Psychological Assessment
1995, Vol. 7, No. 3, 286–299

1040-3590/95/$3.00

Factor Analysis in the Development and Refinement of Clinical Assessment Instruments

Frank J. Floyd
University of North Carolina at Chapel Hill

Keith F. Widaman
University of California, Riverside

The goals of both exploratory and confirmatory factor analysis are described and procedural guidelines for each approach are summarized, emphasizing the use of factor analysis in developing and refining clinical measures. For exploratory factor analysis, a rationale is presented for selecting between principal components analysis and common factor analysis depending on whether the research goal involves either identification of latent constructs or data reduction. Confirmatory factor analysis using structural equation modeling is described for use in validating the dimensional structure of a measure. Additionally, the uses of confirmatory factor analysis for assessing the invariance of measures across samples and for evaluating multitrait–multimethod data are also briefly described. Suggestions are offered for handling common problems with item-level data, and examples illustrating potential difficulties with confirming dimensional structures from initial exploratory analyses are reviewed.

Factor analysis is one of the most commonly used procedures in the development and evaluation of psychological measures. Factor analysis is particularly useful with multiitem inventories designed to measure personality, psychopathology, attitudes, behavioral styles, cognitive schema, and other multifaceted constructs of interest to clinical psychologists. Accordingly, every issue of *Psychological Assessment* since the inception of the journal in 1989 has included studies using factor analytic procedures.

Despite the popularity of factor analysis, both the complexity and the flexibility of factor analytic procedures have created considerable ambiguity about appropriate research practices. In some cases, differences in analytic procedures have produced inconsistent findings across studies and thus controversy over substantive issues. Furthermore, the recent increase in the use of confirmatory factor analysis creates the need for more information about the application of these procedures and the interpretation of the findings when they are discrepant from the results of exploratory analyses.

This article summarizes guidelines for using factor analysis in the development and refinement of measuring instruments and for reporting the results of analyses. The article also reviews recent examples of exploratory and confirmatory factor analytic studies and uses these examples to highlight some of the interpretive controversies that can emerge. More important, the examples suggest guidelines for choosing among factor analytic approaches depending on the research goals, the nature of the constructs assessed, and the type of measuring instrument examined.

Factor Analysis: Basic Uses and Methods

When discussing factor analysis, it is worthwhile to distinguish between the uses of factor analysis and the methods or techniques employed. The issue of the uses of factor analysis concerns the goals of the analysis, whereas the issue of methods or techniques involves the particular statistical and mathematical procedures used to attain those goals.

Uses of Factor Analysis

Exploratory approaches. Factor analysis has two general exploratory uses in the analysis of measures to assess psychological constructs; these two uses are aligned with the goals of explanation and data reduction.

The first use, subserving explanation, is to identify the underlying dimensions of a domain of functioning, as assessed by a particular measuring instrument. In this case, an instrument designed to assess a domain of functioning is factor-analyzed to identify separable dimensions, representing theoretical constructs, within the domain. The factor-analytically derived dimensions then serve as subscales for the instrument. The procedure is exploratory because, presumably, the investigator has no firm a priori expectations based on theory or prior research about the composition of the subscales, and thus the analysis is used to discover the latent variables that underlie the scale.

To attain this end, common factor analysis uses the matrix of correlations or covariances among measured variables, either items or subscales, to identify a set of more general latent variables, or factors, that explain the covariances among the mea-

Frank J. Floyd, Department of Psychology, University of North Carolina at Chapel Hill; Keith F. Widaman, Department of Psychology, University of California, Riverside.

Preparation of this article was supported by National Institute of Child Health and Human Development Grants R01 HD24205, K04 HD01023, R01 HD22953, and R01 HD21056. The authors contributed equally to the preparation of this article and are listed in alphabetical order.

Correspondence concerning this article should be addressed to Frank J. Floyd, Department of Psychology, Campus Box No. 3270, University of North Carolina, Chapel Hill, North Carolina 27599-3270. Electronic mail may be sent via Internet to fjfloyd@email.unc.edu.

sured variables. In theory, these latent variables are the underlying causes of the measured variables. The analysis produces factor loadings, which are the regression weights for predicting the measured variables from the latent variables, and also produces correlations among the latent variables, if these are estimated. Factor analysis also partitions the variance of each measured variable into two parts: (a) *common variance,* or variance associated with the latent variables, which is estimated based on variance shared with other measured variables in the analysis, and (b) *unique variance,* which is a combination of reliable variance that is specific to the given measured variable and random error variance in the variable. Thus, common factor analysis can provide valuable insights into the multivariate structure of a measuring instrument, isolating the theoretical constructs whose effects are reflected in responses on the instrument. For example, exploratory factor analysis was used to identify subcomponents of a global measure of marital satisfaction, the Dyadic Adjustment Scale (Spanier, 1976), and to identify the Freedom From Distractibility factor on the Wechsler Intelligence Scale for Children–Revised (Kaufman, 1975). Furthermore, if a researcher wishes to develop scales of items with very clean loading patterns, the results of a factor analysis can be used to reduce the number of items on an inventory by eliminating items that fail to load on any factor or that load at approximately equal levels on two or more factors.

The second, and related, use of exploratory factor analysis is for data reduction, in which a set of measured variables is to be combined into summary indices. The goal is to discover optimal weightings of the measured variables so that a large set of related variables can be reduced to a smaller set of general summary scores that have maximal variability and reliability. A recent example of using factor analysis for the purpose of data reduction is research on self-report coping scales to identify reliable categories of coping practices from reports of specific coping behaviors (Folkman & Lazarus, 1988; Parker, Endler, & Bagby, 1993).

The goal of data reduction is typically achieved by the use of principal components analysis as opposed to common factor analysis. The procedural differences for the approaches are discussed in detail in a later section. It is also important to distinguish these approaches in terms of a core theoretical assumption. In common factor analysis, the factors are estimated to explain the covariances among the observed variables, and the factors are viewed as the causes of the observed variables. In contrast, in principal components analysis (or, more simply, component analysis), the components are estimated to represent the variances of the observed variables in as economical a fashion as possible (i.e., in as small a number of dimensions as possible), and no latent variables underlying the observed variables need to be invoked. Instead, the principal components are optimally weighted sums of the observed variables so, in a sense, the observed variables are the causes of the composite variables.

Confirmatory approaches. In contrast to pure exploration, factor analysis is frequently used to confirm a priori hypotheses of some sort. Beginning with the pioneering, large-scale empirical studies of abilities using factor analysis (e.g., Thurstone, 1938; Thurstone & Thurstone, 1941), researchers have often been able to generate hypotheses regarding the factors that should be represented in a given domain of inquiry. These hypotheses may be based on theory or on results from previous empirical studies. The ensuing factor analysis typically reveals some of the expected constructs, perhaps along with additional, unexpected factors. If these additional factors can be interpreted meaningfully and shown to be replicable, reliable phenomena, their identification can add substantially both to understanding the research domain and to interpreting scores on the measure. This approach is useful to investigators who suspect that a measure designed to assesses one domain or construct actually contains a meaningful dimensional structure and that assessing the separate dimensions would lead to a better understanding of the phenomena in a given psychological domain.

As a confirmatory procedure, factor analysis is primarily a method for assessing the construct validity of measures, not a means for data reduction. Construct validity is supported if the factor structure of the scale is consistent with the constructs the instrument purports to measure. If the factor analysis fails to detect underlying constructs that explain sufficient variance in the measured variables or if the constructs detected are inconsistent with expectations, the construct validity of the scale is compromised.

Techniques of Factor Analysis

Exploratory procedures. Approaching a factor analysis using exploratory procedures involves a series of steps that includes estimating, or extracting, factors; deciding how many factors to retain; and rotating factors to an interpretable orientation. The particular analytic techniques used in a factor analysis appear to be the aspects of a factor analysis that are least understood by practicing scientists, and thus are often reported in insufficient detail in methods sections of journal articles. In the remainder of this article we will discuss various issues related to the use of these methods so that investigators can make more informed choices among options. In addition, we hope to show that the particular analytic techniques used require justification, that there are standards for choosing among these techniques, and that the techniques used in any study deserve to be reported in sufficient detail to enable a full interpretation of the results by interested readers.

Confirmatory procedures. Confirmation of hypothesized factor structures is most adequately established with confirmatory factor analytic techniques; a special case of structural equation modeling approaches (e.g., Bentler, 1989; Jöreskog & Sörbom, 1989). In confirmatory factor analysis, a factor structure is explicitly hypothesized and is tested for its fit with the observed covariance structure of the measured variables. The approach also allows for testing the relative fit of competing factor models. Furthermore, although this approach is primarily useful for confirmation of theories, as with other applications of structural equation modeling, the procedures provide guidelines for "model trimming," or model modification, that can suggest alterations in proposed factor structures. Thus, confirmatory procedures can be used to revise and refine instruments and their factorial structure.

Some extensions of confirmatory factor analysis pertinent to the development and evaluation of measures include investigating measurement invariance across groups or samples (Reise, Widaman, & Pugh, 1993) and analyzing multitrait–multimethod data with confirmatory factor analysis to test the convergent and discriminant validity of trait factors (Cole, 1987; Kenny & Kashy, 1992; Widaman, 1985). A complete description of these applications is beyond the scope of this article, though the procedures will be outlined briefly.

The analytic procedures implied currently by the term *confirmatory factor analysis* were developed largely within the past 20 years (Jöreskog, 1969, 1971; and in early precursors of the LISREL program). Before that time, researchers used exploratory factor analytic techniques for both blatantly exploratory and consciously confirmatory ends. Careful reading of various programs of research will reveal that investigators began studies of a domain with certain hypotheses regarding the factors that would emerge and then revised their theories and their notions regarding the factors in studies that sought, often quite successfully, to reconfirm factor structures in additional samples of participants. Thus, exploratory factor analytic techniques have been used for more than 60 years to achieve both exploratory and confirmatory analytic goals. However, at present, the newer, confirmatory techniques can accomplish the same ends of replicating factor structures and confirming theory and are quite flexible, enabling the testing of many crucial hypotheses that were approachable only in indirect ways with exploratory analytic techniques.

Basic Assumptions and Procedural Guidelines

Basic procedures for conducting factor analytic studies and reporting the findings are given in detail in introductory texts such as Gorsuch (1983), Comrey and Lee (1992), and Harman (1976) for exploratory factor analysis and Hayduk (1987) and Bollen (1989) for confirmatory factor analysis. Also, conceptual issues and procedures specific to measurement development and evaluation in clinical psychology are described by Comrey (1988) and Streiner (1994) regarding exploratory factor analysis and Cole (1987) and Hoyle and Smith (1994) regarding confirmatory factor analysis. Several common concerns and recommendations for practice are outlined here to promote greater accuracy in using procedures and interpreting findings and greater consistency and completeness in reporting results.

Data Type and Quality

The most basic requirement for optimal uses of factor analysis is high-quality data that are measured on interval or quasi-interval scales. There is no substitute for good data. The epithet "garbage in, garbage out" has been directed toward factor analytic investigations perhaps more often than toward studies using other multivariate techniques, and not without some reason. However, we wish to be clear on one point: the "garbage in, garbage out" barb is rightly targeted toward improper uses and, hence, the users of factor analysis, not toward the method of analysis itself. Computer implementations of factor analytic techniques are widespread, and almost any factor structure for a set of data can be interpreted in some fashion by the dedicated scientist. Thus, researchers have been able to feed disparate sets of variables into factor analyses, hoping that the analysis will turn out silk purses from the sow's ears in their data.

The most direct way to ensure the quality of data is through careful item selection and item analyses (see Haynes, Richard, & Kubany, 1995; Smith & McCarthy, 1995). The domain of human abilities served as the first testing ground for factor analysis, and in this domain, the measured variables subjected to analysis have typically been multiitem tests that have well-established psychometric properties (e.g., reliability). In contrast, factor analyses of clinical assessment instruments are often based on item-level data. Because items tend to be less reliable than multiple-item scales, the psychometric properties of items to be factor-analyzed is of great concern. Pilot-testing of items should be performed to ensure that items designed to measure a common construct correlate moderately with one another and with the total scale score. If an item does not correlate at least moderately (e.g., r = .20 or greater) with other items for the construct, then the item will likely perform poorly in a factor analysis.

A second issue regarding measured variables concerns the scale on which scores fall. As noted earlier, scores for measured variables should fall on interval or quasi-interval scales. Clinical assessment instruments frequently consist of Likert-scaled items, with ratings falling on 1- to 5-point or 1- to 7-point scales. Such scales would generally be considered interval or quasi-interval, and factor analyses are frequently performed successfully on such data.

At least two deviations from interval scales are commonly observed in clinical assessment instruments. One of these deviations is the use of dichotomous variables, or variables scored on a 0–1, or yes–no, basis. Problems in factor-analyzing dichotomous variables were recognized in the 1940s, with the identification of "difficulty factors" that were due only to variation in endorsement rates across items and not to the underlying construct. Although dichotomous items can be factor-analyzed using standard techniques, the results may be biased. Adequate analytic solutions to these problems of bias were developed in the 1970s and 1980s, but special programs must be used that implement these techniques. Interested readers are referred to programs such as TESTFACT (Wilson, Wood, & Gibbons, 1991) and NOVAX (Waller, 1994), among others, to arrive at unbiased estimates of factors from dichotomous data. A second way to deal with dichotomous items is to compute sums of two or more similar items. Such item sums, termed *parcels* (Kishton & Widaman, 1994), have scores that fall on greater than dichotomous scales and are thus more immune from the problem of isolating difficulty factors.

The second deviation from interval scales consists of nominal items with three or more categories (nominal items with only two categories reduce to dichotomous items and can be handled as discussed earlier). Items of this sort could be rescored using standard methods of coding, such as dummy or effects coding (see Cohen & Cohen, 1983). However, most methods of coding introduce positive or negative correlations among the coded variables that could distort the factor structure of measures. As a result, inclusion of such variables is fraught with problems, and they should be included only with great caution.

Distributions

Ideally, factor analysis is applied to data that are distributed in multivariate normal fashion. Although difficult to test, multivariate normality is a reasonable assumption only if each variable in an analysis is distributed approximately univariate normal. Multivariate normality is a strict assumption only for certain methods of parameter estimation, such as maximum likelihood. Because principal axes (or least squares), which is by far the most commonly used approach for exploratory factor analysis, does not require the assumption of multivariate normality, the stipulation that data be distributed multivariate normal may strike many investigators as too stringent. Nevertheless, all methods of factor analysis are more likely to yield clearer, more replicable factor patterns if the data conform to multivariate normality. Having stated this, in practice, both exploratory and confirmatory factor analysis appear to be relatively robust against violations of normality (Gorsuch, 1983). However, a Monte Carlo study of confirmatory factor analysis by Hu, Bentler, and Kano (1992) showed that it is difficult to obtain acceptable confirmatory solutions when nonnormal distributions occur along with other violations of assumptions such as small sample size and the nonindependence of variates and error.

A common problem for research with clinical samples is the gross deviations from normal distributions that occur on variables measuring rare events, such as pathological or unusual behaviors. Such data are often the most direct and important indicators of key diagnostic behaviors, yet are likely to produce unstable or uninterpretable factor structures (Waller, 1989). Combining several rare events into a single variable may improve the distribution of scores while preserving the importance of the rare events for producing high scores on the variable. For example, behavioral counts or ratings might include several rare behaviors such as stereotypy, echolalia, hypersensitivity to touch, and so on. These rare events could be combined into a single index in which higher scores reflect the occurrence of more of these unusual symptoms. The distribution of these scores would be more suitable for factor analysis than the scores for the individual rare behaviors. Of course, the improved distribution of the sum of these items is bought at a price: only a single variable reflecting the rare behaviors is now available for analysis.

Another issue that arises when considering distributions of scores on variables is the analysis of ipsative data. Certain investigators (e.g., Goldberg & Digman, 1994) have reportedly used factor analytic techniques with ipsative data and obtained interpretable solutions. These researchers recommend ipsatizing data prior to analysis, at least in the personality domain. However, Dunlap and Cornwell (1994) applied a variety of factor analytic procedures with ipsative data and recommended that factor analysis not be performed on ipsative data because of the dependencies in the data and the resulting artifactual influences on the factor pattern produced. The divergence in recommendations by Goldberg and Digman (1994) and Dunlap and Cornwell (1994) may be due to at least two issues. First, the number of variables analyzed differs dramatically: Goldberg and Digman often based analyses on at least 100 ipsative measured variables, whereas Dunlap and Cornwell analyzed matrices with from 4 to 6 measured variables. As the number of ipsatized measured variables increases, the dependencies between any two of the measured variables decrease. Thus, the distortions noted by Dunlap and Cornwell in small matrices may well have been considerably reduced in the rather larger matrices analyzed by Goldberg and Digman. Second, the two sets of researchers used different analytic techniques: Goldberg and Digman used principal components analysis, whereas Dunlap and Cornwell used both principal components analysis and variants of common factor analysis. One major drawback to the use of ipsative variables is that certain types of common factor analysis cannot be used because of the dependencies in the data. Given these problems, we feel that it is best, at present, to warn against the use of ipsative variables in factor analysis.

Another caution is in order regarding confirmatory factor analysis. In contrast with recommendations stated earlier about Likert-scaled items, experts in confirmatory factor analysis have often argued that variables obtained using Likert rating scales should not be treated as continuous measures, and that many of these scales are not normally distributed. Using standard confirmatory factor methods on such data might well lead to fit statistics that may inaccurately represent the degree of true model fit. Although there is a lack of unanimity among experts on this issue, several options may be taken to investigate the effect of this issue for the fit of a model to a particular set of data. For example, Jöreskog and Sörbom (1989) recommended using polychoric and polyserial correlations to represent the associations of these variables with each other and with other continuous measures. Jöreskog and Sörbom also suggested using the unweighted least squares procedure for obtaining parameter estimates in confirmatory analyses when the distributions are skewed. If these options lead to rather different levels of fit than the standard methods of confirmatory factor analysis, then the alternative methods may be far more appropriate. Others (e.g., Browne, 1984) have developed asymptotically distribution-free methods of estimation with data departing from multivariate normality, methods that appear to work well with very large samples. However, Hu et al. (1992) suggested that this method may fail completely with samples of 250 or fewer, which are more common in clinical studies. Hu et al. described an alternative test statistic, the Satorra–Bentler scaled statistic, which is less likely to reject true models under conditions in which normality assumptions are violated. Finally, Múthen (1988) developed the LISCOMP program for fitting confirmatory factor models using indicators that had noninterval, ordered categorical rating scales. Although many of these sources represent advanced issues in structural modeling, using confirmatory factor analytic techniques with rather nonnormal data may require such special efforts to find models that are appropriate for the data.

Sample Size

Until recently, a general rule of thumb regarding sample size for principal components and common factor analysis has been "the more participants, the better." Explicit guidelines for sample size have always been in flux, passed down from generation to generation of factor analysts in an oral tradition. This tradition held that a subjects-to-variables ratio of 4:1 or 5:1 was sufficient for exploratory factor analysis. More recently, sources on factor analysis have offered variants of the older rules. For example, Gorsuch (1983) stated that there should be at least 5 participants per variable and that a sample size of at least 200 is preferred. Streiner (1994) recommended that adequate solutions will be obtained with 5 participants per variable as long as there are 100 participants in the sample, and with 10 participants per variable when there are less than 100.

Guadagnoli and Velicer (1988) challenged such rules and argued that no sound theoretical or empirical basis exists for across-the-board participant-to-variable ratio recommendations. Instead, their Monte Carlo study suggested that variable saturation with the factors, indicated by the size of the factor loadings, along with the total sample size and the number of indicators per factor were important in determining the stability of factor solutions. Most notably, with factor loadings of .80, solutions were highly stable across replicated samples regardless of the number of indicators, even with as few as 50 participants. When factor loadings were in the .60 range, stable solutions were obtained with sample sizes greater than 150, or with still smaller samples when each component contained at least four variables loading at .60. In general, larger samples of 300–400 were needed when the factor loadings were only .40. However, when at least 10 variables loaded at .40 on each factor, samples of 150 produced accurate solutions. Thus, instead of focusing solely on the participant-to-variable ratio, Guadagnoli and Velicer recommended careful attention to selecting variables that are heavily saturated with the factors, or when factor loadings are weak or unknown, using many measures to represent each construct.

The 5–10 participants per variable guideline is commonly used in confirmatory factor analysis as well, although Jöreskog and Sörbom (1989) suggested that it is best to have 10 participants per parameter estimated. Because the number of parameters estimated in a confirmatory analysis can increase greatly as more variables are added and the model becomes more complex, this recommendation is a sound reason for keeping models simple when using confirmatory factor analysis. There is little justification for using a "the more the better" rule in confirmatory factor analysis. In fact, with very large samples, it would be more useful to subdivide the sample for the purpose of replicating the factor solution rather than conducting one analysis with the entire sample.

Procedures for Exploratory Factor Analysis

Extracting Factors or Components

Approaches to extracting factors. Exploratory factor analysis consists of two general approaches, principal components analysis and common factor analysis (or principal factor analysis). As noted in a previous section, principal components analysis should be used primarily for data reduction, whereas common factor analysis should be used to understand the relations among a set of measured variables in terms of underlying latent variables.

The primary difference between the two approaches is their assumptions about the communalities of the measured variables, or the variance in measures to be represented by the procedure. The *communality* of a variable is the variance that variable shares with the latent variables underlying the set of observed measures. Principal components analysis analyzes the matrix of correlations among measured variables with 1.0s on the main diagonal; as such, component analysis attempts to represent all of the variance of the observed variables. In contrast, common factor analysis analyzes the matrix of correlations among measured variables with communality estimates on the main diagonal. By analyzing this reduced matrix, common factor analysis attempts to represent only the common variance of each variable. This common variance is variance shared with other observed variables as a result of the dependence of the measured variables on the latent variables.

Both principal components analysis and common factor analysis typically use the principal axes (or unweighted least squares) method of estimating factors from the correlation matrix of measured variables to extract components that account for the maximum possible variance in the observed variables. The first component or factor is extracted so that it produces the highest possible squared correlations between the variables and the component or factor and thus maximizes the amount of variance accounted for. Subsequent components are extracted from successive residual matrices after all variance accounted for by previous components is removed. In most cases, the first component accounts for considerably more variance than all subsequent components. The method continues to extract as many components as variables in the analysis until either all of the variance (components analysis) or all of the common variance (common factor analysis) of the measured variable is accounted for. However, usually a truncated solution is used, in which only components that account for meaningful amounts of variance are retained.

It is important to note that, with a set of highly related variables, all components or factors subsequent to the first account for relationships among variables that are independent of the common variance shared by all variables, and thus later components or factors should always be interpreted in light of this fact (Gorsuch, 1983). Questionnaire measures of satisfaction, well-being, or general attitudes (e.g., authoritarianism) frequently fit this situation, particularly when a criterion of high item-total correlations or internal consistency had been used earlier for retaining items on the scale. These scales produce a first component or factor, composed of most of the items, which accounts for the majority of variance among the items, and other weaker components or factors, composed of fewer items, which account for much less variance. For example, component analysis of the Marital Satisfaction Questionnaire for Older Persons (Haynes et al., 1992) produced a primary communication/companionship component composed of 16 items, which accounted for 58% of the variance, and a much weaker second component accounting for 6% of the variance before rotation, with primary loadings by three items assessing satisfaction with sex and affection in the relationship. Thus, the shared variance of these sex/affection items is independent of the variance they have in common with the other satisfaction items. Because sex/affection was considered to be an important, separable component of satisfaction with married life, the component was retained in the scale. However, if the component had been judged to be an extraneous feature, one or more of the items could have been deleted from the scale to remove this extraneous source of variance. Note that although the rotated solution usually shows a more even distribution of variance accounted for among components, nevertheless, as discussed later, the usefulness of the minor, secondary components consisting of few items needs to be demonstrated in studies of predictive validity.

Because common factor analysis uses initial estimates of communalities that more accurately reflect the actual communalities of the variables, common factor analysis produces more accurate final estimates of communality than does principal components analysis. In common factor analysis, estimates of communality are derived from the correlations of each variable with the remaining variables. The three most common methods for estimating communalities are (a) the highest correlation of a variable with any of the remaining variables, (b) the squared multiple correlation (SMC) of a variable with the remaining variables in the analysis, and (c) iterated estimates. The choice between the highest correlation and the SMC as the communality estimate for a variable is simply that, a choice. Perhaps the SMC has a slight edge: the SMC is, theoretically, a lower bound estimate of communality, and SMCs must be used to perform the parallel analysis test for the number of common factors, to be described later. The final option, the use of iterated estimates, requires a decision on the number of factors and some initial estimates of communality, such as SMCs. Then, an iterative sequence ensues in which factors are extracted, communalities are computed and placed back on the main diagonal, and factors are once again extracted. This sequence continues until the communality estimates stabilize. Iterative estimates have the advantage of fitting the data best, and iterative estimates tend to stabilize at the same value regardless of starting value (Widaman & Herringer, 1985), although SMCs and the highest correlation are the most reasonable starting values.

Principal components versus common factors. For many years, the accepted view on the choice between component analysis and common factor analysis was that it was a choice, with little effect on results. Moreover, researchers routinely interpret the results of component analysis as approximations to common factor analysis, and experts have often supported this practice, at least in part. For example, Gorsuch (1983) argued that, with more than 35 variables and communalities that exceed .70, there are negligible differences among communality estimates, and that using principal components with 1.00s in the diagonal of the correlation matrix is sufficient. Furthermore, he pointed out that Monte Carlo studies indicate that interpretation of the factors is unaltered when the number of measured variables is as low as 30 and the communalities are greater than .40. Generally, the greater the number of variables in the analysis and the higher the communalities, the fewer the differences among procedures. However, with relatively few variables in the analysis, Gorsuch recommended common factor analysis to avoid spuriously high factor loadings and misinterpretations of the data.

However, a number of authors have recently questioned whether component analysis and common factor analysis yield similar estimates of parameters in most research situations. For example, Widaman (1993) showed that the number of indicators per factor (rather than the number of observed variables in the analysis) and the level of communality influence the degree of similarity of component and common factor analysis results. As the number of indicators per factor and communality increase, the differences between the two techniques are lessened. However, many research applications report analyses on item-level data, and items tend to have low communalities. If only a small number of items load on each dimension and if the items have relatively low communality, the results of component and common factor analyses may diverge markedly. In these cases, common factor analysis leads to accurate estimates of factor loadings and factor correlations; in contrast, component analysis tends to lead to positive bias in estimates of loadings and corresponding negative bias in correlations among the dimensions. The differences in estimates provided by the two techniques is not trivial. Perhaps most important, estimates based on common factor analysis should generalize well to those obtained using confirmatory factor analytic techniques, whereas inaccurate estimates from component analysis may not be reproduced. Thus, clearly, common factor analysis techniques should be strongly preferred over component analysis techniques for most research applications that attempt to understand a domain of phenomena in terms of a smaller number of underlying, latent variables.

Deciding on the Number of Factors

Rules for deciding on the number of factors to retain in a solution can be divided into three categories: statistical tests, mathematical and psychometric indices, and rules of thumb.

Statistical tests. Statistical tests are not often conducted with exploratory factor analysis because investigators typically use the principal axes (least squares) estimation method, which does not yield any statistical test for model fit. However, statistical tests are provided by other methods of estimation, including maximum likelihood estimation and the methods of generalized least squares and asymptotically distribution free estimation. These statistical tests are usually computed as chi-square tests and represent tests of significant residual covariation among observed measures after extracting a certain number of factors. If the chi-square statistic is significant, there is a statistical basis for rejecting the model in favor of a model with one or more additional factors. The primary problem with statistical tests is their dependence on sample size. With rather large sample sizes (e.g., $N > 500$), a statistical test will frequently suggest too many factors, rejecting substantively adequate factor models because of essentially trivial levels of residual covariation. Conversely, with small sample sizes (e.g., $N < 100$), too few factors may be implied by the test, because of low power to detect practically significant levels of residual covariation.

Mathematical and psychometric criteria. More commonly used are mathematical or psychometric criteria for the number of factors. Indeed, the most frequently used criterion for retaining components in principal components analysis, the eigenvalue > 1.00, or Kaiser–Guttman criterion, is the default in most statistical packages. However, this criterion is probably not optimal in many circumstances. One justification for this rule stems from the goal of data reduction. Each component has an *eigenvalue,* which is the amount of variance accounted for by the component; the sum of all eigenvalues equals the number of variables in a component analysis. Thus, an eigenvalue < 1.00 indicates that a component accounts for less variance than a single variable. Because the goal of a component analysis is to reduce the set of variables, components with eigenvalues < 1.00 do not serve these purposes; thus, only components with eigenvalues > 1.00 are retained. However, a Monte Carlo study by Zwick and Velicer (1986) indicated that the goal of data reduction may be inadequately met by the Kaiser–Guttman rule because, consistently, it greatly overestimated the number of dimensions to retain. In contrast, Cliff (1988) showed that the rule can also underestimate the number of factors in some circumstances. Furthermore, Cliff noted that the frequent claim that an eigenvalue > 1.00 indicates that the component is reliable is erroneous, because component reliability depends on the reliability of the observed measures, not the eigenvalue. Still other problems with mathematical rules exist; for example, Guttman's (1954) derivation of the eigenvalue > 1.00 rule required the presence of a population correlation matrix (as opposed to a sample matrix), which makes the rule inapplicable in any finite research context, at least in principle. Additional information on these mathematical and psychometric rules are found in texts on factor analysis (e.g., Gorsuch, 1983).

Rules of thumb. Several more accurate methods for retaining factors are often grouped under the rubric of "rules of thumb" (see Gorsuch, 1983; Zwick & Velicer, 1986), among which the scree test is readily available, frequently used, and usually provides satisfactory results. The scree test plots the eigenvalues of the unrotated factors on a coordinate plane and examines the slope of the line connecting them. The cutoff for retaining factors is determined as the point at which the slope approaches zero, which indicates a point at which deleting a given factor would no longer result in discarding significant variance. The Cattell-Nelson-Gorsuch (CNG) scree test (Gorsuch, 1983) gives a statistical test for this point, but visual inspection is usually satisfactory. Gorsuch contended that with principal components analysis the scree test usually produces cutoffs near eigenvalue = 1.00. However, factors having eigenvalues considerably less than 1.00 may be retained when using the scree test with common factor analysis. Because of subjectivity in determining the "elbow" in the scree curve, the investigator should examine the impact of various cutpoints. That is, when two or more factors are near the cutoff point, it is useful to examine the interpretability of alternative factor solutions with differing numbers of factors.

Another practical criterion for retaining factors concerns the number of variables that have significant factor loadings on the factor. Obviously, factors with only a single significant loading assess only the specific factor variance associated with that variable. Dimensions with only a single high loading occur with some regularity when using component analysis but rarely occur with the use of common factor procedures. In general, three variables per factor are needed to identify common factors (Anderson & Rubin, 1956; Comrey, 1988). Moreover, because increasing the number of indicators per factor improves factor stability (Guadagnoli & Velicer, 1988), it may be necessary to find additional variables to define factors of interest that initially contain only two or three variables with high loadings.

Two additional tests for the number of factors deserve mention. The first of these is the parallel analysis criterion (see Montanelli & Humphreys, 1976). Here, the eigenvalues from a common factor analysis with SMCs as communality estimates are plotted against estimated eigenvalues from random data. (The Montanelli & Humphreys, 1976, article provides regression equations for estimating the random data eigenvalues.) The maximal number of factors is provided by the point at which the two curves cross; one should not retain a "real data" factor that explains less variance than the corresponding factor from "random data." The second indicator of the number of factors is the reliability coefficient that can be obtained when maximum likelihood estimation is used to estimate common factor models. This coefficient was proposed by Tucker and Lewis (1973) as an adjunct of the chi-square test associated with maximum likelihood estimation and was designed to be less dependent on sample size than was the chi-square index. Tucker and Lewis argued that this reliability coefficient should, in general, have a value of .95 or greater for a factor solution to be considered acceptable. Both the parallel analysis criterion and the Tucker–Lewis index have fared well in Monte Carlo evaluations and should be used more widely, along with the scree test, in determining the number of factors to retain.

Rotating Factors

Following extraction, the retained factors are usually rotated to simple structure to make them more interpretable. Simple structure is achieved when each variable loads highly on as few factors as possible; preferably, each variable will have only one significant or primary loading. Interpretation is also simplified when all factor loadings are positive. Obtaining positive loadings is generally easily done with psychological measures because most measures of attitudes, abilities, and other psychological processes can be reverse-scored if necessary.

The rotation procedure can be either *orthogonal,* in which factors are kept uncorrelated, or *oblique,* in which the factors are allowed to correlate. In exploratory factor analysis, orthogonal rotation using the varimax procedure is most commonly used. This procedure is the default option for most computer programs, and it produces reasonable simple structure in most situations. However, researchers should be encouraged to look at oblique solutions for their data. When using well-formulated rotations such as promax or the Harris–Kaiser orthoblique procedures, the oblique simple structure may be more compelling than the orthogonal solution for the data. The investigator can experiment with allowing various amounts of correlation among the factors, although allowing factors to be highly correlated may run counter to the purposes of identifying latent constructs that are distinct. Moreover, if the factors are virtually orthogonal in a given sample, the oblique rotations will return solutions with essentially orthogonal factors.

Hierarchical Solutions

In many cases, we would expect that one or more general factors underlie and account for the major portion of variance in scores on the measured variables. Hierarchical factor solutions may be the most appropriate model for these data, in which relatively specific first-order factors are grouped together under more general, higher order factors. When a higher order, general factor is expected, the initial factor structure matrix should be submitted to oblique rotation to allow for correlations among the first-order factors. The matrix of correlations among these first-order factors may then be factor-analyzed to identify the higher order factor.

Hierarchical factor analysis is, arguably, a highly underused procedure in developing and evaluating measures in clinical psychology. Hierarchical solutions probably are appropriate for many psychological instruments because most psychological constructs are composed of multiple, correlated facets. Sabourin, Lussier, Laplante, and Wright (1990) presented an illustrative example of using a hierarchical solution to explain how the separate subscales of the Dyadic Adjustment Scale (Spanier, 1976) load on a more general second-order factor, Marital Adjustment. Similarly, Achenbach, Conners, Quay, Verhulst, and Howell (1989) discussed the presence of a general psychopathology factor that influences scores on separate domains of child behavior problems.

Confirmatory Factor Analysis Procedures

Confirmatory factor analysis is a method for evaluating whether a prespecified factor model provides a good fit to the data. It departs markedly from exploratory factor analysis in that it relies heavily on a different method of estimation, the method of maximum likelihood, and on a different set of standards for evaluating the adequacy of factor solutions. In most cases, confirmatory factor analysis will be most useful in the later stages of measure development to refine and improve instruments. Confirmatory factor analysis is most effective when it is used to assess both whether a proposed factor structure adequately fits the data and whether the structure fits as well as and as parsimoniously as other models.

Recently, the wider availability of computer programs such as LISREL (Jöreskog & Sörbom, 1989) and EQS (Bentler, 1989), ongoing methodological advances (Bollen, 1990; Kenny & Kashy, 1992; Reise et al., 1993), and the availability of introductory sources (e.g., Bollen, 1989; Hayduk, 1987) have made confirmatory factor analysis more accessible to measurement researchers. Also, information about the strengths and limitations of confirmatory factor analysis in measurement evaluation research is beginning to accrue.

Model Specification

The initial, key step in confirmatory factor analysis is model specification. Model specification involves the specification of the number of factors that underlie the data as well as which measured variables should load on each of the factors. One question that should be answered in deciding to use confirmatory factor analysis is whether a factor structure can be proposed that has sufficient specificity to expect that confirmation would be obtained. Successful confirmation is more likely when exploratory factor analysis is used in the development of the instrument and its measurement structure. Prior exploratory factor analyses on previous samples can be used to develop the likely confirmatory factor structure for the instrument. For example, exploratory factor analyses can be the basis for eliminating from the instrument any items with substantial loadings on more than one factor, if the researcher wishes to do this; identifying which items should load on more than one factor, if these secondary loadings have substantive interpretations; and identifying pairs of items that share significant measurement variance not accounted for by the latent variables or factors (e.g., Church & Burke, 1994).

In general, factor structures are difficult to confirm when the measured variables are individual items from a questionnaire that is even moderately lengthy, especially if this means that more than five to eight items are free to load on each latent variable. Pairs of items often share variance apart from the variance accounted for by the factors due to item content overlap. Confirmatory models can be specified to include correlated error terms, or correlated residuals, that reflect this shared item variance. But, it is difficult to identify, on an a priori basis, all of the necessary correlated error terms if the indicators in the analysis are items from a long questionnaire that may require many of these correlated error terms. Ironically, although including as many items as possible that are heavily saturated with a factor usually improves the internal consistency and reliability of an instrument and is associated with robust solutions from exploratory factor analysis (Guadagnoli & Velicer, 1988), doing so also increases the potential for correlated error and may make confirmatory factor analysis more difficult. Thus, it may be unreasonable to expect that lengthy questionnaires with many items assessing each factor will show satisfactory solutions when the individual items are submitted to confirmatory factor analysis. However, this is precisely the case in which the use of item parcels (e.g., Kishton & Widaman, 1994) is most appropriate. For example, consider a hypothetical assessment instrument that has a total of 100 items, consisting of 20 items per scale for each of five scales. Specifying five factors each with 20 loadings would require a very large sample size and almost insurmountable difficulties in specifying correlated residuals among pairs of items. A far stronger confirmatory strategy would be to create four five-item parcels for each scale and then specify four parcels to load on each of the five factors.

Although as noted earlier, the use of EFA during initial measurement development may help to specify successful models for CFA, because of certain differences in the analytic techniques of exploratory and confirmatory factor analysis, the conclusions drawn from the same data may disagree. That is, exploratory factor analysis focuses on retaining factors that account for significant amounts of variance in the data, whereas confirmatory factor analysis assesses goodness of fit based on the variance remaining after the factors are taken into account. Thus, although exploratory factor analysis may identify factors that account for significant variance in the data, confirmatory procedures may show that significant additional variance remains. For example, Parker et al. (1993) were unable to confirm their own exploratory solution for the Ways of Coping Questionnaire, even when they kept all procedures equivalent for the samples used in the exploratory and confirmatory analyses. The authors presented an interesting discussion of several characteristics of the questionnaire items that may account for the instability of the factor structure. However, the failure to confirm the exploratory solution may be due to the differences in the criteria for successful exploratory and confirmatory solutions. In short, the different types of analytic techniques are sensitive to different features of the data.

Another problem of differences across techniques is that orthogonal solutions from exploratory factor analyses may not be confirmable with confirmatory factor analysis. On technical grounds, forcing zero correlations among factors can create problems with underidentification in confirmatory analysis

(Kenny & Kashy, 1992). Also, a study by Church and Burke (1994) illustrated that attempting to confirm orthogonal components may be inappropriate for measures of constructs, such as personality traits, in which simple structure is limited. Because facets of personality are nonindependent by nature, individual items or subscales assessing the facets will fail to conform to models requiring that they load only on one factor and take on zero loadings for all other factors. Although successful solutions can be obtained by allowing factors to correlate and items to load on more than one factor, doing so may create elevated correlations between factors that challenge the distinctions proposed by the factor models. This problem is illustrated by a recent exploratory factor analysis of the 1983 version of the Child Behavior Checklist (Macmann et al., 1992), in which analyses were conducted using correlation matrices of scales with overlapping items. In three of the six gender and age groups, a hierarchical solution failed to confirm the distinction between the Internalizing and Externalizing broad-band factors; instead, the analysis suggested that a single higher order factor, General Psychopathology, was sufficient. It is important to note that this illustration is somewhat contrived, because the data and scoring differed from the original factor analysis of the Child Behavior Checklist and did not reflect more recent developments in this measure (Achenbach, 1991). Nevertheless, the authors' discussion of the problem of item overlap is instructive for investigators using instruments containing items that are scored on more than one indicator or items that have substantial secondary loadings even if they are not scored on multiple indicators.

Thus, some caution and necessary modifications may be in order before indiscrimanately applying confirmatory factor analysis to solutions obtained from exploratory procedures. Most notably, it may be impossible to obtain a satisfactory fit when the analysis examines individual items from a relatively lengthy scale or when items are likely to load significantly on more than one factor. With large numbers of items, the use of item parcels should be pursued. On the other hand, confirmatory factor analysis at the item level may be successful when the scale contains relatively few items and when the data are consistent with a relatively simple factor solution, one comprised of either a single factor or a set of highly distinct constructs. Additionally, confirmatory factor analysis may be relatively more successful when it is used to confirm hierarchical solutions using data at the scale level. Scale scores are more reliable than individual items (and thus have smaller uniquenesses), and the use of fewer variables means fewer chances for shared measurement variances that cannot be sufficiently stipulated by theory. Cole (1987) illustrated successful confirmatory solutions of hierarchical models based on scale-level data.

Evaluating the Fit of Confirmatory Factor Models

The adequacy of confirmatory solutions can be assessed in may ways, and investigators typically use multiple criteria for goodness of fit. The first concern is that the solution converges and that all parameter estimates are within acceptable ranges. If these conditions are not met, the model probably contains specification errors and should be respecified. If the solution appears acceptable, the next questions are whether the model accounts for substantial variance in the measured variables and whether alternative, trimmed models provide an equally good or better fit to the data.

Overall model fit can be evaluated with a number of indices of absolute fit, which index the fit of any given model in and of itself. The most common index is the traditional chi-square goodness-of-fit test, which evaluates the significance of unexplained covariances, or covariances among measured variables that are not accounted for by the model. If the chi-square statistic is nonsignificant, the model provides an adequate representation of the data. However, because even very small amounts of residual covariance may be significant with large samples, the test is difficult to pass in analyses based on large samples of participants or in models containing many observed variables. A very promising alternative approach to evaluating model fit is through the use of a chi-square test of close fit (Browne & Cudeck, 1993). Because any structural model we may specify is at best an approximation of reality, Browne and Cudeck argued that models should not be expected to fit the data perfectly, but rather should provide a relatively close fit to the data. The Browne and Cudeck test of close fit appears to perform much better across a range of sample sizes than does the traditional chi-square test of fit. Other indices of absolute fit include the chi-square/degrees of freedom ratio and the Goodness of Fit Index (GFI) and Adjusted Goodness of Fit Index (AGFI); (Jöreskog & Sörbom, 1989), but these have not fared well in Monte Carlo evaluations.

Problems with assessing absolute fit led to an alternative approach to assessing model fit that contrasts the relative fit of the model as compared with a null model (specifying no covariation among variables) and various competing models that are related hierarchically by containing either more or fewer restrictions (Bentler & Bonett, 1980; Hoyle, 1991). On the basis of a Monte Carlo study of a large number of such indices, Marsh, Balla, and McDonald (1988) recommended the use of the Tucker–Lewis index as a measure of relative fit that is relatively independent of sample size. Other useful indices are the comparative fit index (Bentler, 1990) and the relative noncentrality index (McDonald & Marsh, 1990). In addition, the relative parsimony of competing models should be assessed by examining the relative degrees of freedom of the null and competing models (Mulaik et al., 1989). See Bentler (1990), Bollen (1990), Browne and Cudeck (1993), McDonald and Marsh (1990), Mulaik et al. (1989), and Tanaka (1993) for reviews of various available indices of overall fit, relative fit, and parsimony.

Interpreting and Reporting Results of Factor Analyses

For both exploratory and confirmatory factor analysis, the rationale for the procedures followed, the specific techniques used, the criteria guiding the interpretations, and the results from all analyses must be reported in sufficient detail so that researchers can easily understand and replicate all procedures. The following observations are relevant to this endeavor.

Factor Loadings

In exploratory analyses, factor loadings are generally considered to be meaningful when they exceed .30 or .40. When the sample size is between 5 and 10 participants per variable, coefficients at this level and above are usually considered to be significant on practical grounds. Although it may improve the appearance of the obtained simple structure to report only loadings above .30 (or above .40), all factor loadings should be reported to ensure sufficient information for a full evaluation of the results. If an item or items fail to have any substantially high loadings on any factor, these items may be deleted from the analysis and the factor analysis may be recomputed on the remaining subset of items.

Another pragmatic criterion for factor solutions is that the factors should account for a substantial percentage of the total variance of the measured variables. Streiner (1994) suggested that factors should explain at least 50% of the total variance. However, a more reasonable expectation is that the set of factors should explain at least 80% of the estimated common variance (i.e., the sum of the initial communality estimates). If less variance is accounted for, variable communalities are low, in which case it might be possible to eliminate variables with relatively weak factor loadings on all factors to improve the overall factor solution. Otherwise, accounting for relatively little common variance challenges the relative importance of common factors as opposed to the specific factor variance associated with individual variables. The Parker et al. (1993) factor analysis of the Ways of Coping Questionnaire illustrates this problem. The inability to obtain a satisfactory confirmatory solution for the four factors obtained in the exploratory factor analysis could be anticipated from the fact that the factors accounted for a total of only 36.6% of the variance in the exploratory analysis. These results suggest that the use of coping strategies appears to be affected more by characteristics specific to each approach rather than by underlying coping styles that lead people to rely primarily on certain classes of coping responses.

Factor Scores

Factor scores are most reliable and accurate in the sample on which the factor analysis was based when they are computed using factor scoring weights derived from the common factor pattern (Harman, 1976). However, unit weighting for all of the items with significant primary loadings on the factors provides factor scores that are virtually as accurate in the original sample, and these unit weights will work better than the factor scoring weights in any new samples (Gorsuch, 1983). Thus, it is unnecessary to use exact factor scoring coefficients to weight items differentially when estimating factor scores. Instead, factor scores may be estimated by summing the scores for all items or variables with significant and primary loadings on each factor, excluding items that have split loadings across more than one factor. It is important to remember that all items should be on the same scale (e.g., all on similar 1- to 7-point scales) to ensure equal weighting. If the measures loading on a particular factor have very different scales, all variables should be standardized before being summed into factor scores.

A hotly debated topic concerning factor scores is the problem of the indeterminacy of factor scores in common factor analysis. According to the common factor model, there are more factors than observed variables because, in addition to the common factors that account for the covariances among the measured variables, there also are unique factors associated with each of the variables in the analysis. As a result, there is no unique solution for the common factor scores, and the common factor scores can only be estimated. This problem points up one advantage of the component analysis model: even though the component model provides inaccurate estimates of factor loadings and factor correlations, component scores are determinate and may be computed (as opposed to being estimated). The problem of indeterminacy of common factor scores is inversely related to the relative magnitude of common variance; indeterminacy poses greater problems with variables that have low levels of communality. Rozeboom (1988) cast this problem in the larger framework of the indeterminacy of causal models and questioned whether constraints used in common factor analysis to create convenient solutions provide accurate representations of causal relations. Gorsuch (1983) acknowledged the problem of indeterminacy relevant to a factor solution based on any one sample but argued that, in practice, the replication of factor solutions and of relations of factor scores with outside variables across samples suggests that factor scores are reasonable measures of the constructs.

Cross-Validation

Cross-validation is desirable for both exploratory and confirmatory solutions. Ideally, sample sizes are large enough to assign participants randomly to at least two groups, using one group as the derivation sample and the remaining group or groups as the cross-validation sample(s). Of course, random assignment is important so that the groups do not differ on characteristics that might affect the factor structures within the groups. Thus, for example, comparing factor solutions for groups of women and men or for different subtypes of a clinical sample addresses the generalizability of the factor structure across groups, but does not serve as a cross-validation of the factor solution. Nevertheless, group differences in factor solutions are cause for concern. In some cases, factor solutions fail to replicate across groups with known differences, such as clinical and nonclinical samples, because of the restriction of range on the variables within one or both of the groups. In this situation, the factor solution for the entire unselected sample may be most appropriate for determining the composition of subscales. However, the failure to replicate the factor structure across groups with similar variance on the variables, such as groups of men and women, may indicate that the actual structure of the measure (and, by implication, the constructs it assesses) differs for the groups. At this point, the investigator must decide whether to develop separate scales for the groups or whether to focus only on the variables and the features of the construct that have a consistent structure across the groups. Developing separate measures may increase the sensitivity of the measures to within-group variation but complicates or totally eliminates the possibility of cross-group comparisons. For example, the 1983 version of the Child Behavior Checklist (Achenbach & Edelbrock, 1983) included separate factor-analytically derived subscales for boys and girls and for different age groups, which made comparison across gender groups impossible and greatly confused the interpretation of longitudinal data. The newer version of this scale (Achenbach, 1991) includes subscales that are invariant across gender and age.

A potentially useful practice for cross-validation is to conduct exploratory factor analysis with half of the sample and then use confirmatory factor analysis to confirm the factor structure with the other half. However, as discussed earlier and illustrated by the Parker et al. (1993) study with the Ways of Coping Questionnaire, confirmation will be unsuccessful when the exploratory solution fails to account for most of the variance in the data. In cases such as this, in which numerous item-level variables are submitted to exploratory factor analysis that then fails to achieve an acceptable solution in CFA, it may be possible to cross-validate the solution by conducting a second exploratory analysis with an independent sample and comparing the factor solutions using the coefficient of congruence, which assesses the similarity of factor loadings across samples. Nevertheless, because the coefficient of congruence is a liberal test of factor similarity across samples, the failure to obtain acceptable confirmation using confirmatory factor analysis indicates that the scale needs modification to improve the measurement precision of its items.

For many clinical purposes, one setting cannot provide a sample large enough to allow cross-validation within the same study; instead, one must rely on comparisons across studies. Although the psychological literature is replete with examples of successful replications of factor solutions, the ability to replicate factor solutions is subject to the same influences that affect the reliability of measures across time and settings. For example, measures of generalized attitudes or behavioral styles may be easier to replicate than measures of specific behaviors or practices under specific conditions (Cairns & Green, 1979). Thus, factor analyses of behavioral observations or cognitive sampling measures, for example, may fail to replicate across conditions, though the factors show considerable validity under each set of conditions.

Usefulness of Factor Solutions

In most cases, the ultimate criterion for the usefulness of a factor solution is whether the obtained factor scores provide information beyond that obtained from the global score for the entire scale. This criterion is demonstrated by showing both specificity and differential predictive validity for the factors. Unfortunately, most studies using factor analysis focus on internal validity and fail to address prediction to external criteria. An excellent exception is a study by Windle (1992), which illustrated an effective validation of a hierarchical factor solution for the Dimensions of Temperament Survey. Factor scores from adolescents' self-reports on the survey showed significant convergent correlations with their parents' ratings of the same attributes and nonsignificant correlations with the parents' ratings of nonlike attributes. Similarly, in developing the Retirement Satisfaction Inventory, Floyd et al. (1992) demonstrated that four factor-analytically derived scales assessing reasons for retirement showed expected differential associations with preretirement work characteristics and with different aspects of postretirement adjustment.

Reporting Findings

The information that should be given in all reports of factor analyses is summarized in the Appendix. For exploratory analyses, it is important for the purposes of interpreting and replicating the findings to have information about the factor analysis model used, the method for estimating communalities in common factor analysis, the method of extracting factors, the criterion for retaining factors, and the rationale for and method of factor rotation. Exploratory analyses also should include a table that reports factor loadings for all variables on all factors. In this table, the variables are usually grouped by the factors on which they have primary loadings and arranged in the order in which the factors were identified, with the primary loadings set in boldfaced type. Additionally, the table or the text should give eigenvalues and the percentage of variance accounted for by each factor.

In confirmatory factor analysis, the nature of the proposed factor structure must be specified beforehand, including whether the factors are considered to be orthogonal or correlated and whether secondary loadings are expected for any of the items. The results should include standard errors for each of the factor loadings, intercorrelations among the factors and indications of the significance of these correlations, estimates of the percentage of variance accounted for by each factor, and overall model fit statistics. The factor solution can be presented in a table or figure illustrating the structural model, which lists all factor loadings, factor intercorrelations, and disturbance coefficients.

Because many seemingly subjective decisions may be made in reaching a final factor solution in both exploratory and confirmatory analyses, investigators should also make their data available to other investigators who wish to apply different decision rules in reaching a solution. Whenever possible, the complete correlation matrix among the measured variables should be given in either a table or an appendix, along with item means and standard deviations, so that exploratory or confirmatory analyses can be performed on the data by interested readers. If journal space does not allow the reporting of a table of correlations, then the authors should be willing to provide this information on request.

Extensions of Confirmatory Factor Analysis

Before closing, two extensions of confirmatory factor analysis should be mentioned. Although confirmatory factor analysis is a relatively recent development, these techniques are readily adapted to test important measurement-related hypotheses that are testable in only indirect ways with exploratory factor analytic techniques. We hope to spur interest in these topics on the part of the clinical psychology community.

Measurement Invariance

A frequent question that emerges in using psychological measures is whether the factor structure of a scale is applicable across samples of participants who differ in ethnicity, age, or some other characteristic that might affect responses to the measure. The issue is not whether the groups differ in mean level on the constructs assessed by the instrument, but rather whether the instrument measures the same constructs across samples. If the factor structure fails to show invariance across groups, then meaningful comparisons across groups on factor scores are precluded. However, if measurement invariance is established, then group differences on scores reflect accurately the differences on the latent characteristics assessed by the factor. Increasing recognition of cultural, developmental, and contextual influences on psychological constructs has raised interest in demonstrating measurement invariance before assuming that measures are equivalent across groups.

Jöreskog and Sörbom (1989) described procedures for using an extension of confirmatory factor analysis with multiple samples to assess measurement invariance across groups. In the least restrictive case, the investigator examines whether the variables show the same pattern of significant factor loadings across the groups by testing the goodness of fit of a multisample model with similar patterns of fixed and free factor loadings in each group. A more stringent test, and the proper assessment of measurement invariance, is to restrict all common factor loadings (and perhaps even unique variances) to be invariant across the groups. For example, Windle's (1992) evaluation of the Revised Dimensions of Temperament Scale tested a multisample model with this restriction to demonstrate measurement invariance across groups of girls and boys. It is also possible to restrict the model further, constraining the variances and covariances of the factors to invariance across the groups, but these restrictions actually test for differences in the associations among the latent variables (factors) rather than assess measurement invariance per se. Reise et al. (1993) illustrated procedures for testing full and partial measurement invariance across groups and described appropriate hypotheses for testing the fit of competing models. Also, when measurement invariance is only partially confirmed, Reise et al. advocated using partially invariant factor solutions to scale factor scores appropriately within each group. With this procedure, group differences on the latent variables can be examined even when measures are not completely invariant across the groups.

Multitrait–Multimethod Analysis

Another exciting application of confirmatory factor analysis involves evaluating multitrait–multimethod designs. The advantage of confirmatory factor analysis over conventional correlational approaches is that unambiguous statistical tests for both convergent and discriminant validity are available (Widaman, 1985). In a full multitrait–multimethod design, the analysis partitions the communality of variables into variance caused by the trait and the method factors, and it produces disattenuated estimates of all parameters. Rather than examining the simple correlations among measures of the same construct for evidence of convergence, the analysis assesses the association of each measured variable with the latent variable and demonstrates significant convergent validation when the variables obtain significant loadings on the appropriate trait factors. In some cases, it is possible to establish separate factors for each method that are orthogonal to the trait factors. A significant factor loading on the relevant method factor indicates that the variable is significantly influenced by method variance. Finally, discriminant validity is established by testing the hypothesis that the correlations among the trait factors differ significantly from unity, indicating that the traits are empirically discriminable (i.e., are not correlated perfectly).

Cole (1987) presented a convincing argument for the advantages of confirmatory factor analysis with multitrait–multimethod designs, illustrating the application of the procedures with various types of clinical data. However, Kenny and Kashy (1992) and Marsh (1989; Marsh & Byrne, 1993) described common problems encountered with these procedures and suggested practical approaches that both improve the chances of obtaining solutions and help to clarify the interpretation of the findings (for an example of a clinical application, see Greenbaum, Dedrick, Prange, & Friedman, 1994). Of course, to end where we began, the strength of this procedure relies on having quality data in which clearly distinguishable traits are assessed with clearly distinguishable methods.

References

Achenbach, T. M. (1991). *Integrative guide for the 1991 CBCL/4-18, YSR, and TRF profiles.* Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M., Conners, C. K., Quay, H. C., Verhulst, F. C., & Howell, C. T. (1989). Replication of empirically derived syndromes as a basis for taxonomy of child/adolescent psychopathology. *Journal of Abnormal Child Psychology, 17,* 299–323.

Achenbach, T. M., & Edelbrock, C. (1983). *Manual for the Child Behavior Checklist.* Burlington: University of Vermont, Department of Psychology.

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.

Bentler, P. M. (1989). *EQS structural equation program manual.* Los Angeles: BMDP Statistical Software.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107,* 256–259.

Browne, M. W. (1984). Asymptotic distribution free methods in analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cairns, R. B., & Green, J. A. (1979). How to assess personality and social patterns: Observations or ratings? In R. B. Cairns (Ed.), *The analysis of social interactions: Methods, issues and illustrations* (pp. 209–266). Hillsdale, NJ: Erlbaum.

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality and Social Psychology, 66,* 93–114.

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin, 103,* 276–279.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55,* 584–594.

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56,* 754–761.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dunlap, W. P., & Cornwell, J. M. (1994). Factor analysis of ipsative measures. *Multivariate Behavioral Research, 29,* 115–126.

Floyd, F. J., Haynes, S. N., Doll, E. R., Winemiller, D., Lemsky, C., Burgy, T. M., Werle, M., & Heilman, N. (1992). Assessing retirement satisfaction and perceptions of retirement experiences. *Psychology and Aging, 7,* 609–621.

Folkman, S., & Lazarus, R. S. (1988). *Ways of Coping Questionnaire: Research edition.* Palo Alto, CA: Consulting Psychologists Press.

Goldberg, L. R., & Digman, J. M. (1994). Revealing structure in the data: Principles of exploratory factor analysis. In S. Strack & M. Lorr (Eds.), *Differentiating normal and abnormal personality* (pp. 216–242). New York: Springer.

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

Greenbaum, P. E., Dedrick, R. F., Prange, M. E., & Friedman, R. M. (1994). Parent, teacher, and child ratings of problem behaviors of youngsters with serious emotional disturbances. *Psychological Assessment, 6,* 141–148.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103,* 265–275.

Guttman, L. (1954). Some necessary and sufficient conditions for common factor analysis. *Psychometrika, 19,* 149–161.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances.* Baltimore: Johns Hopkins University Press.

Haynes, S. N., Floyd, F. J., Lemsky, C., Rogers, E., Winemiller, D., Heilman, N., Werle, M., Murphy, T., & Cardone, L. (1992). The Marital Satisfaction Questionnaire for Older Persons. *Psychological Assessment, 4,* 473–482.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247.

Hoyle, R. H. (1991). Evaluating measurement models in clinical research: Covariance structure analysis of latent variable models of self-conception. *Journal of Consulting and Clinical Psychology, 59,* 67–76.

Hoyle, R. H., & Smith, G. T. (1994). Formulating clinical research hypotheses as structural equation models: A conceptual overview. *Journal of Consulting and Clinical Psychology, 62,* 429–440.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 122,* 351–362.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: User's reference guide.* Mooresville, IN: Scientific Software.

Kaufman, A. S. (1975). Factor analysis of the WISC-R at 11 age levels between 6 1/2 and 16 1/2 years. *Journal of Consulting and Clinical Psychology, 43,* 135–147.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait–multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165–172.

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54,* 757–765.

Macmann, G. M., Barnett, D. W., Burd, S. A., Jones, T., LeBuffe, P. A., O'Malley, D., Shade, D. B., & Wright, A. (1992). Construct validity of the Child Behavior Checklist: Effects of item overlap on second-order factor structure. *Psychological Assessment, 4,* 113–116.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait–multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335–361.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

Marsh, H. W., & Byrne, B. M. (1993). Confirmatory factor analysis of multitrait–multimethod self-concept data: Between-group and within-group invariance constraints. *Multivariate Behavioral Research, 28,* 313–349.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247–255.

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,* 341–347.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stillwell, C. D. (1989). An evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105,* 430–445.

Múthen, B. (1988). *LISCOMP: Analysis of linear structural equations with a comprehensive measurement model.* Mooresville, IN: Scientific Software.

Parker, J. D. A., Endler, N. S., & Bagby, R. M. (1993). If it changes, it might be unstable: Examining the factor structure of the Ways of Coping Questionnaire. *Psychological Assessment, 5,* 361–368.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566.

Rozeboom, W. W. (1988). Factor indeterminacy: The saga continues. *British Journal of Mathematical and Statistical Psychology, 41,* 209–226.

Sabourin, S., Lussier, Y., Laplante, B., & Wright, J. (1990). Unidimensional and multidimensional models of dyadic adjustment: A hierarchical reconciliation. *Psychological Assessment, 2,* 333–337.

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7,* 300–308.

Spanier, G. B. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family, 38,* 15–28.

Streiner, D. L. (1994). Figuring out factors: The use and misuse of factor analysis. *Canadian Journal of Psychiatry, 39,* 135–140.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–39). Newbury Park, CA: Sage.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs,* No. 1.

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs,* No. 2.

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

Waller, N. G. (1989). The effect of inapplicable item responses on the structure of behavioral checklist data: A cautionary note. *Multivariate Behavioral Research, 24,* 125–134.

Waller, N. G. (1994). NOVAX 1.3: A PC-DOS factor analysis program for polytomous data and mainframe size problems. *Applied Psychological Measurement, 18,* 195–196.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait–multimethod data. *Applied Psychological Measurement, 9,* 1–26.

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research, 28,* 263–311.

Widaman, K. F., & Herringer, L. G. (1985). Iterative estimates of communality: Initial estimate need not affect stabilized value. *Psychometrika, 50,* 469–477.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Chicago: Scientific Software.

Windle, M. (1992). Revised Dimensions of Temperament Survey (DOTS-R): Simultaneous group confirmatory factor analysis for adolescent gender groups. *Psychological Assessment, 4,* 228–234.

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

Appendix

Information to Include in Reporting Factor Analyses

Confirmatory Factor Analysis

- Initial proposed model(s)
  - Number and composition of factors
  - Orthogonal versus correlated factors
  - Secondary loadings, correlated error terms
  - Other model constraints (fixed and free parameters)
- Method of estimation
- Goodness of fit
  - Overall fit
  - Relative fit
  - Parsimony
- Any model modification to improve model fit to data
- Factor loadings (lambda) and standard errors
- Communality (or squared correlations of observed variables with the factors)
- Factor correlations and standard errors (significance)

Exploratory Factor Analysis

- Principal component analysis of common factor analysis
- Initial communality estimates (common factor analysis)
- Method of factor extraction
- Criteria for retaining factors
- Eigenvalues, percentage of variance accounted for by the unrotated factors
- Rotation method (and rationale)
- All rotated factor loadings
- Factor intercorrelations (oblique solutions)
- Variance explained by factors after rotation

Received March 1, 1995
Revision received April 19, 1995
Accepted April 21, 1995 ■