Zwick, WR & Velicer, WF (1986). Comparison of five rules for determining the number of components to retain. Psych. Bulletin, 99, 432-442.

Psychological Bulletin
1986, Vol. 99, No. 3, 432-442


0033-2909/86/$00.75

# Comparison of Five Rules for Determining the Number of Components to Retain

William R. Zwick
Rhode Island Group Health Association

Wayne F. Velicer
University of Rhode Island

The performance of five methods for determining the number of components to retain (Horn's parallel analysis, Velicer's minimum average partial [MAP], Cattell's scree test, Bartlett's chi-square test, and Kaiser's eigenvalue greater than 1.0 rule) was investigated across seven systematically varied conditions (sample size, number of variables, number of components, component saturation, equal or unequal numbers of variables per component, and the presence or absence of unique and complex variables). We generated five sample correlation matrices at each of two sample sizes from the 48 known population correlation matrices representing six levels of component pattern complexity. The performance of the parallel analysis and MAP methods was generally the best across all situations. The scree test was generally accurate but variable. Bartlett's chi-square test was less accurate and more variable than the scree test. Kaiser's method tended to severely overestimate the number of components. We discuss recommendations concerning the conditions under which each of the methods are accurate, along with the most effective and useful methods combinations.

Behavioral scientists often wish to represent a set of observed variables ($P$) by a smaller set of derived variables ($m$). Component analysis and factor analysis are two procedures designed to solve this problem. Such analyses may allow the calculation of $m$ scores to replace the original $P$ observations or provide information for the understanding and interpretation of the original variables. Researchers seeking to summarize the data set must make a number of decisions, including choice of method, choice of rotation, and choice of method of calculating the scores. One of the most critical decisions the applied researcher faces is selecting how many factors or components to retain ($m$). In this article we present the results of a Monte Carlo evaluation of five methods that have been proposed for determining the value of $m$.

The determination of the number of components or factors to retain is likely to be the most important decision a researcher will make. Decisions involving choice of method, type of rotation, and type of score will have relatively less impact because of the demonstrated robustness of results across different alternatives in these areas. Under- or over-extraction will distort subsequent results. The obvious problem of underextraction involves the loss of important information by ignoring a factor or combining it with another factor. The effects of overextraction, followed by rotation, are less well documented but equally important. Comrey (1978) describes some of the dangers, such as minor factors being built up at the expense of major factors and/or the creation of factors with only one high loading and a few low loadings. These are factors that are both uninterpretable and unlikely to replicate. Velicer and Jackson (1985) assert that overextraction is likely the prime reason for discrepancies between factor analysis and component analysis. In view of this decision's importance to the analysis, it is interesting that some recent textbooks provide little or no guidance in this area (Chatfield & Collins, 1980; B. B. Jackson, 1983; Lunneborg & Abbott, 1983).

## Properties of Retained Components

The comparison of methods to determine the number of components to retain requires a description of the qualities desirable in a retained component. A review of the properties of principal components, linked with the goal of data summarization, provides such a description.

### *Number of Substantial Loadings*

If principal-components analysis (PCA) is used to summarize a data set, each retained component must contain at least two substantial loadings. Algebraic (T. W. Anderson & Rubin, 1956) and statistical (Lawley, 1940; Morrison, 1976) examinations of common factor analysis (CFA) agree that at least three variables are required before the first factor can be identified. T. W. Anderson and Rubin (1956) have further demonstrated that each subsequent identifiable factor must contain at least three nonzero loadings. At a sample level, a minimum of at least three significant loadings are required for factor identification.

### *Variance Accounted For*

The variance of each principal component is equal to the eigenvalue of that component. A 1.0-eigenvalue component accounts for as much variance as a single variable. Components with eigenvalues near zero provide no summarizing power. A component with an eigenvalue greater than 1.0 provides more summarizing power than an original variable.

Some of the material included in this article appeared in the first author's doctoral dissertation.

Correspondence concerning this article should be addressed to William R. Zwick, Counseling Department, Rhode Island Group Health Association, 530 North Main Street, Providence, Rhode Island 02904.

### *Component Reliability*

Kaiser (1960) and Kaiser and Caffrey (1965) addressed the issue of component reliability. Kaiser (1960) argued that the reliability of a component will always be nonnegative when the eigenvalue exceeds 1.0. Horn (1969) noted that this approach to reliability includes all $P$ variables regardless of their component loadings. In applied usage, component scores are usually generated as an unweighted sum of those variables with substantial component loadings. Reliability estimates based only on those items contributing to the component score can be quite high even when the component eigenvalue is below 1.0 (Horn, 1969).

The component properties reviewed above can be integrated to present an operational definition of a useful component. Conventional use of PCA as a data reduction technique, combined with algebraic and statistical necessity in CFA, dictate that at the population level, at least three nonzero loadings are required to identify a useful component. At the sample level, three significant and substantial loadings are needed. In order to guarantee nonnegative component reliability, retained components are required to have an eigenvalue greater than 1.0. Therefore, we refer to all components with three or more substantial loadings *and* an eigenvalue of greater than 1.0 as *major components* (MJCs). Such MJCs would probably be of interest to most investigators. Components that have either (a) less than three substantial loadings but an eigenvalue of 1.0 or greater or (b) more than three substantial loadings but an eigenvalue of less than 1.0 may be of interest to some investigators, and we refer to them as *minor components* (MNCs). Finally, components with both less than three substantial loadings and an eigenvalue less than 1.0 should never be retained, and we refer to them as *trivial components* (TCs).

Principal-components analysis (Hotelling, 1933) may be viewed as involving an eigendecomposition of the $P \times P$ sample correlation matrix $\mathbf{R}$,

$$\mathbf{R} = \mathbf{L}\mathbf{D}^2\mathbf{L}, \quad (1)$$

where $\mathbf{D}^2$ is the $P \times P$ diagonal matrix containing the eigenroots of $\mathbf{R}$, and $\mathbf{L}$ is a $P \times P$ matrix containing the corresponding eigenvectors. When PCA is used as a data summary model, only the first $m$ components are retained. The component pattern ($\mathbf{A}$) may be written as

$$\mathbf{A} = \mathbf{L}_m\mathbf{D}_m, \quad (2)$$

where $\mathbf{D}_m$ contains the first $m$ eigenroots and $\mathbf{L}_m$ contains the corresponding first $m$ eigenvectors. Glass and Taylor (1966), Pruzek and Rabinowitz (1981), and Kaiser (1970) have reported on the widespread use of PCA in this manner. Velicer (1974, 1976a, 1977) and Velicer, Peacock, and Jackson (1982) have shown that this use of PCA and CFA results in essentially equivalent solutions.

A second class of procedures, CFA, has also been used to summarize a set of $P$ variables. It is important to note that $m$ is frequently assumed to be known for the derivation of these factor analysis procedures. Sometimes the maximum likelihood test is used to test whether the assumed number of factors is correct.

Because both CFA and PCA are used as data summary techniques, it is important to note some differences between them. The CFA approach requires that $m$ be known prior to the analysis. The value of $m$ is usually determined in one of two ways. It may be determined by applying some rule to a PCA solution, with the result then used in the factor analysis solution, or a maximum likelihood test may be used to evaluate different values of $m$. Unfortunately, the methods applied to the PCA solution often provide conflicting results. Further, D. N. Jackson and Chan (1980) have discussed numerous computational difficulties with the maximum likelihood approach itself. Finally, an indeterminacy has been identified in the solution of the basic factor analysis equation (Guttman, 1954; Schonemann & Wang, 1972; Steiger & Schonemann, 1979). In light of difficulties associated with the requirement that $m$ be known a priori, the indeterminacy of the factor model, the computation problems with factor analysis, the widespread use of PCA, and the general comparability of results across the two methods, we chose to focus on the PCA procedure.

## Determining the Number of Components

A number of rules have been suggested to determine the appropriate number of components to retain (Bartlett, 1950, 1951; Cattell, 1966; Crawford, 1975; Everett, 1983; Horn, 1965; Joreskog, 1962; Kaiser, 1960; Revelle & Rocklin, 1979; Veldman, 1979; Velicer, 1976b). These rules often do not give the same results (A. D. Anderson, Acito, & Lee, 1982; Cattell & Vogelman, 1977; Hakstian, Rogers, & Cattell, 1982; Horn, 1965; Linn, 1968; Zwick & Velicer, 1982). Applied researchers are, therefore, often at a loss as to how to proceed. Conflicting research conclusions can be traced to differing methods of defining the correct number of components.

In this section we describe the five methods to be evaluated. These methods are: (a) Bartlett's chi-square test, (b) Kaiser's eigenvalue greater than 1.0 rule (K1), (c) the minimum average partial rule (MAP), (d) the scree test, and (e) the parallel analysis (PA) method. These methods were selected for inclusion because of their widespread use or their extensive theoretical justification.

### *Bartlett's Test*

Following Lawley's (1940, 1941) test for maximum likelihood factor analysis, Bartlett (1950, 1951) developed an analagous statistical test of the null hypothesis that the remaining $P - m$ eigenvalues are equal. Each eigenvalue is excluded sequentially until the approximate chi-square test of the null hypothesis of equality fails to be rejected. The first $m$ excluded components are retained. The Bartlett test is not commonly available in standard statistical packages.

Bartlett's test appears sensitive to the sample size. Gorsuch (1973) argued that increased power at larger sample sizes could lead to the retention of more components. Horn and Engstrom (1979) suggested changing the alpha level at different sample sizes. However, as the sample size increases, the estimates of population eigenvalues become increasingly more accurate. This increased accuracy leads to smaller observed differences between equal eigenvalues and may appropriately offset the increased power of the Bartlett test when the population eigenvalues are actually equal. Zwick and Velicer (1982) found the Bartlett test somewhat more accurate with large samples than with small samples.

### *Eigenvalue Greater Than 1.0 (K1)*

Perhaps the most popular—certainly the most commonly used—method is to retain the components with eigenvalues greater than 1.0. The K1 method is the default option on many statistical packages (i.e., SPSS-X, SAS, BMDP). Kaiser (1960) developed the rationale for this method, focusing on component reliability, pattern meaningfulness, and Guttman's (1954) work examining the lower bounds for the number of components in image analysis. Gorsuch (1983) noted that many users follow Kaiser (1960) and use the K1 rule to determine the number of components rather than as a lower bound. Mote (1970) and Humphreys (1964) reported that rotation of a greater number of components resulted in more meaningful solutions. They indicated that the K1 rule may sometimes lead to the retention of too few components.

A number of researchers (Browne, 1968; Cattell & Jaspers, 1967; Horn, 1965; Lee & Comrey, 1979; Linn, 1968; Revelle & Rocklin, 1979; Yeomans & Golder, 1982; Zwick & Velicer, 1982) have found that the number of components retained by K1 is often an overestimate. Gorsuch (1983) and Kaiser (1960) report that the number of components retained by K1 is commonly between one-third and one-fifth or one-sixth the number of variables included in the correlation matrix. A Monte Carlo study by Zwick and Velicer (1982) supports this result. This relation of the number of retained components to the number of variables is problematic. The K1 method, although commonly used, is believed by some critics to sometimes underestimate and by many others to grossly overestimate the number of components.

### *Minimum Average Partial (MAP)*

Velicer (1976b) has suggested a method based on the matrix of partial correlations. The average of the squared partial correlation is calculated after each of the $m$ components has been partialed out. When the minimum average squared partial correlation is reached, no further components are extracted. The average squared partial correlation reaches a minimum when the residual matrix most closely resembles an identity matrix. Using this rule, at least two variables will have high loadings on each retained component. Velicer (1976b) points out that the method is exact, can be applied with any covariance matrix, and is logically related to the concept of factors as representing more than one variable. Zwick and Velicer (1982) reported that the MAP rule was more accurate in identifying a known number of components than either the K1 or the Bartlett test rule. Reddon (in press) also reports good performance in a monte carlo evaluation. Reddon (1985) describes some simple FORTRAN subroutines for calculating the MAP rule.

### *Scree Test*

Cattell (1966) described this rule, which is based on a graph of the eigenvalues. The scree test is simple to apply. The eigenvalues are plotted, a straight line is fitted through the $P - m$ smaller values, and those falling above the line are retained. A number of complications may occur, including: (a) a gradual slope from lower to higher eigenvalues with no obvious break point in the line, (b) more than one break point in the line, and (c) more than one apparently suitable line may be drawn through the low values. Horn and Engstrom (1979) have noted the underlying similarity of the logic of Bartlett's chi-square test and the scree test. Both tests are based on an analysis (one statistical, the other visual) of the essential equality of the "remaining" eigenvalues.

The graph for scree inspection is available as an option in SPSS-X. The actual decision must be made by the researcher. Tucker, Koopman, and Linn (1969) found the scree test to be correct in 12 of 18 cases. Cliff (1970) found it to be accurate, particularly if questionable components are included. Cattell and Jaspers (1967) found it was correct in 6 of 8 cases, whereas Cattell and Vogelmann (1977) reported the test to be accurate over 15 systematically differing analyses. Further, Cliff and Hamburger (1967) found more definite breaks with larger sample sizes ($n$ = 400 vs. $n$ = 100), and Linn (1968) concurred with this conclusion. Zwick and Velicer (1982) found the scree test to be most accurate with larger samples and strong components. They found that it was the most accurate of four methods evaluated across many examples of matrices of known, noncomplex structure.

Use of the scree test always involves issues of reliability. Cattell and Vogelmann (1977) and Zwick and Velicer (1982) have reported good interrater reliability both among naive and among expert judges. However, Crawford and Koopman (1979) have reported extremely low interrater reliabilities. The circumstances associated with high and low interrater reliability on the scree procedure have not been identified.

### *Parallel analysis (PA)*

Parallel Analysis (Horn, 1965) is a sample-based adaptation of the population-based K1 rule. Horn (1965) noted that at the population level, the eigenvalues of a correlation matrix of uncorrelated variables would all be 1.0. When samples are generated based on such a matrix, however, the initial eigenvalues exceed 1.0 whereas the final eigenvalues are below 1.0. Horn (1965) suggested that the eigenvalues of a correlation matrix of $P$ random uncorrelated variables should be contrasted with those of the data set in question, based on the same sample size. Components of the matrix of interest, which have eigenvalues greater than those of the comparison random matrix, would be retained. This approach integrates the reliability and data-summarizing emphasis of the population-based K1 rule and the effects of sample size.

Although there has been no published systematic examination of the PA method with PCA, M. B. Richman (personal communication, October 14, 1983) reported a series of simulation studies using this method. He found that PA was very accurate when applied to correlation matricies conforming to the formal factor analytic model. He further reported that PA led to retention of too many components when applied to correlation matricies conforming to the middle model described by Tucker et al. (1969). The method was more accurate in both cases at larger ($n$ = 500) than at smaller ($n$ = 100) sample sizes.

Humphreys and Montanelli (1975) applied PA to principal axis factor analysis and found it was accurate over a range of examples. Montanelli and Humphreys (1976) developed a regression equation that accurately predicts the eigenvalues of random correlation matrices with squared multiple correlations inserted as the diagonal. Green (1983) used this prediction equa-

tion to evaluate the performance of factor analysis of binary items.

## Variables Affecting Decision Methods

Previously reported findings on the performance of each of the decision methods indicate that they are affected by several different factors. These factors include sample size, the number of variables, component saturation, component identification, and the presence of special types of variables.

### *Sample Size*

Depending on the decision method used, it is possible that the number of subjects may affect the accuracy of the decision about the number of components to retain. Sample size is typically determined both by practical applied considerations and the need for accurate estimation of correlations.

### *Number of Variables*

Principal-components analyses often involve 80 to 100 variables. Analysis of 200 variable sets is becoming common. Larger numbers of variables reportedly increase the accuracy of some rules but decrease it for others (Zwick & Velicer, 1982).

### *Component Saturation*

Underlying components made up exclusively of high loadings (e.g., .80) were more likely to be retained by various decision methods than components exclusively made up of lower loadings (e.g., .40; Linn, 1968; Zwick & Velicer, 1982).

### *Component Identification*

The accurate identification of a component may depend on the number of variables that have nonzero loadings on that component. The impact of unequal numbers of variables per component is unclear for any of the rules under discussion.

### *Special Variables*

Complex variables have a nonzero loading on more than one component. Unique variables have only one nonzero loading and no other variable loads substantially on the same component. Component patterns containing both complex and unique variables are believed to occur frequently in applied situations (Tucker et al., 1969). The effect of these types of variables on the various decision rules is unclear.

The robustness of the five rules in question across these variables is the central focus of this study.

## Method

### *Method of Data Generation*

Studies of the effectiveness of the various decision methods may be categorized into one of two types. Historically, the more common type of study used real data representing either new work or classic data sets. These studies used some logical criteria concerning the appropriate number of components and compared the performance of a proposed decision method to the logically determined value (e.g., Cattell, 1966; Horn, 1965; Humphreys & Montanelli, 1975; Velicer, 1976a). In using an arbitrary logical criterion, these studies may have inaccurately estimated the performance of the decision method in question. More recently, studies of decision rule effectiveness have used correlation matrices generated from component structures entirely under the control of the investigator (e.g., A. D. Anderson et al., 1982; Cattell & Vogelman, 1977; Tucker et al., 1969; Zwick & Velicer, 1982). These studies have the advantage of a known criterion against which to measure the performance of the decision methods. They are, however, open to the criticism that the generated matrices, although conforming to a mathematical model, may not represent real data and may thus lead to inappropriate conclusions (Tucker et al., 1969).

The question of a rule's accuracy cannot be examined without a known criterion. Although logical arguments can be mounted to defend the number of components present in some data sets, these arguments are always open to question. For the assessment of the impact of various conditions on a rule's accuracy, generated data of a known number of components is preferable. The issue of generalization to real data sets is an important but separate issue that may be independently addressed in the particular way the data is generated. We used generated data in an approach similar to the "middle model" of Tucker et al. (1969).

### *Procedure*

The number of variables ($P$) to be used was set at 36 and 72. These values represent small and moderately large data sets and accommodate constraints imposed by the selection of the number of components to be included. Larger sets of variables have been shown to have a positive impact on MAP, the Bartlett test (Zwick & Velicer, 1982), and the scree test (Cattell & Vogelman, 1977), and a negative impact on K1 (Zwick & Velicer, 1982).

The sample sizes were selected to reflect common, applied usage. They were set as a function of the number of variables. The lower sample size was set at twice the number of variables. The higher sample size was set at five times the number of variables. The resulting $n$s were 72 and 180 in the cases including 36 variables. Sample sizes of 144 and 360 were selected when 72-variable cases were examined. These appear to bracket much of the range of sample sizes as reported in applied educational and psychological research. Larger sample sizes have been shown to moderately improve the performance of the MAP, scree, and K1 methods (Cattell & Vogelman, 1977; Zwick & Velicer, 1982) and to sometimes improve and sometimes weaken the accuracy of the Bartlett test (Gorsuch, 1983; Zwick & Velicer, 1982).

Three and six major components were built into the population correlation matrix when $P$ equaled 36, and six and nine when $P$ equaled 72. We chose these values to reflect a reasonable range of reported applied usage. They permit a span of an average number of variables per MJC ($P$/MJC) from 6 to 12. Although this $P$/MJC is somewhat high, such values were required to permit variability in the number of variables per component, while still meeting the constraint of at least three substantial loadings in each MJC.

The distribution of $P$/MJC was set as either an equal or an unequal number of variables per MJC. For the unequal number cases, the $P$/MJC number was symmetrically distributed about the mean number of variables per MJC. That is, if $P$ were 36, and $m$ were 6, the average number of variables per MJC would be 6. When $P$/MJC was planned to be unequal, the number of variables per component was 8, 7, 6, 6, 5, and 4. Similar distributions for other combinations of $P$ and $m$ were used.

Component saturation (SAT), the magnitude of the loading of each variable on an MJC, was split between a high of .80 and a low of .50. These values bridge much of the applied range and have been shown (Zwick & Velicer, 1982) to differentially affect four of the decision rules under consideration. Within any one matrix, the component loadings on all major components were equal and either high (.80) or low (.50).

Table 1
*Means of the Difference From the Number of Population Major Components at Different Sample Sizes*

| | | Method | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *P* | *n* | MAP | PA | Scree | K1 | BA | BB | BC |
| | | | | Saturation = .50 | | | | |
| 36 | 72 | −1.08 | −0.05 | 0.50 | 8.32 | −2.87 | −3.92 | −3.98 |
| | 180 | −1.17 | 0.13 | 0.68 | 7.30 | −1.78 | −2.20 | −2.27 |
| 72 | 144 | −0.45 | 0.02 | 1.16 | 17.80 | −0.43 | −1.60 | −1.73 |
| | 360 | −0.43 | 0.07 | 0.46 | 15.42 | 0.40 | −0.13 | −0.22 |
| | | | | Saturation = .80 | | | | |
| 36 | 72 | 0.10 | −0.02 | 0.27 | 1.77 | 0.47 | −0.48 | −0.60 |
| | 180 | 0.0 | 0.0 | 0.23 | 1.32 | 1.23 | 0.68 | 0.62 |
| 72 | 144 | 0.02 | 0.0 | 0.28 | 2.97 | 3.88 | 2.62 | 2.50 |
| | 360 | 0.0 | 0.0 | 0.31 | 2.52 | 5.03 | 4.10 | 3.98 |

*Note.* BA, BB, and BC = three levels of Bartlett's chi-square test (.05, .001, and .0005, respectively). K1 = Kaiser's eigenvalue greater than unity rule. MAP = minimum average partial. *P* = a set of variables. PA = parallel analysis. Scree = Cattell's scree test.

For the purposes of this study, unique variables were defined as variables that do not load at all on either MJCs or TCs in the population structure. Instead, unique variables are the only variables loading on one type of MNC. Such an MNC has a population eigenvalue of 1.0. We define complex variables as those variables that load substantially on an MJC but also load minimally (.20) on a second type of MNC in the population structure.

We constructed specific combinations of these variables. Previous work has indicated that sample size, *P*, and SAT have an impact on some of the decision rules. At each level of *P* and SAT, component patterns were constructed to evaluate the impact of sample size *P*/MJC, and a number of combinations of factors. A "best case" set was defined for comparison purposes. This first level of complexity (a) had an equal number of variables per MJC, no MNCs or TCs present, and thus no complex or unique variables. Five other levels of structural complexity were created for comparison with the best case. This was done by (b) including complex variables to create the MNCs, (c) letting the number of variables per MJC become unequal, (d) including unique variables (as many as there were MJCs) to provide the second type of MNC, (e) including unique and complex variables together to provide MNCs, and, finally, (f) we constructed a level that included an unequal number of variables per MJC and both complex and unique variables. We included the last two levels of complexity to highlight the possible importance of the very common situation of unequal numbers of variables per MJC.

## *Data Generation*

Population correlation matrices were created for each combination of the 6 (Complexity) × 2 (*P*) × 2 (SAT) × 2 (*m*) factors outlined above. Each population correlation matrix was determined as follows. One appropriate population component pattern (**A**) was created in accordance with the number of variables factor (*P*), the level of saturation factor (SAT), the number of components factor (*m*), and the complexity under consideration. Postmultiplying by its transpose resulted in a matrix ($\mathbf{R}^* = \mathbf{AA}'$). Substituting ones in the diagonal produced a population correlation matrix ($\mathbf{R} = \mathbf{R}^* + \mathbf{D}^2$). Five sample correlation matrices were generated from each population matrix (Montanelli, 1975) at each sample size.

Principal-components analysis was performed on each of the resulting 480 ($6 \times 2 \times 2 \times 2 \times 2 \times 5$) sample matrices. The number of components to be retained by each of the four calculable rules (K1, MAP, PA, and Bartlett's test) was determined. We selected three alpha levels to use with Bartlett's test in order to incorporate Horn and Engstrom's (1979) recommendation: BA = .05, BB = .001, and BC = .0005. The PA decision was based on the mean eigenvalues of 50 random correlation matrices at each level of *P* and the sample size.

The scree test was performed on computer generated plots (8 × 14 in.) of the eigenvalues of each of the 480 matrices. These plots were examined by two naive raters trained in the scree method (Cattell & Vogelman, 1977). The graphs were presented to the raters in different mixed orders. If either rater asked a question about a particular plot, both listened to an explanation based on the instructions given by Cattell and Vogelman (1977). Whenever possible, examples from the Cattell and Vogelman (1977) directions were used to clarify questions. Independently, an experienced expert judge, uninformed as to the details of the experiment but fully familiar with the use of the scree test, rated one sample from each of the 96 cells.

The judgment required by the scree method raises the question of rater reliability. The interrater reliability estimates of the mean were calculated for the raters' decisions corrected for the number of raters. The reliability estimates were calculated for each level of complexity, saturation and the number of variables (24 different alpha coefficients). The reliability estimates ranged from .61 to 1.00, with a median value of approximately .88. Generally, reliability was lower for higher levels of complexity and the low (.50) saturation level.

The correlations of the mean of the raters decisions with the expert's judgment ranged from .60 to .90 across the six levels of complexity. The median and mean (Fisher *Z* transformed) correlation of the averaged raters' decision with the expert's judgment were both approximately .80.

## Results

Each decision method leads to an estimate of the number of MJCs to retain. The difference between these rule-determined estimates of MJC (RMJC) and the known population value of the number of MJCs (PMJC) was the primary dependent variable in this study. This difference was computed as *d* = RMJC − PMJC. Positive *d*s indicate overestimations whereas negative *d*s indicate underestimations. A difference of zero indicates a correct estimation of MJC. The standard deviation of the difference is an indication of the methods' consistency. Smaller standard deviations indicate more consistent estimates of MJC. We present the means for each method, under various conditions, next.

The results are presented for the *P* = 36 cases and the *P* = 72 cases in each table. Each level of sample size (Table 1), number

Table 2
*Means of the Difference From the Number of Population Major Components at Different Numbers of Variables per Component*

| | | Method | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *P* | P/MJC | MAP | PA | Scree | K1 | BA | BB | BC |
| | | | | Saturation = .50 | | | | |
| 36 | 6 | −2.27 | 0.05 | 0.65 | 7.10 | −1.50 | −2.67 | −2.77 |
| | 12 | 0.02 | 0.03 | 0.53 | 8.52 | −3.15 | −3.45 | −3.48 |
| 72 | 8 | −0.92 | 0.07 | 1.02 | 15.92 | −0.7 | −1.07 | −1.48 |
| | 12 | 0.03 | 0.02 | 0.60 | 17.32 | 0.13 | 0.37 | 0.47 |
| | | | | Saturation = .80 | | | | |
| 36 | 6 | 0.08 | 0.17 | 0.33 | 1.68 | 2.93 | 2.17 | 2.10 |
| | 12 | 0.02 | 0.0 | 0.17 | 1.40 | −1.23 | −1.97 | −2.08 |
| 72 | 8 | 0.02 | 0.0 | 0.21 | 3.02 | 5.15 | 4.00 | 3.87 |
| | 12 | 0.09 | 0.0 | 0.38 | 2.47 | 3.77 | 2.72 | 2.62 |

*Note.* BA, BB, and BC = three levels of Bartlett's chi-square test (.05, .001, and .0005, respectively). K1 = Kaiser's eigenvalue greater than unity rule. MAP = minimum average partial. *P* = a set of variables. PA = parallel analysis. *P*/MJC = the average number of variables per major component. Scree = Cattell's scree test.

of variables per MJC (Table 2), and pattern complexity (Tables 3 and 4) will be summarized within each level of *P* and SAT. Table 5 presents the proportion of each method's estimates of MJC that deviated a set amount from the population value.

Table 1 presents the means of the difference between each method's estimate of MJC and the known MJC for different sample sizes when $P = 36$ and 72 and the component saturation is .50 and .80. Table 1, therefore, summarizes results collapsed across all six levels of pattern complexity and both levels of the number of variables per MJC to focus on the impact of sample size. Each of the eight rows of differences in Table 1 summarized 60 observations. Tables 1–4 follow essentially the same format. We therefore give a detailed description only for Table 1.

The first row of Table 1 presents the mean difference of each method's estimate of MJC from the known value when $P = 36$, the saturation was .50 and $n = 72$. Under these conditions, the MAP method produced a mean difference score ($d$) of −1.08, an underestimation. The PA method produced a much smaller underestimation, −0.05. The scree (0.50) and K1 (8.32) methods both overestimated the criterion, with K1 providing a very large overestimation. The results of the Bartlett test led to underestimations (−2.87, −3.92, and −3.98). The increase in $n$ from 72 to 180 is reflected in row 2. It appeared to have had minor effects on the MAP, PA, and the SCREE method. The K1 and Bartlett methods show some improvement at the higher sample size. Rows 3 and 4 of Table 1 present the mean differences for each method when $P = 72$ and $N = 144$ and 360. The MAP and PA methods were again minimally influenced by sample size. The increase in *P* improved the performance of MAP, PA, and the Bartlett test while worsening that of scree and, especially, K1. Rows 5–8 of Table 1 present the mean differences for each method when *P* equaled 36 and 72, and the saturation was .80. All the methods showed improvement at this higher level of saturation. Although not present in the tables, it should be noted that the standard deviation of the differences increased for all levels of the Bartlett test (and to a lesser extent, for the K1 rule as well) at higher

Table 3
*Means of the Difference From the Number of Population Major Components at Different Levels of Pattern Complexity*

| | | Method | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Complexity | *P* | MAP | PA | Scree | K1 | BA | BB | BC |
| 1 | 36 | −0.30 | 0.0 | 0.38 | 7.15 | −0.90 | −1.15 | −1.20 |
| | 72 | 0.0 | 0.0 | 0.82 | 15.95 | −0.25 | −0.60 | −0.75 |
| 2 | 36 | −0.50 | −0.10 | 0.88 | 7.35 | −1.00 | −1.40 | −1.45 |
| | 72 | −0.05 | 0.05 | 1.18 | 16.00 | −0.15 | −0.60 | −0.65 |
| 3 | 36 | −0.80 | −0.10 | 0.50 | 7.30 | −1.20 | −1.40 | −1.55 |
| | 72 | −0.95 | −0.20 | 0.40 | 15.90 | −0.60 | −1.00 | −1.10 |
| 4 | 36 | −1.60 | 0.0 | 0.38 | 8.45 | −1.55 | −1.85 | −1.90 |
| | 72 | −0.10 | 0.30 | 1.00 | 17.15 | −0.45 | −0.95 | −1.00 |
| 5 | 36 | −2.05 | −0.25 | 0.58 | 8.20 | −1.30 | −1.85 | −1.90 |
| | 72 | −0.10 | 0.20 | 0.90 | 17.15 | −0.40 | −0.85 | −1.00 |
| 6 | 36 | −1.50 | 0.20 | 0.85 | 8.40 | −1.30 | −1.70 | −1.75 |
| | 72 | −1.45 | −0.10 | 0.35 | 17.50 | −0.70 | −1.20 | −1.25 |

*Note.* Saturation = .50. BA, BB, and BC = three levels of Bartlett's chi-square test (.05, .001, and .0005, respectively). K1 = Kaiser's eigenvalue greater than unity rule. MAP = minimum average partial. *P* = a set of variables. PA = parallel analysis. Scree = Cattell's scree test.

Table 4
*Means of the Difference From the Number of Population Major Components at Different Levels of Pattern Complexity*

| | | Method | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Complexity | *P* | MAP | PA | Scree | K1 | BA | BB | BC |
| 1 | 36 | 0.0 | 0.0 | 0.12 | 0.15 | 0.05 | 0.0 | 0.0 |
| | 72 | 0.0 | 0.0 | 0.12 | 0.0 | 0.05 | 0.0 | 0.0 |
| 2 | 36 | 0.20 | 0.0 | 0.25 | 0.15 | 0.30 | 0.10 | 0.0 |
| | 72 | 0.05 | 0.0 | 0.18 | 0.35 | 1.70 | 0.90 | 0.80 |
| 3 | 36 | 0.0 | 0.0 | 0.10 | 0.10 | 0.05 | 0.0 | 0.0 |
| | 72 | 0.0 | 0.0 | 0.0 | 0.05 | 0.05 | 0.0 | 0.0 |
| 4 | 36 | 0.0 | 0.0 | 0.32 | 2.90 | 3.20 | 2.70 | 2.50 |
| | 72 | 0.0 | 0.0 | 1.05 | 5.40 | 5.85 | 5.30 | 5.20 |
| 5 | 36 | 0.05 | 0.0 | 0.32 | 2.85 | 4.55 | 3.80 | 3.65 |
| | 72 | 0.0 | 0.0 | 0.40 | 5.40 | 7.95 | 7.10 | 6.80 |
| 6 | 36 | 0.05 | −0.05 | 0.38 | 3.10 | 3.45 | 3.00 | 2.90 |
| | 72 | 0.0 | 0.0 | 0.0 | 5.25 | 7.60 | 6.85 | 8.65 |

*Note.* Saturation = .80. BA, BB, and BC = three levels of Bartlett's chi-square test (.05, .001, and .0005, respectively). K1 = Kaiser's eigenvalue greater than unity rule. MAP = minimum average partial. *P* = a set of variables. PA = parallel analysis. Scree = Cattell's scree test.

saturation and at higher sample sizes. The three levels of the Bartlett test also retained more components at a higher sample size. This led to a larger overestimation at BA and a switch from under- to overestimation at BB and BC.

The K1 method performed slightly better at the higher sample size at both levels of component saturation. The MAP and PA methods were minimally influenced by the sample size change at both levels of component saturation and number of variables. When the saturation was .50 and $P = 72$, the scree method showed less overestimation at the higher than at the lower sample size. This effect was not apparent when the saturation was .80.

The role of the number of variables is presented from a different perspective in Table 2, which summarizes the results for each average number of variables per MJC (*P*/MJC) when *P* equaled 36 and 72 and the saturation was .50 and .80. At both levels of SAT, MAP, PA, and scree performed more accurately and consistently when the average number of variables per MJC was higher. The K1 showed an increased overestimation when *P*/MJC increased and the saturation was .50 and a decreased overestimation when *P*/MJC increased and the saturation was .80. The Bartlett test consistently showed a decrease in the number of components retained as *P*/MJC increased, except when $P = 72$ and SAT = .50.

Tables 3 and 4 present a summary of the methods' performance when *P* equaled 36 and the saturation was .50 or .80 at each of six levels of pattern complexity. The results are collapsed

Table 5
*Percentage of Each Method's Estimate Deviating a Set Amount From the Number of Population Major Components*

| | Method | | | | | | |
|---|---|---|---|---|---|---|---|
| Deviation | MAP | PA | Scree | K1 | BA | BB | BC |
| | | | Saturation = .50 | | | | |
| ±3 | 0.0 | 0.0 | 6.7 | 100.0 | 0.8 | 0.0 | 0.0 |
| +2 | 0.0 | 1.2 | 18.0 | 0.0 | 2.9 | 0.4 | 0.0 |
| +1 | 1.6 | 9.2 | 27.5 | 0.0 | 6.2 | 0.8 | 0.7 |
| ±0 | 67.5 | 84.2 | 41.7 | 0.0 | 33.8 | 25.0 | 21.5 |
| −1 | 9.2 | 3.4 | 4.6 | 0.0 | 17.9 | 18.8 | 21.8 |
| −2 | 8.0 | 1.7 | 1.6 | 0.0 | 7.9 | 12.0 | 11.15 |
| −3 | 13.8 | 0.0 | 0.0 | 0.0 | 30.4 | 42.9 | 44.5 |
| | | | Saturation = .80 | | | | |
| ±3 | 0.0 | 0.0 | 2.9 | 43.4 | 42.9 | 32.5 | 31.6 |
| +2 | 0.0 | 0.0 | 3.3 | 6.6 | 2.5 | 7.9 | 8.4 |
| +1 | 2.9 | 0.0 | 21.6 | 6.2 | 11.7 | 2.0 | 2.0 |
| ±0 | 97.1 | 99.6 | 71.2 | 43.8 | 27.1 | 37.5 | 37.5 |
| −1 | 0.0 | 0.4 | 0.0 | 0.0 | 3.4 | 3.4 | 3.4 |
| −2 | 0.0 | 0.0 | 0.4 | 0.0 | 3.4 | 3.8 | 2.1 |
| −3 | 0.0 | 0.0 | 0.4 | 0.0 | 9.2 | 12.9 | 12.0 |

*Note.* BA, BB, and BC = three levels of Bartlett's chi-square test (.05, .001, and .0005, respectively). K1 = Kaiser's eigenvalue greater than unity rule. MAP = minimum average partial. PA = parallel analysis. Scree = Cattell's scree test.

across both levels of sample size and P/MJC so that each level of complexity represents 20 observations. The levels of complexity are defined above.

The range of pattern complexity affected the methods differently. Although the methods tended to perform best at Complexity Level 1, they had different worst cases. In Table 3, when the saturation was .50, the worst cases were MAP at Level 5, $P = 36$; PA at Level 4, $P = 72$; scree at Level 2; K1 and the Bartlett test at Levels 4–6 at both $P = 36$ and 72. As a comparison of Tables 3 and 4 indicates, MAP, PA, scree, and K1 showed substantial improvement at all levels of complexity when the saturation was .80. At this saturation level, PA slightly underestimated at Complexity Level 6. The MAP slightly overestimated at Complexity Levels 2, 5, and 6. At $P = 36$, scree slightly overestimated at all levels of complexity, and Level 6 resulted in its largest overestimation. At $P = 72$, scree overestimated the most at Level 4. The Bartlett test overestimated slightly or not at all at Levels 1–3 but did overestimate at Levels 4–6. The K1 method markedly overestimated at Complexity Levels 4–6. Levels 4–6 all contain unique variables.

A general overview of the performance of the different methods may be gained by calculating the percentage of times each method's estimate deviated a set amount from the criterion. Although P and saturation appear to have had the most substantial impact on the methods, the percentages were computed at each level of saturation only. The effects of P are strongest on K1 and the Bartlett test and are clear from the tables above. The impact of saturation appears to allow a clear differentiation between the remaining methods. Deviations of more than 3 were collapsed for simplicity of presentation. Differences of zero indicate accurate estimates. These percentages are presented in Table 5.

At both levels of saturation PA was clearly the most frequently accurate method, followed by MAP and scree. The tendency of K1 to overestimate was marked. The K1 method never underestimated. The Bartlett test was quite inaccurate and variable at both levels of saturation.

## Discussion

The question of interest in this study was the ability of five decision methods to estimate the number of major components present in the population correlation matrices given only the generated sample matrices. The difference between the estimated number and the defined number of major components served as the primary dependent variable in this simulation study. The standard deviation of the difference scores gave further information about each method's consistency. Finally, the percentage of decisions at specified levels of deviation from the criterion were also calculated.

The five decision rules we used were the eigenvalue greater than 1 rule (K1), Bartlett's test, the scree test, the minimum average partial (MAP) method, and the parallel analysis (PA) method. We examined the performance of the five methods for determining the number of components in 10 samples drawn from each of 48 simulated population correlation matrices over a range of component pattern complexity. The least complex pattern replicated earlier work (Zwick & Velicer, 1982) and included only equal numbers of variables per component and no unique or complex variables. Component pattern complexity was varied by modifying this clear, least complex case with combinations that included: (a) complex variables, (b) equal and unequal numbers of variables per component, and (c) unique variables. We feel that the resultant six levels of complexity cover an adequately wide range to permit a test of the relative strengths and weaknesses of the decision methods examined. We defined major components (MJCs) as those having more than three substantial loadings and an eigenvalue greater than or equal to 1.0 at the population level. Two types of minor components (MNCs) were defined. We feel that these complex patterns expand on the formal model and incorporate cases likely to be encountered in real data analyses.

The K1 rule consistently overestimated the number of major components. It never underestimated. This finding is consistent with those of Cattell and Jaspers (1967), Linn (1968), Yeomans and Golder (1982), and Zwick and Velicer (1982). At a component saturation of .50, the number retained often fell in the one-third to one-half of P range discussed by Gorsuch (1983). As the number of variables increased, so did the number of components retained. The K1 method retained more components when unique variables were included in the population pattern. These findings are clearly contrary to those of Humphreys (1964) and Mote (1970), who concluded that the K1 method sometimes retained too few components. Their data may represent a type of component complexity not included in this study or their original judgments of the number of components in their data sets may have been overestimates. Given the apparent functional relation of the number of components retained by K1 to the number of original variables and the repeated reports of the method's inaccuracy, we cannot recommend the K1 rule for PCA.

The results and conclusions about the K1 rule presented here are consistant both with previous empirical studies and the theoretical criticism of the method. However, our conclusions are in sharp contrast to many recent textbooks in which K1 is either the preferred or the only method discussed (Afifi & Clark, 1984; Everitt & Dunn, 1983; Johnson & Wichern, 1982; Marascuilo & Levin, 1983). For example, Marascuilo & Levin (1983) are typical when they first discuss it with a caution ("In most cases, Kaiser's rule is quite workable, but . . ."; p. 237), but in a later summary, flatly recommend it ("It is one we recommend"; p. 260). The use of the K1 rule as the default value in some of the standard computer packages (BMDP, SPSS-X, SAS) is an implicit endorsement of the procedure, particularly to naive 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.

The Bartlett test's performance was the most variable of those examined. In addition to variability, it was sensitive to a number of influences. Increases in sample size, P, and SAT, as well as the use of conservative alpha levels and the presence of unique variables all lead to the retention of more components. Although examination of different alpha levels led to fewer or greater numbers of components retained, the accuracy and consistency of the method did not appear to be markedly improved by adjusting alpha levels with sample size as was suggested by Horn and Engstrom (1979). Other factors present in this study appear to have

had a greater influence on the performance of the test, across alpha levels, than did sample size alone.

The Bartlett test is accurate in answering statistical questions concerning the equality of eigenvalues (Bartlett, 1950, 1951). Researchers inclined to examine minor components, particularly early in the course of exploratory analysis, may find the method helpful. However, we cannot recommend it as a general method of determining the number of major components to retain. Its tendency to retain both minor and trivial components might reflect the basic logic of the test. Only error should be expected to meet the requirement of equal eigenvalues. However, most researchers would not find minor or trivial components to be consistent with their implicit definition of a factor or component that is worthy of retention. Therefore, although the test works correctly, it may not be appropriate for general applications.

We did not investigate the maximum likelihood test that assumes the factor analysis model. The maximum likelihood test is based on a logic identical to that of the Bartlett test. Empirical investigations have found the same pattern of results with the likelihood test as we have reported with the Bartlett test (Hakstian et al., 1982; M. B. Richman, personal communication, October 14, 1983). Again, the problem may be an inappropriate application of the test, rather than the performance of the test per se.

The scree test had moderate overall reliability when the mean of two trained raters was used. The correlation of the mean of those raters' decisions with an expert judge indicated fair overall agreement. Reports of rater reliability on the scree test have ranged from very good (Cattell & Jaspers, 1967) to quite poor (Crawford & Koopman, 1979). This range may reflect either the training or the task complexity across research projects. The raters in this study showed greater agreement at higher than at lower component saturation levels. Perhaps more importantly, the interrater reliability of the scree test had a fairly wide range across levels of complexity. The moderate reliability of the scree test is very problematic for the applied researcher. Unreliability at this point in the analysis may well expose a study to otherwise avoidable experimenter bias. In any case, applied researchers and reviewers should note that reliability questions always arise in any use of the scree test.

In general, the scree test was more accurate and less variable than either the K1 method or the Bartlett test. It was more accurate and less variable at the higher level of component saturation. Larger sample sizes also improved its accuracy when $P$ equaled 72 and component saturation was .50. Sample size had no appreciable effect at other levels of $P$ or component saturation. This effect of larger sample size is consistent with those reported elsewhere (Cliff & Pennell, 1967; Linn, 1968; Zwick & Velicer, 1982). The accuracy of the scree test was not affected by an increase in the number of variables examined. An increase in the average number of variables per component did not affect its accuracy. In an earlier study (Zwick & Velicer, 1982) with noncomplex matrices, the scree test performed better than MAP when the major components contained six or fewer variables and the saturation was low. This trend can be observed again under more complex conditions.

The scree test tended to overestimate rather than to underestimate when it deviated from the criterion value. As reflected in Table 5, it was accurate about 57% of the time. When it was in error, 90% of the errors were overestimations. The scree test appeared to be most variable at the low level of component saturation or when unique or complex variables were present. Nevertheless, even given its variability and tendency toward overestimation, the scree test seldom led to the retention of more than two components over the criterion value. Many experienced investigators routinely examine one or two components above and below the estimate given by their favorite decision method. This practice, coupled with good judgment concerning interpretability, may often result in appropriate solutions when the scree test is used. This optimistic assertion notwithstanding, the ever-present question of rater reliability, the tendency to overestimate, and the apparent increased variability in the common case of unique or complex variables all argue against the exclusive use of the scree test. Given these drawbacks and the availability of other clearly superior methods, we can no longer recommend the scree test as the method of choice for determining the number of components in PCA.

In general, the MAP method was more often accurate and less variable than the K1, Bartlett, or scree methods. It showed an overall tendency to underestimate the criterion. The MAP method was most accurate at the higher level of component saturation or when the average number of variables per component was large. Its performance was not markedly influenced by sample size within the range examined in this study. The MAP method was quite accurate under many conditions and, when accurate, showed little variability. In cases representing both a low level of saturation and a low number of variables per component, the MAP method consistently underestimated the number of major components. The combination of these influences appears to result in a component that accounts for less variance than those components containing only a unique variable. The MAP method will not retain a unique variable component and therefore, fails to pick up less well-identified major components.

The MAP method gave results within ±1 of the criterion between 78% (component saturation = .50) and 100% (component saturation = .80) of the time. When it was in error, the MAP method tended to underestimate. Approximately 90% of the MAP errors were underestimations.

The MAP method provides an unequivocal stopping point. It is tied to the concept of parsimony by directly rejecting components identified by only one variable. It is quite accurate when component saturation is high or when there is an average of eight or more variables per component. Researchers wishing to ignore relatively small major components should use MAP as a primary method of determining the number of components to retain.

The PA method was consistently accurate. It was typically the most accurate method at each level of complexity examined. The average deviation of PA from the criterion did not exceed 0.30 under any condition examined. The difference scores of PA showed less variability than those of any other rule. Increases in sample size, component saturations, and the number of variables per component improved PA's performance. We also found some evidence of overestimation due to minor components.

Overall, the PA method was the most frequently accurate method examined. It gave results within ±1 of the criterion for between 97% (component saturation = .50) and 100% (component saturation = .80) of the cases examined. When PA was in error, it showed a slight tendency toward overestimation. Approximately 65% of the PA method's errors were overestimations.

The accuracy of the PA method in this study is consistent with the CFA results reported by Humphreys and Montanelli (1975).

A major drawback in the applied use of PA is the necessity of generating a large set of random correlation matrices at the particular combination of $P$ and sample sign under consideration. The prediction equation developed for principal axis factor analysis (Montanelli & Humphreys, 1976) is not appropriate for PCA. However, Bobko and Schemmer (1984) recently proposed a prediction equation appropriate for component analysis. Further evaluation of this and other equations of this type over a wide range of conditions is needed.

In summarizing the results of the present study it is useful to postulate a further division of the MJCs. We label those components that involve a small number of variables and low saturation poorly defined components (PDCs). We label components with either a large number of variables or high saturation as well-defined components (WDFs). Poorly defined components do not possess any "marker" variables, that is, variables with high loadings on that component. Investigators typically use such marker variables as defining variables in interpreting the component. Guadagnoli and Velicer (in press) found that PDCs were unlikely to accurately replicate even in fairly large samples. The combination of the two issues, difficult interpretability and questionable replicability, make the retention of these components problematic.

The two most accurate methods, MAP and PA, provided divergent results primarily when PDCs were present. An a priori decision about whether or not to attempt to extract and retain such components may dictate whether MAP or PA is the method of choice. Lacking such a decision, a researcher may want to examine all solutions in a set bracketed by the MAP and PA estimates. Rotational criteria and interpretability may be the basis for a final decision.

Previous studies have examined subsets of these rules under some of the conditions examined here. In those areas in which the simulated situations were similar, the results of Linn (1968), Humphreys and Montanelli (1975), Cattell and Vogelman (1977), Hakstian et al. (1982), Zwick and Velicer (1982), and Reddon (in press) were confirmed and expanded.

Within the limitations imposed by the simulation approach, the results of this study, paired with previously reported work, permit some conclusions concerning methods of determining the number of components in real data sets. There is no evidence supporting the continued use of K1 or the Bartlett test as exclusive, primary methods to determine the number of major components to retain. These methods should not be used. The scree procedure has been reported to be relatively accurate. This study is consistent with those reports but indicates that the method is too variable and too likely to overestimate to use as the sole decision method. However, the scree test may still be useful for initial estimates or as a complementary method used in conjunction with PA or MAP. The MAP method was generally quite accurate and consistent when the component saturation was high or the component was defined by more than six variables. The MAP did not retain PDCs. The PA method was consistently accurate; it retained PDCs and showed a slight tendency to overestimate. The general application of the PA method is problematic at this time because programs needed for its application are not widely available. Assuming that this problem can be overcome, either PA or MAP is the method of choice, with many situations arising in which both should be used.

## References


Afifi, A. A., & Clark, V. (1984). *Computer-aided multivariate analysis.* Belmont, CA: Lifetime Learning Publications.

Anderson, A. D., Acito, F., & Lee, H. (1982). A simulation study of three methods for determining the number of image components. *Multivariate Behavioral Research, 17,* 493–502.

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 5,* 111–150.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3,* 77–85.

Bartlett, M. S. (1951). A further note on tests of significance in factor analysis. *British Journal of Psychology, 4,* 1–2.

Bobko, P., & Schemmer, F. M. (1984). Eigenvalue shrinkage in principal component-based factor analysis. *Applied Psychological Measurement, 8,* 439–451.

Browne, M. W. (1968). A note on lower bounds for the number of common factors. *Psychometrika, 33,* 233.

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

Cattell, R. B., & Jaspers, J. (1967). A general plasmode for factor analytic exercises and research. *Multivariate Behavioral Research Monographs, 3,* 1–212.

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.

Chatfield, C., & Collins, A. J. (1980). *Introduction to multivariate analysis.* London: Chapman & Hall.

Cliff, N. (1970). The relation between sample size and population characteristic vectors. *Psychometrika, 35,* 163–178.

Cliff, N., & Hamburger, C. D. (1967). The study of sampling errors in factor analysis by means of artificial experiments. *Psychological Bulletin, 68,* 430–445.

Cliff, N., & Pennell, R. (1967). The influence of communality, factor strength and loading size on the sample characteristics of factor loadings. *Psychometrika, 32,* 309.

Comrey, A. L. (1978). Common methodological problems in factor analytic studies. *Journal of Consulting and Clinical Psychology, 46,* 648–659.

Crawford, C. B. (1975). Determining the number of interpretable factors. *Psychological Bulletin, 82,* 226–237.

Crawford, C. B., & Koopman, P. (1979). Note: Inter-rater reliability of scree test and mean square ratio test of number of factors. *Perceptual and Motor Skills, 49,* 223–226.

Everett, J. E. (1983). Factor comparability as a means of determining the number of factors and their rotation. *Multivariate Behavioral Research, 18,* 197–218.

Everitt, B. S., & Dunn, G. (1983). *Advanced methods of data exploration and modeling.* London: Heinemann.

Glass, G. V., & Taylor, P. A. (1966). Factor analytic methodology. *Review of Educational Research, 36,* 566–587.

Gorsuch, R. L. (1973). Using Bartlett's significance test to determine the number of factors to extract. *Educational and Psychological Measurement, 33,* 361–364.

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

Green, S. B. (1983). Identifiability of spurious factors using linear factor analysis with binary items. *Applied Psychological Measurement, 7,* 139–147.

Guadagnoli, E., & Velicer, W. F. (in press). Stability of component patterns: A simulation study. *Psychological Bulletin.*

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19,* 149–162.

Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number of factors rules with simulated data. *Multivariate Behavioral Research, 17,* 193–219.

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

Horn, J. L. (1969). On the internal consistency reliability of factors. *Multivariate Behavioral Research, 4,* 115–125.

Horn, J. L., & Engstrom, R. (1979). Cattell's scree test in relation to Bartlett's chi-square test and other observations on the number of factors problem. *Multivariate Behavioral Research, 14,* 283–300.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24,* 417–441, 498–520.

Humphreys, L. G. (1964). Number of cases and number of factors: An example where *N* is very large. *Educational and Psychological Measurement, 24,* 457.

Humphreys, L. G., & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10,* 193–206.

Jackson, B. B. (1983). *Multivariate data analysis.* Homewood, IL: Irwin.

Jackson, D. N., & Chan, D. W. (1980). Maximum-likelihood estimation in common factor analysis: A cautionary note. *Psychological Bulletin, 88,* 502–508.

Johnson, R. A., & Wichern, D. W. (1982). *Applied multivariate statistical analysis.* Englewood Cliffs, NJ: Prentice-Hall.

Joreskog, K. G. (1962). On the statistical treatment of residuals in factor analysis. *Psychometrika, 27,* 335–354.

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

Kaiser, H. F. (1970). A second-generation little jiffy. *Psychometrika, 35,* 401–415.

Kaiser, H. F., & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika, 30,* 1–44.

Lawley, D. M. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh, 60,* 64–82.

Lawley, D. M. (1941). Further investigation in factor estimation. *Proceedings of the Royal Society of Edinburgh, 61,* 176–185.

Lee, H. B., & Comrey, A. L. (1979). Distortions in a commonly used factor analytic procedure. *Multivariate Behavioral Research, 14,* 301–321.

Linn, R. L. (1968). A Monte Carlo approach to the number of factors problem. *Psychometrika, 33,* 37–71.

Lunneborg, C. E., & Abbott, R. D. (1983). *Elementary multivariate analysis for the behavioral sciences.* New York: North-Holland.

Marascuilo, L. A., & Levin, J. R. (1983). *Multivariate statistics in the social sciences: A researcher's guide.* Monterey, CA: Brooks/Cole.

Montanelli, R. G., Jr. (1975). A computer program to generate sample correlation and covariance matrices. *Educational and Psychological Measurement, 35,* 195–197.

Montanelli, R. G., Jr., & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika, 41,* 341–348.

Morrison, D. F. (1976). *Multivariate statistical methods* (2nd ed.) New York: McGraw-Hill.

Mote, T. A. (1970). An artifact of the rotation of too few factors: Study orientation vs. trait anxiety. *Revista Interamericana de Psicologia, 37,* 61–91.

Pruzek, R. M., & Rabinowitz, S. N. (1981). A simple method for exploratory structural analysis. *American Educational Research Journal, 18,* 173–189.

Reddon, J. R. (1985). MAPF and MAPS: Subroutines for the number of principal components. *Applied Psychological Measurement, 9,* 97.

Reddon, J. R. (in press). Monte Carlo type one error rates for Velicer's partial correlation test for the number of principal components. *Criminometrica.*

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

Schonemann, P. H., & Wang, M. (1972). Some new results on factor indeterminacy. *Psychometrika, 37,* 61–91.

Steiger, J. H., & Schonemann, P. H. (1979). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis in the behavioral sciences* (pp. 136–178). San Francisco: Jossey-Bass.

Tucker, L. B., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34,* 421.

Veldman, D. J. (1979). Simple structure and the number of factors problem. *Multivariate Behavioral Research, 14,* 191–200.

Velicer, W. F. (1974). A comparison of the stability of factor analysis, principal component analysis and rescaled image analysis. *Educational and Psychological Measurement, 34,* 563–572.

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

Velicer, W. F. (1976b). The relation between factor score estimates, image scores and principal component scores. *Educational and Psychological Measurement, 36,* 149–159.

Velicer, W. F. (1977). An empirical comparison of the similarity of principal component, image, and factor patterns. *Multivariate Behavioral Research, 12,* 3–22.

Velicer, W. F., & Jackson, D. N. (1986). *Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure.* Unpublished manuscript.

Velicer, W. F., Peacock, A. C., & Jackson, D. N. (1982). A comparison of component and factor patterns: A Monte Carlo approach. *Multivariate Behavioral Research, 17,* 371–388.

Yeomans, K. A., & Golder, P. A. (1982). The Guttman-Kaiser criterion as a predictor of the number of common factors. *Statistician, 31,* 221–229.

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.

Received October 12, 1984
Revision received October 26, 1985