Practical Assessment, Research, and Evaluation

Volume 18 *Volume 18, 2013* Article 8

2013

Determining the Number of Factors to Retain in EFA: Using the SPSS R-Menu v2 0 to Make More Judicious Estimations

Matthew G. R. Courtney

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 18, Number 8, April 2013 ISSN 1531-7714

Determining the Number of Factors to Retain in EFA: Using the SPSS R-Menu v2.0 to Make More Judicious Estimations

Matthew Gordon Ray Courtney
The University of Auckland (New Zealand)

Exploratory factor analysis (EFA) is a common technique utilized in the development of assessment instruments. The key question when performing this procedure is how to best estimate the number of factors to retain. This is especially important as under- or over-extraction may lead to erroneous conclusions. Although recent advancements have been made to answer the number of factors question, popular statistical packages do not come standard with these modern techniques. This paper details how to program IBM SPSS Statistics software (SPSS) to conveniently perform five modern techniques designed to estimate the number of factors to retain. By utilizing the five empirically-supported techniques illustrated in this article, researchers will be able to more judiciously model data.

Exploratory factor analysis is an established and popular technique in the social and behavioral sciences to model latent factors (Cudeck & MacCallum, 2007). EFA is particularly appropriate for scale development where little theoretical basis exists for specifying the number and patterns of common factors. In this context, it is critical that practitioners extract an appropriate number of factors because this decision has a direct effect on results and subsequent theory development. However, empirically determining the number of factors to retain when performing EFA has been identified as a significant challenge to its successful implementation. Henson and Roberts (2006) reviewed 60 articles that utilized EFA across four prominent journals: *Educational and Psychological Measurement*, *Journal of Educational Psychology*, *Personality and Individual Differences*, and *Psychological Assessment*. They found that of the 60 recent articles, 55 relied on Kaiser's (1960) dubious eigen-value-greater-than-one rule (K1) and Cattell's (1966) scree plot methods, while only four utilized Horn's (1965) highly recommended parallel analysis (PA). Moreover, none of the 60 papers made use of multiple modern techniques in an attempt to find convergence, such as PA and Velicer's (1976) minimum average partial (MAP) procedures. The fact that none of the 60 articles utilized multiple modern techniques is cause for concern given that EFA experts have long recommended such an approach (Gorsuch, 1983; Zwick & Velicer, 1986; Velicer, Eaton, & Fava, 2000; Hayton, Allen, & Scarpello, 2004). Ruscio and Roche's (2012) simulation study made use of several modern techniques to demonstrate the empirical advantage of seeking convergence. The authors assessed the performance of several modern methods, including the use of their own variant of PA, comparison data (CD). The authors explain that "In all 10,000 target datasets, PA and CD agreed with one another by identifying the same number of factors 78.1% of the time. When the two methods agreed, the accuracy rate was 92.2%" (p. 291). The authors also demonstrated that when other modern methods were in agreement, accuracy could be increased even further.

Despite the associated advantages of using multiple modern techniques to estimate dimensionality, easy access to many of these procedures from within popular software programs has been limited. O'Connor (2000) has written syntax for performing MAP and PA for both SPSS and SAS programs, however making use of such code can be time consuming and complicated for practitioners unfamiliar with syntactic functionality. In a previous article in *Practical Assessment, Research and Evaluation,* Ledesma and Valero-Mora (2007) described and illustrated how to perform PA with the ViSta-PARAN program. However, the program did not include any other modern procedures. This article adds to the literature by illustrating how to perform five empirically based procedures from within SPSS's main interface (Basto, 2012). To do this, the present article: 1) briefly reviews the previous-reported K1, scree plot, and Very Simple Structure (VSS) methods for determining the number-of-factors-to-retain, 2) introduces the recommended PA, optimal coordinates (OC), acceleration factor (AF), MAP, and CD procedures made available in SPSS and explains how to modify these procedures for different data situations, 3)

demonstrates how to carry out the recommended procedures in SPSS correctly with an example, and, 4) provides a detailed step-by-step illustration of how to correctly install the SPSS R-menu v2.0.

THE NUMBER-OF-FACTORS-TO-RETAIN QUESTION

Hayton, Allen, and Scarpello (2004) identify three reasons why the decision concerning the number of factors to retain is essential. First, the decision concerning the number of factors to retain appears more important to EFA than extraction and rotation methods because there is evidence of relative robustness across such methods (Zwick & Velicer, 1986). Second, EFA must balance parsimony with sufficiently representing the underlying sets of correlations, so its usefulness depends on its ability to differentiate major from trivial factors (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Lastly, empirical evidence suggests that under- and over-extraction represent substantial errors that can significantly alter the solution and subsequent interpretation of EFA results (Velicer, Eaton, & Fava, 2000). Potentially useful or theoretically interesting scales may be excluded if too few factors are extracted. Conversely, the factor or pattern loadings may appear weak if items that would otherwise cluster together nicely are spread across an artificially large number of subscales. In summary, both under- and over-extraction can be viewed as potentially detrimental to scale development and instrumentation. Therefore the appropriate estimation of the number of factors to retain is of significance to EFA practitioners.

Methods for determining the number of factors to retain

A host of methods have been suggested for determining the number of factors to retain in EFA. In addition, many simulation studies have been carried out to evaluate the comparative efficiency of these methods (Zwick & Velicer, 1986; Garrido, Abad, & Ponsoda, 2011; Garrido, Abad, & Ponsoda, 2012; Basto & Pereira, 2012; Ruscio & Roche, 2012). Simulation studies allow researchers to predetermine the number of underlying factors in each simulated target dataset. Therefore, such studies are able to measure a procedure's comparative efficiency by assessing the accuracy (in percentage) that it correctly estimates the number of factors in all target simulations. Although, when EFA is more commonly used as part of data reduction or theory development, a true number of factors can neither be assumed nor determined. Nevertheless, despite limitations of simulated performance, it is widely recognized as the best approach to determining the real-world practicability of such procedures (Zwick & Velicer, 1986; Garrido, Abad, & Ponsoda, 2011; Garrido, Abad, & Ponsoda, 2012; Basto & Pereira, 2012; Ruscio & Roche, 2012). Therefore, a brief review of the previously used K1, scree, and VSS methods, and their relative performance in simulation studies will be provided. Thereafter, a review of the five more empirically-supported techniques made available in SPSS R-menu v2.0 will be provided in kind.

Kaiser's eigenvalue-greater-than-one rule

The eigenvalue-greater-than-one rule (K1), proposed in 1960 by Kaiser, is the default setting of many statistical packages and is the most well-known and most utilized method in practice. In accordance with this rule, only the factors that have eigenvalues greater than one are retained. Despite its widespread use and simplicity, it is widely agreed that the method is dubious. Fabrigar et al. (1999) identified three general issues concerning the use of this method. First, the method was first proposed for principal components analysis (PCA). In this case eigenvalues were drawn from the correlation matrix with unities at the diagonal. Fabrigar et al. (1999) argue that this method is not valid for EFA where eigenvalues are drawn from a correlation matrix with *communality* estimates at the diagonal. Second, it makes little sense defining a factor with an eigenvalue of 1.01 as major and another of .99 as trivial as suggested by the rule. Third, multiple simulation studies have demonstrated the tendency of this method to over-estimate the number of factors. For example, Ruscio and Roche's (2012) simulation study summarizing the accuracy of various methods across 10,000 target datasets, determined that the K1 rule grossly over-estimated the number of factors and was only correct 8.77% of the time. Despite this, the K1 rule is the default procedure in IBM SPSS Statistics software (SPSS) for determining the number of factors to retain in EFA.

Cattell's Scree test

Another popular method for determining the number of factors to retain is Cattell's (1966) scree test, which involves eye-balling the plot of the eigenvalues for a break or hinge (also referred to as an "elbow"). The rationale for this test is based on the idea that a few major factors will account for the most variance, resulting in a "cliff", followed by a shallow "scree" depicting the consistently small and relatively shallow error variance described by minor factors. Although this test works well with strong factors, it suffers from ambiguity and subjectivity when there is no clear break or hinge in the depicted eigenvalues. Despite suffering from inter-rater reliability bias, simulation suggests that the test can be more accurate and less variable than the K1 method (Zwick & Velicer, 1986). In Zwick and Velicer's (1986)

Monte Carlo evaluation study, the scree test's relative performance was assessed in its ability to determine the correct number of factors in 480 target sample datasets. Based on the mean of two trained performers, the scree test correctly identified the correct number of factors 41.7% of the time, while the K1 was not correct once (0%), incorrectly over-estimating the number of factors in each sample. Although inherently subjective, Cattell's (1966) scree test can be easily carried out in standard versions of SPSS by selecting *Scree plot* in the *Extraction* dialogue box.

Very Simple Structure Criterion

Revelle and Rocklin (1979) proposed using the very simple structure criterion (VSS) for determining the number of factors to extract. Revelle (2011) explained that most EFA practitioners tend to interpret factor output by focusing on the largest loadings on a factor pattern matrix for a variable and ignoring the smaller ones. Revelle and Rocklin's (1979) VSS criterion operationalizes this tendency by assessing the extent to which the original correlation matrix is reproduced by a simplified pattern matrix, in which only the highest loading for each item is retained, all other loadings being set to zero. The VSS criterion for assessing the extent of replication can take values between 0 and 1, and is a measure of the goodness-of-fit of the factor solution. The VSS criterion is gathered from factor solutions that involve one factor (k = 1) to a user-specified theoretical maximum number of factors. Thereafter, the factor solution that provides the highest VSS criterion determines the optimal number of interpretable factors in the matrix. In an attempt to accommodate datasets where items covary with more than one factor (i.e., more factorially complex data), the criterion can also be carried out with simplified pattern matrices in which the highest *two* loadings are retained, with the rest set to zero (Max VSS complexity 2). However, Revelle (2011) explains that simulation studies suggest that the VSS procedure will only work well if the complexities of some of the items are no more than two. In addition, at the time this paper was drafted, no robust simulation research concerning the performance of the VSS criterion relative to other modern procedures could be found. The procedure is part of the SPSS R-menu v2.0.

Optimal Coordinate and Acceleration Factor

In an attempt to overcome the subjective weakness of Cattell's (1966) scree test, Raiche, Roipel, and Blais (2006) presented two families of non-graphical solutions. The first method, coined the optimal coordinate (OC), attempts to determine the location of the scree by measuring the gradients associated with eigenvalues and their preceding coordinates. The second method, coined the acceleration factor (AF), pertains to a numerical solution for determining the coordinate where the slope of the curve changes most abruptly. Both of these methods have out-performed the K1 method in simulation (Raiche, Roipel, & Blais, 2006; Ruscio & Roche, 2012). In the Ruscio and Roche study (2012), the OC method was correct 74.03% of the time rivaling the PA technique (76.42%). The AF method was correct 45.91 % of the time with a tendency toward under-estimation. Both the OC and AF methods, generated with the use of Pearson correlation coefficients, were reviewed in Ruscio and Roche's (2012) simulation study. Results suggested that both techniques performed quite well under ordinal response categories of two to seven (C = 2-7) and quasi-continuous (C = 10 or 20) data situations. Both the OC and AF techniques are part of the new SPSS R-menu v2.0.

Velicer's Minimum Average Partial

Velicer's (1976) MAP test "involves a complete principal components analysis followed by the examination of a series of matrices of partial correlations" (p. 397). The squared correlation for Step "0" (see Figure 4) is the average squared off-diagonal correlation for the unpartialed correlation matrix. On Step 1, the first principal component and its associated items are partialed out. Thereafter, the average squared off-diagonal correlation for the subsequent correlation matrix is then computed for Step 1. On Step 2, the first two principal components are partialed out and the resultant average squared off-diagonal correlation is again computed. The computations are carried out for k minus one step (k representing the total number of variables in the matrix). Thereafter, all of the average squared correlations for each step are lined up and the step number in the analyses that resulted in the lowest average squared partial correlation determines the number of components or factors to retain (Velicer, 1976). By this method, components are maintained as long as the variance in the correlation matrix represents systematic variance, as opposed to residual or error variance. Although methodologically akin to principal components analysis, the MAP technique has been shown to perform quite well in determining the number of factors to retain in multiple simulation studies (Zwick & Velicer, 1986; Garrido, Abad, & Ponsoda, 2011; Ruscio & Roche, 2012).

Various modifications have been proposed to improve the accuracy of the procedure in simulation. The MAP test was revised with the average squared off-diagonal correlation ($MAPr^2$) raised to the fourth power ($MAPr^4$) in 2000 (Velicer, Eaton, & Fava, 2000). Despite the suggested revision, recent research has suggested that the $MAPr^2$ version outperforms the $MAPr^4$ version for continuous data.

For example, under simulation, the MAPr[2] was 65% accurate in determining the correct number of factors with continuous data, whereas the MAPr[4] version was only 55% accurate (Garrido, Abad, & Ponsoda, 2011, p. 560).

Research suggests that the MAP procedure can be modified to accommodate strictly ordinal variables. For such data situations, the correlation matrix generated to perform the MAP test can be created using polychoric correlations (Olsson, 1979), as opposed to Pearson's product-moment correlation coefficients, which have been shown to attenuate the relationship between categorical variables (Babakus, Ferguson, & Joreskog, 1987; Bollen & Barb, 1981). Polychoric correlations rest on the assumption that the observed categories function as proxies for bivariate normal continuous phenomena and have been shown to produce unbiased parameter estimates for EFA and CFA procedures (Flora & Curran, 2004; Holgado-Tello, Chacon-Moscoso, Barbero-Garcia, & Vila-Abad, 2010). A simulation study by Garrido, Abad, & Ponsoda (2011) suggested that, for ordinal variables, using polychoric correlations and the squared partial correlations (MAPρ[4]) leads to more accurate estimations of dimensionality than other variations (e.g., MAPr[4]). The MAP procedure is part of the new SPSS R-menu v2.0.

Horn's Parallel Analysis

Among the many techniques proposed to determine the number of factors to retain, Horn's (1965) Parallel Analysis has emerged as one of the most strongly recommended techniques (Zwick & Velicer, 1986; Fabrigar et al., 1999; Velicer, Eaton, & Fava, 2000; Hayton, Allen, & Scarpello, 2004; Peres-Neto, Jackson, & Somers, 2005; Henson and Roberts, 2006; Ruscio & Roche, 2012; Garrido, Abad, & Ponsoda, 2012). The K1 rule posits that only factors with eigenvalues greater than one should be retained. Horn (1965) argued that the K1 rule was not applicable to sample-based research because its proofs were based on population statistics. Horn argued that because of sampling error in the computation of latent roots, some components from uncorrelated variables in the true population could have eigenvalues over one. Consequently, Horn (1965) proposed the PA method, which takes into account the proportion of variance resulting from sampling error. Thus, PA can be defined as a sample alternative to the K1 rule (Garrido, Abad, & Ponsoda, 2012, p. 2). The PA method is implemented by generating a large number of data matrices from random data. Each matrix is generated in parallel with the real data meaning that matrices with the same number of cases and variables are created. Factors are retained in the real data as long as they are greater than the mean eigenvalue generated from the random data matrices.

Notwithstanding the recommendations to use PA in empirical research, its application is not simple. Different modifications have been proposed to improve the accuracy of the procedure in simulation studies. Recent research by Ruscio and Roche (2012) suggests that PA using principal components extraction, with Pearson product-moment correlations and the mean eigenvalue criterion (PA-PCArm) performed very well across a range of data conditions (C = 2-7, 10, & 20) with a degree of accuracy of 76.42%. Additionally, a recent simulation study by Garrido, Abad, & Ponsoda (2012) suggests that the method for generating the random criterion variables can be improved by using random column permutations of the real data matrix. This modification is more appropriate as it maintains the same level of skewness and number of response categories as those from the real data (Garrido, Abad, & Ponsoda, 2012). In terms of adapting the above PA procedure for ordinal type variables, simulation research by Garrido, Abad, and Ponsoda (2012) suggests that the procedure be carried out using principal components estimation, polychoric correlations, and the mean eigenvalue criterion (PA-PCAρm). The PA procedure is part of the new SPSS R-menu.

Ruscio and Roche's Comparison Data

In 2012 Ruscio and Roche introduced the comparative data (CD) technique in an attempt improve upon the PA method. In describing the method, the authors state that "rather than generating random datasets, which only take into account sampling error, multiple datasets with known factorial structures are analyzed to determine which best reproduces the profile of eigenvalues for the actual data" (p. 258). The authors explain that the strength of the technique is its ability to not only incorporate sampling error, but also the factorial structure and multivariate distribution of the items. Ruscio and Roche's (2012) simulation study determined that the CD technique outperformed all other methods aimed at determining the correct number of factors to retain. In their simulation study, the CD technique, utilizing Pearson correlations accurately predicted the correct number of factors 87.14% of the time. Although, it should be noted that simulated data did not involve more than five factors. Therefore, the applicability of the procedure to estimate factorial structures beyond five factors is yet to be tested.

Like other authors (Goodman & Kruskal, 1954; Bentler, 2005), Ruscio questioned the applicability of polychoric correlations given the assumption of underlying normality

(personal communication, 29 November, 2012). To accommodate ordinal variables without requiring normality, Ruscio suggested the use of Spearman rank-order correlations for the CD procedure (CDr_s) and has demonstrated that this approach may be a more appropriate estimator for ordinal data situations (Ruscio, 2012). The CD procedure is part of the new SPSS R-menu version 2.0.

To summarize, eight procedures aimed at determining the number of factors to retain have been discussed. Despite the use of the K1 rule as the default value in some standard computer packages (SPSS, SAS), its relative performance is very poor and is not recommended. The scree test, although having demonstrated moderate performance, largely depends on the ability of the rater and suffers from inherent inter-rater reliability and subjectivity. Thus, the scree test is also not recommended here. The VSS criterion, despite its apparent pragmatism and inclusion in the SPSS R-menu v2.0, lacks the empirical support necessary for recommendation. Based on empirical research aimed at determining the relative performance of several techniques, the CD, PA, OC, MAP, and AF procedures are recommended here. A summary of the five suggested techniques presented in this article, alongside a review of their estimated accuracy and recommended modification for ordinal data, are presented in Table 1.

Table 1. Summary of Modern Techniques for Determining Number of Factors to Retain in EFA

Modern Techni-que	Standard for all Data Types	% Accuracy	Bias in simulation	Recommended version for ordinal data
CD	CDr	87.14	Slight under-extraction	CDr_s
PA	PA-PCArm	76.42	Unbiased	PA-PCAρm
OC	OCr	74.03	Slight under-extraction	Not established
MAP	$MAPr^2$	59.6	Moderate under-extraction	$MAP\rho^2$
AF	AFr	45.91	Substant-ial under-extraction	Not established

Note: Accuracy and Bias estimates taken from Ruscio & Roche's (2012) simulation study (p. 289). Although the OC and AF procedures may be carried out with Spearman or polychoric correlations in the SPSS R-menu v2.0, such modifications are not established.

MAKING USE OF THE SPSS R-MENU V2.0

R is an open source statistical software program for statistical and graphical computing. It offers an enormous range of statistical procedures written by contributors from all over the world. In January 2012, Basto and Pereira wrote an article in the *Journal of Statistical Software* entitled *An SPSS R-Menu for Ordinal Factor Analysis*. Their paper explained how practitioners could use the SPSS interface to essentially outsource more sophisticated procedures to R and have R report back in regular SPSS output. Although the authors mentioned several software packages and plugins to install the R-menu (v1.0) in SPSS, navigating the appropriate websites and correctly installing the software can be complicated. For this reason, a detailed step-by-step illustration of the installation of version 2.0 is provided in the final section of this article. Readers are advised to carefully follow these instructions to ensure the latest R-menu is correctly installed.

Since its introduction in January 2012, the SPSS R-menu has undergone two major upgrades with respect to its ability to appropriately estimate the number of factors in a given matrix. First, of current relevance to the MAP and PA procedures, v2.0 is now able to more astutely carry out estimates for matrices that use combinations of ordinal and continuous variables. Basto (2012a) names the new correlation matrix option "heterogeneous" reflecting the integration of polychoric (ordinal-ordinal), polyserial (ordinal-continuous), and Pearson (continuous-continuous) correlations in the one matrix. Second, Basto and Pereira (2012a) have built the CD procedure into version 2.0.

The following is a step-by-step illustration explaining how to make use of the SPSS R-menu v.2.0 to perform the CD, PA, OC, MAP and AF procedures. The example, from a real sample of 484 survey participants, uses 14 total variables (10 ordinal and four continuous). The dataset has no missing values.

Step 1: Setting up MAP, PA, OC, and AF Procedures

a. Install the SPSS R-menu v2.0 by following the eight steps described in the final section of this paper, *INSTALLING R, AND THE SPSS R-MENU V2.0.*

b. To start, open the SPSS dataset of interest.

c. Thereafter, go to *Analyze* → *Dimension Reduction* → *ORD R Factor v2.0* to open the *R Factor v2.0* dialogue box.

d. Thereafter, click the dialogue box of relevance (*N. Factors: MAP-VSS-PA-OC-AF*) and set up the

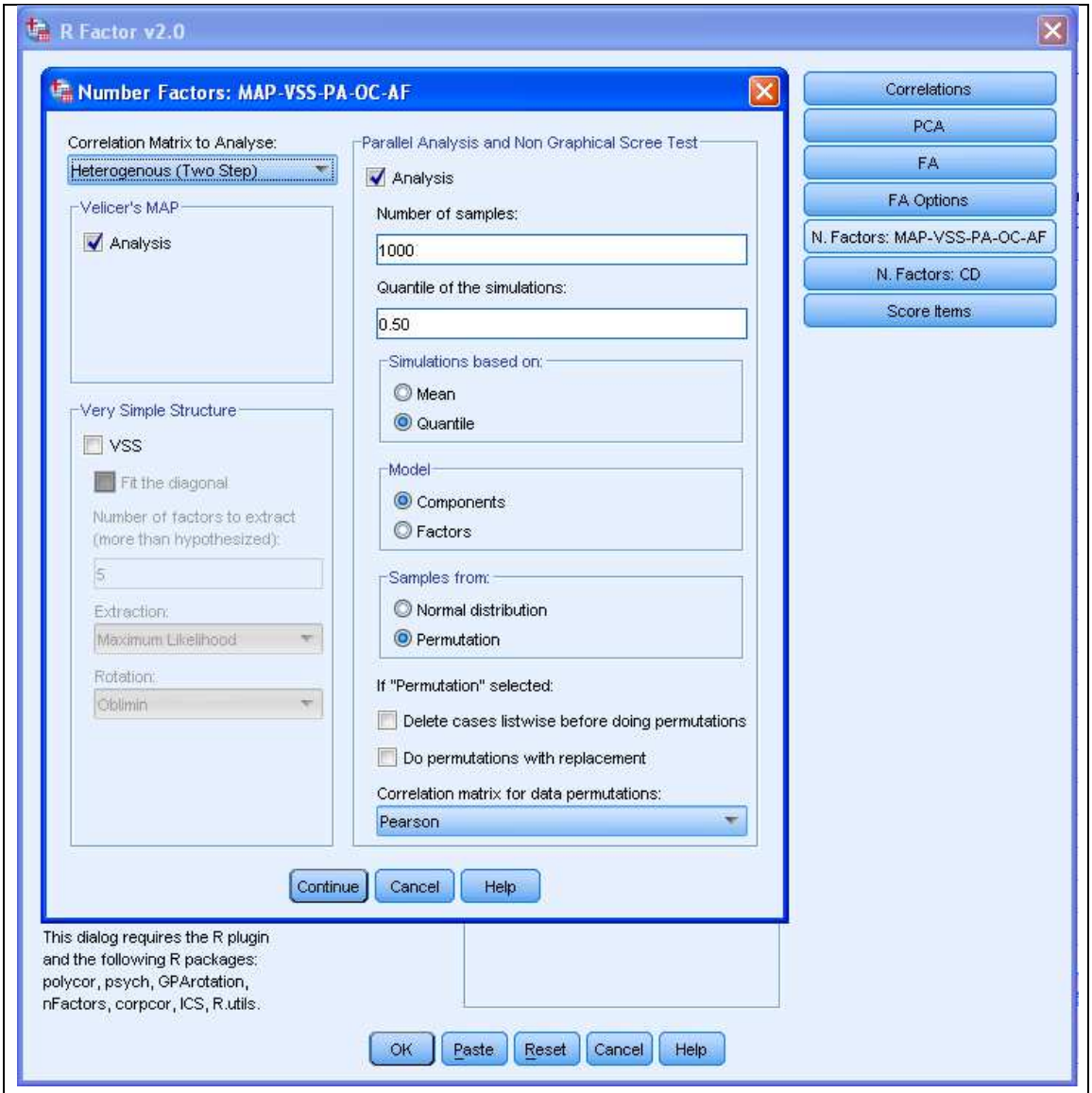

Figure 1. Setup for MAP, PA, OC, and AF Procedures for Heterogeneous Data

procedure in accordance with the screen shot in Figure 1.

e. With reference to the current dataset, select the *Heterogeneous (Two Step)* estimation method to generate the correlation matrix. This accommodates the combination of ordinal and continuous variables. Although the *Heterogeneous (Max. Lik.)* estimation method could have been selected, this would be computationally time-intensive. In regards to this decision, Olsson (1979) has demonstrated that the difference between the two-step and maximum likelihood method is negligible. Therefore, the two-step method is adopted for the purpose of computational convenience.

f. Thereafter, select the *Velicer's MAP* checkbox. This means that both MAP[2] and MAP[4] procedures will also be carried out on the heterogeneous matrix.

g. Check the The *Parallel Analysis and Non Graphical Scree Test* checkbox. In accordance with O'Connor

(2000), input 1000 sample datasets to be generated for the PA procedure. Additionally, set the quantile to 0.50 (median eigenvalue criterion), set the simulations to be based on the *Quantile* and ensure the model uses *Components* (PCA) for the extraction method (Garrido, Abad, & Ponsoda, 2012).

h. After the *If "Permutation" selected* script, leave the *Delete cases listwise before doing permutations* box unchecked. For the dataset in question, this makes no difference as no missing data exists. However, if missing data did exist, it is recommended that this box be checked to minimize the confounding potential of missing data on sample permutations.

i. Leave the *Do permutations with replacement* box unchecked to provide truly random sample datasets inline with bootstrapping methods (M. Basto, personal communication, November 2012).

j. Set the *Correlation matrix for data permutations* to *Pearson*. This is computationally more efficient and makes little difference, as the initial correlation matrix analyzed in the procedure is already heterogeneous. At this point press *Continue* to save the setup for the four procedures.[1]

Step 2: Set up CD Procedure

a. With the full *ORD R Factor v2.0* dialogue box open, select the *N. Factors: CD* dialogue box and set up the procedure in accordance with Figure 2.

b. Check the *Perform analysis only till nonsignificant improvement* box. Doing this saves time as the calculations are only carried out until the optimal number of factors is reached. If unchecked, the procedure is performed until the *Largest number of factors* chosen is reached.

c. Within the *Missing Data* dialogue area, select either option. For the dataset in question, select *Delete cases listwise before doing the analysis.* (however, this makes no difference as no missing data exists). If missing data did exist, choosing the same option would also be most appropriate as the *Keeping missing data* option potentially limits the randomness of the forthcoming sample datasets to be generated.

d. To best deal with the ordinal and continuous data conditions, select *Spearman* from the *Correlations employed* options.

e. Choose the *Largest possible number of factors* expected from the matrix in accordance with what you believe is the theoretical maximum number in the dataset.

f. Based on simulations by Ruscio and Roche (2012, p. 288), set the *Size of finite population of comparison data* to 10,000. Set the *Number of samples drawn from each population* to 500. And, set the *Alpha level when testing significance of improvement by adding factor* to 0.3.

g. By pressing *Continue*, the five procedures are set up ready to execute.

Step 3: Run the Five Procedures

a. After setting up the procedure as above, simply press *OK* on the main *R Factor v2.0* dialogue box to carry out the five procedures. However, SPSS syntax is limited in that an error can occur in very large datasets if the names of the variables, listed in series, happen to exceed 251 characters (Error # 6892). In this case, take the following course of action explained in b, c, d, and e, below.

b. At the bottom of the *R Factor v2.0* dialogue box, press *Paste*, as opposed to *OK*.

c. After pressing *Paste*, the syntax window appears as depicted in the Figure 3 screenshot.

d. The operations to be carried out are summarized on the left of the syntax window. By clicking on the red *BEGIN PROGRAM* text in the left summary panel, the problematic code automatically appears in the main window. The red code identifying the matrix data variables (separated by spaces) is displayed in the main panel. Simply place cursor at the end of the problematic variable(s) (see red arrow, Figure 3) and press "enter" to break the line, essentially spreading the line down the page, to resolve this issue (the problematic red code should change to black).

e. To run the entire code, simply press *Run* → *All* from within the syntax window.

[1] The current setup calls for the OC and AF procedures to be carried out based on the heterogeneous correlation matrix. Simulation studies concerning the validity of this modification have not yet been undertaken. Users, therefore, may want to re-run the procedure with Pearson-only correlations.

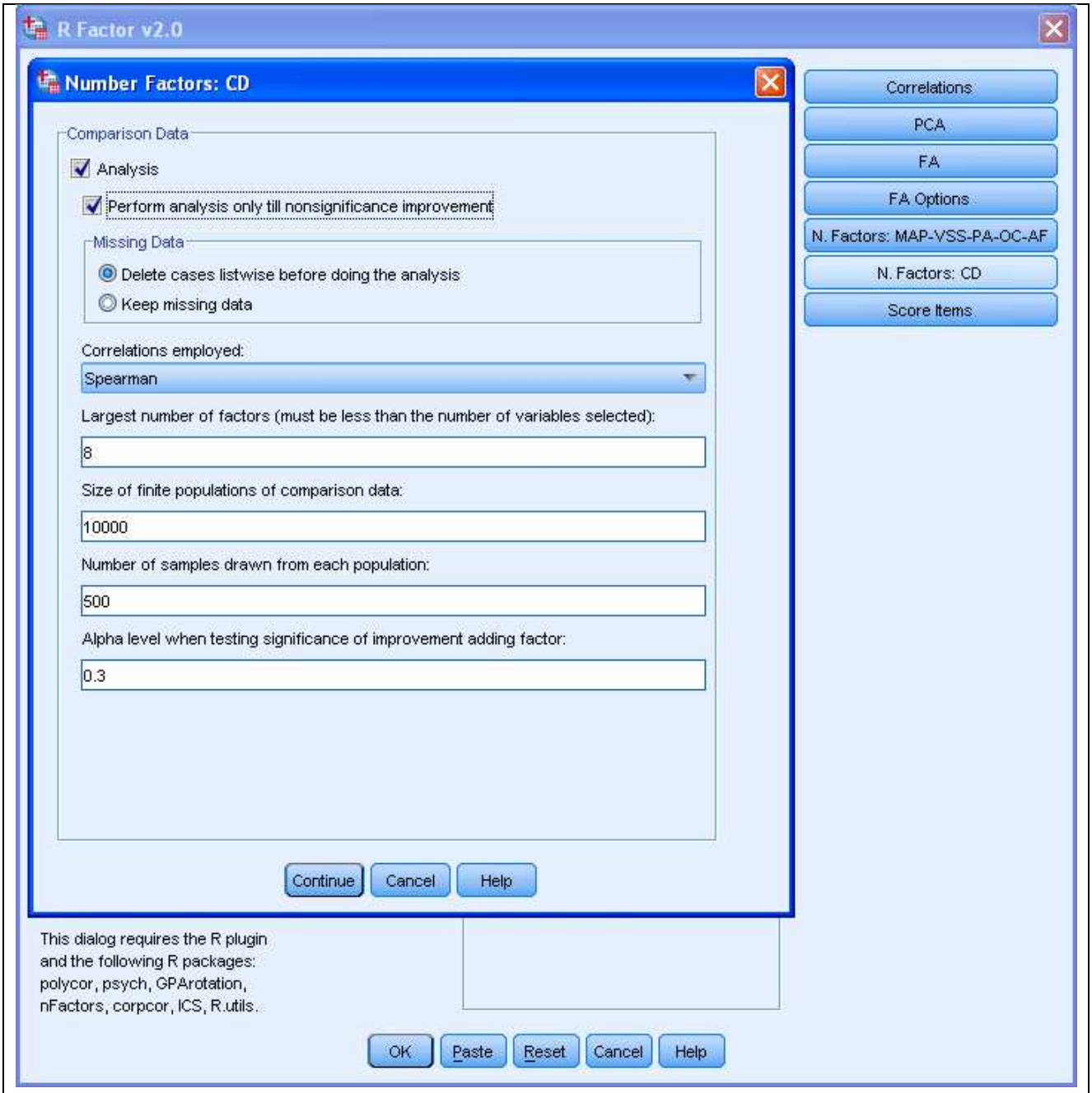

Figure 2. Setup for CD Procedure for Heterogeneous Data

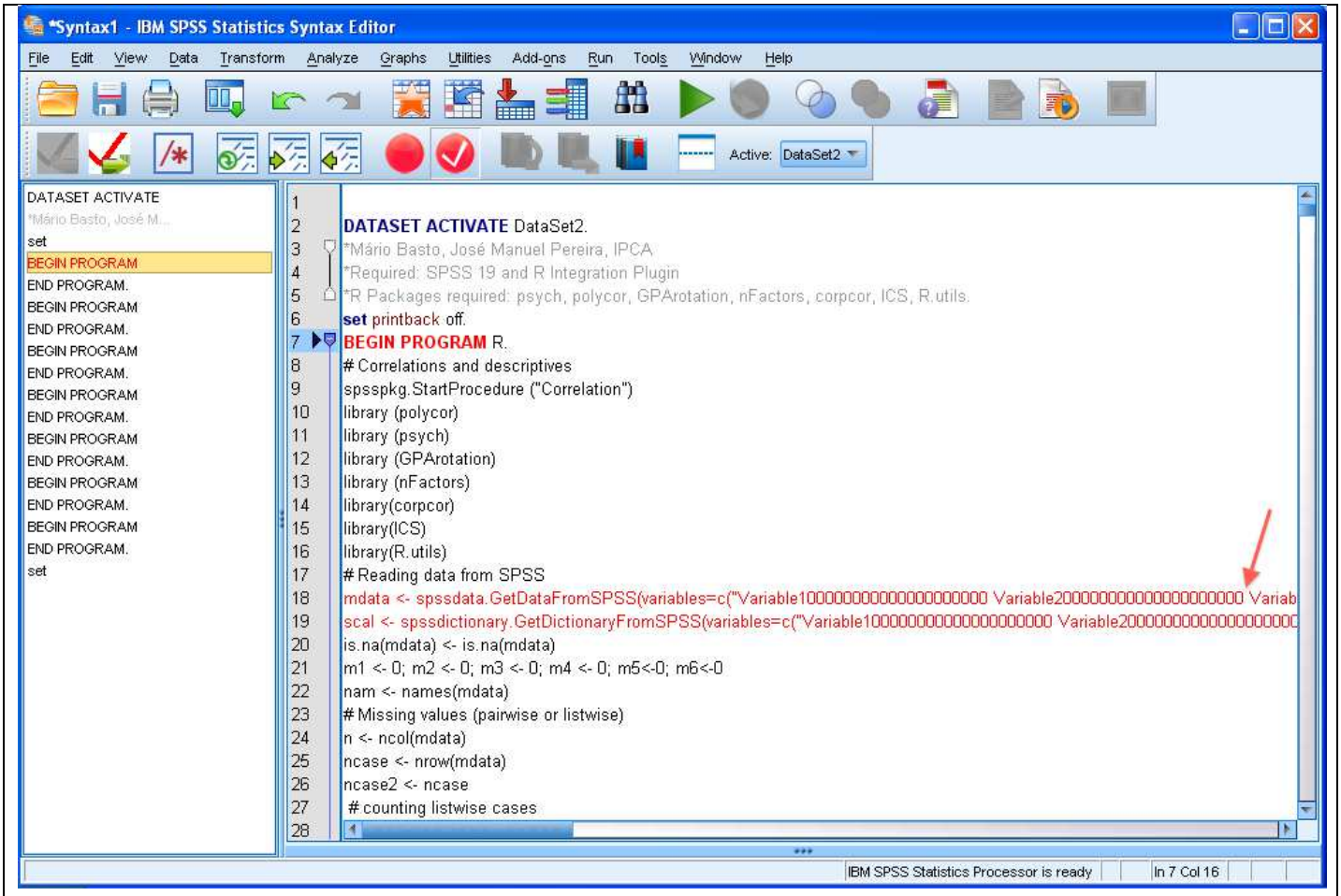

Figure 3: Dealing with 251+ Characters of Variables

Step 4: Interpret the SPSS Output and Make Decision

a. To identify the recommended factor numbers from the five procedures, identify the following SPSS output tables and graphs depicted in Figures 4 through 9. In Figure 4, the MAP output gives a distinct 3rd step minimum squared average partial correlation of .039 suggesting three factors (see highlight on Figure 4 and the suggested number of factors in Figure 5).

Velicer's MAP values

	Squared average partial correlations	4th average partial correlations
0	.202	.076
1	.089	.020
2	.064	.012
3	.039	.008
4	.054	.014
5	.072	.015
6	.090	.025
7	.118	.034
8	.148	.067
9	.165	.081
10	.234	.142
11	.364	.258
12	.481	.363

Figure 4. MAP Partial Correlation Output

Velicer's Minimum Average Partial Test

	Velicer's Minimum	
	Minimum	Components to retain
Squared MAP	.039	3
4th power MAP	.008	3

Figure 5: Velicer's MAP Output

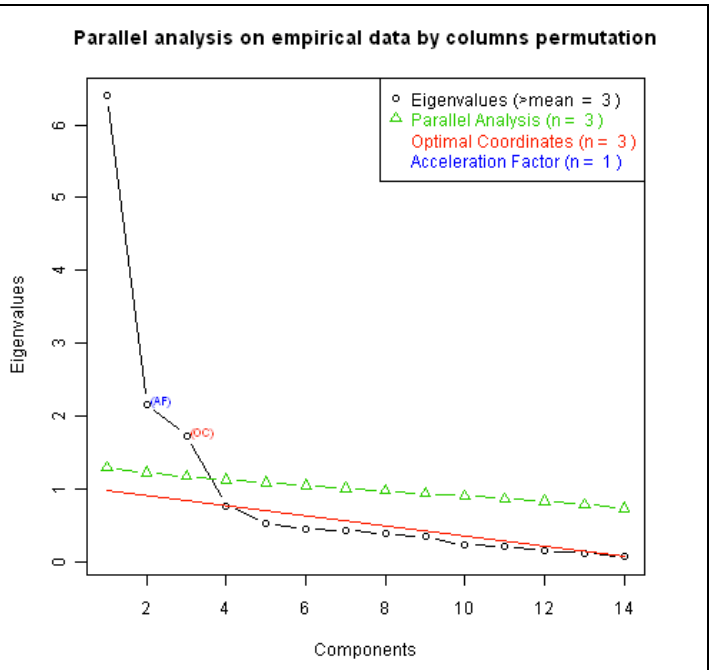

Figure 6: Eigenvalue, PA, OC, and AF Output

Fit to Comparison Data

Number of factors	RMSR Eigenvalue	p-value
1 factor	.435	NA
2 factor	.311	.000
3 factor	.079	.000
4 factor	.084	.943
5 factor	.082	.149
6 factor	.076	.003
7 factor	.071	.002

Figure 7. Fit to Comparison Data

Comparison Data

Correlations	Number of factors to retain
Spearman	3

Figure 8: CD Factor Number Output

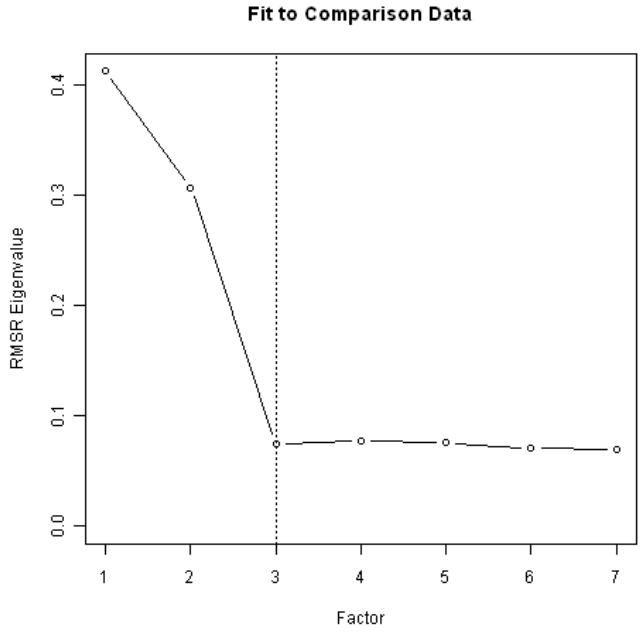

Figure 9. Fit to Comparison Data Plot

In Figure 6, against the plot of eigenvalues, the PA and OC procedures both estimate three factors, while the OC estimate, being biased to under-extraction, estimates one factor. Figure 7, depicts the fit to comparison data and suggests that moving from one to two, and from two to three factors provides a statistically significant improvement in fit ($p < .001$), while moving from three to four factors provides a statistically insignificant improvement (*p-value* = .943), suggestive of three factors. Figure 8 depicts the CD factor number estimation. Figure 9 provides a useful graphical illustration of the tabulated data allowing users to see the associated steep slopes of improvement between factor solutions.

b. Table 2 provides a review of the average time it took to carry out the procedures on the given sample. The procedures in Table 2 in italics represent more stringent but timely forms of the PA and CD procedures.

Table 2. Time to Complete Factor Number Procedures

Procedure	Setup	Minutes	Seconds
MAP	Standard	0	12
PA-OC-AF	PA: 1000 samples	0	15
PA-OC-AF	*PA: 10,0000 samples*	*0*	*35*
CD	Pop: 10,000; Samples: 500	0	51
CD	*Pop: 100,000; Samples: 5,000*	*10*	*05*

Note: Procedures carried out on sample dataset (N = 484, 14 variables) as illustrated in Figures 1 and 2. Procedures run individually on a 2.66 GHz PC. Estimates based on average of five tests.

c. Attempt to make a decision based on the conversion of these five modern techniques. As four out of the five techniques suggest three factors, three factors would be the most appropriate number of factors to go with in regards to the current dataset.

d. In the event of divergence (for example, estimations of 2, 2, 3, 4 & 4), users are well-informed to check for close calls across all five procedures to make a final decision. In the event of a close call in the PA procedure (i.e., a close distance between the green line and the last retained factor), users may increase the *Number of Samples* to 10,000 to more stringently carry out the procedure. Similarly, for the CD procedure, if the significance-of-improvement (i.e., p-value) of including another factor is borderline, users can increase the *Size of Finite Population* to 100,000 and *The Number of Samples Drawn from Each Population* to 5,000 and re-run the procedure (see Table 2, procedures in italics, for estimated increased computational time). If divergence still exists, users may make a final decision based on the relative significance of each estimate. For example, if PA suggests two factors, and the CD procedure suggests three, but CD's plot of fit values is very flat from two to three factors (i.e., with a merely significant p-value), it would be very reasonable to select two factors to retain. If results are still inconclusive, users are encouraged to rely on more heavily on the techniques with a proven track record across multiple data situations in multiple simulation studies, such as the PA and MAP procedures.

Final Thoughts

The SPSS R-Menu v2.0 provides for a range of modern methods to deal with the number-of-factors-to-retain problem. Attempting to gather convergent information across these methods, alongside thoughtful consideration of theory, enables practitioners to more judiciously model data. Of course, ensuing decisions pertaining to factor estimation and rotation methods, followed by cross-validation and confirmatory approaches, must also be made carefully. Finally, it is hoped that readers of this paper not only make use of the techniques illustrated in this article but also the much improved EFA functionality associated with installing the SPSS R-menu (v2.0), such as a wider range of estimation and rotation methods.

INSTALLING R, AND THE SPSS R-MENU V2.0

The following is an 8-step guide to installing the SPSS R-Menu v2.0 on a MS Windows operating system. All software needed to install the improved menu is available for free to IBM SPSS Statistics software (SPSS) users. Users are advised to install the latest SPSS fix packs for their version of SPSS

available from their administrator prior to installing this software. To download the SPSS R-menu v2.0, users will have to select the appropriate software for their computer operating system (e.g., Windows 32 or 64 bit), and their version of SPSS 19, 20, or 21.

Basto (personal communication, December 12, 2012) provides the following advice for those running Windows Vista or Windows 7: One needs to disable user account control and restart one's computer before installing *R*, the *Essentials for R*, *Python Essentials*, and *Net Plugins*. This can be done as follows:

Vista users: *Control Panel* → *Add or remove user accounts* → *Guest Account* (for example) → *Go to the main User Accounts page* → *Change security settings* → Uncheck *Use User Account Control (UAC) to help protect your computer* → *OK* → *Restart.*

Windows 7 users: *Control Panel* → *System and Security* → *(Action Center) Change User Account Control settings* → Move slider to *Never notify* position → Enter *Password* → Restart one's computer in order for the *User Account Control* to be turned off.

STEP 0: [UPDATED November 2015] It is recommended that users make use of the more recent SPSS 21, 22, or 23 versions.

For SPSS 22, download and install R 2.15.0 here

For SPSS 23, download and install R 3.1.x here

Thereafter, users of SPSS 21, 22, or 23 can also skip the eight steps below by downloading and installing the R-Factor v2.4 *spe* file (extension bundle) found here. The file should not be unzipped and installed as it is: "Utilities > Extension Bundles > Install local extension bundle..." (Users need to ensure that they are connected to the Internet whilst installing the bundle to ensure all packages are automatically downloaded). Restart your version of SPSS to enjoy new functionality. You do not need to have the R program open to use the SPSS R-menu. Happy factor number estimating.

STEP 1: Download and install the appropriate version of R (necessary, even if you have another version already installed):

SPSS 19: R 2.10.1 here

SPSS 20: R 2.12.1 here

SPSS 21: 2.14.2 here

STEP 2: Complete IBM registration here and sign in.

STEP 3: Download and install the appropriate *Essentials for R* from here: Note: Two download options eventually become available. The *Download using http* seems to function better that *Download using Download Director* on most browsers. Find and download appropriate *Essentials for R* file, e.g., *SPSS_Statistics_REssentials_19002_win32.exe.*

STEP 4: For SPSS 19, download and install appropriate *Python Essentials* and *Net Plugins* from here. For SPSS 20, download and install *Python Essentials* and *Net Plugins* from here. For SPSS 21, Python Essentials and Net Plugins are included with the installation media and download (you need to insert the SPSS 21 CD to install these). It is recommended that users install the essentials and plugins to make full use of the R-menu v2.0.

STEP 5: Now you need to download the special *spd.* file (Basto, 2012). Version 2.0 is available from here:

STEP 6: To install the *spd.* file, open SPSS. Go to *Utilities*→ *Custom Dialogues* → *Install Custom Dialogue*… and install the *R-Factor v2.0.spd* file.

STEP 7: Thereafter, you will now need to download several R packages. This is easily done via a proximal mirror. Open R and go to *Packages* → *Set CRAN mirror*… Choose the region closest to you. Thereafter, go to *Packages* → *Install package(s)*… The packages that you will need to install are listed below:

psych, polycor, ICS, nFactors, GPArotation, corpcor, and *R.utils.*

There is no need to save the *work image* upon installing the packages.

STEP 8: After successfully installing the packages, the SPSS R-Menu v2.0 should now be setup. Restart your version of SPSS to enjoy new functionality. You do not need to have the R program open to use the SPSS R-menu. Happy factor number estimating.

REFERENCES

Babakus, E., Ferguson, C. E., & Joreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, 24, 222-228.

Basto, M. (2012). SPSS R-Menu Files. *Sourceforge.net.* Retrieved December 11, 2012, from http://sourceforge.net/projects/spssrmenu/

Basto, M., & Pereira, J. M. (2012). An SPSS R-Menu for Ordinal Factor Analysis. *Journal of Statistical Software*, 46(4), 1-29.

Bentler, P. M. (2005). *EQS 6 Structural Equations Program Manual.* Encino, CA: Multivariate Software.

Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review*, 46, 232-239.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.

Cudeck, R. & MacCallum, R. C. (Eds) (2007). Factor analysis at 100: Historical developments and future directions. Mahwah, NJ: LEA.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 3, 272-299.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466-491.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's Minimum Average Partial Factor Retention Method with Categorical Variables. *Educational and Psychological Measurement*, 71(3), 551-570.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2012). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, in press. Epub ahead of print retrieved December 10, 2012. doi:10.1037/a0030005.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49, 132-169.

Hayton, J. C., Allen, D. G., Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, 7, 191-205.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66, 393-416.

Holgado-Tello, F. P., Chacon-Moscoso, S., Barbero-Garcia, I., & Vila-Abad, E. (2010). Polychoric verses Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, 44, 153-166.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185.

IBM Corporation (2010). *IBM SPSS Statistics 19*. IBM Corporation, Armonk, NY: IBM.

IBM Corporation (2011). *IBM SPSS Statistics 20*. IBM Corporation, Armonk, NY: IBM.

IBM Corporation (2012). *IBM SPSS Statistics 21*. IBM Corporation, Armonk, NY: IBM.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational & Psychological Measurement*, 20, 141-151.

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out Parallel Analysis. *Practical Assessment, Research & Evaluation*, 12(2), 1-11. Retrieved April 10, 2013, from http://pareonline.net/getvn.asp?v=12&n=2

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, 32, 396-402.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443-460.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49, 974-997.

Revelle, W., Rocklin, T. (1979). Very simple structure – alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, 14(4), 403-414.

Revelle, W. (2012). Package 'psych': Procedures for psychological, psychometric, and personality research. Evanston, Illinois. R package version 1.2.8, retrieved December 10, 2012, from http://personality-project.org/r/psych.manual.pdf.

Raiche, G., Roipel, M., & Blais, J. G. (2006). Non graphical solutions for the Cattell's scree test. Paper presented at The International Annual Meeting of the Psychometric Society, Montreal. Retrieved December 10, 2012, from http://www.er.uqam.ca/nobel/r17165/RECHERCHE/COMMUNICATIONS/2006/IMPS/IMPS_PRESENTATION_2006.pdf .

Ruscio, J. (2012). EFA with Comparison Data (R). *Tcnj.edu.* Retrieved December 11, 2012, from http://www.tcnj.edu/~ruscio/EFA%20Comparison%20Data.R

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using

comparison data of a known factorial structure. *Psychological Assessment*, 24(2), 282-292.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, 321-327.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.). *Problems and Solutions in Human Assessment: Honoring Douglas N. Jackson at Seventy* (pp. 41-71. Boston, MA: Kluwer Academic.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99, 432-442.

Acknowledgements

This author would like to thank Professor Mario Basto (Polytech Institute of Cavado and Ave), Professor John Ruscio (The College of New Jersey), and Associate Professor Gavin Brown (The University of Auckland) for their on-going support and comments concerning an earlier draft of this paper.

IBM Copyrighted Material

Citation:

Courtney, Matthew Gordon Ray (2013). Determining the Number of Factors to Retain in EFA: Using the SPSS R-Menu v2.0 to Make More Judicious Estimations. *Practical Assessment, Research & Evaluation*, 18(8). Available online: http://pareonline.net/getvn.asp?v=18&n=8

Corresponding Author:

Matthew Courtney (Office H504)
Department of Learning, Development and Professional Practice
Faculty of Education
The University of Auckland
74 Epsom Avenue
Auckland 1023
New Zealand
matty_courtney [at] hotmail.com