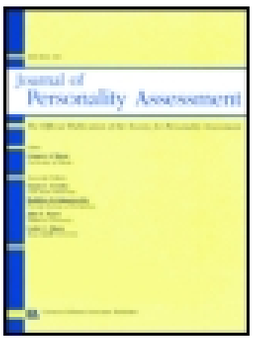

Routledge
Taylor & Francis Group

Journal of Personality Assessment

ISSN: 0022-3891 (Print) 1532-7752 (Online) Journal homepage: https://www.tandfonline.com/loi/hjpa20

Exploratory Factor Analysis: Its Role in Item Analysis

Richard L. Gorsuch

To cite this article: Richard L. Gorsuch (1997) Exploratory Factor Analysis: Its Role in Item Analysis, Journal of Personality Assessment, 68:3, 532-560, DOI: 10.1207/s15327752jpa6803_5

To link to this article: https://doi.org/10.1207/s15327752jpa6803_5

JOURNAL OF PERSONALITY ASSESSMENT, 1997, 68(3), 532–560

Exploratory Factor Analysis: Its Role in Item Analysis

Richard L. Gorsuch
Graduate School of Psychology
Fuller Theological Seminary

The special characteristics of items—low reliability, confounds by minor, unwanted covariance, and the likelihood of a general factor—and better understanding of factor analysis means that the default procedure of many statistical packages ("Little Jiffy") is no longer adequate for exploratory item factor analysis. It produces too many factors and precludes a general factor even when that means the factors extracted are nonreplicable. More appropriate procedures that reduce these problems are presented, along with how to select the sample, sample size required, and how to select items for scales. Proposed scales can be evaluated by their correlations with the factors; a new procedure for doing so eliminates the biased values produced by correlating them with either total or factor scores. The role of exploratory factor analysis relative to cluster analysis and confirmatory factor analysis is noted.

Exploratory factor analysis (EFA) has been widely used as a technique to develop scales and subscales. For example, using the terms "factor analysis" and "item" for a *Psychological Abstracts* search for 1990–1995 suggested that 400 to 500 studies used factor analysis with items; of these about 75% used EFA.

The several situations in which EFA may be helpful with items are explored in the first section of this article, where its relation to classical, simple scale development is outlined. A brief comparison of the role of EFA in relation to cluster and confirmatory factor analysis (CFA) is also made.

The second major section of this article, "Decision Making for Item Factor Analysis," addresses the practical details of the use of factor analysis with item data. It begins with a discussion of the special problems of factoring items and why the standard factor analytical procedure—principal components extraction with Varimax rotation (the standard default option in most computer packages and referred to as "Little Jiffy")—can produce misleading results with items. This section then speaks to several practical questions: What type of sample of items?

How large a sample size? Is component or common factor analysis best? How many factors? Which rotation? How, then, are items selected for scales? We shall find that the last 20 years have clarified some of these questions so that some answers given in, for example, Gorsuch (1974) and Nunnally (1967) are now out-of-date and can no longer be recommended. This section is designed to merge the worthwhile older conclusions with more recent knowledge. The last section of this article is a set of conclusions.

FACTOR AND ITEM ANALYSES

The goal of item analysis is to select those items that are most related to the construct. This goal is aided by evaluating how each item relates to its own construct, as well as how it relates to other associated or similar constructs. Each construct is understood to be unidimensional. The desired result is that all items measuring the same construct are scored together to give the best estimate of each person's score on that construct.

The purpose of factor analysis is to identify the fewest possible constructs needed to reproduce the original data. Mathematically, it seeks the set of equations that maximize the multiple correlations of the factors to the items. The equation for each item is:

$$i_1 = p_{1A} A + p_{1B} B + p_{1C} C + \ldots + u_1 \qquad (1)$$

where i is the response to Item 1; A, B, and C are the factor scores (the "..." indicating there may be more factors), the ps are the weights used to best reproduce the original standardized Item 1 responses, and u is the residual for Item 1 when the fit is not perfect. There are as many equations as there are items. Note the emphasis on reproducing the original item responses; all equations for factoring data are derived from the aforementioned equation (Gorsuch, 1983) and so are directly linked to the original data. This is an important theoretical difference from, for example, cluster analysis.

The relations of each variable to each of the factors tells whether the item is related to only one of the factors (constructs) or to more than one. Those items most clearly related to only one factor can then be recommended as a scale for the construct underlying that factor. Using the results of the factor analysis helps achieve the goals of item analysis in several ways. First, it provides information on the constructs. Is there only one? If so, then a single factor will, assuming no methodological problems, result. Is there more than one construct among the pool of items? If so, there will be one factor for each construct. The correlations of each item with each factor allow for selection of the best items.

Common methods for evaluating an item are special cases of factor analysis. Correlating each item with the total score from the set of items and then selecting those items with the highest item-total correlations is a special case of factor analysis. The total score is the general factor—technically a centroid factor with the items weighted as a function of their standard deviations—and the item-total correlations are the *p*s in Equation 1.

As an illustration of item analysis with one factor, analyses of the first six items of the Beck Depression Scale (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) were computed from Lee's (1995) data. The first six items were chosen so that the total score from all items could be considered the item domain. Item domain correlations are therefore available against which to compare the several procedures. These correlations are in the first column of Table 1.

The second and third columns of Table 1 give the usual item-total and item-remainder correlations. The former are, as expected, inflated because each item is correlated with its own error as well as the common variance among the items. The item-remainder are too low because only five items do not give a good, reliable estimate of the domain.

The last two columns of Table 1 give component and common factor correlations. Component analysis gives inflated loadings (see also Snook & Gorsuch, 1989), and so is not recommended. Common factor analysis gives the best estimate of the domain correlations.

Another special case in traditional item analysis occurs when a set of ability items, for example, are scored to measure two constructs, verbal and numerical abilities. To evaluate the items, each item is correlated with the sum of the verbal items and the sum of the numerical items. The items selected for a revised scale are those that correlate well with one of the scores but not the other (e.g., verbal items correlating with the verbal scale score but not the numerical scale score). This is technically a multiple group factor analysis with one centroid factor for each

TABLE 1
Item Analysis Correlation by Traditional and Factor Analyses (One Dimension)

				Factor Analysis	
Item	*Item-Domain*	*Item-Total*	*Item-Remainder*	*Component*[a]	*Common*
1.	.55	.63	.45	.65	.55
2.	.49	.51	.36	.55	.44
3.	.56	.69	.46	.68	.58
4.	.64	.71	.49	.70	.61
5.	.45	.59	.41	.61	.49
6.	.44	.61	.35	.55	.43

[a]Also known as "Little Jiffy," and the default in most statistical packages.

construct. Equation 1 could be solved by using multiple regression to estimate the item scores from the verbal and numerical factor scores.

Both of the aforementioned examples defined the factors by adding together a preidentified set of items (e.g., the verbal items or the numerical items). In traditional item analysis, the definition of the factor is only as good as the judgment of the investigator as to which items are to be added together. At this point, factor analysis offers an improvement. Instead of basing the factors on investigator judgment, it bases each factor on a set of highly correlated items. Hence, misjudgments about what items measure are less likely to distort the operationalization of the construct. Additionally, new constructs may emerge that the investigator did not realize were being measured. Such new constructs might be variations of the original constructs by which the item pool was built or contaminants that may otherwise have gone unnoticed. In either case, the investigator needs this information to properly evaluate the items.

For example, most scales are designed to measure just one construct, to be homogeneous. It is often incorrectly assumed that a measure of internal consistency (e.g., coefficient alpha) provides a means to address this question. However, it is easy to design a set of items that measure more than one independent construct and yet produce a coefficient alpha of .9 or better when scored as if they measured only one construct. Factor analysis provides a better means of examining scale homogeneity.

In traditional item analysis, potential scales are evaluated by trying out subsets of items. Thus the three items with the highest correlations are scored and correlated with the total item set, then the best four items, and so on. However, this common procedure has the same problem as item-total correlations. The fact of the item being in both the scale and the total produces an inflated correlation. Consider an extreme case where the item-domain correlations are all 0.0; a 3-item scale correlates .72 with the total from six items and a 4-item scale correlates .82. The latter is higher than the former solely because the fourth item is included in the total score, and all correlations appear important because of the built-in bias to which item-total and scale-total correlations are heir.

A new factor analytic procedure (Gorsuch, 1997) is now available for the problem just noted. It uses the same common factor approach that makes the item-common factor correlations better estimates of the item-domain correlations than either item-total or item-remainder correlations, but the correlations of proposed scales with the factor(s) are unaffected by any item being both in a proposed scale and in the factor analysis. Items too exploratory to be in the original total/remainder scores can be evaluated by this extension analysis without the bias favoring the items in the total score.

EFA can also be used with the confirmatory and cluster techniques discussed later. Roth and Roychoudhury (1991) did so and noted that many of the conclusions of factor analysis and clustering were the same, but they had more confidence that the results were independent of the methodology by including the factor analysis.

CONFIRMATORY FACTOR ANALYSIS

CFA requires clear predictions as to which factors exist, how they relate to the variables, and how they relate to each other. Without such predictions, exploratory analyses are needed. Of course, factor analysis can be used as a purely exploratory technique. This can occur when a set of items are developed to represent an area of interest with the sampling being broad and without specific subareas. The question addressed is whether one scale, one scale that can be subscaled, or a set of scales is needed in the area. Techniques recommended later (but not the "Little Jiffy" solution) can readily decide between such options.

EFA can also be used as an adjunct to CFA. Exploratory methods may be valuable in a preliminary study to focus hypotheses for the confirmatory analyses. It may also be useful as a follow-up in a confirmatory structural modeling analysis. Because of the stringent requirements for excellent hypotheses and data (see Gorsuch, 1996), many confirmatory structural model analyses fail to provide clear results—the chi-square of the residuals is still significant, an improper solution occurs, or the correlations between factors are too high. Often adjustments are subsequently made to the structural equations model to "cure" the problem, but such exploratory adjustments are dangerous for two reasons. First, the probabilities are no longer accurate but can be much smaller than they should be due to capitalization on chance. Second, MacCallum (1986) found that with *N*s of 300 and only one or two adjustments needed, only half of the adjustments were toward the population values. The results were worse with *N*s less than 300. (Rumor has it that most published CFAs have had adjustments made and so should have no significance levels reported; most of these adjustments do not seem to be reported.) In the situation when the CFA does not work, an EFA is an appropriate alternative to attempting to adjust the confirmatory model.

Note that confirmatory analysis comes in several types. An argument can be made that confirmatory multiple group factor analysis is one of the most useful approaches for items (Bernstein, Jaremko, & Hinkley, 1994, 1995; Bernstein & Keith, 1991; Gorsuch, 1974, 1983; Nunnally & Bernstein, 1994). It is much less prone to the problems noted earlier with confirmatory structural equations modeling.

EFA is also used for confirmatory purposes. First, it is conceptually a "multi-tailed" test as compared to CFA being conceptually a "one-tailed" test. If the same factors appear again in another exploratory analysis, they would surely—except for the limitations of structural modeling CFA—occur in a confirmatory analysis. Second, the test of whether a particular factor replicated is given only indirectly in most confirmatory structural modeling factor analyses. Generally there is no way of determining the degree of replication of any one factor. But the correlation of each exploratory factor as found in the first EFA with each factor in a second exploratory analysis is readily calculated, tested for significance, and interpreted (Gorsuch, 1996, in press).

CLUSTER ANALYSIS

Although cluster analysis is sometimes used instead of factor analysis, the limitations of cluster analysis in comparison to factor analysis are several. First, it has no direct link to the raw scores like factor analysis has (see Equation 1). Instead it is generally—there are many variants—a method of grouping objects that have high coefficients of similarity. The coefficients could be correlations or distance measures. This provides versatility, but also moves a step away from the data. Earlier, the basic equation of factor analysis was shown to be directly based on the item scores, just as is multiple regression, analysis of variance, and other common statistical procedures.

Second, cluster analysis has no built-in method for identifying the most parsimonious set of clusters. A four-cluster solution may be found when there are only three factors, making the data look more complex than would a factor analysis.

Third, the usual requirement that an object be either in or out of a cluster oversimplifies most results. Instead, a factor analysis provides the correlation of each item with each factor, and the complexity of the item (as determined by its pattern of loadings across factors) is useful in item rewriting or selection (see later). This information is lost in a cluster analysis.

DECISION MAKING IN ITEM FACTOR ANALYSIS

Special Problems of Item Analyses

Given that traditional item analyses are special cases of factor analysis, why not just use the same methods of factor analysis for items as are used for scales? This is certainly done, but the methods of factor analysis were developed for analysis of scales. Items have different properties that require special factor analytic techniques to avoid misleading results, as is shown in following examples (cf. Cattell, 1957, for an early discussion of the problems.)

Items differ from scales in four ways that influence a factor analysis. The differences and their impacts are:

1. Items have lower reliabilities than scales. This is a major reason why we develop scales. Adding together a set of items with low reliabilities averages out the error of the items while combining the shared variance. For example, consider a 20-item scale with an internal consistency reliability of .8. Using the Spearman–Brown correction formula to "work backwards," the reliability of a 1-item scale is estimated to be .17! A population correlation of .9 for the latent measure perfectly measured would have a population correlation of .81 for the scale, but only .37 for the "average" item. If the N were 100, the observed correlation for the scale could be .7 and .9 in two samples, whereas the same observed correlation for

the average item would, due to lower reliabilities, be .27 and .47. The variation of observed correlations across typical samples would be approximately 12% for the scale but approximately 27% for the item, for a *N* of 100. This means item correlations are lower and more difficult to work with than scale correlations. Because items correlate lower than scales, the factors are weaker with a higher percentage of the covariations being error.

2. Items often contain confounding variance in addition to the construct being measured. For example, one item of the BDI addresses the symptom of crying. This item thus contains possible gender specific information that may be unrelated to the question of whether the person is depressed. Almost every item of every scale has a word or phrase that may be interpreted by someone in an idiosyncratic manner. Scales avoid these minor confounds by adding across the confounds found in multiple items to average them out.

3. Item distributions often differ from each other, in part deliberately. These differences may be in either skew or kurtosis. To measure all variations of a construct well, it is necessary to have items with means that vary across values of the response range. The items with means at the extremes usually have skewed distributions. The skewed distributions impact the correlations because a correlation can only be high and positive if the two items being correlated have the same distributions. Different distributions among the items, therefore, reduce the correlations among the items. Scales sum across items, thus averaging the distributions. As a result, scales generally show higher intercorrelations than items do.

Factors produced by different distributions have been traditionally called *difficulty factors*. For example, in analyses of ability scales, it is commonly found that the least difficult items and the most difficult items form separate factors. The factors result from the fact that easy items are positively skewed (and difficult items negatively skewed) and so correlate more highly than they do with difficult items not sharing that skew.

4. Item scores are almost always a set of ordered categories rather than a continuous value. The categories may be "correct" or "incorrect," as in ability tests, or a range of responses as in a 5- or 7-category response scale ranging from *strongly disagree* to *strongly agree*. Because there are not a large number (e.g., 100) of points on the response scale, item scores can only reflect major differences among individuals on the underlying construct. Any 50 people with an item score of 5 will still have some differences on the underlying construct, and another item may distinguish between them. These two items then correlate lower than they would be if both were measured as continuous variables. (Because total scale scores reflect the sums of the component item scores, scales have many more possible scores and so function more like continuous variables.)

The number of response options has an impact on item distributions. When only two responses are allowed—correct or incorrect, or yes/no—the skew can be great.

This problem is not eliminated by the use of tetrachoric correlations, which introduce new problems. (See Comrey & Lee, 1992, and Nunnally & Bernstein, 1994, who recommend regular correlations over tetrachoric or biserial correlations.) With multiple-point response scales (e.g., 7 points ranging from *strongly disagree* to *strongly agree*), item means can vary with a greater likelihood of a less skewed distribution, all other things being equal, but skew problems often exist.

Bernstein and Teng (1989) discussed the problems of item distribution and item categorization and their implications for item factor analysis. They also discussed another potential problem. As more categories to the response option are added, a distribution difference that was previously obscured may be produced and reduce the correlation between items. For example, one item may have a 50/50 split by a normal distribution being cut in half. Another may have a 50/50 split by being cut at the median but be heavily skewed or have a nonnormal kurtosis. The two category versions of the items can correlate 1.0 but multiple category items (and continuous items) cannot correlate 1.0. Nevertheless, more categories of response do provide more information and can be generally expected to give higher correlations.

The impact of different distributions in item analysis can be easily demonstrated by factoring a perfect Guttman scale of 20 items using a Yes/No answer format. In such a scale, each item does exactly what it is supposed to do: it separates the respondents into two halves based on the underlying latent variable. Each item is perfectly reliable. The items also form a perfect Guttman scale in the sense that (a) each item has a different mean so that they, as a set, measure the entire range and (b) they are cumulative (i.e., if a person answers "Yes" to Item 3, then that person will have answered "Yes" to Items 1 and 2 also). The data set consists of five people at each of the possible scores, from 0 (no item answered "Yes") to 20 (all items answered "Yes"), for a total N of 105.

The results of the standard "Little Jiffy" analysis are presented in Table 2. The items are ordered from the highest percentage to the lowest percentage answering "Yes." Four uncorrelated factors were extracted despite the fact that each item measures exactly the same thing! The problem is that the conditions needed for "Little Jiffy" to be applicable were ignored. The "Little Jiffy" procedure is applicable to highly reliable variables with the same distributions with strong correlations from several uncorrelated domains. The different distributions of items in a perfect Guttman scale result in items of similar distributions correlating more highly together than they do with items of dissimilar distributions. These different patterns of correlations were interpreted by "Little Jiffy" to mean that different factors were being measured. Note that each item loads the same factor as other items with similar means and therefore similar distributions, producing technical rather than substantive factors.

Each of the following sections explain how to take item characteristics into account when conducting an item factor analysis. General presentations of factor

TABLE 2
"Little Jiffy" of a Perfect Guttman Scale

	Factors			
Items	*1*	*2*	*3*	*4*
1.	.82	.06	.13	−.02
2.	.87	.26	.11	.03
3.	.78	.46	.09	.07
4.	.64	.63	.09	.10
5.	.49	.76	.12	.12
6.	.35	.83	.18	.12
7.	.24	.86	.27	.11
8.	.15	.84	.38	.09
9.	.10	.78	.49	.08
10.	.07	.69	.61	.08
11.	.06	.58	.71	.09
12.	.06	.47	.80	.12
13.	.08	.35	.85	.18
14.	.10	.24	.86	.27
15.	.11	.15	.83	.38
16.	.11	.09	.74	.52
17.	.09	.07	.61	.66
18.	.06	.07	.44	.80
19.	.02	.10	.23	.88
20.	−.03	.12	.04	.82

analysis that are consistent with this exposition have been presented by Gorsuch (1983) and Nunnally and Bernstein (1994).

Sample of Items

From the aforementioned discussion, it is apparent that items that are more like scales are better for factoring. They should be highly reliable, measure only one construct, and have the same distributions as produced by a continuous response scale. But if we had perfect items, who would need scales? The point is that the items should be selected to be closer to the ideal rather than farther from it. The easiest component to manipulate in this regard is the scale for responses. With the possible exception of ability measures, or to meet needs for ease of administration or scoring, response options limited to two or three categories should be replaced with seven or more response categories. This is particularly important if the items are selected, as they should be, so that the means are systematically different.

It is always wise to pretest a set of possible items; if an item shows a strange distribution, it can then be rewritten to avoid the response category/distribution

interaction noted by Bernstein and Teng (1989). The item selection has a major impact on what type of rotation is appropriate, as discussed later.

Not all the items need be in the factor analysis itself. Only those items that clearly help define the domain are in the factor analysis. If desired, the analysis may be extended to additional items not in the factor analysis. Extension analysis (Gorsuch, 1997) gives the correlations of the nonfactored items with the factors. For example, the factor analysis may be restricted to the classical set of items used in past literature to evaluate replicability of past factors. Then a new set of items are also evaluated to improve measurement by replacing some of the original items or by adding more items to increase the length of the scales. The new items would be in the extension analysis.

Sample of Respondents

The two issues to be confronted with regard to a sample of participants are the type of respondent and the size of the sample. The sample should consist of people similar to those with whom the scale will be ultimately used. For evaluating anxiety in everyday life, the sample should be of people in everyday life. For evaluating depression in patients being treated in an outpatient mental health clinic, the sample should consist of mental health clinic outpatients. Lee (1995), for example, sampled Korean students because the scales he evaluated are frequently used in student counseling centers.

Any analysis is enhanced if the sample has a wide variety of people. There should be many who would score low on the proposed scale(s) and many who would score high. The sample need not closely represent any clearly identified population so long as those who would score high and those from that population who would score low are well represented.

The sample size needed was in former times given as a function of the number of items (e.g., 10 cases for every item). This was a recommendation proposed largely out of ignorance rather than theory or research. We now have information for EFA from several sources suggesting that the "10-case-per-item" rule sets the sample size above the minimum needed. Cattell (personal communication, 1950) found the same factors in a small sample size as in a larger sample size even when the smaller sample size was less than twice the number of variables. Guadagnoli and Velicer (1988) used a simulation study to evaluate the stability of results across several conditions. They concluded that a N of 150 was sufficient for up to 40 or 50 variables (they did not investigate more variables), which is a ratio of 3 to 1 rather than the 10 to 1 formerly suggested. Reddon (1990) evaluated the sample size necessary to be able to extract any factors from a correlation matrix and reached a similar conclusion.

The general current conclusion is that the sample size needed is a function of the stability of a correlation coefficient without any correction needed for the number of variables, at least for situations of up to 40 or 50 variables. The stability

of a correlation is a function of the square root of the sample size. To reduce by half the impact of error, the square root of the sample size would need to double. Thus if the N were 100, the error would be halved with a N of 400. To halve it again would require a N of 1600. A better sense of the needed precision can be gained from considering the expected size of the correlations among items. Note that if two items both are loaded by the same factor at .5, then their expected correlation is .25. Hence a sample size for which a .25 would be highly significant should be chosen; this is approximately 125. This is similar to the N of 150 that the aforementioned studies suggest is probably sufficient if there are clear sets of highly significant correlations.

The lower the expected correlations, the greater the sample size needed for the factors to stand out from the error variance. If loadings of .4 are of interest, then the expected correlation between two variables loaded by the same factor is .16 and a N of 300 is needed. For most item analyses of previously untested items, the traditional item analysis recommendation of 300 is also a good one for item factor analysis.

What of item analyses with more items than 50? The crucial issue here is not so much the number of items but rather the number of decisions that will be made and the level of selection at which they will be made. Picking the 10 items with the highest correlations with a factor has little capitalization on chance with a total item pool of only 11, but considerable capitalization on chance with an item pool of 100 from which to choose the 10. This requires an increase in the sample size; large sample sizes always help.

There is a check on whether the sample size is clearly too small for the factors extracted: Bartlett's significance test (Gorsuch, 1983). It is based on the eigenvalues of the correlation matrix and tests whether a residual matrix is significant after a given number of factors has been extracted. Any factor of interest should be highly significant by this test. Most item factor analyses with at least 200 to 300 cases can be assumed to be significant but the test should be computed on smaller sample sizes as a safeguard. Borderline sample sizes should not be rejected out-of-hand, for capitalization on chance is influenced by other factors than just the sample size. For example, if the item-factor correlations are high, a smaller sample size will give higher quality results than if the correlations are low. Both Type I and Type II error rates should be considered important. The final criterion for the needed sample sizes, however, is that of replicability. With borderline sample sizes, decent cross-validation of the factor structure needs to be demonstrated.

Component Analysis Versus Common Factor Analysis

Within the factor analysis paradigm one can distinguish between principal component and common factor analysis. Gorsuch (1983) proved that they vary only in that common factor analysis has the last element of Equation 1, an error term, and

principal component analysis does not. Using the latter assumes that the variables are conceptualized as reproduced perfectly by the factors. Perfect reproduction generally means that the variables are almost perfectly reliable and correlate highly with at least one other variable. From our theoretical analysis of items, we concluded that items are unlikely to be highly reliable and may contain more than one factor. Error is also introduced by distribution and categorization problems. Hence, including the last element of Equation 1 is appropriate for item analysis, and so common factor analysis is preferred.

There is considerable support for using principal component analysis despite its theoretical inappropriateness (cf. the special issue of *Multivariate Behavioral Research*, Vol. 25, No. 1, 1990). The reasons appear to be four-fold. First, common factor analysis has a technical problem: There is no unique set of factor scores that can be calculated from a common factor analysis, but there is from a component analysis. This is less relevant to item factor analysis because the purpose is to select a subset of items to score each factor, and that does give a unique set of factor scores for common factor analysis.

A second reason principal components are occasionally recommended is that they are easier to compute. Of course with modern personal computers, this is rarely a problem. But such was not the case when "Little Jiffy" was introduced by Kaiser (personal communication, 1960). The computer he was using was Illiac I, which was considerably smaller and slower than the first of the desktop personal computers. But with the second generation of computers—only somewhat larger and faster than the original personal computers and considerably smaller and slower than any personal computer now manufactured—Kaiser recommended common factor analysis (Kaiser, 1970). The common factor analyses used as illustrations in this article took very little time—perhaps 1 min each—on a personal computer.

Third, some comparisons of principal component and common factor analysis use maximum likelihood common factors with highly iterated communalities, a type of common factor analysis that shows an occasional problem. But this is a type of common factor analysis I have never recommended (Gorsuch, 1974, 1983, 1990). These problems do not occur with principal axis common factor analysis with two or three iterations for communalities. Empirical analysis shows that such common factor analysis produces factor loadings not significantly different from population values whereas component analysis gives inflated loadings (Snook & Gorsuch, 1989). Hence, this reason for preferring principal components appears to be outdated.

A fourth reason is that the results are about the same regardless of whether principal component or common factor analysis is used. This is true, but only when the variables correlate highly with each other or there is a large number of variables per factor. Item factor analysis always has a hard time meeting the first condition, but if 50 or so items are being factored with only five or so factors rotated, it may meet the second condition.

Note that Equation 1 is the general case, and principal component analysis is a special case in which the u is zero. Thus principal component analysis is a special case of common factor analysis. If a component analysis is appropriate, common factor analysis will produce it. But the reverse is not true: without the u in the equation, the principal component analysis procedures force a common factor situation as much as possible into a component model. Gorsuch (1983) provided examples where this forcing distorts loadings (also see Snook & Gorsuch, 1989). No such conditions exist for common factor analysis.

Consider the example in Table 3. All the items have some significant correlation except for the last item, which is a random variable and has no significant correlation with any of the four variables. In the common factor analysis, it is obvious that the item should not be used. In the principal component analysis it has a high loading of .60, leading to the conclusion that the item should be used. Because the loading is based on nonsignificant correlations, however, it will neither replicate nor help that scale.

An analogy to the choice between principal component and common factor analysis is item-total versus item-remainder correlations. The item-total correlation is inflated because the item's own error variance is part of the total score; item-remainder correlations are less misleading. Mistakes will seldom be made with item-remainder correlations, but will occasionally occur with item-total correlations. In like manner, principal component analysis will occasionally produce an erroneous judgment as in Table 3, but such errors seldom occur with common factor analysis.

Number of Factors

A major decision affected by the special conditions of item factor analysis is the number of factors. This is particularly relevant with "Little Jiffy" because it uses

TABLE 3
Example of Component Analysis Making a Nonsignificant Item Appear Significant

	Component		*Common Factor*	
Item	*1*	*2*	*1*	*2*
1.	.08	.82	.14	.58
2.	.08	.82	.14	.58
3.	.70	.33	.57	.28
4.	.81	.00	.58	.06
Random variable	.60	.01	.32	.09

Note. $N = 90$. The shrunken squared multiple correlation for the random variable was .04. None of its correlations with the other items were significant.

the "roots greater than 1" (abbreviated as R > 1) criterion. For this criterion all of the characteristic roots of the correlation matrix are extracted. The number of these roots greater than 1.0 is the number of factors to extract. The rationale for R > 1 is based on the roots of a population correlation matrix with no factors. In that case, all roots will be 1.0. If there is a factor in the population matrix, the root for the factor will be greater than 1.0; because the roots must total to the number of variables, the other roots with no factors will be less than 1.0. So the number of factors in the population is the number of roots greater than 1.0. This analysis is for the population matrix; with no factors, all correlations are 0.0 and all roots 1.0. But with a sample, there are chance correlations. In this case, the average number of R > 1 will be one half the number of variables (i.e., $.5v$ where v is the number of variables). Hence the use of the R > 1 criterion produces many factors for a set of random variables.

In a well-designed factor analysis, the factors are clearly larger than random factors and so their roots are clearly larger than the other roots. Let's assume that we have 10 variables per factor, so the population matrix would have $.1v$ roots meeting the R > 1 criterion. But as the correlations among the variables drop due to poor quality variables, the number of roots is a comprise between random data, which has $.5v$ R > 1 factors, and the population number of factors, which is $.1v$ R > 1 factors for this example. While the situation is more complex than can be described here, in general the observed R > 1 ranges from being accurate to being an overestimate of the number of factors. In simulations the R > 1 criterion has been found to give reasonable results with good data but to overestimate the number of factors with data having lower correlations. As noted earlier, item factor analyses have lower correlations than well-designed studies of scales. This is due to their lower reliabilities, among other problems. Hence, the R > 1 criterion will produce too many factors for items.

The problem of the R > 1 criterion was obvious in early research using item factor analyses. Gorsuch (1968) reduced the number of factors by running an initial analysis with the R > 1 criterion, rotating the factors to Varimax, and counting the number of factors that had at least three items salient on the factor. (*Salient* is defined as a loading that is greater than |.4| and is the highest loading for the variable.) This was used as an estimate of the number of factors. Also some random variables were included, and the factoring was stopped when a factor appeared with a high loading by a random variable. Gorsuch (1974, 1983) continued to recommend counting the number of factors with three or more salient loadings to establish the number of item factors.

The Gorsuch recommendation to restrict the number of factors to those having three salient variables was, it can now be seen, useful for two reasons. First it helped correct for the R > 1 tendency to indicate too many factors due to the low correlations among items. But it also served another goal as well. One of the problems with items noted earlier is that they have reliable variance that is unique

to a couple of items but not to a general construct. For example, two items may be about one's family life, and so vary together in addition to the variation from the construct being measured. These items would therefore load a small factor together, but it would be a trivial factor. (*Trivial* is defined as lacking salient variables, both by few items loading the factor and also by most of its items having higher loadings with other factors.) Keeping only the larger factors eliminates these trivial factors that occur among items.

Other criteria may also be useful for item analysis. These include parallel analysis, the scree test, and, if the sample is large enough, separate factoring in two halves of the sample to evaluate the number of factors that replicate across random samples (each subsample would need to have at least 150 cases). No systematic evidence is available on the use of these procedures for estimating the number of item factors.

Additional evidence supports two general conclusions that appear applicable to item factor analysis. First, it is better to overfactor than underfactor (Fava & Velicer, 1992; Wood, Tataryn, & Gorsuch, 1996). Thus, keeping an additional factor or two is not likely to be a problem. Second, extracting too few factors can radically change one or more factors while extracting an additional factor when "in the right range" leaves the earlier ones unchanged.

In practice, it is recommended that several factor analyses be computed. The first one would be based on $R > 1$ and then, in the rotated solution, the number of trivial factors noted. Then the analysis is rerun with the trivial factors dropped. The two solutions would be compared. If any factor in the second analysis changed dramatically, then the analysis would be redone with an additional factor or two. With contemporary computers, this can be done in a half hour or so of the investigator's time.

Rotation

Simple structure rotation. The basic principle of factor rotation is *simple structure*. This seeks to minimize the number of variables that load on a factor—and so keep the factor simple—and to minimize the number of factors both loading the same variables—and so keep the solution simple. This means that the ideal variable loads on only one factor. And the rotation procedures are designed to rotate to a solution in which each factor has several loadings for variables not loading other factors. Hence programs rotating to simple structure are most appropriate when the items clearly load several different factors. The rotation procedures work well, perhaps too well. They almost always produce a solution in which the factors each have a set of items loading them high that do not load other factors. Even if a solution exists in which all items load the same factor, that solution will be avoided if at all possible so that all factors can also have high-loading variables that do not load another factor.

Many factor analyses of scales use a sample of scales for which simple structure rotation is appropriate. Each scale is expected to correlate high with some scales but low with the other scales in the analysis. This is appropriate for simple structure rotation.

Items, however, may not be sampled so that each item is expected to correlate with only a few other items. Instead, many item sets consist of items expected to correlate well with all the other items, not just some. Items on the BDI are all expected to measure depression, and so all items should correlate with every other item. Hence, the sampling of items is radically different than the sampling that fits simple structure rotation. When the sampling does not match simple structure, the programs do the best they can to produce a simple structure fit. Therefore the results look like simple structure even if every item primarily measures one factor—such as depression—but also measures some another trivial factor.

A good example of what simple structure does is shown with the perfect Guttman scale in Table 2. Instead of all loading the same factor—which they do in the underlying model—they load four factors that show simple structure. Our interpretation would go astray here unless we remember this is a function of the rotation technique, not of the items' substantive content.

Overcoming simple structure bias. How, then, can we test if the items do all relate to one theoretical factor? There are two methods in theory but only one in practice. In theory the initial unrotated factor can be examined. It almost always produces a general factor. However, it is as biased toward a general factor as simple structure rotation is biased away from it. It also ignores the other factors loading these items which may give information about biases among the items. Instead of allowing these to shift to separate factors, the first unrotated factor has as much of all the variance, including the bias variance, in the general factor as is possible. As a result, this method is seldom used in practice.

In practice, the method to test for a general factor underlying the item set is to extract one or more higher order factors. This begins by avoiding any rotation that restricts the factor intercorrelations (note that Varimax rotation restricts them to being uncorrelated). The correlations among the rotated factors are used as the correlations for a second factor analysis. The original factors—called primary factors—thus become the "variables" for the second, or higher order, factor analysis. This procedure tests whether the primary factors are correlated and, if so, how those correlations are structured. Items can then be correlated with the higher order factors by extension analysis (Gorsuch, in press). If most items correlate well with the higher order factor, then it can be considered a general factor.

Consider, for example, the perfect unidimensional Guttman scale discussed earlier that produces four primary factors instead of one. The results of an item factor analysis leaving the correlations among the factors unrestricted is presented

in Table 4. These factors correlate, and a higher order factor can be extracted. All of the four primary factors are loaded by the higher order factor and all items correlate with it. Our interpretation is simple: All items relate to the single, higher order factor, which is the general factor we know underlies all the items. In addition, different distributions cause subsets of items to correlate together as seen in the primary factors. This then produces a complete understanding of how the items relate to each other.

The importance of a higher order factor analysis to evaluate whether there is a general factor is also shown with analyses of anxiety. Gorsuch (1966) rotated the previously found factors of the Test Anxiety Questionnaire without, as had been

TABLE 4
Item Factor Analysis of a Perfect Guttman Scale

	Factor				
	Primary				
	1	*2*	*3*	*4*	*Higher Order*
Primary factor					
1.	1.00	.52	.27	.18	.64
2.	.56	1.00	.57	.27	.81
3.	.27	.57	1.00	.58	.81
4.	.18	.27	.56	1.00	.68
Item					
1.	.65	.29	.17	.09	.40
2.	.87	.44	.24	.13	.56
3.	.90	.58	.29	.17	.66
4.	.85	.71	.33	.21	.71
5.	.77	.82	.39	.24	.74
6.	.67	.89	.45	.26	.77
7.	.58	.93	.53	.28	.78
8.	.50	.93	.61	.30	.79
9.	.44	.90	.70	.32	.79
10.	.39	.84	.78	.35	.79
11.	.35	.77	.85	.39	.79
12.	.32	.69	.90	.44	.79
13.	.30	.60	.93	.51	.79
14.	.28	.52	.93	.59	.78
15.	.26	.44	.89	.68	.77
16.	.24	.38	.81	.77	.74
17.	.21	.33	.71	.85	.71
18.	.17	.28	.58	.90	.65
19.	.13	.24	.43	.87	.56
20.	.09	.17	.29	.65	.40

Note. Items are ordered from the "easiest" to the "hardest."

done in the previous analyses, restricting the factor correlations. He found the factors to be highly correlated and discovered a higher order factor that matched the author's intent of measuring test anxiety. The trait or state items from the State–Trait Anxiety Inventory (STAI: Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983) have often been factored. Despite a reliability of .9, two factors emerge from these items. When rotated to simple structure, half of the items are loaded by each factor. One of the primary factors has all the items that are scored in the positive direction (e.g., "Yes" to "Are you anxious?"), and the other has all the reversed-scored items (e.g., "No" to "Are you calm?"). If the factor analysis is stopped at this point, there appears to be no such thing as anxiety. If the analysis is continued without restricting the rotation, then a higher order, general factor is found that is the trait of anxiety. The scale includes items scored in both the positive and reverse-scored directions.

It follows that it is critical to note that simple structure bias against a general factor requires an unrestricted rotation to allow compensation for the bias. Restricting the rotation to uncorrelated factors, as Varimax does, precludes any general factor. Varimax is the worst method for item factor analysis because there is no way to overcome the simple structure bias, a bias that is present when the items come from the same domain (e.g., are all ability items, motivational items, or depression items). It should also be noted that nonrestricted solutions—such as Oblimin or Promax—will give uncorrelated factors when that provides a reasonable solution (they are only called *oblique* rotations because they may give correlated factors; Varimax is called *orthogonal* because it only gives uncorrelated factors).

Using unrestricted rotation, the two primary factors and the general factor structure of the STAI have been replicated. For example, Chan (1989) replicated these factors in Chinese, Lee (1995) in Korean, and Courelli (1991) in Greek monolingual respondents.

The BDI has often been factored with a variety of techniques, and comparing the results here with other studies helps to evaluate the impact of different procedures on a practical level. Table 5 presents the results of a BDI item factor analysis from Lee's (1995) work examining a Korean translation in a monolingual sample. The $R > 1$ criterion indicated six or seven factors, but only four meet the criterion of having four items with their highest loading on the factor in an initial rotated solution. Communality estimates were squared multiple correlations with two iterations. The factors were rotated by Promax (using Varimax for the target matrix and setting k to 4). As can be seen in Table 5, the factors—when rotated without any restriction for uncorrelated or for correlated factors—were clearly correlated, and so a higher order factor was extracted and its correlations with the items computed. Table 5 shows that the primary factors all correlate highly with the higher order factor, as do almost all the items. Comparing Table 5 with Table 4 shows the BDI general factor to have more items with the highest correlation being with the general factor in the BDI than with the general factor of the perfect Guttman scales.

TABLE 5
Item Factor Analysis of the Beck Depression Inventory

	Factor			
	Primary			
	1	*2*	*3*	*Higher Order*
Primary factor				
1.	1.00	.57	.70	.87
2.	.57	1.00	.45	.45
3.	.70	.45	1.00	.73
Item				
1.	.48	.28	.52	.49
2.	.45	.31	.45	.45
3.	.42	.26	.61	.49
4.	.61	.25	.58	.55
5.	.28	.28	.51	.41
6.	.28	.24	.44	.38
7.	.51	.24	.66	.54
8.	.47	.27	.56	.50
9.	.49	.40	.48	.50
10.	.50	.34	.41	.46
11.	.51	.33	.45	.48
12.	.55	.27	.33	.45
13.	.44	.26	.35	.41
14.	.50	.27	.45	.48
15.	.71	.40	.49	.58
16.	.25	.50	.20	.28
17.	.53	.35	.36	.46
18.	.32	.62	.23	.34
19.	.19	.50	.16	.24
20.	.38	.48	.25	.35
21.	.50	.38	.31	.44

There is clearly strong evidence for a single general depression score for the BDI. This evidence would not have been found if only Varimax had been used for the rotation. The primary factors may be technical factors, just as in the case of the Guttman scale of Table 4.

Technical factors need not replicate. They may be a function of idiosyncrasies of some trivial characteristic of the population sampled. We have found this to be so in factor analyses of the BDI in samples from different populations. Factoring the BDI items is factoring a set of items that were selected by Beck et al. (1961) to have a general factor, depression, and that would, we hope, function much the same across cultures. Chan (1989) collected data from Hong Kong Chinese, and com-

pared the factors with the results of other analyses. She found several primary factors and a second-order factor with which all the items correlated. Thus there appeared to be some technical factors but items also measured a general factor, depression. The comparisons with other studies showed that the primary factors did not replicate any previously found; however, the general, higher order factor replicated quite well.

The check for replication is to use the past study's factors as hypotheses for multiple group factors (Gorsuch, 1983). Multiple group factor analysis is a type of CFA in which each factor is defined by a weighted subset of the variables. The correlations of these multiple group factors with the exploratory factors is the degree to which the prior study's factors replicate. (Note that coefficients of congruence can be misleading and are to be avoided.) The UniMult statistical package (Gorsuch, 1994) has multiple group factor analysis as an option and computes the correlations with other factors from the same data as well. If such a program is not available, an "almost as good" procedure is to sum the items salient for each factor of the prior factor analysis as one set of scales and sum items for the factors of the exploratory analysis as a second set of scales, and then compute the correlation between these two sets of scales.

We have related Lee's (1995) factors using UniMult to the factors resulting from other factor analyses. One set of factors was with another Asian sample, Chinese (Chan, 1989). Another was with Courelli's (1991) Greek elderly people. A third was with the three-factor structure Byrne (1994) used with CFAs. Table 6 shows the correlations of the factors of other studies with those of Lee.

TABLE 6
Correlations of Korean Factors With Other Analyses

	Korean Factor			
	Primary			
Factor	*1*	*2*	*3*	*Higher Order*
Chinese (Chan, 1989)				
1.	.57	.75	.43	.55
2.	.68	.35	.79	.68
3.	.72	.32	.82	.71
Greek (Courelli, 1991)				
1.	.82	.38	.82	.77
2.	.80	.34	.61	.69
U.S. confirmatory factor analysis (Byrne, 1994)				
1.	.87	.34	.81	.79
2	.85	.34	.80	.78
3.	.64	.88	.39	.60

The most striking aspects of Table 6 are twofold. First, none of the primary factors replicate. They all have odd mixtures of correlations but do not correlate highly with just one factor from the other study. This was also found in comparisons that Chan (1989) and Courelli (1991) made with other factor analyses in their samples. It may be that the primary factors are technical factors specific to a sample or are culturally specific factors. Either way, they do not replicate in these samples. Subscoring the BDI therefore appears to be of limited value, applicable only within one subculture, and of little use in evaluating the general applicability of the BDI. Second, there is a strong general factor that does replicate well. Even the factors restricted to being uncorrelated in an American sample (by using Varimax) were highly correlated when measured in a new sample such as Chan's, Courelli's, or Lee's. Additionally, all of the primary factors correlate well with the second-order factor of depression, as do the items. The results point clearly to the conclusion that the BDI measures one construct. While the item correlations will show different minor patterns that either vary by subculture or are nonreplicable, the strongest and most consistent replication is of the general depression factor.

Note how wrong we would have been if only "Little Jiffy" had been run. It would, by the uncorrected $R > 1$ criterion, have given too many factors and then, by the restrictions of Varimax rotation, have suggested that there were a set of uncorrelated factors with no general depression factor. And, the Varimax factors are unreplicable. The conclusion, based on poor methodology, would have been that the BDI had no replicable structure and so was meaningless. The correct conclusion is that the BDI has a strong general factor just as Beck et al. (1961) aimed it to have, and this general factor is highly replicable even across cultures.

Other examples of nonrestricted rotation in item factor analysis with higher order factors can be found. For example Dura, Bernstein, and Kiecolt-Glaser (1990) analyzed dementia items. Johnson and Johnson (1993) analyzed items regarding school life (although they probably should also have used confirmatory multiple group analysis to test the original author's subscales). Procter's (1993) brief presentation of factor analysis for attitude scales assumed higher order analysis. All found higher order factors meaningful.

In item factor analysis, "Little Jiffy" is not just a "personal choice"; it is a way of guaranteeing that any general, replicable factor is lost. In the case of BDI analyses by "Little Jiffy," the several samples would have strongly suggested that there are no replicable factors for the BDI when the opposite is true.

Interpretation of Factors

Before construing an item factor as substantive, the alternative interpretation that it is a technical factor must be ruled out. This happens by rejecting a set of alternative technical factor hypotheses:

1. The first alternative hypothesis is that the factor is nonreplicable. Using multiple group factor analysis to relate the factors to those found by others is helpful. However, this assumes that the previous analysis was an appropriate item factor analysis.

2. The second alternative hypothesis is that the factor is too small to be considered further. Assuming the number of factors has been examined as noted earlier, there are several salient items to measure it. Nevertheless, it could still be unmeasurable. This is found by scoring all factors as if they were scales (see selecting items later), computing the reliabilities, and correlating the scales. A usable scale has a reasonable reliability and has no correlation with another factor scale approaching either's reliability.

3. The third alternative hypothesis is that the factor is caused by items having different distributions. These are classically called difficulty factors because they were first found in ability items. The less difficult items had negative skew and correlated well with each other, and the more difficult items had positive skew and correlated well together. But the phenomena applies to any data set, just as in the aforementioned Guttman scale example. It may, for example, be the reason that Ross, Joshi, and Currie (1991) found a factor for common, everyday disassociative states and other factors for uncommon, rare disassociative experiences. To evaluate this interpretation, the distribution of each item should be examined. If it is the same among items of the same factor but differs from items not related to the factor, then it is probably a distribution factor.

4. The fourth alternative hypothesis is response style. Do all the items for the factor require the same response set? This could be as simple as the acquiescence response set possible for half of the STAI items, but it could also be more complex. Perhaps the items all require the same radical judgment be made. If that is a response style factor, then the items for other factors will not require a radical statement be made. Chan (1991) found order effects among the items of personal distress that may be a shift in response set while answering the items.

5. The fifth alternative hypothesis is that the items are not linear combinations of the factors. For example, an adjective list for rating one's roommate on dominance may include the terms "domineering," "reasonable person," and "a doormat for others." But these items are curvilinearly related to the latent variable of dominance-submission. The "reasonable person" phrase would be checked only if neither of the other two items had been picked. Note how different this is from the Guttman scale, with each item linearly related to the factor. A person has a moderate score if and only if all items lower on the scale were also selected. For these three phrases to be linearly related to the same factor, the "reasonable person" could only be checked if "a doormat for others" were also checked. And "domineering" could only have been checked if both of the other items were checked. When items are nonlinearly related to factors, they can be analyzed conceptually and statistically as "unfolding" (see van Schuur & Kiers, 1994). Probably such factors correlate,

leading to a single higher order factor, but that has not yet been shown. It has been shown that items curvilinearly related to factors produce an extra primary factor.

A useful procedure to aid in evaluating whether the possible technical factors may be substantive factors is to relate them to other variables. If the factors function as only technical factors and not substantively, then they should have the same beta weight as do the other substantive factors in a multiple regression to other relevant variables. This is tested by summing across the possible substantive factors and entering that sum first into a hierarchical multiple regression to the dependent variable. Then the technical factors are, as the second step in a hierarchical regression analysis, added to the multiple regression and tested as a set. If they are only technical factors, they function with the same weight; adding them together and then entering the total gives each the same weight. Entering them at the second step would not add significantly to the prediction. But if they function as separate substantive factors, then they would have different weights. In that case, the separate factors would be most predictive when weighted separately, and Step 2 would add significantly to Step 1. Gorsuch and McPherson (1989) found the positively and negatively worded items measuring intrinsic religious commitment to form two factors. The multiple regression, however, found no significant increase in predictive power by considering them two substantive factors rather than just one. Roberts, Lewinsohn, and Seeley (1993) also found positive and negative factors among loneliness items but, again, found the same pattern of relations to other variables.

If it is unlikely that the factor is a technical factor, then a substantive interpretation is warranted. These are generally based on the item content, but it is useful if other checks are made. Such checks are dependent on other data being available. If relevant data are available, they can be analyzed for increased understanding.

Indirect Item Factor Analysis

The problems of items—low reliabilities, varying distributions, noncontinuous response formats—have long been known. Cattell (1957) confronted these problems (and the problem of low computer capacity) and used "item parcels," or miniscales, to reduce them. He formed his items into groups that were all of one kind, and scored each set of items for one variable for the factor analysis. These miniscales were then factored and the items correlated back with the factors. Comrey and Lee (1992) advocated this procedure also.

The miniscale approach reduces all the problems inherent in factoring items:

1. The reliability of a miniscale is higher than that of an item.
2. The items for a parcel can be selected to have the range of means/distributions desired in the final scale. When summed together, each miniscale would have a more normal distribution.

3. The miniscales have a wider possible range. If each item is scored on a 3-point scale and there are five items, then the scale ranges from 5 to 15.
4. The idiosyncratic content that leads to two items loading a separate factor is averaged out across the miniscales, and so methods such as the R > 1 number of factors criterion are more applicable. There can still be ample basis for multiple factors; Cattell found 16 personality factors using this method.

Why, then, is this procedure not more widely used? Perhaps because it is not widely known, or perhaps because it requires the additional effort of building the parcels and then relating the items to the factors in addition to the factor analysis. Many scales have too few items for this type of analysis. Finally, investigators may be uncomfortable with determining which items are "alike" on an a priori basis and so should be in the same parcel.

The procedure recommended earlier of extracting higher order factors from primary factors is, in a sense, an empirically based miniscale assignment and factoring procedure. The primary factors group together those items that correlate highest, thus forming them into implicit miniscales. Factoring the primary factors for second-order factors is equivalent to factoring miniscales. The higher order factor approach has the advantage of less work and being more empirical than the parcel approach to building miniscales. Because it does minimize subjective judgments, it should also be more replicable. The drawback is that it does not reduce distribution problems as well as miniscales. Instead, the items with the same distributions will be placed together. Overall, this does not seem to be a serious handicap.

Selecting Items for Scales

Pattern or structure or weight matrix? Whether the items are factored directly or indirectly, the result is a set of coefficients relating each item to each factor, including any higher order factors. There are several measures of relation between the items and the factors. First are the factor-item correlations; the matrix containing all these correlations is called the *factor structure*. Second are the standardized weights to be used to reproduce the item scores from the factor scores, called the *factor pattern*. Third are the standardized weights to be used to estimate factor scores from item scores, called the *factor weight matrix*.

Early tradition was to use the factor pattern (or the reference vector structure). The pattern is a matrix of the weights to be used to reproduce the variable standard scores from the factor standard scores (the reference vector structure contains the correlation of each item with each factor with the other factors partialled out). This was done primarily because of the limits of the computers which made calculating the actual correlations of the items with the factors difficult.

The correlations of each item with each factor—the factor structure—is an appropriate statistic to use to select items. It is either the equivalent of the item-total correlation if it is a component analysis, or the equivalent of the item-remainder or item-domain coefficient if it is a common factor analysis. The correlation reflects directly how the item will function as part of a scale, and the correlations of that item with other factors indicate how it will contribute to the relations of its scale to other scales based on these factors (the pattern only indirectly reflects the first and can be misleading about the second).

The other possible matrix to use to select items is the weight matrix for scoring the factors. The values in this matrix give the beta weights to be used with each item standard score to best estimate the factor standard score. It is the same as if one ran a multiple regression with each of the scale's items as predictors and the total scale as the dependent variable, and then picked items based on their beta weights. But beta weights are not used in item selection for two reasons. First, they are unstable and highly influenced by what other items are in the equation, including items not selected. Second, direct comparisons between beta weights are dangerous because each has a separate standard error; a beta weight of .3 may be significant for Item 1 while a beta weight of .4 may be insignificant for Item 2. How then can we decide which item should be on the scale? Third, multiple regression devalues redundancy and values differential contributions to the estimate. Hence two items that correlate highly together will not both be given high beta weights even if both correlate highly with the factor. Instead, both will be given moderately low weights or one will be weighted high and the other low. However redundancy is highly desirable in the items for a scale (and increases internal consistency reliability). Correlations have none of these problems.

In a comparison of using the correlations and using the weight matrix, ten Berge and Knol (1985) found that the full multiple regression weights lead to the highest correlation with the factors but that using the correlations lead to the highest internal consistency reliability. This is to be expected because the evaluations were in the original sample. Research in factor scoring (Gorsuch, 1983) has, however, found that the regression weights generalize poorly compared to just summing the items, as has been found in other areas. Summing the salient items is the most generalizable procedure. The ten Berg and Knol results would, in cross-validation, find the high internal consistency generalizing but not the highest correlation. Hence these results are in keeping with the recommendation to use the item-factor correlations.

How high a correlation? The same answer is given here as in any item selection based on correlations. First, the item should have a highly significant correlation. But note that the significance level should be smaller than might be normally used for two reasons: (a) factor-item correlations are suspected to have a higher standard error than regular correlations and (b) as in any item selection, many

correlations are computed but only a few are chosen, thus leading to capitalization on chance.

Second, the item-factor correlation should be high enough that, given the items already chosen, it will add to the reliability and validity of the scale rather than reduce it. If the first 10 items correlate at least .6 and the next .3, adding that 11th item may do more harm than good. It appears that the best procedure is to try several sets of items and evaluate what happens to the correlation of the resulting scale with the factor by extension analysis (Gorsuch, in press).

How are the correlations with other factors used? The use of the other factors depends on whether they are technical factors or substantive factors. If technical factors exist, it is often impossible to find items that load only the substantive factor and no technical factors. So we counterbalance the technical factors so that the test score is not particularly associated with any of the technical factors. For the technical factors of different distributions of Guttman scale items, we counterbalance by including an equal number of items at each level of response frequency. For the STAI, we counterbalanced by including equal numbers of positive and negative items. (Note that the reliability of a counterbalanced scale is best defined by a parallel form coefficient, not the usual coefficient alpha. Using the latter is, in effect, assuming that no technical factors exist.)

The correlations of substantive item factors need to be monitored in light of the items already selected for each substantive factor and the correlations among the factors. The items selected should, as a group, have a set of loadings with other factors so that the focal factor scale will have the appropriate correlations with the other factor scales. For example, if the first three items correlate with another factor higher than their factor does with the other factor, the scale will correlate higher with the other factor than the factor analysis suggests it should. Hence the next item should have a much lower correlation with the other factor. This strategy also applies if the first items correlate much too low with the other factor.

Remember that, in selecting items as a function of the correlation with other substantive factors, items have attenuated correlations. Due to their low reliabilities they correlate lower with everything than do scales (see earlier example). Hence the individual items should correlate very low with the other substantive factors because the resulting scale will correlate higher than the items. Generally the lowest possible correlation of items with other factors is sought.

Testing proposed scales. How well did the item selection work? Are the correlations with the factor the scale should measure high and the correlations with other factors low? Would another subset of items perform better?

There are two methods to provide correlations of proposed scales with the factors. The first—and no longer recommended—is to compute estimated factor scores and then correlate these with the scores for each proposed scale. The problem is that the factor scores contain nonfactor covariance among the items, leading to inflated values (just as item-total correlations are inflated).

The second method is Gorsuch's extension analysis (Gorsuch, 1997, which also shows prior extension procedures have the same problems as item-total correlations and estimated factor scores). In that procedure, the factors are located within the hyperspace defined by the items being factored. The factor analysis is then extended to the proposed scales by locating them in the item hyperspace also. The correlations of each proposed scale with the factors are computed. As shown in Gorsuch (1997), this procedure gives appropriate, noninflated correlations.

CONCLUSIONS

Items have lower reliabilities than scales. They can also be characterized by a greater degree of idiosyncratic content such as response bias (which is averaged out in a good scale). Hence, establishing the number of factors by the $R > 1$ criterion—whose rationale is based on an error-free matrix—gives too many factors. If component analysis is used—which has no error term in the model—the loadings will be inflated. Using multiple methods to determine the number of factors and extracting factors by a method allowing community estimates are recommended.

If the items are all from one conceptual domain or a set of conceptually related domains, restricting the rotation to uncorrelated factors (as in Varimax) obscures the general factor or broader constructs the test may measure. It also allows technical factors—factors occurring for methodological reasons—to obscure substantive factors. Using a nonrestricted rotation and extracting higher order factors, if any, allows for narrow to general factors to emerge and for constructing scales that minimize technical factors.

Proposed scales can be evaluated by a new extension analysis (Gorsuch, 1997), which gives better estimates of scale-domain correlations than either traditional item analysis or other factor analytic procedures.

"Little Jiffy," the default in many statistical packages, is designed for high reliability scales sampling several unrelated substantive domains. It is therefore seldom appropriate for item factor analysis; if it is used with items, it generally produces too many factors and prevents broad or general factors from being identified. An author's hypothesis that the items measure one construct will almost always be rejected solely because "Little Jiffy" was used even when the hypothesis is warranted; in such cases the "Little Jiffy" factors may be unreplicable. Using the procedures recommended earlier allows an unbiased test for a general or other broad factors and produces factors more likely to replicate.

REFERENCES

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571.

Bernstein, I. H., Jaremko, M. E., & Hinkley, B. S. (1994). On the utility of the SCL–90–R with low-back pain patients. *Spine, 19,* 42–48.

Bernstein, I. H., Jaremko, M. E., & Hinkley, B. S. (1995). On the utility of the West Haven–Yale Multidimensional Pain Inventory. *Spine, 20,* 956–963.

Bernstein, I. H., & Keith, J. B. (1991). Reexamination of Eisen, Zellman, and McAlister's Health Belief Model Questionnaire. *Health Education Quarterly, 18,* 207–220.

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.

Byrne, M. B. (1994). *Structural equation modeling with EQS/Windows: Basic concepts, applications, and programming.* Thousand Oaks, CA: Sage.

Cattell, R. B. (1957). *Personality and motivation structure and measurement.* Yonkers, NY: World Book.

Chan, M-Y. A. (1989). *Development and evaluation of a Chinese translation of the State–Trait Anxiety Inventory and the Beck Depression Inventory.* Unpublished doctoral dissertation, Graduate School of Psychology, Fuller Theological Seminary, Pasadena, CA.

Chan, J. C. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement, 51,* 531–540.

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

Courelli, P. S. (1991). *A screening battery of tests for the detection of dementia among Greek elderly.* Unpublished doctoral dissertation, Graduate School of Psychology, Fuller Theological Seminary, Pasadena CA.

Dura, J. R., Bernstein, R. A., & Kiecolt-Glaser, J. K. (1990). Refinements in the assessment of dementia-related behaviors: Factor structure of the memory and behavior problem checklist. *Psychological Assessment, 2,* 129–133.

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

Gorsuch, R. L. (1966). The general factor in the Test Anxiety Questionnaire. *Psychological Reports, 19,* 308.

Gorsuch, R. L. (1968). The conceptualization of God as seen in adjective ratings. *Journal for the Scientific Study of Religion, 7,* 56–64.

Gorsuch, R. L. (1974). *Factor analysis.* Philadelphia: Saunders.

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

Gorsuch, R. L. (1990). Common factor analysis vs. component analysis: Some well and little known facts. *Multivariate Behavioral Research, 25,* 33–39.

Gorsuch, R. L. (1994). *UniMult guide.* Pasadena, CA: UniMult.

Gorsuch, R. L. (1996). *Relating factors across studies.* Manuscript submitted for publication.

Gorsuch, R. L. (in press). New procedure for extension analysis in exploratory factor analysis. *Educational and Psychological Measurement, 57.*

Gorsuch, R. L., & McPherson, S. (1989). Intrinsic/extrinsic measurement: I/E–Revised and single-item scales. *Journal for the Scientific Study of Religion, 28,* 348–354.

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

Johnson, W. L., & Johnson, A. M. (1993). Validity of the quality of school life scale: A primary and second-order factor analysis. *Educational and Psychological Measurement, 53,* 145–153.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika, 35,* 401–415.

Lee, C-K. E. (1995). *Evaluation of a Korean translation of the State–Trait Anxiety Inventory and the Beck Depression Inventory.* Unpublished doctoral dissertation, Graduate School of Psychology, Fuller Theological Seminary, Pasadena, CA.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100,* 107–120.

Nunnally, J. (1967). *Psychometric theory.* New York: McGraw-Hill.

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory.* New York: McGraw-Hill.

Procter, M. (1993). *Measuring attitudes.* In N. Gilbert (Ed.), *Researching social life.* London: Sage.

Reddon, J. R. (1990). The rejection of the hypothesis of complete independence prior to conducting a factor analysis. *Multivariate Experimental Clinical Research, 9,* 123–129.

Roberts, R. E., Lewinsohn, P. M., & Seeley, J. R (1993). A brief measure of loneliness suitable for use with adolescents. *Psychological Reports, 72,* 1379–1391.

Ross, C. A., Joshi, S., & Currie, R. (1991). Dissociative experiences in the general population: A factor analysis. *Hospital and Community Psychiatry, 42,* 297–301.

Roth, W. M., & Roychoudhury, A. (1991). Nonmetric multidimensional item analysis in the construction of an anxiety attitude survey. *Educational and Psychological Measurement, 51,* 931–942.

Snook, S. C., & Gorsuch, R. L. (1989). Common factor analysis vs. component analysis. *Psychological Bulletin, 106,* 148–154.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State–Trait Anxiety Inventory for adults.* Palo Alto, CA: Consulting Psychologists Press.

ten Berge, J. M. F., & Knol, D. L. (1985). Scale construction on the basis of components analysis. A comparison of three strategies. *Multivariate Behavioral Research, 20,* 45–55.

van Schuur, W. H., & Kiers, H. A. (1994). Why factor analysis often is the incorrect model for analyzing bipolar concepts, and what model to use instead. *Applied Psychological Measurement, 18,* 97–110.

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.

Richard L. Gorsuch
Graduate School of Psychology
Fuller Theological Seminary
Pasadena, CA 91101

Received December 23, 1996