Psychological Assessment
2012, Vol. 24, No. 2, 282–292

1040-3590/11/$12.00 DOI: 10.1037/a0025697

Determining the Number of Factors to Retain in an Exploratory Factor Analysis Using Comparison Data of Known Factorial Structure

John Ruscio and Brendan Roche
The College of New Jersey

Exploratory factor analysis (EFA) is used routinely in the development and validation of assessment instruments. One of the most significant challenges when one is performing EFA is determining how many factors to retain. Parallel analysis (PA) is an effective stopping rule that compares the eigenvalues of randomly generated data with those for the actual data. PA takes into account sampling error, and at present it is widely considered the best available method. We introduce a variant of PA that goes even further by reproducing the observed correlation matrix rather than generating random data. Comparison data (CD) with known factorial structure are first generated using 1 factor, and then the number of factors is increased until the reproduction of the observed eigenvalues fails to improve significantly. We evaluated the performance of PA, CD with known factorial structure, and 7 other techniques in a simulation study spanning a wide range of challenging data conditions. In terms of accuracy and robustness across data conditions, the CD technique outperformed all other methods, including a nontrivial superiority to PA. We provide program code to implement the CD technique, which requires no more specialized knowledge or skills than performing PA.

Keywords: exploratory factor analysis, number of factors, parallel analysis, comparison data, Kaiser criterion

Exploratory factor analysis (EFA) is performed routinely to study the latent factors that underlie scores on a larger number of measured variables or items. This data-analytic tool is especially popular in the development and validation of assessment instruments. Henson and Roberts (2006) found that EFA is used especially frequently in the research areas of measurement and assessment, education, and personality, and they focused their review of published EFAs on the journals *Psychological Assessment*, *Educational and Psychological Measurement*, *Journal of Educational Psychology*, and *Personality and Individual Differences*. Whereas confirmatory factor analysis is used to test the fit of specified structural models, EFA often is used when there is little or no a priori justification for specifying a particular structural model. For example, rather than specifying the number of latent factors underlying a set of item responses, one might proceed in a more exploratory manner by determining the number of factors empirically. Along with many other investigators, Henson and Roberts noted that this task remains one of the most significant challenges to implementing EFA successfully.

In many ways, the content, meaning, and psychometric properties of scales and subscales depend on the care with which one determines the number of factors underlying a set of item responses, with theoretical and practical costs falling on those who make suboptimal choices. For example, a researcher attempting to create internally consistent scales might make poor decisions regarding which items to retain, remove, or revise if the number of factors is not determined correctly. The internal consistency of one or more subscales might appear weak if items that would hang together nicely are spread across an artificially large number of subscales. Alternatively, potentially useful scales, or even theoretically interesting new constructs, could be missed altogether if too few factors are included in the data analysis to allow a subset of items to reveal an otherwise hidden pattern in the data. Researchers developing or validating assessment instruments would benefit in a number of ways from the application of reliable methods for determining how many factors to retain in EFA.

Many techniques are available to address this challenging problem. Some of the better known methods are the scree test (Cattell, 1966), Kaiser's (1960) eigenvalue-greater-than-1 rule, and parallel analysis (Horn, 1965). Each of these, along with some more recently developed tests, begins with the calculation of eigenvalues from an item correlation matrix.[1] We explain the rationales under-

This article was published Online First October 3, 2011.

John Ruscio and Brendan Roche, Department of Psychology, The College of New Jersey.

This research builds upon work that the second author performed as his senior honors thesis under the supervision of the first author. We would like to express our thanks to the other thesis committee members, Jason Dahling and Jean Kirnan, for helpful comments and suggestions as this project developed.

Correspondence concerning this article should be addressed to John Ruscio, Department of Psychology, The College of New Jersey, P.O. Box 7718, Ewing, NJ 08628. E-mail: ruscio@tcnj.edu

[1] Exploratory factor analyses can be performed using item covariance matrices. Referring to the special case of item correlation matrices, the covariation matrix based on standardized variables, simplifies some of the details we present. For example, the familiar Kaiser criterion identifies the number of factors as the number of eigenvalues greater than 1 when performed using item correlation matrices; with covariance matrices for unstandardized items, different thresholds are required.

lying procedures for examining patterns of eigenvalues, each of which specifies a stopping rule to determine the number of factors (see Velicer, Eaton, & Fava, 2000, for an excellent overview). Other methods compare the fit of structural models with varying numbers of factors. If one conceptualizes model comparison sufficiently broadly, this includes Velicer's (1976) multiple average partial procedure as well as maximum-likelihood approaches that yield fit indices such as the Akaike information criterion (AIC; Akaike, 1974), Bayseian information criterion (BIC; Schwarz, 1978), and chi-square tests. In addition to reviewing and evaluating methods that have fared well in prior research (e.g., Fabrigar, Wegener, MacCallum, & Strahan, 1999; Velicer et al., 2000; Zwick & Velicer, 1986), we present and test a new method that bridges the eigenvalue-pattern and model-comparison families of techniques.

Making the decision about how many factors to retain when one subsequently performs an EFA does not necessarily mean that one has to use the same methods as when the EFA itself is performed. For some of the techniques that we describe and study, determining the number of factors to retain constitutes a decision one can make as a preliminary step without having performed EFA. Thus, we review and test techniques traditionally used prior to a principal components analysis as well as those traditionally used prior to or as part of an EFA, including some techniques that require maximum-likelihood estimation methods and a new procedure that relies on the generation and analysis of comparison data with known factorial structure. In short, we did not set out to study EFA itself, but a decision that may or may not be made prior to performing an EFA.

Examining the Pattern of Eigenvalues for an Item Correlation Matrix

A good place to start in understanding conventional methods used to determine the number of factors to retain is with an examination of the k eigenvalues for an item correlation matrix, where k is the number of items. Each eigenvalue represents the share of the total item variance that can be captured using one linear combination of the items, and the sum of the eigenvalues equals k. For example, if the first (largest) of the eigenvalues for 10 items is 5.00, this means that a single linear combination of items can capture 50% of the total item variance. To the extent that a small number of eigenvalues are relatively large and most are relatively small, this pattern suggests that a small number of linear combinations can be used to capture much of the total item variance. These linear combinations correspond to the latent factors underlying the item responses. Thus, many methods for determining the number of factors to retain in EFA are based on an examination of the pattern of eigenvalues.

Cattell's (1966) scree test is a graphical method in which the k eigenvalues are plotted in descending order, and a graph constructed in this way is called a *scree plot*. The scree test is performed by searching for an "elbow" in the plot, or an abrupt transition from large to small eigenvalues. However, there is not always a visual elbow on the scree plot, in which case the test requires a difficult subjective judgment of where a line should be drawn to determine the number of factors to retain. Gorsuch (1983) noted that scree plots can be ambiguous due to the presence of more than one discontinuity in the graph or the lack of a visible discontinuity. Figure 1 shows two illustrative scree plots for data sets with $N = 200$ cases and $k = 10$ items, each created using a structural model with $F = 2$ orthogonal (uncorrelated) latent factors. (For the moment, ignore the solid line and dotted curve on each plot.) In the first data set, the first two factors captured more than 40% of the total item variance, and the fact that the first two eigenvalues are considerably larger than the remaining eight eigenvalues is easy to see; this illustrates an unambiguous scree plot. In the second data set, the first two factors captured just under 30% of the total item variance, and the discontinuity between the first two and remaining eight eigenvalues is more difficult to discern; this illustrates an ambiguous scree plot, one in which the signal of genuine latent factors is hard to detect against the noise of normal sampling error. Because scree plots are often ambiguous, objective criteria or "stopping rules" have been developed to determine the number of factors to retain.

The Kaiser criterion is the most commonly used stopping rule (Henson & Roberts, 2006), and it is the default on programs such as SPSS. The rule sets the threshold between large and small values at an eigenvalue of 1, the arithmetic mean of the eigenvalues (recall that these sum to k for k items). Each eigenvalue greater than 1 is interpreted as representing a factor, and each value below 1 is not. The basis for the rule is that a factor should not account

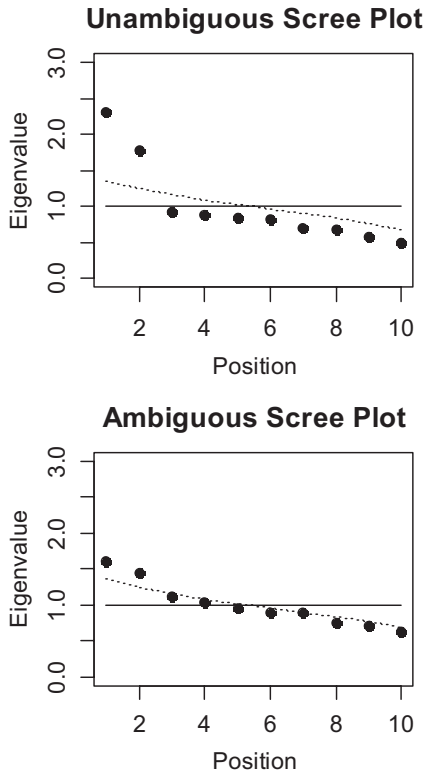

Figure 1. Scree plots for two illustrative data sets with $N = 200$ cases and $k = 10$ items, each created using a structural model with $F = 2$ orthogonal (uncorrelated) factors. In the first data set, the first two factors captured more than 40% of the total item variance. In the second data set, the first two factors captured just slightly less than 30% of the total item variance. In each plot, the solid line at $y = 1$ represents the Kaiser criterion and the dotted curve shows the parallel analysis reference eigenvalues.

for less total item variance than a single item (Kaiser, 1960). However, this criterion fails to take into account normal sampling error. As Turner (1998) noted, the difference between eigenvalues of 1.01 and 0.99 is negligible and may be due to nothing more than sampling error, yet the former value will be considered indicative of a factor while the latter value will not. Moreover, sampling error will yield some eigenvalues greater than 1 even when there are no latent factors. Unsurprisingly, this criterion has been found to identify too many factors (Fabrigar et al., 1999). In Figure 1, the solid lines at $y = 1$ graphically depict the Kaiser criterion. For the unambiguous plot, this stopping rule correctly identifies two factors; for the ambiguous plot, this rule incorrectly identifies five factors. This is consistent with the EFA literature demonstrating that the Kaiser criterion tends to overextract, or identify too many factors.

The statistical technique of parallel analysis (PA; Horn, 1965) simulates normal sampling error to generate a series of reference eigenvalues. To perform PA, one calculates eigenvalues for randomly generated data sets of the same dimensions (N cases with k items). These data are drawn at random from a population with no factors, for which the item correlation matrix is an identity matrix. After calculating eigenvalues for each data set, these are averaged at each position. The number of factors is estimated as the largest value j such that the first j eigenvalues for the actual data exceed those for the random data. The rationale for PA is that the eigenvalues for the random data differ from 1 due to sampling error alone, so eigenvalues that exceed these represent meaningful factors rather than trivial values. Whereas the Kaiser criterion constitutes a flat threshold at an eigenvalue of 1, PA yields a decreasing series of threshold eigenvalues. In Figure 1, the dotted curves graphically depict the criteria derived using PA. For both plots, this stopping rule correctly identified two factors. PA is supported by stronger empirical evidence than other methods for determining the number of factors to retain (Fabrigar et al., 1999; Velicer et al., 2000), and interest in PA is due in large part to a simulation study by Zwick and Velicer (1986). Though PA is considered the method of choice among methodologists, it is not included in common statistical software packages such as SPSS, where the default stopping rule is still Kaiser's criterion.[2]

Raiche, Riopel, and Blais (2006) presented two more objective solutions to the problem of interpreting a scree plot. The optimal coordinates (OC) and acceleration factor (AF) techniques can be used to identify an elbow in a scree plot. To perform the OC technique, one calculates a series of linear equations to determine whether observed eigenvalues exceed the predicted values. Specifically, the equation for the jth position connects the points at positions $j + 1$ and k, where k is the total number of items. When the eigenvalue observed at position j exceeds the value predicted by an extrapolation with this linear equation, it represents an increase that may be an elbow in the scree plot. One searches for the largest value of j such that the observed eigenvalue still exceeds the value predicted by the equation calculated for that position; this is interpreted as the elbow and constitutes the number of factors predicted by the OC technique. Raiche et al. (2006) allow the maximum number of factors to be capped at the value identified by the Kaiser criterion or by PA. Because we found the latter to be a more effective limit in preliminary testing, we use that exclusively. For both illustrative data sets, the OC technique correctly identified two factors.

Using the AF method, one locates the elbow by determining where the slope of the graph changes most sharply. At each position j, the acceleration factor a is calculated as the change in slope:

$$a = (eig_{j+1} - eig_j) - (eig_j - eig_{j-1})$$

This equation only includes eigenvalues because consecutive eigenvalues on a scree plot are evenly spaced one unit apart. Once the position j with the largest value of a is located, the number of factors is identified as $j - 1$. As with the OC technique, the value of j is capped at the number of factors identified by PA. For both illustrative data sets, the AF technique correctly identified two factors.

Comparing the Fit of Structural Models With Differing Numbers of Factors

Rather than examining the pattern of eigenvalues, several additional methods for determining the number of factors to retain involve procedures that implicitly or explicitly involve tests of the fit of structural models with varying number of factors. Velicer (1976) introduced a criterion based on the average partial correlations in the correlation matrix after varying numbers of factors have been partialled out. To begin this procedure, one factor is partialled out of the correlation matrix, and the average partial correlation in the new matrix is calculated. A second factor is then partialled out, and the averaged partial correlation is calculated. This proceeds as long as the average partial correlation continues to decrease. The predicted number of factors is equal to the number of factors that were partialled out when the minimum average partial was reached. This technique is known as the MAP—for "minimum average partial"—procedure. Though no statistical tests are performed to quantify model fit, implicitly the MAP procedure involves a sequential search for a parsimonious structural model. Velicer, Eaton, and Fava (2000) found that the MAP procedure performed well, substantially better than the Kaiser criterion. The MAP procedure correctly identified two factors for the first illustrative data set but incorrectly identified one factor for the second data set.

An approach that more explicitly tests model fit is to use maximum-likelihood procedures to estimate the parameters of a series of structural models and then calculate information criteria to assess the goodness of fit of each model. Popular indices include the AIC (Akaike, 1974) and the BIC (Schwarz, 1978). In the context of EFA, one can fit models that include one latent factor, two latent factors, three latent factors, and so forth, estimating the loading of each observed variable onto each latent factor with no additional constraints on the models. For each model, calculating fit begins with the maximized value of likelihood function L for the estimated model, and this is transformed by multiplying the natural logarithm of L times -2 to obtain $-2\ln(L)$; smaller values represent better fit. The AIC and BIC build upon this foundation in ways that penalize models that are less parsimonious. Allowing more factors increases the number of free parameters to be esti-

[2] For principal components analyses, syntax is available to perform PA in SPSS and other popular programs (O'Connor, 2000). This is available online at https://people.ok.ubc.ca/brioconn/nfactors/nfactors.html

mated and, hence, the ability to reproduce the observed correlation matrix. The AIC adds a penalty of $2p$, where p is the number of free parameters in the model, and the BIC adds a penalty of $\ln(N) \times p$, where N is the total sample size:

$$AIC = -2\ln(L) + 2p$$

$$BIC = -2\ln(L) + ln(N) \times p$$

Thus, the BIC includes a stronger penalty for allowing additional free parameters than the AIC, as $\ln(N)$ will be much larger than two in most model-fitting contexts. When they do not converge on the same structural solution, the AIC tends to retain a less parsimonious model than the BIC (see Burnham & Anderson, 2004, for an excellent overview of these criteria). The AIC correctly identified two factors for both of the illustrative data sets. The BIC correctly identified two factors for the first data set but incorrectly identified one factor for the second data set.

Another criterion based on maximum-likelihood factor analysis is to calculate chi-square values for models with increasing numbers of factors, stopping when the chi-square value is no longer statistically significant ($p > .05$). One begins with the simplest model (one factor) and rejects it only if the chi-square test suggests poor fit. In that case, a model with two factors is estimated, and a new chi-square test performed. This continues until a model is not rejected, at which point the number of factors is identified. The chi-square test correctly identified two factors for the first illustrative data set but incorrectly identified one factor for the second data set.

Generating and Analyzing Comparison Data With Known Factorial Structure

We attempted to improve upon the performance of PA, which appears to be the best supported approach available, by introducing a technique by which one creates and analyzes comparison data (CD) with known factorial structure to determine the number of factors to retain. Rather than generating random data sets, which only take into account sampling error, multiple data sets with known factorial structures are analyzed to determine which best reproduces the profile of eigenvalues for the actual data. Varying the number of factors in populations of comparison data and drawing random samples from each population incorporates sampling error as well as factor structure in a model-comparison approach. Like the model-testing methods described previously, this is an exploratory technique in which full structural models need not be specified. Rather, one only needs to decide how many factors to allow in each of a series of models to determine which achieves the best fit. Likewise, no special expertise is required to use this method because it is fully automated.

The CD method was designed to overcome two weaknesses of PA. The first weakness stems from the fact that PA compares each observed eigenvalue to a reference eigenvalue by using the latter as a threshold. Each reference value is conventionally calculated as the mean eigenvalue for that position over a series of replication samples. Sampling error can cause an observed value to exceed a reference value, and Glorfeld (1995) suggested that this might explain the tendency for PA to overextract. To address this bias, Glorfeld proposed using the eigenvalue at a high percentile of the distribution—rather than the mean—to obtain the reference value for each position. For example, to reduce a tendency of PA to overextract, Glorfeld suggested using the eigenvalue at the 95th percentile as the reference value. Though this would reduce errors of overextraction, it would come at the cost of errors of underextraction. Under the conditions studied by Weng and Cheng (2005), they found that using the eigenvalue at either the 95th or 99th percentile minimized total errors. How one calculates reference values clearly has implications for the magnitude and direction of bias with regard to overextraction or underextraction, but it does not solve the fundamental problem that constrains the accuracy of PA. Whether an observed eigenvalue happens to fall slightly above or below the reference value matters a great deal, and this renders PA vulnerable to the vicissitudes of sampling error, especially when observed and reference values are close to one another. It is precisely when the pattern of eigenvalues is most ambiguous that investigators rely most heavily on objective techniques such as PA.

A second weakness of PA is that when one or more factors are present, the reference eigenvalues obtained via PA provide a biased frame of comparison. Turner (1998), drawing from the earlier work of Harshman and Reddon (1983), explained that the source of this bias is the fact that the k eigenvalues must sum to k. After a factor has been identified by an eigenvalue that exceeds its reference value, the reference values for subsequent positions are not adjusted downward to take into account the large eigenvalue that has already been observed. For example, consider a sample with $k = 10$ items drawn from a population in which two factors exist, one of which explains 50% of the total item variance and the other explains an additional 10%.[3] The expected eigenvalues are 5.00, 1.00, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50, and 0.50. The expected reference values obtained using PA are 1.00 at each position. All eigenvalues are subject to normal sampling error. It is unlikely that the first factor would escape detection, but the second factor could easily be missed: The expected eigenvalue of 1.00 equals the expected reference value of 1.00, leaving matters entirely up to sampling error. Even once a first factor has been identified, the fact that it accounts for 50% of the total item variance is ignored rather than used to adjust all subsequent reference eigenvalues downward. This introduces a bias toward underextraction, which Turner (1998) speculated might tend to offset the bias described earlier, causing PA to overextract. The only circumstance in which all reference values obtained through PA would represent appropriate thresholds for comparison is when no factors exist, and the only source of variance among eigenvalues is normal sampling error. However, it would be unusual in the extreme for an investigator to perform EFA in the absence of even a single factor.

Turner (1998) proposed a sequential modification of PA in which the reference eigenvalues are recalculated after each factor

[3] More generally, suppose F factors exist and that they collectively account for a proportion S of the total item variance. The first F eigenvalues are expected to sum to $S \times k$, and each subsequent eigenvalue is expected to be $(1 - S) \times k/(k - F)$. Comparing expected values beyond the first F values reveals that for any $S > 0$, $(1 - S) \times k/(k - F) > 1.00$, the expected reference values for PA.

identified is removed. Each time an observed eigenvalue exceeds its corresponding reference value, this identified factor is included in a model of the data and PA is performed again to obtain a new reference value for the next position. Rather than using PA as a single-iteration frame of reference, Turner's modification essentially builds a model with a sufficient number of factors to account for the pattern of observed eigenvalues. Our own model-comparison technique builds on this foundation not only by considering models of successively increasing complexity, but also by calculating a measure of fit rather than applying a threshold to overcome the first weakness of PA described previously. The CD method builds a model of the data by incrementing the number of factors until the full pattern of observed eigenvalues is reproduced well. Whereas small differences between observed and reference eigenvalues can make a big difference in the results of PA, this should not be the case for the CD technique because fit is calculated for the full series of eigenvalues rather than compared at a single position via a threshold. To perform the CD method, all that one would provide to the algorithm are the empirical data for which one wishes to determine how many factors to retain (referred to as the *target data set*) and the maximum number of factors to consider (e.g., one might choose to test the fit of models with from one to five factors).

Figure 2 provides a schematic diagram of the CD method for an analysis that considers the possibility of up to four factors underlying the item responses in the target data set. The procedure begins by asking whether samples drawn from a two-factor population of comparison data better reproduce the eigenvalues observed for the target data than do samples drawn from a one-factor population of comparison data. If not, this provides a one-factor estimate. If so, the procedure continues by asking whether samples drawn from a three-factor population of comparison data better reproduce the eigenvalues observed for the target data than do samples drawn from a two-factor population of comparison data. If not, this provides a two-factor estimate. If so, the procedure continues once more. This iterative technique proceeds through questions shown in boxes toward the right until either a new population fails to improve fit, in which case the previous population provides the estimated number of factors, or the maximum number of factors to consider is reached with improved fit, in which case this maximum provides the estimated number of factors. At each stage, steps required to generate and analyze the necessary comparison data of known factorial structure are depicted in the boxes toward the left.

The CD method combines features of those methods that examine the pattern of eigenvalues for an item correlation matrix and those methods that compare the fit of structural models with varying numbers of factors. What we propose is a model-comparison technique that relies on tests of the fit between observed and reproduced eigenvalues. In addition to blending these approaches, the CD method may reduce the chances of obtaining certain kinds of artifactual factors. For example, when binary items vary in their endorsement rates, spurious "difficulty factors" can emerge (McDonald & Ahlawat, 1974). More generally, if items in one subset are positively skewed and items in another subset are negatively skewed (e.g., if all item distributions are skewed but a subset of the items are reverse scored), two factors can emerge even if the items actually represent a single construct. When generating each population of comparison data for analysis in the CD method, not only is the item correlation matrix reproduced via a known

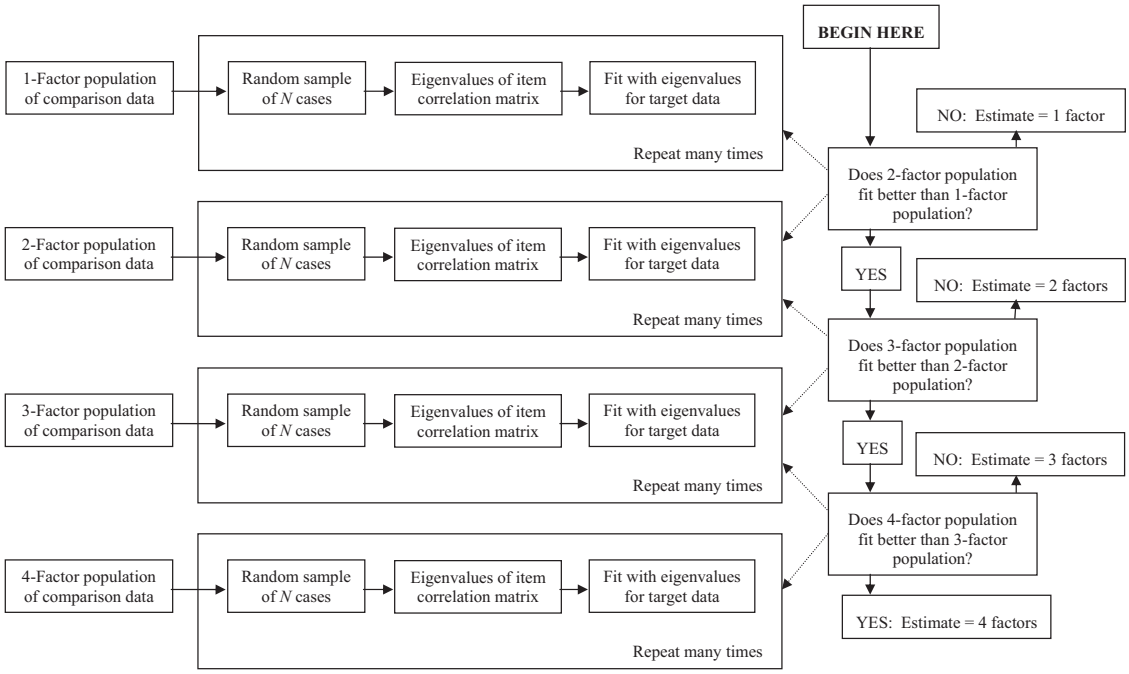

Figure 2. Schematic diagram for comparison data method to determine the number of factors to retain. Structural models with from one to four factors are considered.

factorial structure, but also the multivariate distribution of item responses is reproduced. This holds the distributional characteristics of the data constant across populations, which should reduce the chances of identifying spurious factors.

Novel as this CD method is in the EFA literature, something analogous has proven highly effective in taxometric analysis, where the goal is to distinguish categorical from continuous latent variables (e.g., Ruscio & Kaczetow, 2009; Ruscio & Marcus, 2007; Ruscio, Ruscio, & Meron, 2007; Ruscio, Walters, Marcus, & Kaczetow, 2010). This technique has also been applied to a factor-analytic procedure that examines factor score density plots (Ruscio & Walters, 2009). Building on the use of CD in taxometric analyses, Ruscio and Kaczetow (2009) discussed the potential utility of a similar model-comparison approach using CD to determine the number of factors in EFA. For both of our illustrative data sets, the CD method outlined here correctly identified two factors.

The Present Study

We examine the performance of all nine methods described in the previous sections. This includes four methods based on examining the pattern of eigenvalues for item correlation matrices—Kaiser criterion, PA, and two methods for objectively interpreting scree plots that have not been studied systematically (OC and AF)—as well as four methods based on testing model fit—the MAP procedure, maximum-likelihood information criteria (AIC and BIC), and chi-square tests—and the newly introduced CD method. Each technique was evaluated across a wide range of data conditions designed to be realistic for EFA research, yet sufficiently challenging to avoid ceiling effects in any technique's performance. The focal comparison was between the new CD method and PA, the current state of the art. The other methods were included to place the PA and CD results in the broadest context possible, as well as to study some techniques that have not received as much empirical attention.

Method

Because we wanted this simulation study to span a wide range of data conditions and because performing the full array of methods was computationally intensive, a fully crossed factorial design was not a feasible option. Instead, a random-sampling design was used such that for each target data set, data conditions were determined by drawing a series of parameters at random from specified ranges of values. For example, the first target data set contained $N = 412$ cases and $k = 31$ items that loaded onto a single factor; the second target data set contained $N = 348$ cases and $k = 47$ items that loaded onto two correlated factors. Additional details regarding the data conditions are provided in the following sections. Sampling data conditions at random from specified ranges, rather than insisting on a fully crossed factorial design, provided coverage of the full parameter space in a computationally feasible design.

The study design included a wide range of conditions designed to span those that might be encountered in actual research data but that would provide a sufficient challenge to avoid ceiling effects in accuracy in the present study. Sample size (N) ranged from 200 to 1,000, the actual number of factors (F) ranged from one to five, and the number of variables (k) ranged from 15 to 60. For each sample, factors were either correlated or uncorrelated (orthogonal) and the number of ordered categories (C) for each variable ranged across values of 2, 3, 4, 5, 6, 7, 10, and 20. These response scales range from dichotomous items ($C = 2$) through common Likert scales ($C = 3$–7) to levels that approximate continuous scales fairly well ($C = 10$ or 20). These data conditions are summarized in Table 1.

In total, 10,000 target data sets were generated and analyzed. Each consisted of a data matrix of N cases by k items, not simply an item correlation matrix. A random series of data parameters was drawn, and the data-generation program of Ruscio and Kaczetow (2008) was used to create the target data set; in this program, a factor model, not a component procedure, is used to generate data. The factor loadings used to create a correlation matrix for each target data set were sampled at random from ranges of values shown in Table 1. These loadings were developed through a trial-and-error process in which only PA was performed, and its accuracy never approached 0% or 100% too closely, thereby leaving room for other techniques to perform even better or worse. Once the loadings were sampled for an item, they were randomly assigned to factor numbers (see Table 2). For example, if the three loadings for an item in orthogonal data set were .30, .00, and .00, the loading of .30 had an equal chance of being applied to Factor 1, 2, or 3, with the .00 loadings applied to the other two factors. New loadings and a new random assignment of loadings to factors were generated for each item in turn. Thus, the number of items loading most highly onto each factor was approximately equal for each data set, varying due to normal sampling variation.

Nine techniques for determining the number of factors to retain were analyzed: the Kaiser criterion, parallel analysis (PA), optimal coordinates (OC), acceleration factor (AF), the minimum average partial (MAP) procedure, Akaike information criterion (AIC), Bayesian information criterion (BIC), chi-square tests, and comparison data (CD). We performed sufficient data analysis to implement each technique, but no more. For example, if a technique required only the calculation of eigenvalues and the application of a threshold (e.g., the Kaiser criterion) to determine the number of factors to retain, that was all that was done. If a technique required a series of maximum-likelihood factor analyses to determine the number of factors (e.g., for the AIC, BIC, and chi-square methods), this was done. To save computing time, we did not perform

Table 1
Summary of Data Conditions

Factor	Range of values
Sample size (N)	200–1,000
Number of factors (F)	1–5
Number of variables (k)	15–60
Item response scale (C)	2, 3, 4, 5, 6, 7, 10, or 20 ordered categories
Type of factor structure	Correlated or orthogonal

Note. Sample size, number of factors, number of variables, item response scale, and type of factor structure were sampled at random from uniform distributions spanning the ranges listed above, as were each item's loadings. Once the loadings were sampled for an item, they were randomly assigned to factor numbers.

Table 2
Summary of Factor Loadings

	Number of factors				
Factor type	1	2	3	4	5
Correlated factors					
Loading 1	.30–.70	.30–.70	.30–.70	.30–.70	.30–.70
Loading 2	—	.20–.50	.20–.50	.20–.50	.20–.50
Loading 3	—	—	.00–.30	.00 to .30	.00–.30
Loading 4	—	—	—	.00 to .20	.00–.20
Loading 5	—	—	—	—	.00–.10
Orthogonal factors					
Loading 1	.10–.35	.15–.40	.20–.45	.25–.50	.30–.55
Loading 2	—	.00	.00	.00	.00
Loading 3	—	—	.00	.00	.00
Loading 4	—	—	—	.00	.00
Loading 5	—	—	—	—	.00

Note. Once the loadings were sampled for an item, they were randomly assigned to factor numbers. New loadings and a new random assignment of loadings to factors were generated for each item in turn. Thus, the number of items loading most highly onto each factor was approximately equal for each data set, varying due to normal sampling variation.

maximum-likelihood factor analyses for models with more than $F + 3$ factors. For example, if $F = 5$, models with more than eight factors were not considered.[4]

To implement the CD technique that we introduced earlier, we perform the following steps:

1. Generate a finite population of comparison data ($N = 10{,}000$) using the GenData program (Ruscio & Kaczetow, 2008). This reproduces the item correlation matrix for the target data set using an iterative procedure and reproduces the multivariate distribution of item responses by reproducing each item's marginal distribution using standard bootstrap methods. The first time this step is performed, a structural model with one factor is used; in subsequent iterations, larger numbers of factors are used. Factor loadings are determined by the GenData program to reproduce the item correlation matrix for the target data set as well as possible; correlated factors are allowed if that is helpful.

2. Draw a random sample of N cases from the population generated in Step 1.

3. Calculate the eigenvalues of the item correlation matrix for the sample drawn in Step 2.

4. Calculate the root-mean-square residual (RMSR) eigenvalue by comparing the eigenvalues obtained in Step 3 with those for the target data set.

5. Repeat Steps 2 through 4 a total of 500 times, saving the distribution of 500 RMSR fit values.

6. Return to Step 1, increasing the number of factors in the structural model by one. Repeat Steps 2 through 5 for this new model, then proceed to Step 7 to compare the new distribution of RMSR fit values with those for the previous model.

7. Compare the 500 RMSR fit values for the new factor model with the 500 RMSR fit values from the previous model. The first time this step is reached, the fit values for a two-factor model are compared with those for a one-factor model. This comparison is made using a nonparametric test (Mann–Whitney U) with a liberal alpha level of .30.

A. If the new model (e.g., with two factors) does not provide statistically significantly lower RMSR fit values than the previous model (e.g., with one factor), the previous model is retained, and the CD technique halts (e.g., the number of factors to retain is determined to be one).

B. If the new model provides statistically significantly lower RMSR fit values than the previous model, return to Step 6.

In short, new populations of comparison data are generated using an increasing number of factors. The process continues as long as the addition of a new factor significantly improves the reproduction of the eigenvalues observed for the target data set. Preliminary testing on independent samples suggested that 500 replication samples and a liberal alpha level of .30 worked well in Steps 5 and 7, respectively, although varying either of these values did not affect the accuracy of the CD technique much.[5]

The entire study, including all data generation and factor analysis, was performed with programs written in the R language for statistical computing (R Development Core Team, 2011). R is available as a free download, and the GenData program developed by Ruscio and Kaczetow (2008)—with notes on how to adapt it for other uses—is available online at http://www.tcnj.edu/~ruscio/taxometrics.html. Programs specific to our study are available on

[4] It is possible that the AIC or BIC criteria or chi-square tests might have identified even more than $F + 3$ factors, but studying the amount of overextraction in such cases was not deemed worth the additional computing time. Any bias introduced in favor of these methods was not sufficient to change the conclusions drawn from the results of this study.

[5] In preliminary tests, the CD method's performance varied modestly across alpha levels ranging from .05 to .50, peaking at $\alpha = .30$. Whereas .05 is standard in statistical testing, in this context, the low Type I error rate (overextraction) was accompanied by a very high Type II error rate (underextraction). Using $\alpha = .30$ allowed more Type I errors but with sufficiently fewer Type II errors that the total number of errors was minimized. Because we see no reason to privilege either type of error in this context, we followed the results of the preliminary testing and used an unconventionally liberal alpha level. Users uncomfortable with this decision can provide their own alpha level when running our program to alter the relative risks of identifying too many or too few factors.

request for readers interested in replicating or extending the results reported. We performed all factor analyses using unadjusted correlation matrices.

Results

Table 3 shows the percentage of samples for which the number of factors was overestimated, identified correctly, or underestimated for each of the techniques included in this study. CD had the highest accuracy rate (87.1%), surpassing the MAP procedure (59.6%) and even PA (76.4%). In addition to scoring the most correct identifications, the CD technique rarely over- or underextracted by more than one factor when it erred. As expected, the Kaiser criterion had a very low overall accuracy rate (8.8%).

The information criteria—AIC and BIC—had accuracies of 72.1% and 59.6%, respectively. The BIC had a strong tendency to underestimate the number of factors, with just one instance of overestimation out of all 10,000 samples. The AIC rivaled the accuracy level of PA. Chi-square tests achieved an accuracy of 58.9%. The objective tests based on scree plots—OC and AF—had accuracies of 73.8% (rivaling that of PA) and 43.8%, respectively. Like the BIC, AF tended to underestimate the number of factors; often it estimated only one factor. As a consequence, it was 100% accurate for data sets with only one factor, but much less so for data sets with more than one factor.

In addition to scoring the percentage correct, we examined the bias and precision of each technique. Bias was calculated as the average deviation score (predicted – actual number of factors) to quantify the extent to which a method over- or underestimated the number of factors. Precision was calculated as the average absolute deviation score to quantify the typical magnitude of error for a method. A summary of these results is presented in Table 3. PA was unbiased (mean bias [M_B] = 0.00); its errors were evenly distributed between over- and underextraction. CD showed a small bias toward underextraction (M_B = −0.07), as did OC (M_B = −0.12), whereas MAP and AF were more substantially biased in this direction (M_B = −0.77 and −1.30, respectively). As expected based on their differing penalties for less parsimonious models, the AIC tended to overextract (M_B = 0.20) and the BIC tended to underextract (M_B = −0.63). Chi-square tests tended to overextract (M_B = 0.37). The Kaiser criterion dramatically overextracted (M_B = 7.16).

CD was the most precise method (mean precision [M_P] = 0.14), followed by three techniques that were approximately as precise as one another (PA, M_P = 0.30; AIC, M_P = 0.31; OC, M_P = 0.39). Using chi-square tests was less precise (M_P = 0.60). Because each of the remaining four methods tended to err in a consistent fashion, its precision was approximately equal to its bias (BIC, M_P = 0.63; MAP, M_P = 0.77; AF, M_P = 1.30; Kaiser criterion, M_P = 7.19).

In the final series of analyses, the accuracy of the PA and CD methods was examined across levels of the design factors (see Figure 3). For both techniques, accuracy declined with more factors, with fewer items, and with smaller sample sizes. The fact that accuracy was higher for the data sets with correlated rather than orthogonal factors is probably an artifact of the experimental design; the loadings were selected to avoid ceiling effects, not to afford a well-controlled test of whether it is easier to identify the number of orthogonal or correlated factors. From binary items through quasi-continuous items with 20 ordered categories on their response scales, there was little difference in accuracy levels across item types for either technique. Most important for the purpose of this study is that CD outperformed PA in all but one data condition. The only exception was observed for data sets with five factors. In this situation, PA performed slightly better (74.2%) than CD (72.7%), though this was not a statistically significant difference (z = 0.59, p = .555). Even though PA scored more direct hits, its errors were sufficiently large that on balance its precision (M_P = .33) was no better than the precision of the CD technique (M_P = .31) for the 2,000 samples with F = 5. With this

Table 3
Summary of Accuracy for Analyses of the 10,000 Target Data Sets

	Methods testing the pattern of eigenvalues				Methods testing the fit of competing structural models				
Variable	Kaiser	PA	OC	AF	MAP	AIC	BIC	χ^2	CD
Maximum deviation	25	8	7	2	0	3[a]	0	3[a]	2
Minimum deviation	−2	−3	−5	−4	−4	−3	−4	−3	−3
Factor									
≥ +3 (%)	74.86	0.92	0.65	0.00	0.00	0.29	0.00	4.59	0.00
+2 (%)	7.34	2.07	1.81	0.01	0.00	2.76	0.00	8.05	0.02
+1 (%)	7.67	7.92	7.78	0.12	0.00	19.39	0.00	18.65	3.64
Correct (%)	8.77	76.42	74.03	45.91	59.60	72.65	60.45	58.88	87.14
−1 (%)	1.31	10.56	10.07	15.01	18.54	4.27	22.95	8.35	8.02
−2 (%)	0.05	1.93	2.92	13.70	11.04	0.61	10.89	1.34	1.12
≤ −3 (%)	0.00	0.18	2.74	25.25	10.82	0.03	5.71	0.14	0.06
Bias (*M*)	7.16	0.00	−0.12	−1.30	−0.77	0.20[a]	−0.63	0.37[a]	−0.07
Precision (*M*)	7.19	0.30	0.39	1.30	0.77	0.31[a]	0.63	0.60[a]	0.14

Note. Accuracy was assessed with deviation scores calculated as the predicted minus actual number of factors. Maximum and minimum deviation scores represent the largest errors for each technique when it overextracted or underextracted the correct number of factors, respectively. Percentages for how often each technique erred by one, two, or three factors are listed. Bias was computed as the average deviation from the correct number of factors, and precision was calculated as the average absolute deviation. Kaiser = Kaiser criterion; PA = parallel analysis; OC = optimal coordinates; AF = acceleration factor; MAP = minimum average partial; AIC = Akaike information criterion; BIC = Bayesian information criterion; CD = comparison data.

[a] The maximum overextraction might have been larger had this technique been allowed to consider models that exceeded the correct number of factors by more than three. Because overextraction was the norm for this technique, its bias and precision may have been underestimated.

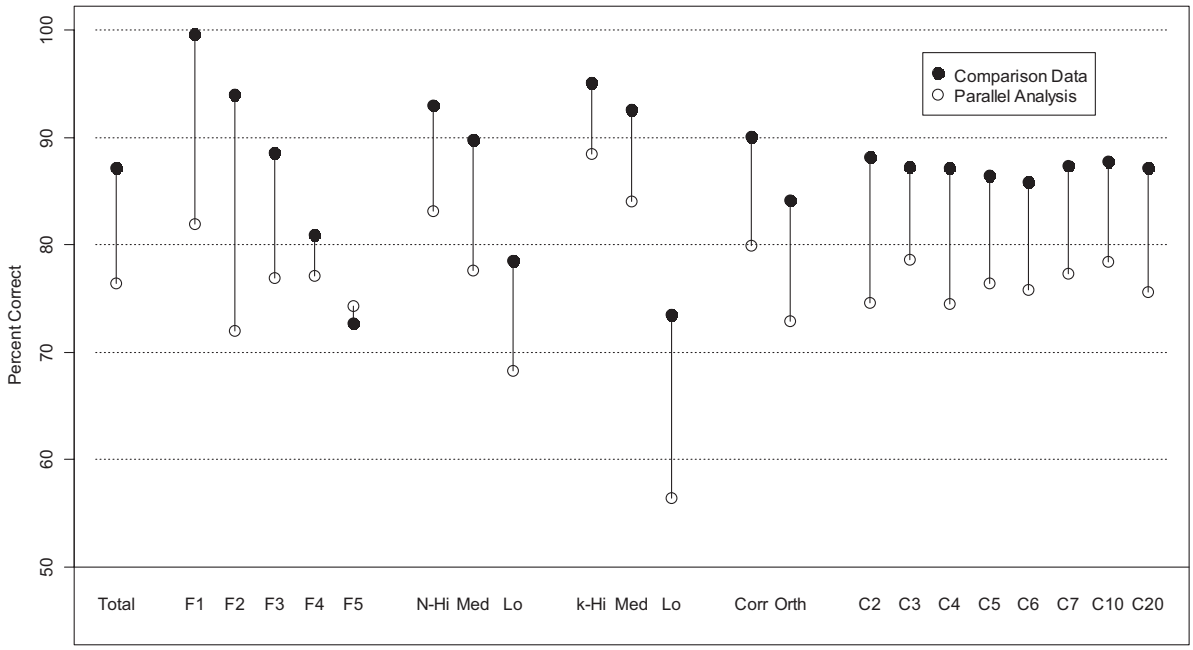

Figure 3. Results for analyses of the 10,000 target data sets using several techniques, broken down by factors in the study design. Design factors appear in the following order: number of factors (F from 1 to 5), sample size (N divided into three ranges of high, medium, and low), number of items (k divided into three ranges of high, medium, and low), correlated (Corr) versus orthogonal (Orth) factors, and number of ordered categories (C = 2, 3, 4, 5, 6, 7, 10, or 20). Vertical lines highlight the differences in accuracy between comparison data and parallel analysis.

partial exception noted, the present data support CD as at least as effective a method as PA for determining the number of factors to retain across the wide range of challenging data conditions included in this study.

Discussion

Determining the number of factors to retain in EFA remains an important yet difficult problem. Examinations of published research (Fabrigar et al., 1999; Henson & Roberts, 2006) suggest that many researchers continue to use methods that are known to perform poorly. For example, Henson and Roberts (2006) found that more than half of all studies that they reviewed (57%) used the Kaiser criterion, and only 7% used PA. PA was first introduced more than 40 years ago (Horn, 1965), and it has outperformed the Kaiser criterion in many studies since that time (for overviews of the evidence, see Fabrigar et al., 1999; Velicer et al., 2000; Zwick & Velicer, 1986). Despite the increasing availability of user-friendly software and sufficient computing power to perform PA, it is seldom used. Perhaps the most significant reason that the Kaiser criterion remains the tool of choice in EFA is that statistical packages such as SPSS continue to maintain this as the default setting (Thompson & Daniel, 1996).

In the present study, substantial differences in performance were observed across techniques. Among the techniques evaluated in this study, the CD method performed the best in terms of accuracy and precision, and it was nearly unbiased. This technique over- or underextracted by more than one factor only about 1% of the time. Only PA was less biased, though its accuracy and precision were lower than those for CD by a nontrivial amount. PA over- or underextracted by more than one factor about 5% of the time. Results were poorer in terms of accuracy, precision, or bias for each of the other methods. For example, accuracy was rather poor for the MAP, AF, and BIC techniques, each of which exclusively or almost exclusively underestimated the number of factors when they erred. Even though chi-square tests were less biased, accuracy was not much better. OC and AIC rivaled the accuracy achieved by PA, but there appears to be no reason to prefer either of these methods to the established standard of PA or the newly developed CD. The Kaiser criterion performed poorly, as the literature suggests. In fact, the Kaiser criterion predicted as many as 26 factors—and that occurred three times, each for a data set with one factor.

The conditions set forth in the study span a wide range of situations that researchers might encounter, but we caution against the generalization of the accuracy levels observed here. We were reluctant to decrease levels of study design factors such as sample size or the number of variables below realistic values. To avoid ceiling effects in accuracy, however, we allowed factor loadings to range down to levels that would be considered fairly low. Because PA has performed very well in prior simulation studies (e.g., Velicer et al., 2000; Zwick & Velicer, 1986), we wanted to avoid ceiling effects when comparing the newly proposed CD method with the PA standard. This required especially challenging data conditions. Whereas the accuracy rates observed in the main study might appear unusually low in comparison with other simulation

studies, that was by design; we do not believe that the present findings call into question the absolute or relative accuracy rates of any methods under other data conditions.

To provide some sense for how much accuracy might improve under more favorable data conditions, particularly with larger factor loadings, we generated an additional 1,000 target data sets using the same randomly sampled data conditions as in the main study, with one exception: All ranges of factor loadings were increased by .10 (e.g., a range of .30–.70 became .40–.80, a range of .10–.35 became .20–.45). The CD method and PA were performed for these data sets. The accuracy rates were 92.1% for the CD method and 82.5% for PA, a substantial and statistically significant difference ($z = 2.30$, $p = .021$). For each subset of 200 data sets with between one and five factors, the CD method was more accurate than PA.

In the present study, we did not vary distributional characteristics that might have led to the identification of spurious factors (e.g., subsets of binary items with distinct difficulty levels). Future research might test whether using the CD method provides some protection against such artifacts.

A discouraging finding of Henson and Roberts' (2006) review of published EFAs is that 55% of the studies that they reviewed reported only one criterion for determining the number of factors. When performing EFA, researchers can use more than one method to check whether the results are in agreement. This could be especially helpful under certain data conditions, such as when an investigator suspects there may be a large number of factors. In the present study, this is where PA and CD performed least satisfactorily. However, PA and CD both predicted five factors for 1,297 target data sets; this was correct in 1,277 instances (98.5%). In all 10,000 target data sets, 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%. Including additional methods (e.g., requiring that PA, CD, AIC, and OC agree) can increase accuracy even further (e.g., to 95.9% in the present study), but this applies only to instances in which agreement was obtained (e.g., 57.5% of the time in the present study). There is a trade-off such that when one uses multiple methods to compare structural models, agreement is not guaranteed, but when results are in agreement, they are highly likely to be accurate (Ruscio et al., 2010). Experts on EFA (e.g., Gorsuch, 1983) have long recommended the use of multiple methods to determine the number of factors to retain, and the present findings demonstrate how effective this practice can be.

Especially when relying on the scree test (Cattell, 1966) or when drawing conclusions from one or more methods of dubious utility (e.g., the Kaiser criterion), efforts to determine the number of factors to retain can be plagued by difficult subjective judgments. Using one or more of the empirically supported methods developed and tested in recent decades can help researchers model their data more accurately and ultimately to develop more reliable and valid assessment instruments. In particular, we recommend that researchers take advantage of PA as a starting point, perhaps supplemented by CD as proposed here. Though these methods will disagree some of the time, the magnitude of disagreement between these techniques is likely to be small. In the present study, they agreed within $\pm$ 1 factor nearly 96% of the time. In the event of a disagreement, one could choose to rely on PA given its demonstrated track record of success in multiple simulation studies or on CD given its superior performance in the present study. We believe the present data provide tentative, but sufficient, support for using CD in EFA and that further study of this new approach is warranted. To facilitate study or application of CD, the first author has made the R code needed to perform it available on his website (http://www.tcnj.edu/~ruscio/taxometrics.html). This allows investigators to implement the CD technique with no special knowledge of how to generate or analyze comparison data, just as no special knowledge of how to generate random data sets or calculate eigenvalues is required to perform PA using a program written for that purpose.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723. doi:10.1109/TAC.1974.1100705

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33,* 261–304. doi:10.1177/0049124104268644

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245–276. doi:10.1207/s15327906mbr0102_10

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, 4,* 272–299. doi:10.1037/1082-989X.4.3.272

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selection the correct number of factors to retain. *Educational and Psychological Measurement, 55,* 377–393. doi:10.1177/0013164495055003002

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Harshman, R. A., & Reddon, J. R. (1983, May). *Determining the number of factors by comparing real with random data: A serious flaw and some possible corrections.* Paper presented at the annual meeting of the Classification Society, Philadelphia, PA.

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. doi:10.1177/0013164405282485

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185. doi:10.1007/BF02289447

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141–151. doi:10.1177/001316446002000116

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27,* 82–99. doi:10.1111/j.2044-8317.1974.tb00530.x

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, and Computers, 32,* 396–402. doi:10.3758/BF03200807

Raiche, G., Riopel, M., & Blais, J.-G. (2006, June). *Nongraphical solutions for the Cattell's scree test.* Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada.

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

Ruscio, J., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research, 43,* 355–381. doi:10.1080/00273170802285693

Ruscio, J., & Kaczetow, W. (2009). Differentiating categories and dimensions: Evaluating the robustness of taxometric analysis. *Multivariate Behavioral Research, 44,* 259–280. doi:10.1080/00273170902794248

Ruscio, J., & Marcus, D. K. (2007). Detecting small taxa using simulated

comparison data: A reanalysis of Beach, Amir, and Bau's (2005) data. *Psychological Assessment, 19,* 241–246. doi:10.1037/1040-3590.19.2.241

Ruscio, J., Ruscio, A. M., & Meron, M. (2007). Applying the bootstrap to taxometric analysis: Generating empirical sampling distributions to help interpret results. *Multivariate Behavioral Research, 42,* 349–386. doi:10.1080/00273170701360795

Ruscio, J., & Walters, G. D. (2009). Using comparison data to differentiate categorical and dimensional data by examining factor score distributions: Resolving the mode problem. *Psychological Assessment, 21,* 578–594. doi:10.1037/a0016558

Ruscio, J., Walters, G. D., Marcus, D. K., & Kaczetow, W. (2010). Comparing the relative fit of categorical and dimensional latent variable models using consistency tests. *Psychological Assessment, 22,* 5–21. doi:10.1037/a0018259

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. doi:10.1214/aos/1176344136

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. doi:10.1177/0013164496056002001

Turner, N. E. (1998). The effect of common variance and structure pattern on random data eigenvalues: Implications for the accuracy of parallel analysis. *Educational and Psychological Measurement, 58,* 541–568. doi:10.1177/0013164498058004001

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41,* 321–327. doi:10.1007/BF02293557

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.

Weng, L., & Cheng, C. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement, 65,* 697–716. doi:10.1177/0013164404273941

Zwick, W. R., & Velicer, W. F. (1986). Factors influencing five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432–442. doi:10.1037/0033-2909.99.3.432

Received January 12, 2010
Revision received July 5, 2011
Accepted July 20, 2011 ■