Question Number Two: How Many Factors?

Fara Goodwyn

Texas A&M University

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 3, 2012.

Abstract

Exploratory factor analysis involves five key decisions. The second decision, how many factors to retain, is the focus of the current paper. Extracting too many or too few factors often leads to devastating effects on study results. The advantages and disadvantages of the most effective and/or most utilized strategies to determine the number of factors to extract will be explored. Equipped with this knowledge, researchers can thoughtfully select the best strategies rather than relying solely on customary practice.

Keywords: factor analysis, Bartlett’s chi-square test, eigenvalue greater than 1.0 rule, scree plot, parallel analysis, minimum average partial, bootstrap factor analysis

Question Number Two: How Many Factors?

One goal of factor analysis is to reduce the large number (e.g., hundreds) of variables to "a more parsimonious set of factor scores that can then be used in subsequent analyses" (Thompson, 2004, p. 5). There are five consecutive decisions in creating these smaller latent constructs:

1. Which matrix of association coefficients should be analyzed?
2. How many factors should be extracted?
3. Which method should be used to extract the factors?
4. How should the factors be rotated?
5. How should factor scores be computed if factor scores are of interest?

(Thompson, 2004, p. 27).

The current paper focuses on the second question, arguably the most important of the five (Mumford, Ferron, Hines, Hogarty, & Kromrey, 2003; Stellefson & Hanik, 2008; Zwick & Velicer, 1986). Mumford et al., (2003) explains, "Since the number-of-factors decision is made prior to the factor rotation stage, it subsequently impacts the result of the factor analysis, such as rotated factor patterns, factor score estimates, and the interpretability of the factors" (p. 2).

The fewest number of factors that can be extracted is one. Table 1 presents an intervariable correlation matrix containing correlation coefficients with values of 1.0 or -1.0. (i.e., variables are perfectly correlated) for the variables clean, organized, and messy. The r^2 between every pair of variables is 100% (Thompson, 2004). Researchers extract one factor in this scenario, because one underlying construct, general cleanliness, explains the scores of all three measured variables.

Table 1

Intervariable Correlation Matrix of Perfectly Correlated Variables

	Clean	Organized	Messy
Clean	1.0	1.0	-1.0
Organized	1.0	1.0	-1.0
Messy	-1.0	-1.0	1.0

The largest number of factors that can be extracted is equal to the total number of variables. Table 2 presents an intervariable correlation matrix containing correlation coefficients with values of 0.0 in every off-diagonal entry. This matrix is an identity matrix. There are no real factors (i.e., combined variables), only the perfectly uncorrelated variables clean, musical, and pet owner. The r^2 between every pair of measured variables is 0% (Thompson, 2004). This heuristic example demonstrates that a clean person may or may not be musical and may or may not own pets. There is absolutely no correlation between these three variables. Researchers would extract the total number of variables in this unlikely case.

Table 2

Intervariable Correlation Matrix of Perfectly Uncorrelated Variables

	Clean	Musical	Pet Owner
Clean	1.0	0.0	0.0
Musical	0.0	1.0	0.0
Pet Owner	0.0	0.0	1.0

Researchers using real data sets typically retain a number of factors somewhere between the lower and upper limits described above. Although there are problems associated with extracting too few or too many factors, extracting too few factors leads to greater inaccuracies due to the loss of critical information. Underextraction occurs when a factor combines with other factors or fails to be extracted altogether (Fava & Velicer, 1992; Zwick & Velicer, 1986). While extracting too many factors keeps critical information intact, it causes researchers to disproportionately consider minor factors over more influential ones (Zwick & Velicer, 1986).

Numerous strategies have been developed to determine the optimal number of factors to retain. This paper presents (a) Bartlett's chi-square test, (b) eigenvalue greater than 1.0 rule, (c) scree plot, (d) parallel analysis, (e) minimum average partial, and (f) bootstrap factor analysis. Strategies are reviewed for their effectiveness in determining the optimal number of factors to retain and their accessibility in common statistical software packages (e.g., SPSS, SAS). The first six variables from Holzinger and Swineford's (1939) data set are used to demonstrate three of the six strategies (eigenvalue greater than 1.0 rule, scree plot, parallel analysis).

Bartlett's Chi-square Test

The chi-square test is a test of statistical significance not commonly found in statistical analysis software packages (Zwick & Velicer, 1986). The SPSS syntax necessary to run Bartlett's chi-square test can be found in Appendix A. The null hypothesis is the correlation matrix equals an identity matrix. If the null hypothesis is rejected, factors are extracted sequentially. After the first factor is extracted, the null hypothesis is again tested. This process continues until a remaining residual correlation matrix equals an identity matrix (i.e., no information remains; Thompson, 2004; Zwick & Velicer, 1986).

The same problems associated with sample size in traditional statistical significance testing exist when statistical significance testing is applied to factor retention. Because large sample sizes are used in factor analysis, trivial factors are often deemed statistically significant with this strategy. Real data, with reasonable sample sizes, will never produce correlation matrices equal to an identity matrix (Thompson, 2004). Not surprisingly, Bartlett's chi-square test is increasingly accurate the larger the sample sizes becomes (Zwick & Velicer, 1982).

Eigenvalue Greater Than 1.0 Rule

The eigenvalue greater than 1.0 rule (also known as K1 rule, Kaiser rule, and Kaiser-Guttman rule) retains all factors with eigenvalues greater than 1.0. The logic contends that factors worthy of retaining should, at a minimum, have more variance than any of the original measured variables comprised in the factor. That is, since a single measured variable has a maximum eigenvalue of 1.0, a factor with an eigenvalue greater than 1.0 should have more predictive power than any of the measured variables alone (Zwick & Velicer, 1986). Due to the effects of sampling error, a researcher could choose to retain a factor with an eigenvalue less than 1.0 or reject a factor with an eigenvalue greater than 1.0 (Thompson, 2004).

The eigenvalue greater than 1.0 rule should be used with caution due to the documented potential to overestimate and underestimate results (Zwick & Velicer, 1986). Although the rule is known to be flawed, it is the most common strategy used by researchers. In fact, it is the default strategy in many statistical analysis software packages (e.g., SPSS, SAS; Thompson & Daniel, 1996; Zwick & Velicer, 1986). Zwick and Velicer (1986) highlight the dangers of using this strategy,

> The use of the K1 rule as the default value in some of the standard computer packages (BMDP, SPSS, SAS) is an implicit endorsement of the procedure, particularly to naïve users. This pattern of explicit endorsement by textbook authors and implicit endorsement by computer packages, contrasted with empirical findings that the procedure is very likely to provide a grossly wrong answer, seems to guarantee that a large number of incorrect findings will continue to be reported. (p. 439)

Using SPSS syntax (see Appendix B), two factors are retained using the eigenvalue greater than 1.0 rule. Table 3 presents the results.

Table 3

Eigenvalue Greater Than 1.0 Rule SPSS Output

	Initial Eigenvalues		
Factor	Total	Percent of Variance	Cumulative Percentage
1	**2.182**	36.369	36.369
2	**1.701**	28.355	64.725
3	.744	12.392	77.117
4	.700	11.674	88.791
5	.406	6.774	95.565
6	.266	4.435	100.000

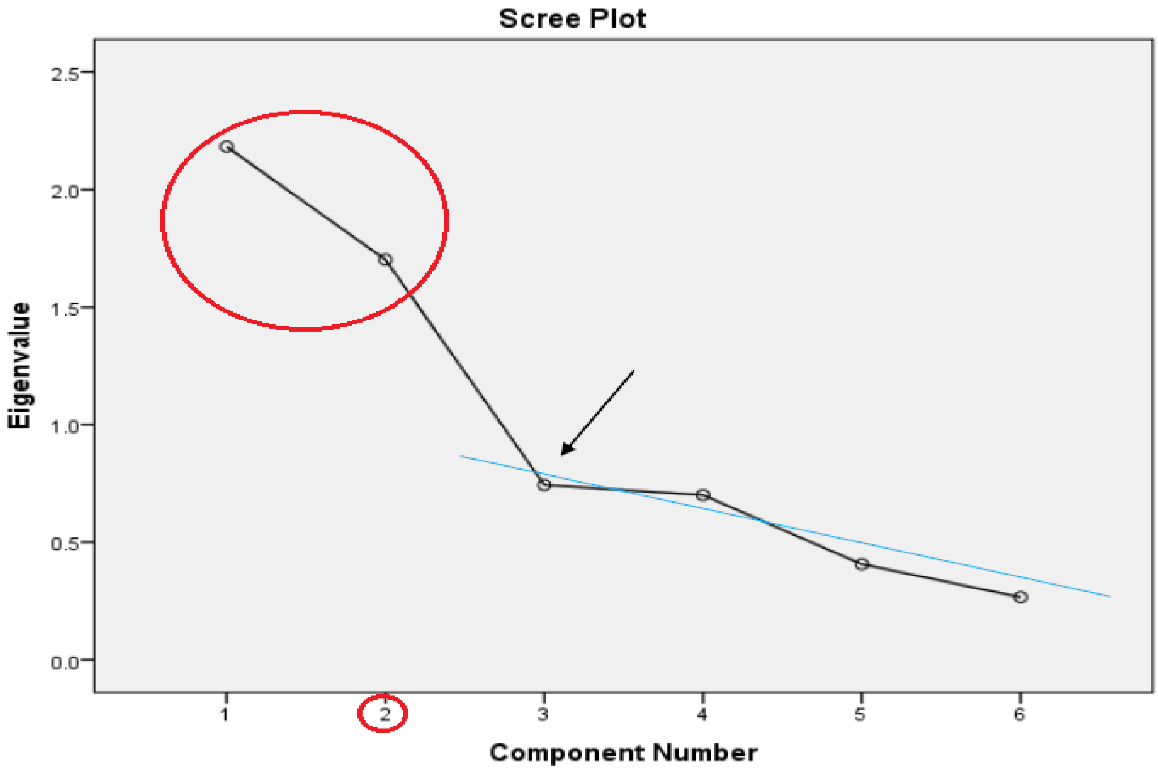
Scree Plot

The scree plot is a graphical test available in SPSS and SAS based on eigenvalues. Scree literally refers to "the line of rubble and boulders which forms at the pitch of sliding stability at the foot of a mountain" (Cattell, 1966, p. 249). Trivial factors are analogous to scree and should be discarded. Nontrivial factors are analogous to mountains and should be retained (Thompson, 2004).

Trivial and nontrivial factors are determined by visually analyzing a line graph (i.e., scree plot) of eigenvalues corresponding to sequentially extracted factors. As seen in Figure 1, the vertical axis of the scree plot represents eigenvalues and the horizontal axis represents factor or component sequence numbers. A line connects all plotted eigenvalues. Because sequentially extracted factors have successively smaller eigenvalues, a downward-sloping (i.e., mountain-like) plot is created. A "pencil test" is invoked by positioning a straight edge or pencil on the eigenvalues forming a near straight line (e.g., blue line in Figure 1; Cattell & Vogelman, 1977). The arrow in Figure 1 points to the "elbow" or position at which the scree begins. Factors with eigenvalues above the straight line and to the left of the elbow are retained. Factors with eigenvalues on or near the straight line (i.e., scree) are discarded. Two factors are retained in this heuristic example.

Because visual analysis is subjective and because not all scree plots have such an obvious, singular elbow, varying numbers of factors may be retained by different researchers examining the same plot (Thompson, 2004; Zwick & Velicer, 1982, Zwick & Velicer, 1986). Therefore, the scree plot should never be the only method utilized when deciding how many factors to retain (Zwick &Velicer, 1986).

Figure 1. Scree Plot SPSS Output



Parallel Analysis

Parallel analysis accounts for sampling error, making it one of the most accurate factor retention strategies. Sampling error is accounted for by comparing eigenvalues from a correlation matrix of original data to eigenvalues from a correlation matrix of randomly ordered variables of identical sample size (Thompson, 2004; Zwick & Velicer, 1986). This comparison is made because randomly ordered scores create a correlation matrix approximating an identity matrix – eigenvalues remain just above and below 1.0 due to sampling error (Horn, 1986; Stellefson & Hanik, 2008). Eigenvalues are exactly 1.0 when derived from a population correlation matrix created from randomly ordered, uncorrelated variables. Factors corresponding with eigenvalues of the original data set that are larger than factors corresponding with eigenvalues of the randomly ordered data set are retained (Horn, 1965; Zwick & Velicer, 1986).

O'Connor (2000) provides SPSS and SAS syntax for parallel analysis. Researchers only need to modify the syntax with their specifications for numbers of cases, variables, and datasets. The remaining syntax remains unaltered. Table 4 presents eigenvalues created from syntax running 100 cases, 6 variables, and 500 data sets. Two factors are retained in this parallel analysis heuristic example.

Parallel analysis steps performed by the syntax include creating a random data set using real data on measured variables with the same rank (i.e., same number of rows by same number of columns). Each column is randomly ordered separately. Eigenvalues are then calculated and aligned. The eigenvalues generated from real data, "real eigenvalues," are aligned parallel to the eigenvalues generated from randomly ordered data, "fake eigenvalues."

Table 4

Parallel Analysis SPSS Output

Number	Actual Data Eigenvalue	Random Order Eigenvalue
1	2.182	1.343271
2	1.701	1.172250
3	.744	1.043149
4	.700	.931626
5	.406	.818577
6	.266	.691127
Sum	6.0 (5.999)	6.0

Minimum Average Partial (MAP)

Although MAP is one of the most accurate strategies, researchers must use syntax to execute MAP because it is not a default strategy in statistical software packages (O’Connor, 2000; Zwick & Velicer, 1986). O’Connor (2000) provides the syntax necessary to run MAP for both SPSS and SAS.

The rationale of MAP and the processes performed by the syntax commands are described as follows. MAP determines the number of factors to retain by examining the correlation matrix. “Statistically, components (or factors) are retained as long as the variance in the correlation matrix represents systematic variance. Components are no longer retained when there is proportionately more unsystematic variance than systematic variance” (O’Connor, 2000, p. 397).

In order to examine variance, a factor is removed from the original matrix of association. The values above and below the diagonal of the reproduced, partial correlation matrix are then squared. The squared values are added together and this sum is divided by the number of squared values in the partial correlation matrix. The next calculation involves removing two factors from the original matrix of association before performing the calculations described above. This process of cumulatively removing factors repeats until all potential factors have been removed and all calculations are performed. Finally, the averaged and squared partial correlations are vertically aligned. The number of factors retained corresponds to the number of factors removed which produced the lowest average squared partial correlation, unless the average squared coefficient from the original matrix is a smaller value. No factors are retained if original matrix results in lowest value (O'Connor, 2000).

Bootstrap Factor Analysis

The bootstrap, when utilized in factor analysis, can also determine how many factors to retain. The syntax for this computer-intensive strategy is available at http://www.coe.tamu.edu/~bthompson/datasets.htm (Zientek & Thompson, 2007). Although a heuristic example is not provided, the basic steps and logic behind bootstrap factor analysis are described below.

After randomly sampling (with replacement) data from a mega-file created from concatenated data (see Thompson, 2006), data are rotated to best-fit positions ensuring all results are in a common factor space. Then eigenvalues and standard deviations are computed for each factor over repeated samples (e.g., 5000). This creates empirically estimated sampling distributions of eigenvalues and standard deviations. Next, sampling distribution data are utilized to compute mean eigenvalues and standard errors (SEs) of the estimates. Finally, confidence intervals (CIs) are created using these values (Zientek & Thompson, 2007).

Unlike the eigenvalue greater than 1.0 rule, bootstrap factor analysis accounts for sampling error by examining mean eigenvalues near 1.0 *and* the width of CIs. Large SEs result in wider CIs indicating less precise estimates. A factor with a mean eigenvalue 1.15, SE 0.20 is less likely to be retained than a factor with the same mean eigenvalue, SE 0.05. Typically, researchers only retain factors with CI lower limits above 1.0 or CI upper limits above 1.0 (Zientek & Thompson, 2007).

Conclusion

Selecting the number of factors to retain is the most important decision a researcher makes and should be done thoughtfully. Researchers can be more confident when the results from several factor retention strategies agree. Heuristic examples for three of the six strategies (eigenvalue greater than 1.0 rule, scree plot, and parallel analysis) were provided. Each strategy resulted in retaining two factors. When researchers use real data, with far more than six variables in the analysis, results may vary.

References

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

Cattell, R. B., & Vogelman, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research, 12,* 289-325.

Fava, J. L., & Velicer, W. F. (1992). The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research, 27,* 387-415.

Holinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution* (No.48). Chicago: University of Chicago Press.

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

Mumford, K. R., Ferron, J. M., Hines, C. V., Hogarty, K. Y., & Kromrey, J. D. (2003). *Factor retention in exploratory factor analysis: A comparison of alternative methods.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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.

Stellefson, M., & Hanic, B. (2008). *Strategies for determining the number of factors to retain in exploratory factor analysis.* Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans.

Thompson, B. (2004). Exploratory and confirmatory factor analysis: understanding concepts and applications. Washington, DC: American Psychological Association.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56,* 197-208.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: The Guilford Press.

Zientek, L. R., & Thompson, B. (2007). Applying the bootstrap to the multivariate case: Bootstrap component/factor analysis. *Behavior Research Methods, 39,* 318-325.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research, 17,* 253-269.

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

Appendix A

SPSS Syntax: Bartlett’s Chi-Square Test

```
FACTOR
  /VARIABLES t1 t2 t3 t4 t5 t6
  /MISSING LISTWISE
  /ANALYSIS t1 t2 t3 t4 t5 t6
  /PRINT initial extraction correlation kmo
  /EXTRACTION PC
  /ROTATION NOROTATE
  /METHOD=CORRELATION.
```

Appendix B

SPSS Syntax: Eigenvalue Greater Than 1.0

```
FACTOR
  /VARIABLES t1 t2 t3 t4 t5 t6
  /MISSING LISTWISE
  /ANALYSIS t1 t2 t3 t4 t5 t6
  /PRINT initial extraction correlation
  /CRITERIA mineigen(1) ITERATE(25)
  /EXTRACTION PC
  /ROTATION NOROTATE
  /METHOD=CORRELATION.
```