Psychological Assessment
2000, Vol. 12, No. 3, 287–297

1040-3590/00/$5.00 DOI: 10.1037//1040-3590.12.3.287

Factor Analysis and Scale Revision

Steven P. Reise
University of California, Los Angeles

Niels G. Waller
Vanderbilt University

Andrew L. Comrey
University of California, Los Angeles

This article reviews methodological issues that arise in the application of exploratory factor analysis (EFA) to scale revision and refinement. The authors begin by discussing how the appropriate use of EFA in scale revision is influenced by both the hierarchical nature of psychological constructs and the motivations underlying the revision. Then they specifically address (a) important issues that arise prior to data collection (e.g., selecting an appropriate sample), (b) technical aspects of factor analysis (e.g., determining the number of factors to retain), and (c) procedures used to evaluate the outcome of the scale revision (e.g., determining whether the new measure functions equivalently for different populations).

Personality measurement by self-report questionnaire is a thriving enterprise of critical importance to theory development and testing in many psychological disciplines such as clinical psychology. At least three journals focus on statistical analyses of questionnaire data: *Psychological Assessment, Journal of Personality Assessment,* and *Assessment.* Many of the articles in these journals use exploratory factor analysis (EFA) and, oftentimes, the factor analytic findings are used to guide scale revision. In this article, we review methodological issues that arise in the application of EFA to the scale revision and refinement process.

This article begins with an overview of several issues pertinent to the application of EFA in scale revisions. Two of the topics we address are the hierarchical nature of psychological constructs and the motivations for revising a scale. Methodological issues that arise in the context of applying EFA to scale revision are then covered. Specifically, we address issues that arise prior to data collection (e.g., selecting an appropriate sample), technical aspects of factor analysis (e.g., determining the number of factors to retain), and procedures used to evaluate the outcome of the scale revision (e.g., determining whether the new measure functions equivalently for different populations). We conclude by highlighting two additional topics: dimensionality and scale score interpretation, and principal components versus factor analysis.

Inevitably, any review will omit important topics. This review is no exception, and to clarify our points of deemphasis we note the following. First, our review is primarily concerned with EFA, and we do not provide a thorough consideration of alternative multivariate models that may be helpful during scale revision, such as principal-components analysis (see Velicer & Jackson, 1990) or multidimensional scaling (Davison, 1994). Second, we discuss the use of confirmatory factor analysis (CFA) in scale revision only sparingly because more extended treatments of this topic are available elsewhere (see Finch & West, 1997, or Floyd & Widaman, 1995). Finally, although this point may seem obvious, we assume that the construct being measured is appropriate for a factor analytic representation. That is, we only consider situations in which dimensional latent variables (factors) account for indicator (item) correlations. We do not consider modeling emergent variables (Bollen & Lennox, 1991; Lykken, McGue, Tellegen, & Bouchard, 1992), latent types (Strube, 1989; Waller & Meehl, 1998), or multifaceted trait concepts (Carver, 1989; Hull, Lehn, & Tedlie, 1991).

Overview: Factor Analysis as Applied to Scale Revision

> If the misapplication of factor methods continues at the present rate, we shall find general disappointment with results because they are usually meaningless as far as psychological research interpretation is concerned. (Thurstone, 1937, p. 73)

The goal of scale revision is to improve the psychometric properties—and ultimately the validity—of individual-differences measures. Here we use the term validity to imply that a measure (a) has item content and a corresponding factor structure that is representative of and consistent with what is currently known regarding a construct, (b) has a factor structure that is replicable and generalizable across relevant populations, and (c) has a clearly interpretable (i.e., univocal) and relatively precise scaling of individuals along one or more common dimensions. It is these three objectives that we have in mind in this article. Note that these objectives are consistent with Loevinger's (1957) notions of substantive and structural validity.

Several review articles have recently addressed the application of EFA to personality and psychopathology test data (e.g., Comrey, 1988; Finch & West, 1997; Goldberg & Digman, 1994).

Steven P. Reise and Andrew L. Comrey, Department of Psychology, University of California, Los Angeles; Niels G. Waller, Department of Psychology, Vanderbilt University.

Correspondence concerning this article should be addressed to Steven P. Reise, Franz Hall, Department of Psychology, University of California, Los Angeles, California 90095. Electronic mail may be sent to reise@psych.ucla.edu.

Although such reviews have appeared in the literature for over half a century, many of the warnings expressed in these reviews have fallen on deaf ears. Stated bluntly, much factor analytic research is neither informative nor trustworthy. Beyond commonly observed design problems such as the use of idiosyncratic or small samples, the use of mixed populations (e.g., men and women, psychotic and nonpsychotics) in a single analysis, and one-shot scale revision or validity studies, a chief culprit behind poor applications of EFA is the overreliance on the default options found in many statistical packages. For instance, the default options in the widely popular Statistical Package for Social Sciences (SPSS) provide principal-component analyses, the eigenvalues greater than 1 rule to retain factors, and varimax rotation to simple structure. All three of these options are potentially troublesome if a researcher is looking for a generalizable factor structure that will replicate across samples. In other words, all three options can hinder a researcher's aim to create a better measure of a psychological construct. Gorsuch (1997) has recently expressed a similar view, and he has noted that "the default procedure of many statistical packages . . . is no longer adequate for exploratory item factor analysis" (p. 532).

The Construct Hierarchy

The number of psychological constructs that can be proposed and assessed is infinite. For this reason, theories that offer "maps" of the personality terrain, such as the five-factor model of personality trait structure, have been well received among researchers (Ozer & Reise, 1994). We cite the five-factor model not because we believe it to be revealed truth or the best map of the normal-range personality domain but rather because it so clearly makes explicit the situation researchers face when proposing new constructs and trying to develop factor-analytic-based measures of them. Specifically, psychological constructs have a hierarchical structure such that different constructs have different levels of conceptual breadth (Comrey, 1988; Guilford, 1975; John, 1990; West & Finch, 1997).

The level of conceptual breadth distinction has implications for the predictive ability of factor-analytic-based measures (Mershon & Gorsuch, 1988; Ozer & Reise, 1994). Measures of higher order dispositions provide optimal prediction of heterogeneous/complex criteria, whereas narrow-band measures are most efficacious in predicting a specific criterion. The construct hierarchy also has profound ramifications for the application of EFA to scale revision. In fact, in revising a scale, it is arguable that the most important decision a researcher confronts is where the measure is to be located in the construct hierarchy (see Smith & McCarthy, 1995, p. 303). This decision will influence the researcher's choices regarding the type of items that are written, how many items of each type are included on the revised scale, how many factors will be extracted, how those factors will be rotated, and ultimately how the resulting factor solution will be judged and interpreted. Throughout this article, we frequently reference the construct hierarchy and its implications for EFA.

Motivations for Scale Revision

Before considering specific issues that arise in the application of EFA to scale refinement, let us consider some underlying motivations. First, a primary reason for scale revision is that the scale's psychometric properties are deemed inadequate. For example, a scale may not provide scores that have an appropriate degree of internal consistency reliability. Second, research may demonstrate that an existing measure has a factor structure that is not generalizable across different samples (e.g., men and women) or has a factor structure that is "not as advertised" by the original authors (e.g., a purported three-dimensional measure is really four dimensional).

A third reason for scale revision is inadequate construct representation (West & Finch, 1997). Two distinct problems with an existing measure can be isolated. First, it may be recognized that a scale is missing an important aspect of a construct. For example, if a traditionalism scale did not include items pertaining to views on religious authority (arguably an important aspect of this construct), a revision might be undertaken to address this deficit. Second, research may demonstrate that a measure is not tapping a researcher's intended dimension. This can occur when a measure does not have the desired location within some conceptual framework, be that the five-factor model (Goldberg, 1992) or some other map of personality or psychopathology.

In the planning stage of a scale revision, writing new items is highly influenced by the particular motivations underlying the revision. Nevertheless, despite the heterogeneous psychometric and substantive reasons for revising a scale, there are two guiding principles that are frequently suggested if a researcher plans to use EFA to develop, refine, and evaluate the new measure. First, the researcher should develop a clearly articulated plan regarding the need for the revised instrument. There are several questions that should be explicitly addressed: What construct is this scale trying to measure? Why do researchers need to measure this construct? At what level of the construct hierarchy is this measure? How is this measure different from other competing measures? Second, the scale developer should conduct a systematic series of studies by using large samples of respondents. Factor-analytic-based scale revision should be an iterative process where data inform construct definition and refinement (see Comrey, 1978; Tellegen & Waller, in press).

Selection of Variables (Items)

Referring to item selection, Goldberg and Digman (1994) recently noted that "This is by far the single most important decision to be made in any investigation, and it should be guided by theory and/or the findings from past research" (p. 218). In a scale revision context, there is an existing item pool that ostensibly represents the construct or constructs of interest. Yet, by definition, this pool is deemed inadequate and hence a revision is called for. In writing new items, it is important for the researcher to plan ahead and anticipate what the final factor solution will look like (Comrey, 1978). For example, several questions should be considered before writing that first new item, such as how many dimensions are there and to what degree will these dimensions correlate?

As for specific suggestions for constructing new item pools, we offer the following. First, the new item pool should be substantively overinclusive (Loevinger, 1957). For example, a researcher revising an achievement motivation scale might write new items that tap theoretically guided or empirically verified aspects of the achievement construct. Note that it will always be easier to subtract rather than to add items at some future point in the scale

refinement process. Second, for each construct, researchers might consider writing multiple sets of content homogeneous items that tap different aspects of the target construct. For example, Comrey (1970) used 40 homogeneous sets of four items to measure the eight primary personality dimensions of the Comrey Personality Scales (CPS).

The creation of an overinclusive item pool with items tapping different aspects of a construct will achieve two objectives. First, subsequent factor analyses will be better able to provide empirical evidence of what type of item content belongs in the construct and what belongs somewhere else (Tellegen & Waller, in press). Note that this empirical parsing process is most effective when other measures of related constructs are included in the data analysis (Clark & Watson, 1995, p. 314). For illustration, imagine that a researcher factor analyzed the six facet scales of the Revised NEO Personality Inventory (NEO–PI–R; Costa & McCrae, 1992) Neuroticism scale. He or she would likely obtain a clear univocal structure. However, if a researcher factor analyzed all 30 facet scales on the NEO–PI–R simultaneously, he or she would find that the facet of Angry–Hostility is not a univocal marker of Neuroticism but rather is multidimensional with a large secondary negative loading on Agreeableness. Thus, it is important to collect data not only on the revised scale but also on marker variables that can be used to map the scale to established frameworks. By including marker variables, researchers can also investigate the discriminant and convergent validity of the revised scale.

An overinclusive item pool with multiple items representing different aspects of a construct also achieves a second goal, which is the creation of facet scales, item parcels (Cattell & Burdsal, 1975; Kishton & Widaman, 1994), homogeneous item composites (Hogan, 1983), or factored homogeneous item dimensions (FHID; Comrey, 1984, 1988). Many authors who have written on factor analysis and scale construction or revision emphasize the importance of creating multi-item homogeneous item clusters. In this article these homogeneous clusters are called *facets*. Essentially, facets are item sets with similar content that tap into narrow-band constructs and are expected to display high item correlations. Facets are also expected to display better reliability and distributional properties than single items.

Many popular personality measures make extensive use of facets. As noted previously, each of the eight dimensions of the CPS (Comrey, 1970) is marked by five FHIDs (i.e., facets). Moreover, each of the 11 personality dimensions of the Multidimensional Personality Questionnaire (MPQ; Tellegen, 1982) contains multiple facets of content homogeneous items. Finally, the NEO–PI–R (Costa & McCrae, 1992) uses six relatively homogeneous eight-item facet scales to tap the higher order constructs of the five-factor model. Facets are often conceptualized as being at the lower end of the construct hierarchy. Researchers who are primarily interested in higher order constructs should not neglect the use of facets because facets can serve as the building blocks from which higher order dimensions may emerge in EFA.

We now turn to the issue of item format (i.e., dichotomous or polytomous). The linear factor analysis model assumes that variables are measured on continuous, interval-level scales. Statistical tests for the number of dimensions, such as those available with maximum-likelihood estimation, assume that the data are multivariate normal (Hu, Bentler, & Kano, 1992). It is well known that dichotomously scored items cannot meet these conditions. Furthermore, the use of dichotomous item response formats can cause serious distortions in the correlation matrix. Even if two items measure the same construct, the phi coefficient may be low if the response proportions differ markedly (Lord & Novick, 1968, p. 347). Because EFA is often based on correlations, any distortions in the correlations can result in misleading EFA findings (Comrey, 1978). Moreover, item-level factor analysis of dichotomous items can lead to the identification of spurious (non-content-based) factors caused by nonlinearities in the relationship between the item and the latent variable (Gibson, 1959, 1960; Gourlay, 1951; McDonald & Ahlawat, 1974; Waller, 1999; Waller, Tellegen, McDonald, & Lykken, 1996).

With the above considerations in mind, many authors suggest that test developers create multipoint rating scales (e.g., 5-, 7-, or 9-point ratings) if they plan to conduct item-level factor analysis (see Comrey, 1978, 1988; Comrey & Montag, 1982; Goldberg, 1981; Goldberg & Digman, 1994). Polytomous items not only make item-level factor analyses more feasible but they are also expected to be more reliable and have greater variance than dichotomous items. Thus, two potential advantages to multipoint ratings are better psychometric properties of the resulting scale and the avoidance of problems that are inherent in the factor analysis of dichotomous items (Comrey, 1988; Comrey & Montag, 1982). Yet, multipoint rating formats can be problematic for at least two reasons. First, they may facilitate individual differences in the willingness to endorse the extremes of a rating scale (Chen, Lee, & Stevenson, 1995; Hamilton, 1968; Loevinger, 1957). Second, they may call for examinee distinctions that simply cannot be made and thus add nuisance variance that ultimately detracts from scale validity (Clark & Watson, 1995, p. 312). Dichotomous response formats, however, are efficient (i.e., examinees can respond quickly to many items), and they potentially hold individual differences in scale use to a minimum.

Long-standing arguments in favor of multipoint rating scales versus dichotomous items are less salient because many problems inherent in factor analyzing dichotomous items can now be adequately addressed with modern statistical procedures. For example, computer programs are now available to conduct full-information item factor analysis (Bock, Gibbons, & Muraki, 1988; Wilson, Wood, & Gibbons, 1984) or nonlinear factor analysis (Fraser, 1986; Fraser & McDonald, 1988) on binary items. See Steinberg and Jorgensen (1996) and Waller et al. (1996) for applications of these procedures. New computer programs have also been developed for factor analyzing large data sets of polytomous and dichotomous items (Muthen & Muthen, 1998; Waller, 1995). Given these new developments, we cannot unambiguously recommend multipoint items over dichotomous items.

Selection of the Sample of Respondents

Several issues warrant attention when considering sample characteristics of a factor analytic study. The first issue concerns sample size. Numerous rules of thumb for the minimum sample size needed to obtain a robust factor solution are offered in research articles and textbooks (Goldberg & Digman, 1994; Guadagnoli & Velicer, 1988; Velicer & Fava, 1987, 1998). Many of these rules stem from accumulated experience or from findings of Monte Carlo simulations. Nevertheless, readers who delve into this literature soon learn that the advice given in these sources has low

internal consistency. For instance, Gorsuch (1983) suggested that no fewer than 100 individuals should be included in a factor analytic study, whereas Goldberg and Digman (1994) recently suggested that between 500 and 1,000 respondents are required.

Commenting on the various rules of thumb from the pundits, MacCallum, Widaman, Zhang, and Hong (1999) recently concluded that "common rules of thumb regarding sample size in factor analysis are not valid or useful" (p. 96). MacCallum et al. (in press) took a fresh approach to the problem and used factor analytic theory (MacCallum & Tucker, 1991) to show that it is impossible to derive a minimum sample size that is appropriate in all situations. By using theoretical arguments and empirical evidence, these authors demonstrated that the minimum sample size needed to accurately recover a population factor pattern is a function of several variables including the variables-to-factor ratio, the average communality of the variables, and the degree to which the factors are overdetermined (defined, in part, by the number of variables that load on each factor). When communalities are high (>.6) and the factors are well defined (have many large loadings), sample sizes of 100 are often adequate. However, when communalities are low (e.g., when analyzing items), the number of factors is large and the number of indicators per factor is small, even a sample size of 500 may not be adequate.

In addition to sample size, a second issue that warrants consideration is sample heterogeneity. In terms of identifying replicable factors, researchers should assemble samples with sufficient examinee representation at all levels of the trait dimensions. In other words, there should be many examinees at all trait levels in order to accurately estimate the population item intercorrelations. One consequence of this rule is that using the standard pool of undergraduates may be suitable when undergraduates manifest sufficient heterogeneity with respect to trait standing. On some constructs, such as extraversion or agreeableness, this seems reasonable. For other constructs, however, such as uni-polar depression or psychotic ideation, undergraduates may not be an appropriate respondent pool to accurately map the factor space of clinical assessment scales.

Factor Extraction

The goal of factor extraction is to identify the number of latent dimensions (factors) needed to accurately account for the common variance among the items. Despite this relatively clear criterion, the issue of how many factors to extract and retain for rotation has been a source of contention for years (see Fava & Velicer, 1992a, and Wood, Tataryn, & Gorsuch, 1996, for reviews). If too few factors are extracted, a researcher may miss important distinctions among the items, and the subsequently rotated solution may be distorted in nonsystematic ways (see Comrey, 1978). However, if too many dimensions are retained, some rotated factors may be ill defined with only one or two salient loadings. Note that over 30 years ago, Comrey (1967) proposed an analytic rotation strategy specifically designed to address these issues.

Although there are many rules of thumb and statistical indices for addressing the dimensionality issue in EFA, no procedure seems entirely satisfactory. One point seems clear, however. Studies (Fava & Velicer, 1992a; Wood et al., 1996) that have empirically compared the effects of under- and overextraction on the factor recovery of known population structures generally agree that it is preferable to extract too many factors rather than too few. For instance, on the basis of a highly ambitious Monte Carlo study, Wood et al. recently concluded that "(a) when underextraction occurs, the estimated factors are likely to contain considerable error; [and] (b) when overextraction occurs, the estimated loadings for the true factors usually contain substantially less error than in the case of underextraction" (p. 354). Ultimately, however, the decision about the number of factors to retain has to be supported by evidence. Several of the procedures summarized below can be used to provide such evidence.

One of the more popular guides for investigating matrix dimensionality (i.e., how many factors to retain) is the scree test (Cattell, 1966). To conduct a scree test, a plot is created with the number of dimensions on the *x*-axis and the corresponding eigenvalues (percentage of variance accounted for by a dimension) on the *y*-axis. The objective of the scree plot is to visually locate an elbow, which can be defined as the point where the eigenvalues form a descending linear trend (see Bentler & Yuan, 1998, for statistical tests of linear trends in eigenvalues). An example scree plot that demonstrates this concept is provided in Figure 1A. To construct this scree plot, we simulated the item responses for 15 items and 200 individuals. The data were generated to fit a three-factor model. We also fit a linear regression line to the smallest 12 eigenvalues of the correlation matrix. The obtained regression line has been superimposed on the scree plot to illustrate the concept of the scree "elbow." Notice that after the third eigenvalue there is a strong linear (descending) trend in the remaining eigenvalues. This trend provides mathematical support for a three-factor solution for these data.

As illustrated in Figure 1B, a scree plot can be augmented by conducting a parallel analysis (Drasgow & Lissak, 1983; Horn, 1965; Longman, Cota, Holden, & Fekken, 1989; Montanelli & Humphreys, 1976; Zwick & Velicer, 1986). In a parallel analysis, random data sets are generated on the basis of the same number of items and persons as in the real data matrix. Then the scree plot of the eigenvalues from the real data is compared with the scree plot of the eigenvalues from the random data. The point where the two plots meet provides the researcher with a good idea of the absolute maximum number of factors that should be extracted. The logic underlying parallel analysis is simple; a researcher should not extract a factor from the real data that explains less variance (i.e., has a smaller eigenvalue) than a corresponding factor in the simulated random data.

Generating eigenvalues from random data matrices has become increasingly easy and convenient in recent years as powerful computer software has become widely available. For instance, the R programming language, which can be freely downloaded from the following Web site http://lib.stat.cmu.edu/R/CRAN/#source, is an ideal package for calculating eigenvalues. Once the package has been installed (versions for PCs, Macs, Unix, and Linex machines are available) the following code can be easily modified to generate eigenvalues from random data matrices. We have written this code to generate 100 data matrices for 200 participants and 15 variables (to correspond to the example in Figure 1). The eigenvalues from the 100 data matrices were averaged to yield robust estimates for the parallel analysis. Simple modifications of the code could be added to simulate Likert item responses or other response formats.

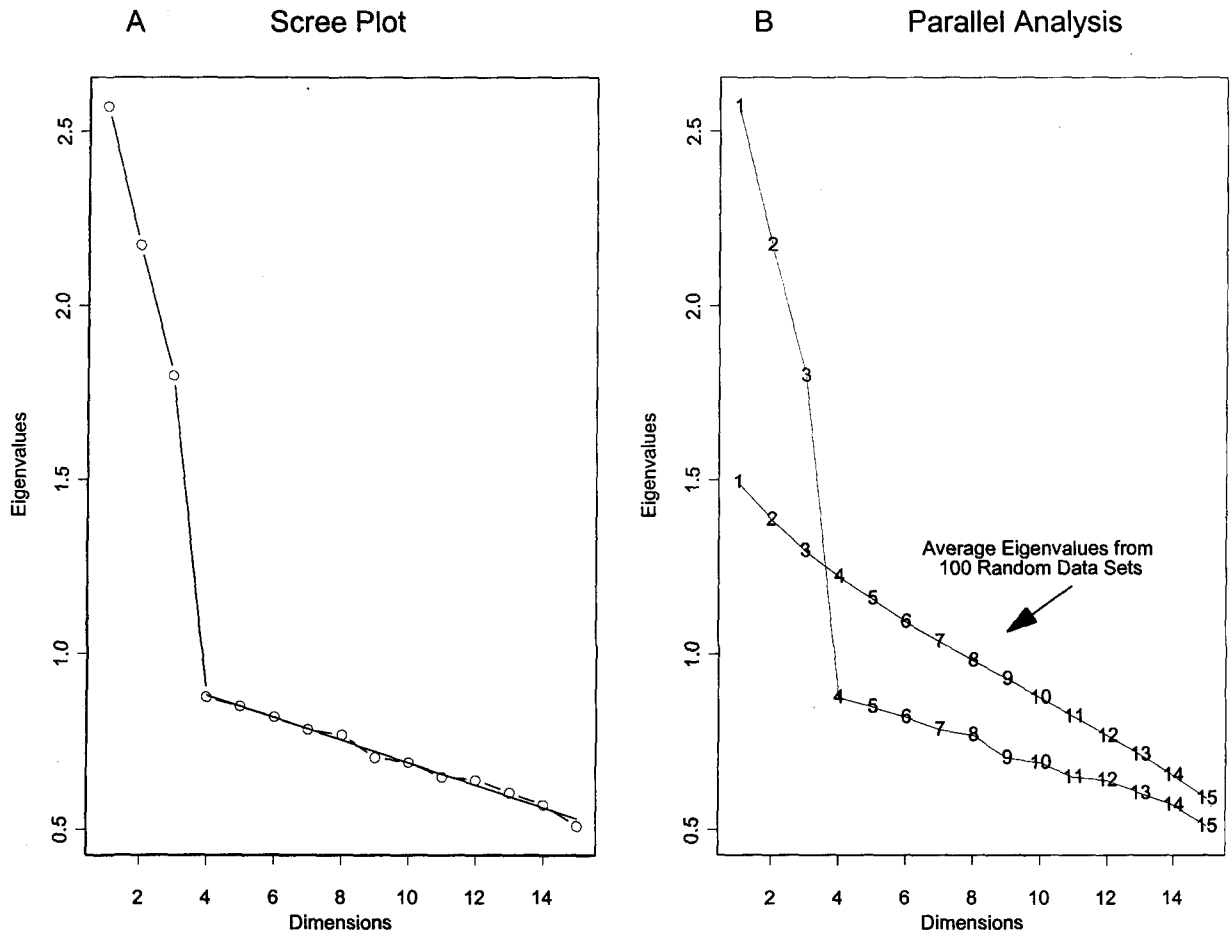

Figure 1. Panel A: A scree plot. Panel B: A parallel analysis plot.

```
random.eig <- matrix(0, nrow = 100, ncol = 15)
   for (i in 1:100) {
      random.data <- matrix(rnorm(200 * 15),
        nrow = 200, ncol = 15)
      random.eig[i, ] <- eigen(cor(random.
        data)) $values
   }
   average.eig <- apply(random.eig, 2, mean)
```

There are many other guidelines for deciding on the number of factors to extract. For example, some researchers prefer to extract as many factors as necessary for the factors to account for a certain percentage of total (e.g., 50%; Streiner, 1994) or common (e.g., 80%; Floyd & Widaman, 1995) variance. Another common, but highly misguided, rule is the eigenvalue greater than 1.0 rule. Unfortunately, this rule is the default in many statistical programs. The rule is problematic because its logic is based on principal-components (see final section) analysis (rather than factor analysis), and more important, research shows that it consistently leads to the retention of too many factors (Zwick & Velicer, 1986). Although it is generally preferable to retain too many factors rather than too few, there is no psychometrically justifiable reason to base overextraction on the eigenvalue greater than 1.0 rule. The number of eigenvalues greater than 1.0 is highly influenced by the number of variables in the factor analysis. Because the size of an eigenvalue has nothing to do with the reliability of a factor (Cliff, 1988), the eigenvalue greater than 1.0 rule cannot be recommended. Lee and Comrey (1979) provided an illustrative example of problems that occur when relying on the eigenvalue greater than 1.0 rule for factor extraction.

Finally, one of the older but still useful methods of determining the number of factors is to compute the difference between the elements in the original correlation matrix and the model reproduced correlation matrix given a certain number of extracted factors. This is known as residual analysis. Under appropriate assumptions (e.g., multivariate normality) certain factor analytic procedures, such as the maximum-likelihood technique, provide a statistical test of a residual matrix (Lawley, 1940) for determining the appropriate number of factors. In general, this approach to choosing a factor solution is problematic. Specifically, the chi-square test has large statistical power, and the reliance on this test in moderate to large samples results in the retention of too many factors of dubious substantive value (Hu et al., 1992; Montanelli, 1974). We suggest that researchers forgo the chi-square test of dimensionality. As an alternative, we suggest that researchers concentrate instead on the examination of simple plots of the

residuals (see also Hattie, 1985). Early factor analysts routinely plotted residuals to look for a normal distribution centered at zero.

Rotation of Factors

The initial factor extraction in an EFA produces orthogonal variables that are often not readily interpretable. Thus, after the initial extraction, researchers typically rotate the factor pattern to a psychologically interpretable position. The rotation algorithms that are most often used in psychology (e.g., Varimax, Promax, and Oblimin) attempt to orient the factors to maximize a criterion known as simple structure (Thurstone, 1947). Simply stated, simple structure implies that items load highly on one or perhaps two factors and have near zero loadings on the remaining factors. Rotated simple structure solutions are often easy to interpret, whereas the originally extracted (unrotated) factors are often difficult to interpret. Simple structure rotations, such as Varimax, however, are not guaranteed to find the most psychologically defensible placement of factors. This is especially true when the scale items do not correspond to a simple structure arrangement. For example, simple structure maximizing rotations (e.g., Varimax and Oblimin) are not appropriate when analyzing tests that were developed to represent circumplex models of personality (Wiggins, 1980), psychopathology (Becker, 1998; Gurtman & Balakrishnan, 1998), or vocational interests (Tracey & Rounds, 1993).

In an orthogonal rotation, such as Varimax, the factors are not allowed to correlate. In oblique rotations, such as Promax or Oblimin, the factors are allowed to correlate. Judging from the literature, many researchers prefer orthogonal rotations because of the simplicity of interpretation. However, there are compelling reasons to consider oblique rotations. Foremost among these is that oblique rotation methods produce orthogonal solutions if an orthogonal solution is appropriate. In other words, oblique rotation methods do not constrain the factors to be uncorrelated.

There are at least five additional reasons to consider oblique rotations. First, if the different factors are postulated to be aspects of a more general construct, then lower order factors can themselves be factored to obtain a higher order general factor. Second, oblique rotations will always meet the simple structure criterion better than orthogonal rotations. Third, some research supports a slight superiority of oblique rotations in terms of factor replicability (Dielman, Cattell, & Wagner, 1972; Gorsuch, 1970). Fourth, it might be unreasonable to assume that any set of psychological variables are truly uncorrelated, and thus oblique rotations may represent a more realistic modeling of psychological phenomena (see Loo, 1979). Finally, the use of an oblique rotation means that the correlations among the factors will also be estimated rather than fixed to zero as in an orthogonal rotation. In turn, the factor correlations may provide substantively valuable information.

There is more to the rotation problem than deciding between two general classes of mathematical algorithm. For instance, a question that should always be considered before using any rotation method is, do the data meet the assumptions of simple structure? As noted in Comrey (1978), "mathematical algorithms designed to approximate simple structure work well only in situations properly designed for the application" (p. 648). In fact, none of the simple structure rotation methods will be able to identify an interpretable simple structure if many variables are complex (i.e., tap into more than a single trait). For this reason, it is often advised that researchers make every effort to design instruments where each variable will load highly on a single dimension. For example, Comrey (1978) advised, "Ideally, a variable for a factor analytic study should measure one and only one factor in the domain to any substantial degree" (p. 650). This quotation should not be taken to imply that simple structure factor patterns are easy to design (see the next paragraph). On the contrary, true simple structure is often highly elusive (Guilford & Zimmerman, 1963), and we believe that there has been inappropriate attention paid to the assumptions of simple structure. Therefore, we recommend that researchers routinely inspect factor plots for obvious departures from simple structure. These plots can be of tremendous value in suggesting whether simple structure is reasonable within a given domain as well as in suggesting directions for future scale revision.

Designing a measure to yield a clean simple structure is not as easy as the above discussion might suggest. It has been argued that in the personality domain many traits are not univocal markers of a single dimension but instead represent a blend of two or more dimensions (Goldberg, 1993). In fact, one of the conclusions of Church and Burke (1994) was that applying CFA hypothesis tests to personality instruments is difficult because many variables tend to load on more than one dimension. We return to this issue shortly when we consider factor replication. For now, note that when items or facets have complex patterns of loadings, aiming for mathematically based simple structure solutions may not be appropriate. Researchers may need to plot factors and to develop hand rotations (Comrey, 1978; Comrey & Lee, 1992; Gorsuch, 1983) or they may need to consider horizontal aspects of construct representation (Goldberg & Digman, 1994). An example of a horizontal (circumplex) representation of the five-factor model is discussed in Hofstee, de Raad, and Goldberg (1992).

Evaluating the Revision

Evaluating the quality and usefulness of a factor analytically derived instrument is a large topic that we cannot adequately cover in this article. Smith and McCarthy (1995) provided a summary of fundamental psychometric criteria that should be considered at all stages of the scale revision process. These criteria include (a) recognizing a scale's hierarchical structure (i.e., what facets of item content it contains), (b) establishing internal consistency reliability when appropriate, (c) testing of content homogeneity of the facets and ensuring that different aspects of the construct are equally represented in a scale, (d) ensuring that the items discriminate between respondents at the appropriate level of trait intensity, and (e) replication of factor structure across independent samples.

We focus on the fifth criterion, namely, evaluating a proposed factor structure's replicability across samples drawn from the same population and generalizability across samples drawn from potentially different populations (e.g., men and women). When thinking about factor replicability and generalizability, it is useful to consider the concept of test score reliability. Tests do not have reliability, only scores from particular samples do (Lord & Novick, 1968). The reliability of test scores often depends dramatically on sample variability. In a similar way, an instrument's factor structure can change depending on the peculiarities of a particular sample. For example, when groups differ in their factor variances, then the factor correlations are also expected to vary between

groups. In considering generalizability and replicability, this fact must be kept in mind, and researchers need to use statistical methods of factor comparison and equivalence testing that do not confuse group differences in factor structure with group differences in means and variances on the latent variables (Widaman & Reise, 1997).

In replication or generalizability studies there are two basic situations that often arise. In the first situation, data are collected on a given sample and a researcher wishes to evaluate whether the sample factor structure is consistent with a hypothesized structure. This design is often seen in replication studies. In the second situation, data are collected in samples drawn from different populations (e.g., men and women) and the researcher wants to evaluate whether the factor structures are similar or equivalent between groups. This latter situation is frequently referred to as a measurement invariance study in which a researcher wishes to test whether an instrument is measuring the same trait or traits in the same way for two or more groups (Reise, Widaman, & Pugh, 1993). If a factor structure fails to show measurement invariance across groups, then generalizability is compromised and meaningful comparisons across groups on the latent variable are precluded (Widaman & Reise, 1997).

Traditionally, the consistency of a factor structure across samples has been evaluated by computing a coefficient of factor similarity (Everett, 1983; Guadagnoli & Velicer, 1991). Perhaps the most commonly applied index of factor pattern similarity is the coefficient of factor congruence (Wrigley & Neuhaus, 1955). If congruency coefficients are high (e.g., >.90), then a researcher is provided with evidence that the factors are similar (Hurley & Cattell, 1962). However, the reliance on congruency coefficients for determining factor pattern similarity has recognized problems. First, congruency coefficients capitalize on chance and can be large even when random data are fit to a target factor pattern by using oblique Procrustes rotations (Horn, 1967; Horn & Knapp, 1973; Korth & Tucker, 1975). Second, Paunonen (1997) demonstrated that the expected values of congruency coefficients change as a function of various data features such as the number of variables in the analysis and the number of high-loading variables per factor. Third, even if a factor congruency coefficient is high, this does not mean that the same latent variable is being assessed in the two samples. In other words, evidence of factor pattern similarity is not evidence of factor invariance (ten Berge, 1986; Barrett, 1986). For these and other reasons, factor replicability questions are increasingly being addressed with CFA procedures (see Alwin & Jackson, 1981; Byrne & Baron, 1994; Hoyle, 1991).

In CFA, a specific hypothesized factor structure is proposed (including the correlations among the factors) and then statistically evaluated. If the estimated model fits the data, then a researcher concludes that the factor structure replicates. If the hypothesized model does not fit the data, then modification indices (Chou & Bentler, 1990), which are provided by most CFA programs, are used to inform where constraints placed on the factor pattern are causing misfit. These modification indices should be used with extreme caution and should never be relied on as the sole guide to model modification (MacCallum, Roznowski, & Necowitz, 1992) because they often suggest models that perform poorly in cross-validation samples.

In the two-sample situation, multiple-group CFA procedures are used to test for full or partial measurement invariance between the groups (Byrne, Shavelson, & Muthen, 1989; Widaman & Reise, 1997). Multiple-group CFA must be conducted on covariance matrices and never on correlation matrices (Cudeck, 1989), because correlations represent standardized measures of association and thus are expected to differ across groups (because groups may differ in mean level and variance on a latent factor) even when factor pattern matrices (that are based on analyses of covariances) are invariant.

CFA procedures potentially offer researchers a powerful set of tools for evaluating the tenability of a hypothesized factor structure both within and between populations (Floyd & Widaman, 1995). Nevertheless, we caution that there are two potential problems with using CFA to establish replicability or generalizability of a factor structure. First, commonly applied CFA algorithms are based on the assumptions of continuous interval-level measurement on observed variables that are distributed multivariate normal. If a multiple-group CFA is conducted at the item level, especially if the items are dichotomous, interpretation of the statistical results may not be appropriate. New computer programs, such as Mplus (Muthen & Muthen, 1998), that are designed to conduct CFA on dichotomous or polytomous variables may partially mitigate this concern.

A second caution in using CFA procedures is that they seem to work best when the factor structure is clean, that is, when each item loads highly on one and only one factor (simple structure). When Church and Burke (1994) used CFA to evaluate hypothesized factor structures for the NEO–PI–R (Costa & McCrae, 1992) and the MPQ (Tellegen, 1982), for instance, they concluded that "parsimonious personality models are unlikely to meet conventional goodness-of-fit criteria in confirmatory factor analysis, because of the limited simple structure of personality measures and the personality domain itself" (p. 93). For this reason, McCrae, Zonderman, Costa, and Bond (1996) recently argued that CFA tests of hypothesized factor structures can be misleading, and they recommended that factor replicability or generalizability be examined through an orthogonal Procrustes rotation method (Cliff, 1966) and the computation of variable congruence coefficients (Kaiser, Hunka, & Bianchini, 1971). Given the tendency of this methodology to find spurious congruence (Paunonen, 1997), this technique must be used with caution.

Auxiliary Issues

Scale Dimensionality and Scale Score Interpretation

It is axiomatic that a measurement scale should be unidimensional; scale scores should reflect individual differences on a single common dimension (Hattie, 1985; Lumsden, 1961). If a scale is multidimensional (i.e., has multiple correlated dimensions), then not only is the total score more challenging to interpret but different aspects of the scale (e.g., its content facets) may have different correlations with external variables (Zuckerman & Gerbasi, 1977). For these reasons, many researchers have conducted factor analytic studies to demonstrate that an existing scale is not unidimensional as advertised by the original authors. This finding is often quickly followed by a suggestion for the creation of multiple subscales that represent the true nature of the construct domain (see West & Finch, 1997, p. 152, for examples). Although the motivations for these studies are well intentioned, there are

simply too many factor analytic articles demonstrating a scale's lack of unidimensionality, when in fact the scale is perfectly adequate for many applications. Even when a scale is multidimensional a strong common factor can dominate the smaller group factors.

The finding of multidimensionality in essentially unidimensional scales has many sources. First, improper item-level factor analysis of dichotomous items can lead to spurious factors (Waller, 1999). For example, in a factor analysis of phi-coefficients nonsubstantive factors may emerge because of differences in item endorsement rates. Second, standard criteria for determining the number of factors originally developed for scale-level factor analysis typically overestimate the number of factors when used in item-level factor analysis (see Bernstein & Teng, 1989). Third, any scale with a reasonable degree of substantive breadth will include diverse item content. In turn, if such a scale contains two or more items that share variance beyond that caused by the general factor, then small secondary factors (i.e., group factors) can be identified. The only measures that are likely to satisfy the strict mathematical assumptions of unidimensionality (see McDonald, 1981) are measures of conceptually narrow constructs (Cattell & Tsujioka, 1964), such as math self-esteem (rather than general self-esteem) or two-digit addition skill (rather than mathematics ability).

Our main point here is that scales that have a strong common factor should not necessarily be broken up into subscales. This practice creates long, time-consuming batteries and contributes to the seemingly endless proliferation of narrow-band construct measures. More important, Cronbach (1951) demonstrated that if a scale consists of a general factor and several smaller group factors, then it is highly likely that the general factor accounts for the lion's share of the scale score (i.e., total score) variance. In short, the existence of small group factors does not necessarily mean that the total score is a poor indicator of the common trait that runs through the scale items. This question can and should be addressed empirically rather than simply assuming that multiple factors vitiate the scaling of examinees on a common dimension (see McDonald, 1999, p. 89). Instead of routinely calling for essentially unidimensional scales to be fractionated into multiple subscales, we recommend that greater use be made of bi-factor analysis (Gibbons & Hedeker, 1992) and hierarchical factor solutions (see Sabourin, Lussier, Laplante, & Wright, 1990). These procedures explicitly recognize that measures in psychology are often saturated with a strong common dimension even though they may contain several group factors. By using such methods, researchers can discern the predictive validity afforded by a scale's common dimension relative to its group dimensions.

Principal Components Versus Factor Analysis

Common factor analysis and principal-components analysis are not the same thing. Common factor analysis is typically based on a correlation matrix with estimated communalities (percentage of variance a variable shares with the common factors) on the diagonal. The goal of common factor analysis is to extract as many latent variables (factors) as necessary to explain the correlations (common variance) among the items. In factor analysis, the factors are considered to be the hypothetical causes that account for the item intercorrelations. Components analysis, however, is used to create summaries of observed variables, and, as such, principal components are best conceived as the effects rather than the causes of the variable correlations.

Despite the obvious differences between components analysis and factor analysis, the two procedures are often considered equivalent in practice. Goldberg and Digman (1994) best represent a prevailing view in the literature: If the data are well structured, it makes no difference whether a factor or components analysis is used (see also Fava & Velicer, 1992b; Velicer, Peacock, & Jackson, 1982). Although this view is widespread, a recent article by Widaman (1993) has cast doubt on its validity. In particular, Widaman has shown that when the measured variables (items) have low communality and the factors (or components) have few salient loadings, then components analysis and factor analysis results can differ markedly. Specifically, the components analysis overestimates the factor loadings and yields component correlations that are negatively biased (see also Borgotta, Kercher, & Stull, 1986). Interested readers should consult the special issue of *Multivariate Behavioral Research* published in 1990 for more detailed discussion of the factor versus components analysis debate.

References

Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. D. Jackson & E. P. Borgotta (Eds.), *Factor analysis and measurement in sociological research: A multidimensional perspective* (pp. 249–280). Beverly Hills, CA: Sage.

Barrett, P. (1986). Factor comparison: An examination of three methods. *Personality and Individual Differences, 7,* 327–340.

Becker, P. (1998). Special feature: A multifacet circumplex model of personality as a basis for the description and therapy of personality disorders. *Journal of Personality Disorders, 12,* 213–225.

Bentler, P. M., & Yuan, K.-H. (1998). Tests for linear trend in the smallest eigenvalues of the correlation matrix. *Psychometrika, 63,* 131–144.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105,* 467–477.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314.

Borgotta, E. F., Kercher, K., & Stull, D. E. (1986). A cautionary note on use of principal components analysis. *Sociological Methods and Research, 15,* 160–168.

Byrne, B. M., & Baron, P. (1994). The Beck Depression Inventory: Testing and cross-validating a hierarchical factor structure for nonclinical adolescents. *Measurement and Evaluation in Counseling and Development, 26,* 164–178.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Carver, C. S. (1989). How should multifaceted personality constructs be tested? Issues illustrated by self-monitoring, attributional style, and hardiness. *Journal of Personality and Social Psychology, 56,* 577–585.

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

Cattell, R. B., & Burdsal, C. A. (1975). The radial parcel double factoring design: A solution to the item-vs-parcel controversy. *Multivariate Behavioral Research, 10,* 165–179.

Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity versus homogeneity and orthogonality in test scales. *Educational and Psychological Measurement, 24,* 3–30.

Chen, C., Lee, S., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science, 6,* 170–175.

Chou, C., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. *Multivariate Behavioral Research, 25,* 115–136.

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.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319.

Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika, 31,* 33–42.

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

Comrey, A. L. (1967). Tandem criteria for analytic rotation in factor analysis. *Psychometrika, 32,* 143–154.

Comrey, A. L. (1970). *Manual for the Comrey Personality Scales.* San Diego, CA: Educational and Industrial Testing Service.

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

Comrey, A. L. (1984). Comparison of two methods to identify major personality factors. *Applied Psychological Measurement, 8,* 397–408.

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.

Comrey, A. L., & Montag, I. (1982). Comparison of factor analytic results with two-choice and seven-choice personality item formats. *Applied Psychological Measurement, 6,* 285–289.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105,* 317–327.

Davison, M. L. (1994). Multidimensional scaling models of personality responding. In S. Strack & M. Lorr (Eds.), *Differentiating normal and abnormal personality* (pp. 196–215). New York: Springer.

Dielman, T. E., Cattell, R. B., & Wagner, A. (1972). Evidence on the simple structure and factor invariance achieved by five rotational methods on four types of data. *Multivariate Behavioral Research, 7,* 223–242.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent-dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68,* 363–373.

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

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

Fava, J. L., & Velicer, W. F. (1992b). An empirical comparison of factor, image, component, and scale scores. *Multivariate Behavioral Research, 27,* 301–322.

Finch, J. F., & West, S. G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality, 31,* 439–485.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286–299.

Fraser, C. (1986). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Center for Behavioral Studies, The University of New England, Armindale, New South Wales, Australia.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267–269.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423–436.

Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika, 41,* 385–404.

Gibson, W. A. (1960). Nonlinear factors in two dimensions. *Psychometrika, 25,* 381–392.

Goldberg, L. W. (1981). Unconfounding situational attributions from uncertain, neutral, and ambiguous ones: A psychometric analysis of descriptions of oneself and various types of others. *Journal of Personality and Social Psychology, 41,* 517–552.

Goldberg, L. W. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4,* 26–42.

Goldberg, L. W. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26–34.

Goldberg, L. W., & 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. (1970). A comparison of Biquartimin, Maxplane, Promax, and Varimax. *Educational and Psychological Measurement, 30,* 861–872.

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

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68,* 532–560.

Gourlay, N. (1951). Difficulty factors arising from the use of the tetrachoric correlations in factor analysis. *British Journal of Statistical Psychology, 4,* 65–72.

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

Guadagnoli, E., & Velicer, W. F. (1991). A comparison of pattern matching indices. *Multivariate Behavioral Research, 26,* 323–343.

Guilford, J. P. (1975). Factors and factors of personality. *Psychological Bulletin, 82,* 802–814.

Guilford, J. P., & Zimmerman, W. S. (1963). Some variable-sampling problems in the rotation of axis in factor analysis. *Psychological Bulletin, 60,* 289–301.

Gurtman, M. B., & Balakrishnan, J. D. (1998). Circular measurement redux: The analysis and interpretation of interpersonal circle profiles. *Clinical Psychology: Science & Practice, 5,* 344–360.

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin, 69,* 192–203.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

Hofstee, W. K. B., de Raad, B., & Goldberg, L. R. (1992). Integration of the Big Five and circumplex taxonomies of traits. *Journal of Personality and Social Psychology, 63,* 146–163.

Hogan, R. T. (1983). A socioanalytic theory of personality. In M. Page (Ed.), *1982 Nebraska Symposium on Motivation* (pp. 55–89). Lincoln: University of Nebraska Press.

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

Horn, J. L. (1967). On subjectivity in factor analysis. *Educational and Psychological Measurement, 27,* 811–820.

Horn, J. L., & Knapp, J. R. (1973). On the subjective character of the

empirical base of Guilford's structure-of-intellect model. *Psychological Bulletin, 80,* 33–43.

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.

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

Hull, J. G., Lehn, D. A., & Tedlie, J. C. (1991). A general approach to testing multifaceted personality constructs. *Journal of Personality and Social Psychology, 61,* 932–945.

Hurley, J., & Cattell, R. B. (1962). The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science, 7,* 258–262.

John, O. P. (1990). The "Big-Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality theory and research* (pp. 66–100). New York: Guilford Press.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23,* 187–200.

Kaiser, H. F., Hunka, S., & Bianchini, J. C. (1971). Relating factors between studies based upon different individuals. *Multivariate Behavioral Research, 6,* 409–422.

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.

Korth, B., & Tucker, L. R. (1975). The distribution of chance congruence coefficients from simulated data. *Psychometrika, 40,* 361–372.

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

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

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694.

Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989). A regression equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. *Multivariate Behavioral Research, 24,* 59–69.

Loo, R. (1979). The orthogonal rotation of factors in clinical research: A critical note. *Journal of Clinical Psychology, 35,* 762–765.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lumsden, J. (1961). The construction of unidimensional tests. *Psychological Bulletin, 58,* 121–131.

Lykken, D. T., McGue, M., Tellegen, A., & Bouchard, T. J. (1992). Emergenesis: Genetic traits that may not run in families. *American Psychologist, 47,* 1565–1577.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504.

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin, 109,* 502–511.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84–99.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., & Bond, M. H. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology, 70,* 552–566.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34,* 100–117.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27,* 82–99.

Mershon, B., & Gorsuch, R. L. (1988). Number of factors in the personality sphere: Does increase in factors increase predictability of real-life criteria? *Journal of Personality and Social Psychology, 55,* 675–680.

Montanelli, R. G., Jr. (1974). The goodness of fit of the maximum-likelihood estimation procedure in factor analysis. *Educational and Psychological Measurement, 34,* 547–562.

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.

Muthen, L. K., & Muthen, B. O. (1998). *Mplus: The comprehensive modeling program for applied researchers.* Los Angeles, CA: Author.

Ozer, D. J., & Reise, S. P. (1994). Personality assessment. *Annual Review of Psychology, 45,* 357–388.

Paunonen, S. V. (1997). On chance and factor congruence following orthogonal Procrustes rotation. *Educational and Psychological Measurement, 57,* 33–59.

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.

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.

Steinberg, L., & Jorgensen, P. (1996). Assessing the MMPI–based Cook-Medley Hostility scale: The implications of dimensionality. *Journal of Personality and Social Psychology, 56,* 1281–1287.

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

Strube, M. J. (1989). Evidence for the type in Type A behavior: A taxometric analysis. *Journal of Personality and Social Psychology, 56,* 972–987.

Tellegen, A. (1982). *Brief manual for the differential personality questionnaire.* Unpublished manuscript, University of Minnesota, Twin Cities Campus.

Tellegen, A., & Waller, N. G. (in press). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In S. R. Briggs & J. M. Cheek (Eds.), *Personality measures: Development and evaluation* (Vol. 1). Greenwich, CT: JAI Press.

ten Berge, J. M. F. (1986). Rotation to perfect congruence and the cross validation of component weights across populations. *Multivariate Behavioral Research, 21,* 41–64.

Thurstone, L. L. (1937). Current misuse of the factorial methods. *Psychometrika, 2,* 73–76.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.

Tracey, T. J., & Rounds, J. B. (1993). Evaluating Holland's and Gati's vocational-interest models: A structural meta-analysis. *Psychological Bulletin, 113,* 229–246.

Velicer, W. F., & Fava, J. L. (1987). An evaluation of the effects of variable sampling on component, image, and factor analysis. *Multivariate Behavioral Research, 17,* 193–209.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3,* 231–251.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25,* 1–28.

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.

Waller, N. G. (1995). *MicroFACT 1.0: A microcomputer factor analysis program for ordered polytomous data and mainframe sized problems.* St. Paul, MN: Assessment Systems Corporation.

Waller, N. G. (1999). Searching for structure in the MMPI. In S. Embretson & S. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 185–217). Mahwah, NJ: Erlbaum.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua.* Thousand Oaks, CA: Sage.

Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality, 64,* 545–576.

West, S. G., & Finch, J. F. (1997). Personality measurement: Reliability and validity issues. In R. Hogan, J. Johnson, & S. Briggs, *Handbook of personality psychology* (pp. 143–164). San Diego, CA: Academic Press.

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

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. Bryant, M. Windle, & S. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Wiggins, J. S. (1980). Circumplex models of interpersonal behavior. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 1, pp. 265–294). Beverly Hills, CA: Sage.

Wilson, D., Wood, R., & Gibbons, R. D. (1984). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Mooresville, IN: Scientific Software.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods, 1,* 354–365.

Wrigley, C. S., & Neuhaus, J. O. (1955). The matching of two sets of factors. *American Psychologist, 10,* 418–419.

Zuckerman, M., & Gerbasi, K. C. (1977). Belief in internal control or belief in a just world: The use and misuse of the I-E scale in prediction of attitudes and behavior. *Journal of Personality, 45,* 356–379.

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.

Received July 9, 1999
Revision received October 5, 1999
Accepted October 25, 1999 ■