of Approximation (RMSEA). CFI is frequently used as it is known to perform well even when the sample size is small (Tabachnick and Fidell, 2007). It

is based on the assumption that all latent variables are uncorrelated and compares the postulated model to the null model. The CFI values can range from 0.0 to 1.0 where values close to 1.0 represent a good fit. To avoid misspecified models it is generally advised for models to achieve CFI over 0.90 which represents a good fit with some authors arguing that the threshold should lie at 0.95 (Hu and Bentler, 1999). TLI measures a relative reduction in misfit per degree of freedom (Tucker & Lewis, 1973) with threshold values of 0.90 indicating adequate fit and 0.95 indicating good fit (Chavez et al., 2019). On the other hand, the RMSEA represents a so-called "badness-of-fit measure" resulting in lower values for a better fit. It measures discrepancy due to the approximation with values below 0.06 representing an acceptable model (Shi et al., 2019).

Results

In this section, we provide a step-by-step instructions on how to run the CFA analysis with corresponding R code provided in the supplementary material to this paper (https://data.mendeley.com/datasets/schrvg5c88/draft?a=fd0b4bc8-f1d5-429e-9b8d-903a0e342a66). The step-by-step instructions are provided to allow reader to follow the supplementary material, especially the R code to run the CFA. On the other hand, we would like to demonstrate that using R is not only simple, but also allows much better reproducibility compared with frequently used point-and-click software.

As mentioned in the introduction it is recommended to use one of the IDEs to write and run R code. RStudio which is a free and open source software can be used for this purpose. With R and RStudio installed, one can start the data analysis. In our case, we provide an example of running a CFA using an R code file that can be found at the Mendeley data repository (https://data.mendeley.com/datasets/schrvg5c88/draft?a=fd0b4bc8-f1d5-429e-9b8d-903a0e342a66).

In the initial step, we read the data from the comma separated values (csv) file using the *read.csv* command in R. Additional to a very basic csv format, R can read multiple different file formats, including the Microsoft Excel, SPSS, SAS and similar formats which allows easier transition from different statistical analysis programs.

After reading the data, we can check whether the data are loaded by using a command *str* which prints the summary information on the structure of the data just loaded. This way it is possible to print the type and a few example values for each variable. In the case of TEIQue-SF example data provided with this paper, we observe that our data consists of 938 samples with 36 variables (30 of those are TEIQue-SF scale variables).

After removal of items 3, 18, 14 & 29 that were considered as a 'general' factor by Petrides (2006), we are left with 26 items that should represent 4 factors: well-being, self-control, emotionality and sociability (Petrides and Furnham, 2006). Now we can use a simple exploratory visualization of the data to check the correlations between the variables and potentially already see the grouping of the variables in four factors. This can be done using the R command *corrplot* (Figure 1).

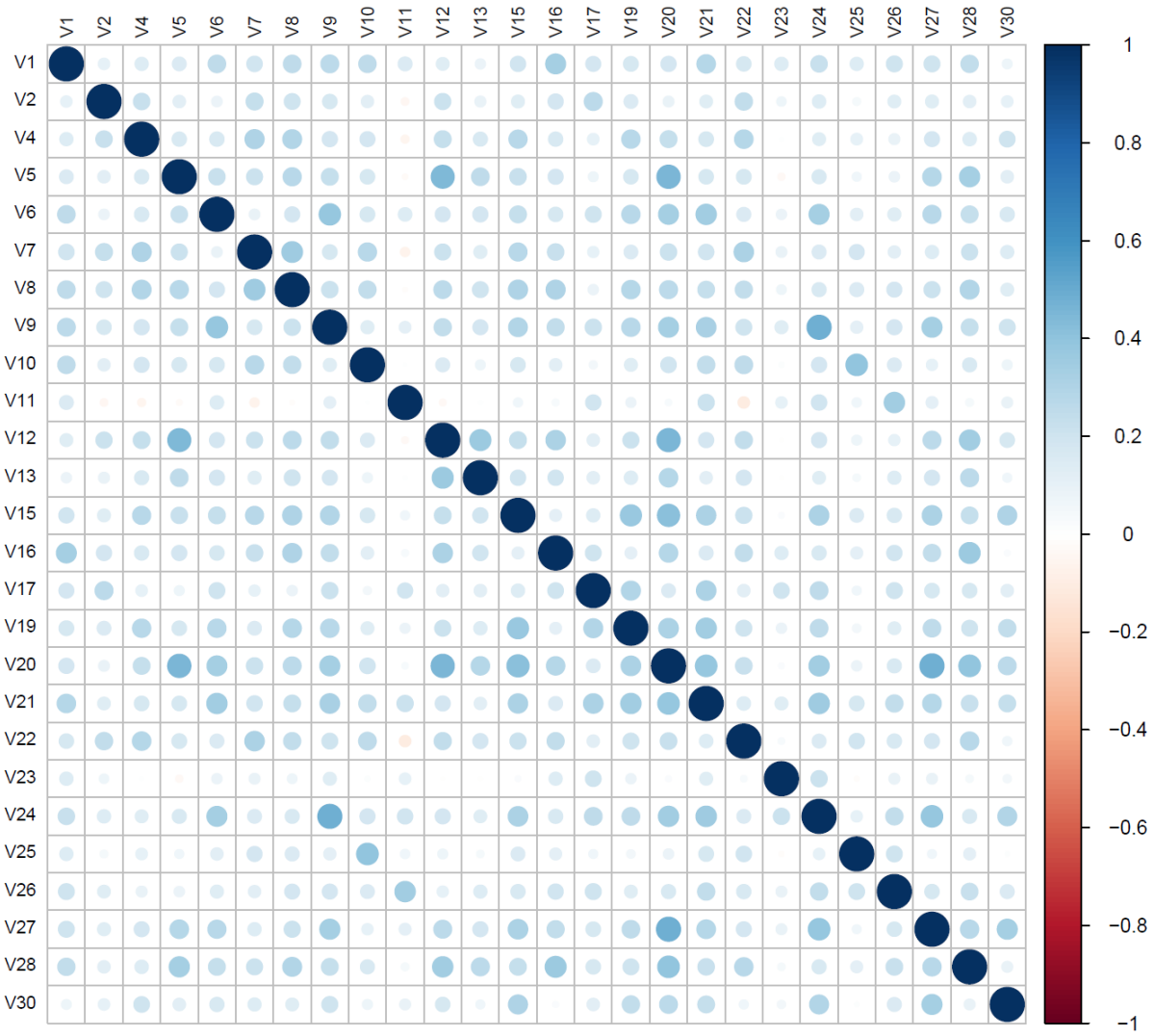

Figure 1. Correlation plot obtained by an R command *corrplot*

The CFA can now be used to test whether our data supports the four-factor structure proposed by Petrides and Furnham (2016). To do this, we have to specify the *laavan* specific model where each line represents a single latent factor as follows:

```
model <- 'Confidence =~ V30 + V15 + V19 + V24 + V27 + V21 + V9 + V6

          Connection =~ V12 + V5 + V28 + V13 + V16

          Uncertainty =~ V7 + V10 + V22 + V25 + V8 + V4 + V2

          Empathy =~ V11 + V26 + V17 + V1 + V23'
```

In the next step, a model can be fitted to the TEIQue-SF data using the command *model.fit* from the *laavan* package and followed by printing a summary of the CFA results after fitting the model to the data. The *laavan* command summary provides a very extensive list of the CFA results with many details. However, most of the users will be interested in some of the basic CFA measures such as CFI, TLI or RMSEA that can be obtained by a simple command *fitMeasures* as demonstrated in the supplementary materials in this paper.

The fit of our four-factor model resulted in RMSEA = 0.058, CFI = 0.818 (NB: these values are identical to those obtained by Snowden et al. (2015)), TLI = 0.797 and SRMR = 0.054 (NB: these values were not reported by Snowden et al.). Although some of the results like CFI or TLI which are below generally accepted threshold of 0.9, point at weak fit, one should be careful when relying on the CFA threshold values as described by Perry et al. (2015). As a final step of a CFA we use *semPaths* command from the semPath package in R which can be used to visualize normalized values for all items and corresponding four factors (Figure 2).

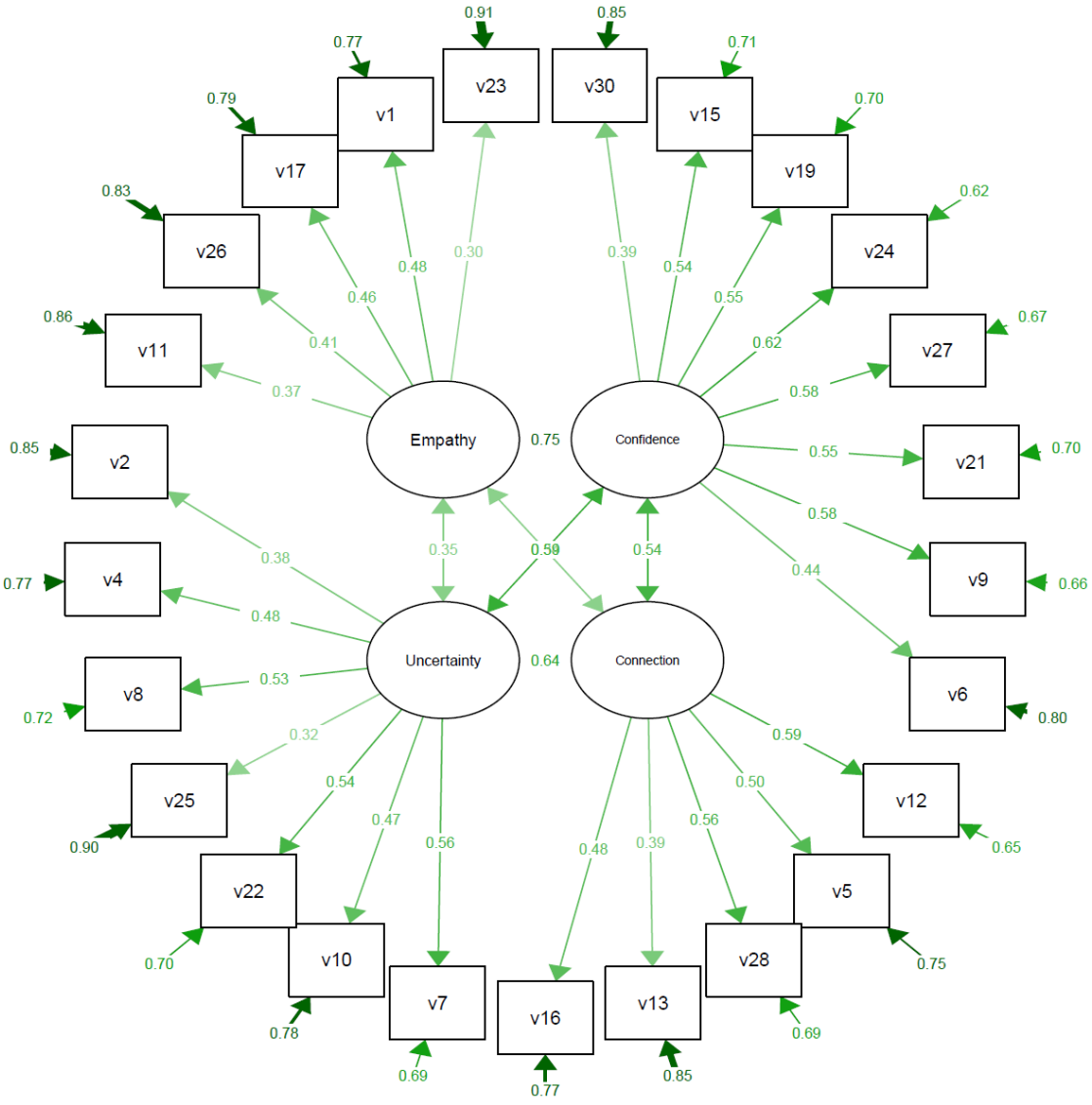

Figure 2. Diagram representing path diagram of all items with corresponding normalized values

Discussion

This study aimed to demonstrate the utility of R for research in nursing and to do this we chose to use the method of CFA. Specifically, we decided to replicate a published study using the same set of data from the TEIQue-SF analysed and published by Snowden et al. (2015) where a confirmatory factor analysis was conducted using AMOS. ~~Essentially, replicating the analysis with R has produce identical results in terms of the CFI and the RMSEA.~~ We only ran

the initial confirmatory analysis of the data which, in fact, did not support a good fit of the data to the proposed model. We did not continue—as did Snowden et al.— to conduct a subsequent exploratory phase where the model was restricted by identifying and systematically correlating error terms with the highest modification indices until the data fitted the model. ~~We consider that the initial steps have been replicated and thus that the utility of the R programme has been demonstrated.~~

Limitations

Our study had the advantage of a pre-existing analysis to guide our efforts. Thus, we were relying on a previous exploratory phase followed by an exploratory phase. We did not have to manipulate the data to achieve a suitable structure to analyse. Therefore, this is not an entirely 'real-world' example whereby we may have no initial indication of the latent structure of a database and would have to conduct a great many more preliminary steps.

Conclusions

~~Using a public domain, open source statistical package (R) we have demonstrated that the software, at least for the kind of analysis we conducted here, provides reliable results compared with one of the standard commercial packages for the same analysis.~~ Apart from the expense of obtaining commercial statistical software packages and the annual upgrades and licensing issues; R proves to be an economical and easily available programming language to conduct the type of multivariate analysis that is becoming routine in nursing research. Naturally, R requires some additional expertise that is not required to use commercially available packages; nevertheless, we believe that it is well within the grasp of nursing researchers to acquire that expertise and we strongly advocate that they do. We also advocate that, in universities generally and in schools where learning R can be helpful, that more classroom teaching is provided.

Thereby they will gain from the vast analytical capacity of R and will also make a contribution, along with the many other disciplines that use it, to the development of R.

References

Acock, A. C. (2005). SAS, Stata, SPSS: A Comparison. The Journal of Marriage and Family, 67(4).

Beaujean, A. A. (2013). Factor Analysis using R. Practical Assessment, Research & Evaluation, 18(4), 1-11.

Chavez, C., Rodriguez, M. C., Vue, K., & Cabrera, J. (2019). A Validity Argument Sensitivity Analysis of Social and Emotional Learning Measures with Few Items.

de Smith, M. J. (2018). Statistical Analysis Handbook: A Comprehensive Handbook of Statistical Concepts, Techniques and Software Tools. Retrieved 3 15, 2019, from https://www.statsref.com/StatsRefSample.pdf

Freudenthaler H.H., Neubauer A.C., Gabler P., Scherl W.G. & Rindermann H. (2008) Testing and validating the trait emotional intelligence questionnaire (TEIQue) in a German-speaking sample. Personality and Individual Differences, 45(7), 673–678. doi:10.1016/j.paid.2008.07.014.

Graham, J. M., Guthrie, A. C., & Thompson, B. (2003). Consequences of not interpreting structure coefficients in published CFA research: A reminder. Structural Equation Modeling, 10(1), 142-153.

Hu, L.T. and Bentler, P.M. (1999), "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives," Structural Equation Modeling, 6 (1), 1-55.

IBM. (2019). Retrieved 3 15, 2019, from https://www.ibm.com/analytics/spss-statistics-software

Nakazawa, M. (2011, June 27). R practice: Factor analysis. Retrieved March 13, 2019, from http://minato.sip21c.org/swtips/factor-in-R.pdf

Oberski, D. (2014). laavan. survey: An R package for complex survey analysis of structural equation models. Journal of Statistical Software, 57(1), 1-27.

Ozgur, C., Kleckner, M., & Li, Y. (2015). Selection of Statistical Software for Solving Big Data Problems: A Guide for Businesses, Students, and Universities. SAGE, 1-12.

Paura, L., & Arhipova, I. (2012). Advantages and Disadvantages of Professional and Free Software for Teaching Statistics. Information Technology and Management Science, 15, 9-14.

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: Caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. Measurement in Physical Education and Exercise Science, 19(1), 12-21.

Petrides, K. V. & Furnham, A. (2006). The role of trait emotional intelligence in a gender-specific model of organizational variables. Journal of Applied Social Psychology, 36, 552-569.

Petrides K.V. (2009) Psychometric properties of the Trait Emotional Intelligence Questionnaire (C. Stough, D.H. Saklofske, & J.D. Parker). Advances in the assessment of emotional intelligence. New York: Springer. pp. 85–101. doi: 10.1007/978-0-387-88370-0_5

R Development Core Team. (2011). R: A language and environment for statistical computing. Retrieved 3 15, 2019, from http://www.R-project.org

Rietvel, T., & van Hout, R. (1993). Statistical Techniques for the Study of Language and Language Behaviour. Berlin: Mouton de Gruyter.

Rossiter, D. G. (2017). Tutorial: An axample of statistical data analysis using the R environment for statistical computing. Retrieved 3 15, 2019, from http://www.css.cornell.edu/faculty/dgr2/teach/R/R_corregr.pdf

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. Educational and Psychological Measurement, 79(2), 310-334.

SAS Institute. 2011. The SAS system for Windows. Release 9.2. SAS Inst.., Cary, NC

Smyth, R., & Johnson, A. (n.d.). Factor Analysis. Retrieved 3 13, 2019, from https://www.uwo.ca/fhs/tc/labs/10.FactorAnalysis.pdf

Snowden, A., Watson, R., Stenhouse, R. and Hale, C., 2015. Emotional Intelligence and Nurse Recruitment: Rasch and confirmatory factor analysis of the trait emotional intelligence questionnaire short form. Journal of advanced nursing, 71(12), pp.2936-2949.

STATA. (2019). Retrieved 3 15, 2019, from https://www.stata.com/

Tabachnick, B.G. and Fidell, L.S. (2007), Using Multivariate Statistics (5th ed.). New York: Allyn and Bacon.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Venables, W. N., Smith, D. M., & R Core Team. (2019, 3 11). An Introduction to R. Retrieved 3 15, 2019, from https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

Yong, A. G., & Pearce, S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. Tutorials in Quantitative Methods for Psychology, 9(2), 79-94.

Zellner, K., & Boerst, C. J. (2007). Statistics Used in Current Nursing Research. Journal of Nursing Education, 46(2).