Factor Analysis in the 1980's and the 1990's: Some Old Debates and Some New Developments[1]

James H. Steiger

1 Introduction

A number of years have passed since my active involvement (Steiger & Schönemann, 1978; Steiger, 1979a, 1979b) in the "factor indeterminacy controversy." In that period, factor analysis has become more firmly entrenched than ever as a multivariate method, enjoying wide use as a data reduction technique, as a "measurement model" for true scores in structural modeling, and as an exploratory data analysis tool. At the same time, the typical computing environment has changed radically, making the cost of execution time and computer memory a negligible issue in factor analytic practice.

Though many would argue that research interest in factor analysis has declined substantially since the heyday of Thurstone and Tucker, there have been some interesting new lines of technical development, most of which have come laterally from the closely related field of structural equation modeling. There have also been some new twists on the old debates about the relative merits of factor analysis and component analysis. In this paper, I will attempt to review these developments and place them in perspective.

This perspective is, of course, a personal one, even in matters of notation. One person might call the matrix **F** a "factor pattern." Another might call it a "factor matrix." To reduce potential communication problems, I begin in Section 2 with a review of basic theory and terminology, that will hopefully make this account reasonably self-contained and comprehensible. In Section 3, I attempt to summarize significant developments in the last 15 years. At this stage, I try to be descriptive rather than judgmental and my account emphasizes more recent developments. For an alternative view of developments from 1940 until the mid-eighties, see Mulaik (1986).

Finally, in Section 4, I summarize, freely injecting my views and biases on the state of the art, and speculating about future trends and important unanswered questions.

I must apologize in advance for concentrating, in this account, on traditional linear factor analysis and component analysis. Space considerations made me choose not to attempt to discuss interesting work by Ramsay, Takane, McDonald, Bartholomew, and others on non-linear (and other) variations on the factor-analytic theme. I was also forced to omit a review of the profound, and extremely interesting discussion in *Behavioral and Brain Sciences* of Jensen's (1985) article on black-white intelligence differences. This article, the commentaries, and the followup discussion in *Multivariate Behavioral Research* (Volume 27, Number 2), especially "Guttman's last paper" (Guttman, 1992) should be considered required reading for the sophisticated social scientist.

---

[1] This article appeared in Ingwer Borg and Peter Ph. Mohler (Eds.), *Trends and Perspectives in Empirical Social Research.* Berlin: Walter de Gruyter, 1994.

2 Some Theoretical Background

Before summarizing key developments in the last two decades, I begin by establishing a common notation and theoretical background, reviewing some of the basic ideas underlying factor analysis, and some of the motivations for undertaking it. I will employ random vector notation, and represent matrices and vectors in boldface.

2.1 The Common Factor Model and Its Uses

Consider the p random variables in the random vector $\mathbf{y}$. These represent variables that the data analyst wishes to represent with an m-factor common factor model. The common factor analysis model states that the p observed variables can be expressed as linear functions of m unobserved ("latent") variables called *common factors*, and that if this is done in the least-squares linear regression sense, i.e., we predict the variables in $\mathbf{y}$ from these common factors with multiple linear regression weights, *the resulting residuals will be uncorrelated*.

Algebraically, we say that

$$\mathbf{y} = \mathbf{F}\mathbf{x} + \mathbf{e}, \text{ with} \tag{1}$$

$$E(\mathbf{x}\mathbf{x}') = \mathbf{P},\ E(\mathbf{x}\mathbf{e}') = \mathbf{0},\ E(\mathbf{e}\mathbf{e}') = \mathbf{U}, \tag{2}$$

where $\mathbf{U}$ is a diagonal, positive-definite matrix. $\mathbf{F}$ is the *common factor pattern*, $\mathbf{P}$ the matrix of factor correlations, $\mathbf{U}$ contains the *unique variances* of the variables on its diagonal. If $\mathbf{P}$ is an identity matrix and the factors are uncorrelated, we say that the common factors are *orthogonal*, otherwise they are *oblique.*

Before exploring the algebra for such a model, we might quickly review some reasons for considering it important. There are many reasons for wanting to fit a common factor model to a set of variables. Here we will consider 4 common, somewhat interrelated ones: (1) The *partial correlation* rationale; (2) The *random noise* rationale; (3) The *true score* rationale, (4) the *data reduction* rationale.

The Partial Correlation-Explanation Rationale

Suppose we were to keep records of every house fire in Canada over a period of 5 years. Among the variables we record are the number of fire trucks sent to each fire, and the damage in dollars done by the fire. After 5 years, we would certainly see a high positive correlation between these two variables. All other things being equal, sending fire trucks to a fire is associated with a higher damage level. Does this mean that more fire trucks cause more damage, or that we should send fewer trucks if we want fires to cause less damage? Of course not. There is a "third variable," a "common cause" (size of fire) which causes more trucks to be sent, and more damage to be done. If we "partial out" this variable from the first two, the resulting residuals will be uncorrelated. We say that the partial correlation between number of trucks sent and fire damage, with size of fire partialled out, is zero. In this case, we might say that "size of fire explains the correlation between fire trucks and fire damage."

This idea leads to the following notion. If the partial correlations among the variables in set **y** with those in set **x** partialled out are zero, then in some sense the variables in **x** *explain*, or *account for* the correlations among the variables in **y.**

With this rationale, we view the "common factors" in **x** as the underlying common causes of the variables in **y.**

The Random Noise Rationale

In some situations, it is reasonable to hypothesize a physical process that involves several underlying sources of variation that are polluted by random noise. A classic example might be EEG responses to carefully timed standardized auditory signals, recorded at several sensors. It may be that each sensor will pick up output from several unified, consistent sources within the brain, but that these signals will also include random, uncorrelated electrical noise. In this case, the underlying sources are the "common factors" in **x**, the observed signals are recorded at **y**.

The True Score Rationale

In psychometrics, we commonly measure attributes with devices that are assumed to be degraded by random error. In particular, classical *true score theory* postulates measurements that involve an underlying true score component, and a random error component. If we measure the same ability with several items, this turns out to be a special case of the common factor model. What we are really interested in is the underlying true scores on the variables of interest.

The distinction between the observed scores on measures of a trait, and the underlying trait itself, can be especially crucial when we seek to establish linear regression relations among variables that have varying amounts of error variance. Observed correlations can be *attenuated* by unreliability, and so the regression relations among the unreliable measures of a set of traits can mislead one about the relations among the traits themselves.

Because of this problem, it is common to try to estimate regression relationships between the common factors underlying a group of measures, rather than the measures themselves.

The Data Reduction Rationale

In many situations, it is computationally inconvenient to operate with a large number of measures. We seek to reduce the number of measures, while simultaneously classifying them into groups, and increasing the reliability of what they measure.

This *data reduction* rationale for factor analysis is a major use for factor analytic technology. We factor analyze a group of items to discover the major sources of variation underlying them, and to find out which items are related to which sources. The resulting information allows us to parcel items into groups, to gain a better understanding of the structure underlying our items, and refine our measures of the sources of variation.

2.2 Some Algebraic Implications and Indeterminacy Problems

The model of equations (1) and (2) is sometimes referred to as the "factor model at the random variable level." If this model fits the data, then a simple consequence is that

$$\mathbf{\Sigma} = E(\mathbf{y}\mathbf{y}') = \mathbf{F}\mathbf{P}\mathbf{F}' + \mathbf{U}. \tag{3}$$

This equation, sometimes called the "fundamental theorem of factor analysis," allows one to test whether the *m*-factor model is tenable by examining whether a $\mathbf{U}$ can be found so that $\mathbf{\Sigma} - \mathbf{U}$ is Gramian and of rank *m.* The early factor analysts, especially Spearman, found this notion almost magical. You can test whether a (hopefully small) set of *m* variables explaining the variation in $\mathbf{y}$ (in the partial correlation sense described above) *could exist*, without ever observing such variables directly. Moreover, you could examine the linear regression relations between $\mathbf{y}$ and the unobserved, hypothetical $\mathbf{x}$ by matrix factorization of $\mathbf{\Sigma} - \mathbf{U}$. The idea is indeed fascinating, and it is easy to understand why Spearman and Thurstone found variants of it so compelling.

There were two elements of the factor model that, if identified, could provide substantial practical benefits. $\mathbf{F}$, by revealing the regression relationships between the observed variables and the more fundamental factors that generate them, could provide information about the structure of the variables being investigated. $\mathbf{x}$ would provide scores on the factors. So, for example, if the factor model fit a set of mental ability tests, one could determine a small set of underlying mental abilities that explain a larger number of tests, *and* the ratings of the test takers on these fundamental abilities.

Unfortunately, it turned out that there was a hierarchy of indeterminacy problems associated with the above algebra. Rather than discuss the problems in the clear, systematic way that simple accuracy would seem to demand, authors committed to the common factor model have generally omitted at least one, or described them in obscure, misleading cliches. I describe them here, and urge the reader to compare my description with treatments of the factor model found in standard texts.

Identification of $\mathbf{U}$.

There may be more than one $\mathbf{U}$ that, when subtracted from $\mathbf{\Sigma}$, leaves it Gramian and of rank *m.* This fact, well known to econometricians, and described with considerable clarity and care by Anderson and Rubin (1956), is not described clearly in several factor analysis texts. One reason for the confusion may be that necessary and sufficient conditions for identification of $\mathbf{U}$ have never been established, and there are a number of incorrect statements and theorems in the literature. There are some well-known conditions when $\mathbf{U}$ is not identified (described by Anderson and Rubin). For example, $\mathbf{U}$ is never identified if $p = 2,\ m = 1$ or if $p = 4,\ m = 2$. On the other hand, if the number of variables is sufficiently large relative to the number of factors so that $(p-m)^2 > (p+m)$, $\mathbf{U}$ will almost certainly be identified.

Rotational Indeterminacy of $\mathbf{F}$.

Even if $\mathbf{U}$ is identified, $\mathbf{F}$ will not be if $m > 1$. Suppose we require *m* orthogonal factors. If such a model fits, then infinitely many $\mathbf{F}$ matrices will satisfy $\mathbf{\Sigma} - \mathbf{U} = \mathbf{F}\mathbf{F}'$, since $\mathbf{F}\mathbf{F}' = \mathbf{F}_1\mathbf{F}_1'$ so long as $\mathbf{F}_1 = \mathbf{F}\mathbf{T}$, for any orthogonal $\mathbf{T}$**.** Thurstone "solved" this very significant problem with the *simple structure* criterion. Development of "machine rotation" methods and digital computers elevated factor analysis from the status of an esoteric technique understood and practiced by a gifted elite, to a technique accessible (for use and misuse) to virtually anyone. Perhaps lost in the shuffle was the important question of why one would expect to find simple structure in many variable systems.

Factor Indeterminacy

If the first two problems are overcome, a third one remains. Specifically, the factors **x** are not uniquely defined, even if **U** and **F** are. To simplify the discussion, suppose the factors are orthogonal, and $\mathbf{P} = \mathbf{I}$, an identity matrix. Then any **x** constructed via the formula

$$\mathbf{x} = \mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{y} + \mathbf{Q}\mathbf{s} = \hat{\mathbf{x}} + \boldsymbol{\mathcal{E}}, \tag{4}$$

where **s** is *any* arbitrary random vector satisfying

$$E(\mathbf{y}\mathbf{s}') = \mathbf{0},\ E(\mathbf{s}\mathbf{s}') = \mathbf{I}, \tag{5}$$

with **Q** an arbitrary Gram-factor satisfying

$$\mathbf{Q}\mathbf{Q}' = \mathbf{I} - \mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F} \tag{6}$$

will fit the common factor model. Once **F** is known, **Q** can be constructed easily via matrix factorization methods. **s** is a completely arbitrary random vector in the space orthogonal to that occupied by **y.** Equation (4) shows that common factors are not determinate from the variables in the current analysis. There is an infinity of possible candidates for **x.** Each has the same "determinate" component $\hat{\mathbf{x}}$, but different "arbitrary component" $\boldsymbol{\mathcal{E}}$. These candidates for **x** each have the same covariance relationship with **y**, but possibly differ substantially from each other.

It is important to understand what factor indeterminacy is, and what it is not. It is not a "sampling" problem. The above equations demonstrate it exists at the population level. Several authors have advanced rationales that attempt to make factor indeterminacy appear less a problem than it is. These rationales involve assuming that the *m* factors present in the current *p* variables will remain for an infinite domain of additional variables. Interestingly, such authors make no mention of how such infinite domains can be defined in a non-circular fashion. Obviously one can construct them, using the above algebra, but this seems of little relevance to the situation typically encountered by the data analyst. Clearly, the ability to make circular assumptions about the nature of variables one has not yet observed breaks new statistical ground, and offers interesting possibilities for the solution of other seemingly intractable problems in statistics. I comment on such trends in more detail elsewhere (Steiger, 1990a).

Mulaik (1990), perhaps sensitive to the circularity in the "infinite domain" approach, attempts to add an reassuring slant by describing a sequence of studies, in which new variables are added to see whether the factors found in previous studies remain supported. It is difficult to imagine why Mulaik thinks this approach offers any solution. Suppose one's latest study (the one with the largest set of variables, including as subsets the variables in previous studies) fails to reject a hypothesis of no difference in factor loadings from previous studies. At this point, one has "determined" the common factors precisely to the extent that the *present study* has determined them. In other words, factor indeterminacy still exists, exactly as described in equation (4). The indeterminacy formula, applied to the best data set yet observed, determines factor indeterminacy. In practice, of course, there is usually only one study, the one currently in hand.

Now, suppose, the latest study *fails* to verify the factors "found" in previous studies. The disappointed factor analyst has many possible excuses including (1) there was a failure to sample the latest set of variables from the "right domain," (2) there was a lack of factorial invariance

across the (inevitably different) populations sampled, (3) there were artifacts of sampling variability, etc. Most of these options do not lead to testable hypotheses, and hence are of dubious value.

It seems Mulaik's (1990) defence, on close examination, has no substance. Factor indeterminacy exists in the "here and now" most data analysts are forced to operate in. Factor indeterminacy only disappears in a kind of factor analytic never-never land. As Steiger and Schönemann (1978) pointed out, proponents of the common factor model trying to discount indeterminacy have produced a long tradition of such theorizing, much of which has been embraced with a startling lack of critical analysis by those seeking any port in a data analytic storm.

2.3 Regression Component Models

A "regression component" model (Schönemann and Steiger, 1976) can be defined that is very similar in structure to the common factor model, but has few of its indeterminacy problems. I use the term "components" to stand for any set of variables expressible as linear combination of the variables in **y.** In a regression component system, let

$$\mathbf{y} = \mathbf{F}\mathbf{x} + \mathbf{e} \tag{7}$$

and stipulate that **F** is a least-squares linear regression pattern, as in factor analysis, but require **x** and **e** to be *components*. Specifically, let

$$\mathbf{x} = \mathbf{B}'\mathbf{y}, \text{ and } \mathbf{e} = (\mathbf{I} - \mathbf{F}\mathbf{B}')\mathbf{y}, \text{ with } E(\mathbf{x}\mathbf{e}') = 0. \tag{8}$$

Then for any set of "defining weights" **B**, **F** is automatically specified as

$$\mathbf{F} = \mathbf{\Sigma}\mathbf{B}(\mathbf{B}'\mathbf{\Sigma}\mathbf{B})^{-1}. \tag{9}$$

Note that the regression component and common factor models have much in common. The main distinction is that components are required to "remain within the test space," while common factors must, except in degenerate special cases, "go beyond the test space" with an element that is totally arbitrary. As a consequence, the residual variables in **e** cannot remain uncorrelated in component analysis

Principal components analysis is perhaps the best known special case of a regression component system. In this case **B** contains eigenvectors of $\mathbf{\Sigma}$.

2.4 Statistical Estimation in Factor Analysis

The above theoretical discussion is in terms of population constructs like $\mathbf{\Sigma}$. In practice, one seldom observes the population covariance matrix, but is forced instead to deal with a sample covariance matrix **S** based (ideally) on a sample of *N* independent observations on **y.** In this case, all the problems discussed above remain, and the additional problem of dealing with sampling variability is added. Although maximum likelihood estimation had been available in theory since Lawley's (1940) classic work, practical problems in implementation forced factor analysts to rely on more reliable estimation techniques with less computational difficulties. These simpler

methods did not, unfortunately, possess the desirable statistical properties of their more complex counterparts. Consequently, until key contributions by Karl Jöreskog (1967) paved the way for maximum likelihood estimation, factor analysis lacked full status as an analytic method with associated statistical tests and confidence intervals.

By the 1970's factor analysis entered a new era when advances in numerical analysis made maximum likelihood and generalized least squares estimation fully practical and widely available. If one estimates the common factor model by minimizing a specific *discrepancy function*, one obtains estimates with well-defined properties, and one also obtains a "goodness-of-fit" statistic with an asymptotic chi-square distribution. Let $\hat{\mathbf{F}}$, $\hat{\mathbf{P}}$, and $\hat{\mathbf{U}}$ be estimates of the corresponding population quantities, and define

$$\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{F}}\hat{\mathbf{P}}\hat{\mathbf{F}}' + \hat{\mathbf{U}}, \text{ and } \hat{\mathbf{E}} = \mathbf{S} - \hat{\boldsymbol{\Sigma}}. \tag{10}$$

Then the *maximum likelihood* (ML) discrepancy function is defined as

$$\mathrm{F}_{\mathrm{ML}}\left(\mathbf{S}, \hat{\Sigma}\right) = \ln\left|\hat{\Sigma}\right| - \ln|\mathbf{S}| + \mathrm{Tr}\left(\mathbf{S}\hat{\Sigma}^{-1}\right) - p \tag{11}$$

The generalized least squares (GLS) discrepancy function is

$$\mathrm{F}_{\mathrm{GLS}}\left(\mathbf{S}, \hat{\Sigma}\right) = \tfrac{1}{2}\,\mathrm{Tr}\left[\left(\mathbf{S} - \hat{\Sigma}\right)\mathbf{S}^{-1}\right]^2 \tag{12}$$

Estimates that minimize these discrepancy functions are referred to, respectively, as *maximum likelihood estimates* and *generalized least squares estimates*. As a byproduct of the estimation procedure, one may obtain estimated standard errors, and hence confidence intervals, for the elements of $\hat{\mathbf{F}}$, $\hat{\mathbf{P}}$, and $\hat{\mathbf{U}}$. Moreover, $(N-1)\mathrm{F}$ has an asymptotic chi-square distribution when population fit is perfect, and the distributional assumptions are met. Thus, the hypothesis that the m-factor model fits perfectly may be subjected to a statistical test.

3. Key Developments from 1980-1994.

Section 2 highlighted the theoretical ideas that were "the state of the art" around 1980. These ideas, of course, still form the key background for factor analytic theory, but there have been some interesting new developments since then. In this section, I describe what I view to be the most significant work in the 1980's and 1990's.

3.1 Emergence of Confirmatory Factor Analysis

In the above discussion, we concentrated on the "unrestricted" factor analysis model. In practice, the user decides on the number of factors, obtains a preliminary "unrotated" **F**, then rotates it to orthogonal or oblique simple structure using one of a dozen or so techniques commonly implemented in computer software. In many cases, the user enters the factor analytic process with firm notions about the general simple structure that **F** should have. In this case, the user may

wish to test the hypothesis that **F** actually has that simple structure. Before the advent of "true" confirmatory factor analysis methods, users would obtain an unrotated **F**, then subject it to "Procrustes" rotation methods, which attempt to rotate **F** as closely as possible to a target matrix with the hypothesized structural form. Jöreskog (1969) developed a general approach to confirmatory factor analysis that allowed some elements of **F, P,** and **U** to be specified (often to zero, when predicting simple structure in **F**), while maximum likelihood estimates were obtained for the parameters that remained free to vary. Subsequently, Jöreskog extended this idea to his well-known LISREL model and associated computer program. More recently, several computer programs have implemented special interfaces to make specification of even very large confirmatory factor analysis models especially easy. The *Windows* version of EQS 4.0 includes a mouse-driven interface that makes construction of confirmatory factor models extremely easy and straightforward. LISREL VIII includes a new simplifying command language, SIMPLIS. My own program, SEPATH, part of the program *Statistica/W,* has a "Confirmatory Factor Wizard" that takes the user, step by step, through the construction of a confirmatory factor model. There is virtually no typing, and any confirmatory model, even with large numbers of variables can be specified in a minute or two.

3.2 Asymptotically Distribution-Free Estimation and Related Robust Methods

ADF Methods

In order to yield a chi-square test statistic with the correct asymptotic distribution, both the ML and GLS discrepancy functions require the assumption that **S** has a Wishart distribution, which is closely linked with (though not synonymous with) the assumption of multivariate normality. Of course, this assumption is violated frequently in practice.

Browne (1982, 1984) pioneered the development of *asymptotically distribution free* (ADF) methods, which require only that the population multivariate distribution have finite 8th-order moments. Browne's work was important in two respects. First, it drew attention to the fact that maximum-likelihood covariance-structure analysis methods, including factor analysis, were not necessarily robust to departures from multinormality, and were particularly sensitive to kurtosis in the observed variables. Second, it provided a possible avenue for improving the situation. Browne showed that, if the non-redundant elements of the matrix $\hat{\mathbf{E}}$ in equation (10) are placed in the vector $\hat{\mathbf{e}}$, and an ADF Weight Matrix **W** is defined as a consistent estimate of the asymptotic covariance matrix of the non-redundant elements of **S**, the ADF discrepancy function

$$F_{ADF} = \hat{\mathbf{e}}'\mathbf{W}^{-1}\hat{\mathbf{e}} \tag{13}$$

will have an asymptotic chi-square distribution. This made it possible, at least in theory, to conduct statistical tests with non-normal data.

ADF procedures have seldom been used in practice, although it seems that the assumption of multivariate normality is frequently contestable with data in the behavioral sciences. One reason for the lack of popularity of ADF procedures is that they were not implemented in widely available computer software like LISREL VI, EzPATH 1.0, or COSAN.

There are other serious practical problem with ADF estimation procedures. First, **W** can, in practice, be a very large matrix, thus imposing practical limits on the size of the problem which can be processed. Second, the elements of **W** require estimates of second and fourth-order moments of the manifest variables. Such estimates have large sampling variability at small to moderate sample sizes. Consequently, one might expect the chi-square test statistic (and associated estimates) based on ADF estimation to converge somewhat more slowly to their asymptotic behavior than comparable normal theory estimation procedures.

The few, and very limited Monte Carlo studies that have investigated this issue have tended to confirm this expectation. The situation appears, at first glance, to be far worse than one might expect. For example, Hu, Bentler, and Kano (1992) found that, with 15 variables and 3 "strong" factors, the ADF test statistic, with nominal $\alpha = .05$,. rejected a true null hypothesis *every time* with $N = 150$ and 92% of the time with $N = 250$ with simulated multivariate normal data! Indeed, in the situation studied by Hu, Bentler, and Kano, acceptable performance required samples of between 2500 and 5000 observations. Such sample sizes are seldom found in social science research.

Unfortunately, more restricted variants of ADF theory, such as the robust methods for use with elliptical distributions (Browne, 1982, 1984; Bentler, 1983; Shapiro & Browne, 1987) have also proven disappointing in Monte Carlo studies (Harlow, 1985; Hu, Bentler, & Kano, 1992).

The Satorra-Bentler Scaled Chi-square Statistic

One potentially very important development in robust testing gives some cause for optimism. Satorra and Bentler (1986) proposed an alternative approach for correcting the ML test statistic under conditions of non-normality. This approach involves estimating a single multiplicative scale correction factor from sample second and fourth order moments, and the gradient of the discrepancy function. Hu, Bentler, and Kano (1992) found in their Monte Carlo study that this statistic substantially outperformed the ADF statistic, and converged to reasonably accurate Type I error performance with sample sizes around 500. They did not report power performance of the statistics. A number of other approaches to asymptotically distribution free test statistics were summarized and tested by these authors, but the scaled chi-square statistic seemed more effective in Type I error rate control over a broader variety of conditions than any of the others. If future Monte Carlo work confirms and extends the impressive results reported by Hu, Bentler, and Kano, the Satorra-Bentler procedure could assume a key role in significance testing in factor analysis.

Theoretical Work on the Robustness of Covariance Structure Models

The difficulties involved in ADF estimation have led a number of authors to examine theoretical conditions under which normal-theory methods might be asymptotically robust. The theory (e.g., Amemiya, 1985; Amemiya & Anderson, 1990; Anderson & Amemiya, 1988; Browne, 1985, 1987; Browne & Shapiro, 1988; Mooijaart & Bentler, 1991) relies on assumptions about the distributional form of the underlying latent variables. Though this work is technically deep and quite interesting, it is not yet clear to this author how it leads to practical improvement in robust testing and estimation in the context of general exploratory and confirmatory factor analysis.

3.3 New Approaches to the Statistical Assessment of Model Fit

A key problem in exploratory factor analysis is determining *m,* the number of common factors. Prior to 1980, procedures for trying to accomplish this included a variety of quasi-statistical rules like the "Scree test" of Cattell, the "eigenvalues greater than one" rule, or "eyeballing" residuals in $\hat{\mathbf{E}}$ and ceasing to add factors when they became "sufficiently" small. With the advent of maximum likelihood factor analysis, statistical procedures became available. Although Akaike (1973, 1983, 1987) and Schwarz (1978) provided general rationales for selecting among nested models, and these methods were applicable to factor analysis, they were generally ignored. More commonly, the statistical testing procedure was the "sequential chi-square test." One began by computing the chi-square statistic with $m = 1$. If the null hypothesis of perfect fit was rejected, one tried $m = 2$. This continued until the chi-square test failed to reject.

Numerous authors pointed out that this logic was essentially flawed, because, for any population Σ (other than one constructed as a numerical example directly from the common factor model) the *a priori* probability is essentially 1 that the common factor model will not fit perfectly so long as degrees of freedom for the chi-square statistic were positive.

In essence, then, population fit for a covariance structure model with positive degrees of freedom is never really perfect. Testing whether it is perfect makes little sense. It is what statisticians sometimes call an "accept-support" hypothesis test, because accepting the null hypothesis supports what is generally the experimenter's point of view, i.e., that the model does fit.

Accept-support hypothesis tests are subject to a host of problems. In particular, of course, the traditional priorities between Type I and Type II error are reversed. If the proponent of a model simply performs the chi-square test with low enough power, the model can be supported. As a natural consequence of this, hypothesis testing approaches to the assessment of model fit *should* make some attempt at power evaluation. Steiger and Lind (1980) demonstrated that performance of statistical tests in common factor analysis could be predicted from a noncentral chi-square approximation. A number of papers dealing with the theory and practice of power evaluation in covariance structure analysis have been published (Matsueda & Bielby, 1986; Satorra and Saris, 1985; Steiger, Shapiro, & Browne, 1985). Unfortunately, power estimation in the analysis of a multivariate model is a difficult, somewhat arbitrary procedure, and such power estimates have not been reported in most published studies.

The main reason for evaluating power is to gain some understanding of precision of estimation in a particular situation, to guard against the possibility that a model is "accepted" simply because of insufficient power. An alternative (and actually more direct) approach to the evaluation of precision is to *construct a confidence interval on the population noncentrality parameter* (or some particularly useful function of it). This approach, first suggested in the context of covariance structure analysis by Steiger and Lind (1980) offers two worthwhile pieces of information at the same time. It allows one, for a particular model and data set, to express (1) how bad fit is in the population, and (2) how precisely the *population* badness of fit has been determined from the *sample* data.

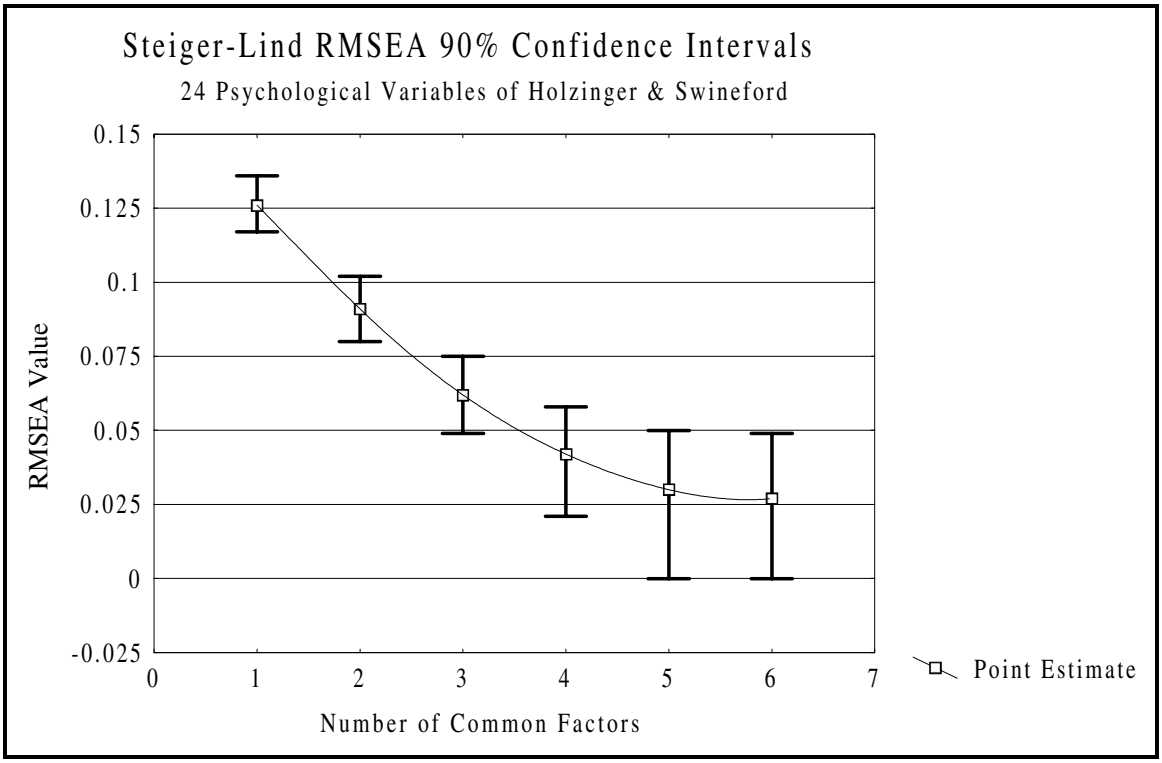
Steiger (1989, 1990b) implemented three noncentrality-based indices of fit in the computer program EzPATH, including the index originally proposed by Steiger and Lind (1980). All these indices can be computed with confidence intervals. Alternative noncentrality-based indices have been proposed by McDonald (1989) and Bentler (1989). Recent revisions of LISREL and CALIS have incorporated Steiger's RMS index, which Browne and Cudeck (1992) also recommend in a recent article.

A rather different approach, involving a cross-validation rationale, was introduced by Cudeck and Browne (1983), and extended in subsequent articles by Browne and Cudeck (1989, 1992). Bandalos (1993) compared the performance of the Browne-Cudeck (1989) single-sample index with the criterion of Akaike (1973) and an alternative information-theoretic criterion proposed by Bozdogan (1987).

Still another alternative line of research led to the development of a surprisingly large number of other fit indices. Bentler and Bonnet (1980), stimulated by concerns about the deficiencies in the sequential chi-square approach, introduced a fit index that compared the badness of fit of the model being tested with that of a "null" baseline model. James, Mulaik, and Brett (1982) corrected this statistic for model parsimony, and numerous other quasi-statistical indices followed. None has a well-established statistical basis. Indeed, Steiger (1989) and Maiti and Mukherjee (1990) demonstrated that two of the most popular quasi-statistical indices, the GFI and AGFI of Jöreskog and Sörbom (1984), could be viewed as badly biased estimators of equivalent population quantities estimated correctly (with a confidence interval) by two of Steiger's (1989) indices.

As an example of the advantages offered by the noncentrality interval estimation approach, consider the classic 24 psychological tests of Holzinger and Swineford (1939). The maximum likelihood "sequential chi-square" approach with $\alpha = .05$ yields 7 factors, as the probability level of the test with 6 factors is .0496. In Figure 1, we plot the 90% confidence intervals and point estimates for the RMSEA index of fit, as generated by the structural modeling module SEPATH of *Statistica for Windows.*

Figure 1. Choosing the number of common factors with the RMSEA index.



Following the generally accepted guideline (Steiger, 1989; Browne & Cudeck, 1992) of .05 for good fit, we find that fit is not adequate until 4 factors are extracted, improves slightly for 5

factors, and remains almost the same for 6 factors. Remember, the RMSEA index compensates for model complexity, so small increments in fit resulting (inevitably) from adding a factor do not improve the index as much as they might improve the chi-square probability level. In this example, the RMSEA index declares fit acceptable at 4 factors, while the chi-square probability level is .002. One might entertain a 5 factor solution, but 6 factors is clearly overfactoring. The moderate width of the confidence intervals indicates that quality of fit is somewhat uncertain, due to the relatively small sample size (145) relative to the number of variables. Larger sample size would be reflected in narrower, more precise confidence intervals.

It is important to note how the RMSEA technique eliminates many of the apparent paradoxes of statistical testing in factor analysis. For example, the RMSEA confidence interval in the 4-factor case clearly misses zero, but is within an acceptable range. Thus, while fit is not perfect (reflected in the chi-square probability of .002 for the hypothesis of perfect fit), it is good. A larger sample size would have made this interval more precise. Thus, *large sample size works for the experimenter with this technique*. It is quite common, using the RMSEA approach, to conclude that population fit is almost certainly not perfect, but almost certainly is very good.

3.4 Renewal and Extension of the "Factor-Component" Debate

A significant source of controversy among factor analysts is the choice of a basic model. Some advocate the common factor analysis model, often for the reasons discussed in Section 2 above, while others stress the benefits of a component approach and usually advocate a principal components analysis. A broad view of the parameters of the ongoing debate can be obtained by scanning the articles in the January 1990 edition of *Multivariate Behavioral Research*, which featured an article by Velicer and Jackson (1990), 10 commentaries, and a rebuttal by the authors.

One fact that clearly emerges from the *Multivariate Behavioral Research* volume is that the debate is far from trivial, and far from decided. Another noteworthy aspect of the volume is the diversity of points of view of the psychometricians, statisticians, social science researchers, and philosophers of science who wrote the commentaries. It is obviously impossible to do justice to several hundred pages of writing from such a knowledgeable group, but I will again inject some personal biases by reviewing some of the key points made by the various authors.

Arguments in Favor of Common Factor Analysis

(1) *In some cases, it is the "right model."* Consider again the discussion in Section 2.1. It would seem that, if your model assumes three correlated signals emanating from three common sources, with normally distributed random noise degrading these signals, you should be fitting the common factor model to your data, since it agrees precisely with the hypothesized physical process. Why fit a model that fails to agree?

(2) *Factors "go beyond the test space," while components, as linear combinations of the observed variables, are confined to the test space. Hence, factor analysis can discover latent variables that component analysis cannot.* This view is one of the most frequently expressed rationales for choosing factor analysis over component methods. It is repeated, for example, by Mulaik (1990), whose 1972 textbook *The Foundations of Factor Analysis* remains, in my opinion, the clearest and most important account of factor analytic methodology ever written.

(3) *Factors generalize beyond the set of variables currently analyzed.* If some large domain of variables is explained by m factors, any subset of the variables can be explained by at most m

of the *same* factors. In other words, factor loadings can remain consistent across subsets of variables that fit those factors. A similar property does not hold for component loadings.

(3) *Factor analysis includes component analysis as a special case.* As unique variances go to zero, factors become components. So component analysis is a special case of factor analysis. Why not use the more general model?

(5) *Factor analysis yields statistical tests.* The availability of a wide range of asymptotic distributional results and related statistical procedures makes factor analysis much more useful than component analysis as an analytic technique.

Arguments in Favor of Component Analysis

(1) *It is the "right model," because it leads to the correct way of identifying and naming the major sources of variation in your data.* Suppose you are a gourmet cook, and a friend ask you to identify "what's in" your favorite dishes. Suppose you have a list of ingredients that are common to all the dishes. By listing the amount of each ingredient in each dish, and how they are combined, you convey the essence of each dish's meaning. One would expect similar principles to hold in the realm of multivariate modeling. Specifically, understanding a common factor would seem to require knowledge of what it is composed of. Equation (4) shows that everything we really know about the factors in $\mathbf{x}$ can be expressed as a linear function of $\mathbf{y}$**.** By examining the linear weights that relate $\mathbf{x}$ to $\mathbf{y}$, we can understand the "recipe" for $\mathbf{x}$. Hence, to name our common factors, we really should be looking at the linear weights that produce $\hat{\mathbf{x}}$ from $\mathbf{y}$. Specifically, we should be examining the matrix $\mathbf{F}'\mathbf{\Sigma}^{-1}$. What people do, in practice, is, in essence, examine $\mathbf{F}'$ instead to try to understand what variables are common to each factor. The patterns of large and small values in these two matrices tends to be the same, especially with strong simple structure, but they need not be the same in common factor analysis.

There is no such confusion in component analysis. Component analysts acknowledge that components are identified by how they are composed from the manifest variables. The content of the matrix $\mathbf{B}'$ in Equation (8) determines this, and components may readily be named by examining this matrix.

(2) *Common factors are outside the test space, but the vector* $\boldsymbol{\varepsilon}$ *that that takes them there is arbitrary. Discovering a factor pattern is not the same as discovering a factor. Hence, the only determinate part of a common factor is, indeed, a component.* The view that "going beyond the test space" is a theoretical advantage for common factors is based on a long-standing, fundamental confusion shared by many proponents of factor analysis. These proponents apparently fail to distinguish between discovering the regression weights for predicting $\mathbf{y}$ from $\mathbf{x}$, and identifying $\mathbf{x}$ itself. At any point in time, your knowledge of a factor must be based on the variables you have factor analyzed. This is expressed clearly in Equation (4), which shows that everything you know about $\mathbf{x}$ can be expressed as two random vectors, one that can be specified precisely, and one that is essentially arbitrary. You know some places where the $\boldsymbol{\varepsilon}$ of Equation (4) cannot be. (It cannot project into the test space.) However, you cannot specify where it is. Thus, what you *know* about the common factors is actually a vector of components, the $\hat{\mathbf{x}}$ of Equation (4), which can be written in the form $\mathbf{B}'\mathbf{y}$.

The examples of generalizability offered by common factor proponents always involve the article of faith that, somehow, the additional variables you have not yet factor analyzed will be fit by the same common factors that fit your current set of tests. The mathematical demonstrations of generalizability are all based on this supposition, which one might well accord zero prior probability. The more realistic expectation, which we would expect to be true with probability

near 1 for "real world" situations, is as follows: (1) The tests you may add to your test battery in the future will almost certainly not fit the same common factors you have now. (2) When you add these tests to the current ones, the common factors for the combined battery will not be the same as for the current battery. This simple reality seems to have eluded such writers as Widaman (1993), who devotes nearly 50 pages to examining the unrealistic alternative scenario.

(3) *Numerical calculations for component methods are much easier and more reliable than corresponding factor analysis calculations.* Factor analysts frequently base their demonstrations of the difference between component analysis and factor analysis on calculations performed on "population" matrices where the common factor model just happens to fit perfectly. In practice, of course, such population matrices do not exist, because the factor model does not fit perfectly in the population or in the sample. Moreover, with rare exceptions, we are usually treating the current data as a sample, rather than a population. In such cases, it makes sense to have at our disposal relevant information about the computational performance of a proposed method.

It is well known that, for example, principal component analysis calculations proceed much faster than the equivalent calculations in maximum likelihood factor analysis. What is less well-understood is how unstable and unreliable factor analysis computations can become when sample sizes are high and unique variances are low. In such cases, the prior probability of encountering a Heywood case (i.e., negative estimated element of $\mathbf{U}$) may be disturbingly high.

(4) *The news is starting to come in about non-robustness of factor analytic testing, and the performance problems of maximum likelihood factor analysis. Most of the news is not good.* As Hu, Bentler, and Kano (1992) put it, "Can Test Statistics in Covariance Structure Analysis be Trusted?" Statistical testing in factor analysis can be quite non-robust to precisely the kinds of data (categorical, high kurtosis, etc.) encountered in social science research. The advantages of such statistical testing, in the absence of reliably robust methods, may be illusory.

(5) *Factor and component methods usually yield results with no important differences.* As numerous authors have pointed out, there exist well-defined situations where factor loadings and component loadings differ. Such situations usually involve hypothetical "population" covariance matrices that fit the common factor model perfectly, and are somehow known, as opposed to population matrices that do not fit perfectly, and analyses based on random samples from such a population. Other authors, most noticeably Velicer and coworkers (e.g., Velicer, 1977; Velicer & Fava, 1987; Fava & Velicer, 1992) have examined a broad range of situations and found that usually there is no substantial difference between solutions produced by the two methods.

As the reader can see, it is far from clear that the factor component debate has been settled, or is likely to be settled any time in the near future.

3.5 Changes in the Computing Environment and the Development of "User-Friendly" Software

In 1980, microcomputers were in their infancy. By 1983, the IBM PC had been released, and the ensuing decade saw an explosive rise in the access to microcomputing power. As the power to price ratio grew, the emphasis in statistical computing turned toward increased "user-friendliness," and improved access to statistical graphical capabilities. Ultimately, mainframe structural programs like LISREL and EQS migrated to the microcomputer environment, to be joined by competitors like EzPATH designed specifically for microcomputers. All of these programs can perform confirmatory factor analysis easily and efficiently.

As Steiger and Fouladi (1990) predicted, *Microsoft Windows* has become the dominant operating system on personal computers, and has opened up new avenues of development for statistical programmers. Although most popular DOS-based statistical packages have migrated successfully into the *Windows* environment, few have summoned the effort necessary to produce the radical redesign required by a true *Windows* application. As a result, many of the programs currently describing themselves as "*Windows*" programs are that in name only. They have the look and feel of a DOS application that has been set up to run in a window, rather than an application designed from the ground up for *Windows*. At least one program, *Statistica/W*, seems to have broken the mold and made full, interactive use of the graphics technology available in *Windows*. One would hope that others will be quick to follow.

A number of the programs now produce high resolution two and three-dimensional plots of factor loadings. As hardware and software continue to evolve, one looks forward to further advances, such as interactive graphical rotation routines that allow the user to "grab" a factor axis and rotate it with the mouse, and automatic, fully integrated bootstrapping routines. However, one must be patient, and remember the enormous intellectual and financial pressures faced by software developers, who must attempt to implement these changes within an incredibly complex and constantly changing development environment.

In the final analysis, it seems that, except for aforementioned improvements in the quality of graphical output, little has changed in basic exploratory factor analysis software in the last 15 years. On the other hand, there has been explosive growth and improvement in structural modeling software used to perform confirmatory factor analysis.

3.6 Specialized Theory for Factor Analysis of "Categorical" Data

Frequently, especially when developing questionnaire scales, the variables subjected to factor analysis are not continuous, or even close to it. Rather, they are often dichotomous or polychotomous Likert scale items. Suppose that, underlying such categorical data in the random vector $\mathbf{y}$, are variables $\mathbf{y}^*$ that have a multivariate normal distribution with covariance matrix $\mathbf{\Sigma}^*$ fitting the common factor model. It may well be that $\mathbf{\Sigma}$, the covariance matrix for the observed categorical variables will not fit the common factor model, or that the matrix of factor loadings $\mathbf{F}$ obtained from a factor analysis of observations on $\mathbf{y}$ will not correctly mirror the factor loadings $\mathbf{F}^*$ obtained from a factor analysis of observations on $\mathbf{y}^*$.

If one is willing to make strong assumptions (generally multivariate normality) about the distribution of $\mathbf{y}^*$, one may estimate population correlation matrix P^*, and its factor pattern $\mathbf{F}^*$. Generally, the model assumes that the multivariate normal observations on $\mathbf{y}^*$ have been transformed to categorical variates through the application of set of *threshold values* $\alpha_{i,k(i)}$. For example, Lee, Poon, & Bentler (1992) model the process that

$$y_i = k(i) \quad \text{if } \alpha_{\mathrm{i,\,k(i)}} \le y_i^* < \alpha_{i,k(i)+1} \tag{14}$$

for $i = 1, 2, \ldots m$, and $k(i) = 1, 2, \ldots J(i)$, where $\alpha_{i,k(i)}$ is the threshold parameter with $\alpha_{i,1} = -\infty$ and $\alpha_{i,J(i)+1} = +\infty$. Thus, besides the traditional estimation of the elements of $\mathbf{F}$, $\mathbf{P}$, and $\mathbf{U}$, one must also estimate the threshold parameters.

Several competing methods have been proposed for performing this estimation, some quite new. The methods commonly obtain estimates of polychoric and polyserial correlations at one or

more stages of computation, along with an estimate of their asymptotic covariance structure, then employ generalized least squares estimation to obtain parameter estimates and a chi-square statistic.

Muthén (1984) proposed a three-stage estimation procedure that he implemented in the computer program LISCOMP (Muthén, 1987).

Jöreskog and Sörbom (1988a, 1988b) implement a procedure for analyzing categorical variables in their computer programs LISREL VII and PRELIS. Lee, Poon, and Bentler (1992), commenting on LISREL VII, noted "unfortunately, no statistical development of the underlying procedure has been given, and hence, it is impossible to provide a theoretical comparison of our procedure with theirs, except that our procedure is not confined to the analysis of correlation structures."

Lee, Poon and Bentler (1992) describe a two-stage procedure for analyzing structural models with continuous and categorical variables. This work is the culmination of a series of contributions in this area by these authors (Lee & Poon, 1986; Lee, Poon, & Bentler 1990a, 1990b; Poon & Lee, 1987). Lee, Poon and Bentler (1992) pointed out that, due to the particular estimation approach used by Muthén (1984), the chi-square statistic reported by LISCOMP does not have the correct asymptotic distribution. They performed a brief Monte Carlo study that confirmed this, but also showed slightly better (but very similar) performance in the quality of the parameter estimates LISCOMP, as compared to their own procedure.

Jöreskog (1993) describes some interesting new work on the analysis of categorical variables by Quiroga (1992), who, in her doctoral dissertation, has developed an extended notion of the polychoric correlation. In this extended polychoric, the densities of the underlying bivariate distributions are no longer restricted to be bivariate normal, but instead can be modeled as a weighted sum of two quantities — the bivariate standard normal distribution and the product of two univariate skew-normal densities (Azzalini, 1985). Since this density includes the bivariate normal as a special case, it would not be surprising to find that it fits typical data sets better than the more traditional data sets. Full technical details of these promising new methods will hopefully be available soon.

Since categorical data are so popular in social science research, it would seem reasonable to expect a continuation of the substantial research interest devoted to this topic. We will discuss below whether this attention is fully warranted.

3.7 Special Techniques for Correct Analysis of Correlations in Confirmatory Factor Analysis

Problems When a Correlation Matrix is Analyzed Directly

Traditional models and procedures for analysis of covariance structures are based on the assumption that the *sample covariance matrix* is being analyzed. The sampling theory underlying the test statistic and resulting standard error estimates assumes that *all elements* of the input "covariance matrix" **S** are in fact random variables. This is not always convenient. In many situations, the sample covariance matrix is ill-scaled, and convergence of iterative algorithms may suffer as a result. In addition, variables standardized to the same scale (i.e., unit variance) are generally easier to interpret. In some cases, the analyst is working from a very old published article containing the correlation matrix, but not the covariance matrix. These considerations have led many researchers to input sample *correlation* matrices to covariance structure analysis pro-

grams as though they were covariance matrices. Cudeck (1989) points out that this often can lead to incorrect results. In particular, unless the model is invariant under diagonal rescaling, the calculated standard errors will almost certainly be incorrect, and the observed test statistic may also be incorrect.

The reason for this problem is not difficult to grasp. If a correlation matrix is input, the elements of the diagonal are no longer random variables — they are always 1. Clearly, then, when a covariance matrix is replaced by a correlation matrix, a random vector containing $p(p+1)/2$ random variables has been replaced by a random vector with only $p(p-1)/2$ elements free to vary.

This fact was dramatized by Lawley and Maxwell (1971), who gave a numerical example of a confirmatory factor analysis where the estimated standard errors of the loadings when a correlation matrix is analyzed as if it were a covariance matrix can be more than 2.5 times as large as the correct values. Lawley and Maxwell (1971) gave formulas for calculating correct standard errors when a correlation matrix is analyzed. Despite the fact that developers of structural modeling software were well aware of the problem, incorrect values were given routinely by many covariance structure analysis programs when correlation matrices are input as covariance matrices.

Browne's Technique for Correct Analysis of Correlation Structures

A much more general technique than the Lawley-Maxwell approach, which reproduces their results as a special case, was pioneered by Browne (1982), and implemented in the computer program RAMONA (Mels, 1989; Browne & Mels, 1992). This technique combines constrained estimation and a modified structural model to achieve a procedure that yields correct standard errors when a correlation matrix is analyzed. Browne's procedure has also been implemented in SEPATH (Steiger, 1994) the structural equations module that is part of the program *Statistica/W*. Thus, RAMONA and SEPATH eliminate the problem of analyzing a correlation matrix correctly.

4. Challenges and Directions for Future Research

In this section I offer a decidedly personal perspective on research questions that continue to challenge and trouble me. Space considerations restrict me to a subset of the problems I would like to discuss here.

4.1 Are Special Methods for Categorical Data Worth the Trouble?

There seems little doubt that categorization can distort correlation, and hence affect factor analysis solutions. Moreover, considerable effort has been invested in developing methods for dealing with categorical data under some fairly restrictive assumptions. The problem is, can these methods be relied upon to work, and are they worth the trouble?

The first problem that must be faced head on is that we cannot really test the assumption that underlying variables have a particular distribution. We can test for evidence that the observed categorical proportions are in line with restrictive distributional theory (e.g., multivariate normality), and reject that hypothesis if departures from theory are extreme enough. But failure to reject does *not* imply agreement with the assumption. For example, your observed categorical

proportions may fit a multivariate normal model, but they would also fit another model, i.e., a multivariate categorical model with proportions equal to those you just observed. Assuming some kind of underlying continuous distribution obviously requires faith. In many cases, it may indeed be blind faith. This fact often gets lost in the shuffle when one is wading through the difficult technicalities involved in developing these methods.

Generally, we would like our statistical decisions to be guided by a little more than faith. Is there any evidence that these techniques are robust when the underlying distributions are *not* multivariate normal? How powerful are the methods available for detecting non-normality in the presumed underlying variates?

So far, the amount of Monte Carlo evidence amassed concerning these questions would barely fill a thimble. Yet, advanced commercial software performing the methodology has been in circulation for about 7 years.

This raises another point, which is probably worthy of serious research investigation. Suppose we decide that methods specifically designed for categorical variables are, in fact, not robust, and cannot be relied on. What can we do? One suggestion might be, "Stop using categorical variables directly!" In other words, always use composites of several items, so that the total number of possible score values is at least 15. My strong hunch is that, if scales were developed using unit weighting on the basis of ordinary component analysis, and these scale scores were used instead of individual items, that there would indeed be no need for special techniques for categorical variables, because the resulting scores would be "close enough" to continuous variates. (Why doesn't someone develop a truly authoritative Monte Carlo study to investigate this?) What this requires, of course, is that careful attention to measurement considerations *precede* research using the actual measurement. All too often, we see advanced psychometric analyses applied to single item measures that have never been used before, and whose properties are a complete mystery.

In the final analysis, then, I think we need a lot more information, much of which can only come from carefully executed, time-consuming Monte Carlo studies. In that regard, I think Bentler (1989) is to be applauded for building advanced Monte Carlo capability into EQS, and allowing EQS users to investigate for themselves the efficacy of new procedures presented in that software. I have attempted to follow his lead by building extensive, fully integrated Monte Carlo routines into my own software (Steiger, 1994), and I would urge others to do so also.

4.2 What Should We Do About Non-Normality?

We must frequently deal with data that are continuous-nonnormal or categorical. With such data, the classic normal theory methods can give test statistics that are seriously biased. However, discussion in some of the above sections indicates that, although considerable progress has been made in dealing with the problem, it is far from resolved. Particularly depressing is the paucity of Monte Carlo studies attacking basic performance issues. Such studies require huge amounts of computer time, and are made difficult by the commercialized and semi-documented nature of the procedures that, ideally, one would like to compare. (See Section 4.4 below.)

At present, a key question is whether the ADF approach can be "rescued" in some way, or whether rescaled statistics like the Satorra-Bentler scaled chi-square statistic are the best avenue of approach. It appears that the major weakness in the ADF approach is the poor estimation of the weight matrix at small samples. Perhaps alternative estimation approaches like bootstrapping the weight matrix might provide some benefits.

4.3 How Should We Integrate Exploratory and Confirmatory Approaches to Factor Analysis?

Exploratory and confirmatory factor analysis can be viewed as discrete techniques, one to be applied when a domain is first investigated, the other to be applied in testing a well-formulated hypothesis about factorial structure. However in practice the two techniques may be used together in an "exploratory-confirmatory" approach. Jöreskog (1978) illustrated the technique by re-examining the 24 psychological variables of Holzinger and Swineford (1939). In this example, Jöreskog first performed an exploratory maximum likelihood factor analysis. After a rotation to simple structure, he then rotated each factor into congruence with a reference variable by choosing the largest loading on each factor, then rotating obliquely so that row of **F** containing this largest loading had zero loadings on the other factors. Finally, he eliminated "non-significant" loadings by comparing each loading with its estimated standard error, and eliminating loadings that were less than two standard errors in absolute value. The result was a "confirmatory" factor model with simple structure, with a chi-square quite comparable to the original exploratory factor model.

One can, of course, criticize the procedure on numerous statistical grounds. For example, it appears no serious consideration was given to familywise error rates or the problems of *post hoc* selection. The larger question remains: "What is a sound way of integrating exploratory and confirmatory approaches to factor analysis." It is a question that does not seem to have been given the careful attention it deserves.

4.4 How Can We Reconcile Academic and Commercial Considerations in Statistical Software?

Statistical practice is dictated, to a considerable extent, by available statistical software. As statistical software becomes increasingly sophisticated, there will be inevitable pressures to release students in the social sciences from the "burden" of actually understanding the computational aspects of the procedures they perform. One can debate the propriety of this. There is a similar, slightly more subtle trend that I have noticed recently that may interact with the preceding one in some particularly dangerous ways. This is the tendency for producers of statistical software to include "new" features in their products without testing them, and without documenting them properly. This tendency would not be particularly distressing if users were sufficiently sophisticated to understand the dangers of using techniques they don't comprehend. However, there is substantial evidence in the field of structural modeling software that highly questionable, or simply erroneous procedures, once introduced in such software, take on instant legitimacy in the minds of the true believers.

For example, Lee, Poon, and Bentler (1992) state that the "chi-square" statistic computed by LISCOMP (Múthen, 1987) does not have the correct distribution, and support their assertion with Monte Carlo evidence. The Monte Carlo procedure used by Lee, Poon, and Bentler seems straightforward enough. The question remains, was the LISCOMP chi-square not tested prior to release? If not, why not?

Because of the way the software industry now functions, it is far easier for developers of new techniques to test them than it is for the users to verify them. This is because the design of many software programs does not lend itself to simple Monte Carlo analysis. The computational

engines of such software are not in the public domain, and are not available as callable subroutines.

Earlier in this paper, I discussed the work on analysis of categorical data. Suppose, for example, an independent researcher wished to compare the techniques reported in these papers in an authoritative Monte Carlo study. Some of the techniques are available in commercial software, others are not. Some commercial software includes a Monte Carlo interface, other software does not. It would be difficult for such a researcher to write routines to perform the related computations, because most of the published papers are difficult to read, and in several cases key derivative expressions are left out "for brevity." Also, one finds on occasion that key published expressions contain typographical errors of considerable subtlety. This reduces the probability that an independent researcher can successfully perform the required computations. Unless the authors can make callable subroutines available, such a researcher is left with the prospect of performing an *enormous* amount of work to test routines that may be "superseded" by the time a study is ready for publication. This is the prospect that confronts most "outsiders" (i.e., non-developers) who wish to test various procedures. It is not surprising that Monte Carlo evidence has been slow in coming.

These problems occur in an environment where software developers face incredible pressures to produce "new" features in their software, and to keep these developments out of the hands of their competitors. As these pressures continue, we can expect to see more "new" developments that turn out to be incorrect or seriously flawed.

Since software availability will probably drive statistical practice even more in the future than it does today, what steps should we take to prevent serious abuses? Clearly, we do not want software to lag years behind theory, as now seems to be the case with the best-selling general statistics packages. (Just when do these manufacturers plan to implement bootstrapping in the architecture of their programs?) On the other hand, we would like to see greater accountability, and more open documentation, in the commercial software development process. Given the complexity of the procedures now being developed, and the growing demands of users for "friendly" and sophisticated interfaces, I suspect that it will be difficult for independent individuals to develop software that remains "acceptable" to the majority of users. There is simply too much to know, and too little time to learn it. Consequently, I believe, "academic" software may soon become a thing of the past. Would-be independent developers will discover that they have little choice but to join forces with commercial publishers, to obtain the help they need to make their techniques available in a form that will gain them acceptance.

The impact of this trend, and the dangers it poses, are issues that I would like to see discussed more openly, and more carefully, in psychometric journals. At the very least, such journals should avoid subtlety in publicizing serious errors in published software.

5. References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second International Symposium on Information Theory.* Budapest: Akademiai Kiado.

Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute: Proceedings of the 44th Session, Volume 1.* Pages 277-290.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317-332.

Amemiya, Y. (1985). *On the goodness-of-fit tests for linear structural relationships* (Tech. Rep. No. 10). Stanford, CA: Stanford University, Econometrics Workshop.

Amemiya, Y., & Anderson, T.W. (1990). Asymptotic chi-square tests for a large class of factor analysis models. *Annals of Statistics, 18,* 1453-1463.

Anderson, T.W., & Amemiya, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *Annals of Statistics, 16,* 759-771.

Anderson, T.W., & Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: The University of California Press.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics, 12,* 171-178.

Bandalos, D.L. (1993). Factors influencing cross-validation of confirmatory factor analysis. *Multivariate Behavioral Research, 28,* 351-374.

Bentler, P.M. (1989). *EQS Structural equations program manual.* Los Angeles, CA: BMDP Statistical Software.

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

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika, 52,* 345-370.

Browne, M.W. (1982). Covariance Structures. In D.M. Hawkins (Ed.) *Topics in Applied Multivariate Analysis.* Cambridge: Cambridge University Press.

Browne, M.W. (1984). Asymptotically distribution free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62-83.

Browne, M.W. (1985). Robustness of normal theory tests of fit of factor analysis and related models against nonnormally distributed common factors. Paper presented at the Fourth European Meeting of the Psychometric Society and Classification Societies, Cambridge, England.

Browne, M.W. (1987). Robustness of statistical inference in factor analysis and related models. *Biometrika, 74,* 375-384.

Browne, M.W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24, 445-455.*

Browne, M.W., & Cudeck, R. (1992). Alternative ways of assessing model fit. In K.A. Bollen and J. S. Long (Eds.) *Testing structural equation models.* Beverly Hills, CA: Sage.

Browne, M.W., & Mels, G. (1992). *RAMONA user's guide.* Department of Psychology, Ohio State University.

Browne, M.W., & Shapiro, A. (1988). Robustness of normal theory methods in the analysis of linear latent variate models. *British Journal of Mathematical and Statistical Psychology, 44,* 347-357.

Cudeck, R., & Browne, M.W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147-167.

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

Guttman, L. (1992) The irrelevance of factor analysis for the study of group differences. *Multivariate Behavioral Research, 27,* 173-204.

Holzinger, K.J., & Swineford, F.A. (1939). *A study in factor analysis: the stability of a bi-factor solution*. University of Chicago: Supplementary Monographs, No. 48.

Hu, L., Bentler, P.M., & Kano, Y. Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351-362.

James, L.R., Mulaik, S.A., & Brett, J.M. (1982). *Causal analysis. Assumptions, models, and data.* Beverly Hills, CA: Sage Publications.

Jensen, A.R. (1985). The nature of the black-white difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8,* 246-263.

Jöreskog, K.G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32,* 443-482.

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

Jöreskog, K.G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43,* 443-477.

Jöreskog, K.G. (1993). Latent variable modeling with ordinal variables. In K.Haagen, DJ. Bartholomew, & M. Deistler (Editors). *Statistical modeling and latent variables.* London: Elsevier Science Publishers.

Jöreskog, K.G., & Sörbom, D. (1984). *Lisrel VI. Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods*. Mooresville, Indiana: Scientific Software.

Jöreskog, K.G., & Sörbom, D. (1988a). *PRELIS: A preprocessor for LISREL.* Mooresville, IN: Scientific Software.

Jöreskog, K.G., & Sörbom, D. (1988b). *LISREL VII. A guide to the program and applications.* Mooresville, IN: Scientific Software.

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.

Lawley, D.N., & Maxwell, A.E. (1971). *Factor analysis as a statistical method. (2nd. Ed.).* London: Butterworth & Company.

Lee, S.-Y., & Poon, W.-Y. (1986). Maximum likelihood estimation of polyserial correlations. *Psychometrika, 51,* 113-121.

Lee, S.-Y., Poon, W.-Y., & Bentler, P.M. (1990a). A three-stage estimation procedure for structural equation models with polytomous variables. *Psychometrika, 55,* 45-51.

Lee, S.-Y., Poon, W.-Y., & Bentler, P.M. (1990b). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics and Probability Letters, 9,* 91-97.

Lee, S.-Y., Poon, W.-Y., & Bentler, P.M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika, 57,* 89-105.

Maiti, S. S. & Mukherjee, B. N. (1990). A note on the distributional properties of the Jöreskog-Sörbom fit indices. *Psychometrika, 55,* 721-726.

Matsueda, R.L., & Bielby, W.T. (1986). Statistical power in covariance structure models. In N.B. Tuma (Ed.) *Sociological methodology.* Washington, D.C.: American Sociological Association.

McDonald, R.P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97-103.

Mels, G. (1989). A general system for path analysis with latent variables. M.S. Thesis: Department of Statistics, University of South Africa.

Mooijaart, A., & Bentler, P.M. (1991). Robustness of normal theory statistics in structural equation models. *Statistica Neerlandia, 45,* 159-171.

Mulaik, S.A. (1972). *The foundations of factor analysis.* New York: McGraw-Hill.

Mulaik, S.A. (1986). Factor analysis and Psychometrika: major developments. *Psychometrika, 51,* 23-33.

Mulaik, S.A. (1990). Blurring the distinctions between component analysis and common factor analysis. *Multivariate Behavioral Research, 25*, 53-59.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Muthén, B. (1987). *LISCOMP: Analysis of linear structural equations using a comprehensive measurement model.* Mooresville, IN: Scientific Software.

Poon, W.-Y., & Lee, S.-Y. (1987). Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika, 52,* 409-430.

Quiroga, A.M. (1992). *Studies of the polychoric correlation and other correlation measures for ordinal variables.* Ph.D. Dissertation, Uppsala University, Department of Statistics.

Satorra, A., & Saris, W.E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50,* 83-90.

Satorra, A., & Bentler, P.M. (1986). Some robustness properties of goodness of fit statistics in covariance structure analysis. *American Statistical Association 1986 Proceedings of the Business and Economics Sections,* (pp. 549-554). Alexandria, VA: American Statistical Association.

Schönemann, P.H., and Steiger, J.H. (1976). Regression component analysis. *British Journal of Mathematical and Statistical Psychology*, *29*, 175-189.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Steiger, J.H. (1979a). Factor indeterminacy in the 1930's and in the 1970's... some interesting parallels. *Psychometrika*, *44*, 157-167.

Steiger, J.H. (1979b). The relationship between external variables and common factors. *Psychometrika*, *44*, 93-97.

Steiger, J.H. (1989). EZPATH: A supplementary module for SYSTAT and SYGRAPH. Evanston, IL: SYSTAT, Inc.

Steiger, J.H. (1990a). Some additional thoughts on components and factors. *Multivariate Behavioral Research, 25,* 41-45.

Steiger, J.H. (1990b). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioral Research, 25,* 173-180.

Steiger, J.H. (1994). SEPATH - A Statistica for Windows structural equations modeling program. In F. Faulbaum (Ed.) *Softstat '93: Advances in statistical software 4.* Stuttgart: Gustav Fischer Verlag.

Steiger, J.H., & Fouladi, R.T. (1990). Some key emerging trends in statistical and graphical software for the social scientist. *Social Science Computing Review*, 8, 627-664.

Steiger, J.H., and Schönemann, P.H. (1978). A history of factor indeterminacy. In S. Shye (ed.) *Theory construction and data analysis in the behavioral sciences.* San Francisco: Jossey-Bass.

Steiger, J.H., & Lind, J.C. (1980). Statistically-based tests for the number of common factors. Paper presented at the annual Spring Meeting of the Psychometric Society in Iowa City. May 30, 1980.

Steiger, J.H., Shapiro, A., & Browne, M.W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, *50*, 253-264.

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

Velicer, W.F., & Fava, J.L. (1987). An evaluation of the effects of variable sampling on component, image, and factor analysis. *Multivariate Behavioral Research, 22,* 193-210.

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.

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