Psychological Bulletin
1988, Vol. 103, No. 3, 411-423

0033-2909/88/$00.75

Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach

James C. Anderson
J. L. Kellogg Graduate School of Management
Northwestern University

David W. Gerbing
Department of Management
Portland State University

In this article, we provide guidance for substantive researchers on the use of structural equation modeling in practice for theory testing and development. We present a comprehensive, two-step modeling approach that employs a series of nested models and sequential chi-square difference tests. We discuss the comparative advantages of this approach over a one-step approach. Considerations in specification, assessment of fit, and respecification of measurement models using confirmatory factor analysis are reviewed. As background to the two-step approach, the distinction between exploratory and confirmatory analysis, the distinction between complementary approaches for theory testing versus predictive application, and some developments in estimation methods also are discussed.

Substantive use of structural equation modeling has been growing in psychology and the social sciences. One reason for this is that these confirmatory methods (e.g., Bentler, 1983; Browne, 1984; Joreskog, 1978) provide researchers with a comprehensive means for assessing and modifying theoretical models. As such, they offer great potential for furthering theory development. Because of their relative sophistication, however, a number of problems and pitfalls in their application can hinder this potential from being realized. The purpose of this article is to provide some guidance for substantive researchers on the use of structural equation modeling in practice for theory testing and development. We present a comprehensive, two-step modeling approach that provides a basis for making meaningful inferences about theoretical constructs and their interrelations, as well as avoiding some specious inferences.

The model-building task can be thought of as the analysis of two conceptually distinct models (Anderson & Gerbing, 1982; Joreskog & Sorbom, 1984). A confirmatory measurement, or factor analysis, model specifies the relations of the observed measures to their posited underlying constructs, with the constructs allowed to intercorrelate freely. A confirmatory structural model then specifies the causal relations of the constructs to one another, as posited by some theory. With full-information estimation methods, such as those provided in the EQS (Bentler, 1985) or LISREL (Joreskog & Sorbom, 1984) programs, the measurement and structural submodels can be estimated simultaneously. The ability to do this in a one-step analysis approach, however, does not necessarily mean that it is the preferred way to accomplish the model-building task.

In this article, we contend that there is much to gain in theory testing and the assessment of construct validity from separate estimation (and respecification) of the measurement model prior to the simultaneous estimation of the measurement and structural submodels. The measurement model in conjunction with the structural model enables a comprehensive, confirmatory assessment of construct validity (Bentler, 1978). The measurement model provides a confirmatory assessment of convergent validity and discriminant validity (Campbell & Fiske, 1959). Given acceptable convergent and discriminant validities, the test of the structural model then constitutes a confirmatory assessment of nomological validity (Campbell, 1960; Cronbach & Meehl, 1955).

The organization of the article is as follows: As background to the two-step approach, we begin with a section in which we discuss the distinction between exploratory and confirmatory analysis, the distinction between complementary modeling approaches for theory testing versus predictive application, and some developments in estimation methods. Following this, we present the confirmatory measurement model; discuss the need for unidimensional measurement; and then consider the areas of specification, assessment of fit, and respecification in turn. In the next section, after briefly reviewing the confirmatory structural model, we present a two-step modeling approach and, in doing so, discuss the comparative advantages of this two-step approach over a one-step approach.

Background

Exploratory Versus Confirmatory Analyses

Although it is convenient to distinguish between exploratory and confirmatory research, in practice this distinction is not as clear-cut. As Joreskog (1974) noted, "Many investigations are to some extent both exploratory and confirmatory, since they involve some variables of known and other variables of unknown compo-

This work was supported in part by the McManus Research Professorship awarded to James C. Anderson.

We gratefully acknowledge the comments and suggestions of Jeanne Brett, Claes Fornell, David Larcker, William Perreault, Jr., and James Steiger.

Correspondence concerning this article should be addressed to James C. Anderson, Department of Marketing, J. L. Kellogg Graduate School of Management, Northwestern University, Evanston, Illinois 60208.

sition" (p. 2). Rather than as a strict dichotomy, then, the distinction in practice between exploratory and confirmatory analysis can be thought of as that of an ordered progression. Factor analysis can be used to illustrate this progression.

An exploratory factor analysis in which there is no prior specification of the number of factors is exclusively exploratory. Using a maximum likelihood (ML) or generalized least squares (GLS) exploratory program represents the next step in the progression, in that a hypothesized number of underlying factors can be specified and the goodness of fit of the resulting solution can be tested. At this point, there is a demarcation where one moves from an exploratory program to a confirmatory program. Now, a measurement model needs to be specified a priori, although the parameter values themselves are freely estimated. Although this has historically been referred to as *confirmatory analysis,* a more descriptive term might be *restricted analysis,* in that the values for many of the parameters have been restricted a priori, typically to zero.

Because initially specified measurement models almost invariably fail to provide acceptable fit, the necessary respecification and reestimation using the same data mean that the analysis is not exclusively confirmatory. After acceptable fit has been achieved with a series of respecifications, the next step in the progression would be to cross-validate the final model on another sample drawn from the population to which the results are to be generalized. This cross-validation would be accomplished by specifying the same model with freely estimated parameters or, in what represents the quintessential confirmatory analysis, the same model with the parameter estimates constrained to the previously estimated values.

Complementary Approaches for Theory Testing Versus Predictive Application

A fundamental distinction can be made between the use of structural equation modeling for theory testing and development versus predictive application (Fornell & Bookstein, 1982; Joreskog & Wold, 1982). This distinction and its implications concern a basic choice of estimation method and underlying model. For clarity, we can characterize this choice as one between a full-information (ML or GLS) estimation approach (e.g., Bentler, 1983; Joreskog, 1978) in conjunction with the common factor model (Harman, 1976) and a partial least squares (PLS) estimation approach (e.g., Wold, 1982) in conjunction with the principal-component model (Harman, 1976).

For theory testing and development, the ML or GLS approach has several relative strengths. Under the common factor model, observed measures are assumed to have random error variance and measure-specific variance components (referred to together as *uniqueness* in the factor analytic literature, e.g., Harman, 1976) that are not of theoretical interest. This unwanted part of the observed measures is excluded from the definition of the latent constructs and is modeled separately. Consistent with this, the covariances among the latent constructs are adjusted to reflect the attenuation in the observed covariances due to these unwanted variance components. Because of this assumption, the amount of variance explained in the set of observed measures is not of primary concern. Reflecting this, full-information methods provide parameter estimates that best explain the observed covariances. Two further relative strengths of full-information approaches are that they provide the most efficient parameter estimates (Joreskog & Wold, 1982) and an overall test of model fit. Because of the underlying assumption of random error and measure specificity, however, there is inherent indeterminacy in the estimation of factor scores (cf. Lawley & Maxwell, 1971; McDonald & Mulaik, 1979; Steiger, 1979). This is not a concern in theory testing, whereas in predictive applications this will likely result in some loss of predictive accuracy.

For application and prediction, a PLS approach has relative strength. Under this approach, one can assume that all observed measure variance is useful variance to be explained. That is, under a principal-component model, no random error variance or measure-specific variance (i.e., unique variance) is assumed. Parameters are estimated so as to maximize the variance explained in either the set of observed measures (reflective mode) or the set of latent variables (formative mode; Fornell & Bookstein, 1982). Fit is evaluated on the basis of the percentage of variance explained in the specified regressions. Because a PLS approach estimates the latent variables as exact linear combinations of the observed measures, it offers the advantage of exact definition of component scores. This exact definition in conjunction with explaining a large percentage of the variance in the observed measures is useful in accurately predicting individuals' standings on the components.

Some shortcomings of the PLS approach also need to be mentioned. Neither an assumption of nor an assessment of unidimensional measurement (discussed in the next section) is made under a PLS approach. Therefore, the theoretical meaning imputed to the latent variables can be problematic. Furthermore, because it is a limited-information estimation method, PLS parameter estimates are not as efficient as full-information estimates (Fornell & Bookstein, 1982; Joreskog & Wold, 1982), and jackknife or bootstrap procedures (cf. Efron & Gong, 1983) are required to obtain estimates of the standard errors of the parameter estimates (Dijkstra, 1983). And no overall test of model fit is available. Finally, PLS estimates will be asymptotically correct only under the joint conditions of consistency (sample size becomes large) and consistency at large (the number of indicators per latent variable becomes large; Joreskog & Wold, 1982). In practice, the correlations between the latent variables will tend to be underestimated, whereas the correlations of the observed measures with their respective latent variables will tend to be overestimated (Dijkstra, 1983).

These two approaches to structural equation modeling, then, can be thought of as a complementary choice that depends on the purpose of the research: ML or GLS for theory testing and development and PLS for application and prediction. As Joreskog and Wold (1982) concluded, "ML is theory-oriented, and emphasizes the transition from exploratory to confirmatory analysis. PLS is primarily intended for causal–predictive analysis in situations of high complexity but low theoretical information" (p. 270). Drawing on this distinction, we consider, in the remainder of this article, a confirmatory two-step approach to theory testing and development using ML or GLS methods.

Estimation Methods

Since the inception of contemporary structural equation methodology in the middle 1960s (Bock & Bargmann, 1966;

Joreskog, 1966, 1967), maximum likelihood has been the predominant estimation method. Under the assumption of a multivariate normal distribution of the observed variables, maximum likelihood estimators have the desirable asymptotic, or large-sample, properties of being unbiased, consistent, and efficient (Kmenta, 1971). Moreover, significance testing of the individual parameters is possible because estimates of the asymptotic standard errors of the parameter estimates can be obtained. Significance testing of overall model fit also is possible because the fit function is asymptotically distributed as chi-square, adjusted by a constant multiplier.

Although maximum likelihood parameter estimates in at least moderately sized samples appear to be robust against a moderate violation of multivariate normality (Browne, 1984; Tanaka, 1984), the problem is that the asymptotic standard errors and overall chi-square test statistic appear not to be. Related to this, using normal theory estimation methods when the data have an underlying leptokurtic (peaked) distribution appears to lead to rejection of the null hypothesis for overall model fit more often than would be expected. Conversely, when the underlying distribution is platykurtic (flat), the opposite result would be expected to occur (Browne, 1984). To address these potential problems, recent developments in estimation procedures, particularly by Bentler (1983) and Browne (1982, 1984), have focused on relaxing the assumption of multivariate normality.

In addition to providing more general estimation methods, these developments have led to a more unified approach to estimation. The traditional maximum likelihood fit function (Lawley, 1940), based on the likelihood ratio, is

$$F(\theta) = \ln|\boldsymbol{\Sigma}(\theta)| - \ln|\mathbf{S}| + \mathrm{tr}[\mathbf{S}\boldsymbol{\Sigma}(\theta)^{-1}] - p \qquad (1)$$

for p observed variables, with a $p \times p$ sample covariance matrix $\mathbf{S}$, and $p \times p$ predicted covariance matrix $\boldsymbol{\Sigma}(\theta)$, where θ is the vector of specified model parameters to be estimated. The specific maximum likelihood fit function in Equation 1 can be replaced by a more general fit function, which is implemented in the EQS program (Bentler, 1985) and in the LISREL program, beginning with Version 7 (Joreskog & Sorbom, 1987):

$$F(\theta) = [\mathbf{s} - \sigma(\theta)]'\mathbf{U}^{-1}[\mathbf{s} - \sigma(\theta)], \qquad (2)$$

where $\mathbf{s}$ is a $p^* \times 1$ vector (such that $p^* = p(p+1)/2$) of the nonduplicated elements of the full covariance matrix $\mathbf{S}$ (including the diagonal elements), $\sigma(\theta)$ is the corresponding $p^* \times 1$ vector of predicted covariances from $\boldsymbol{\Sigma}(\theta)$, and $\mathbf{U}$ is a $p^* \times p^*$ weight matrix. Fit functions that can be expressed in this quadratic form define a family of estimation methods called *generalized least squares* (GLS). As can be seen directly from Equation 2, minimizing the fit function $F(\theta)$ is the minimization of a weighted function of the residuals, defined by $\mathbf{s} - \sigma(\theta)$. The likelihood ratio fit function of Equation 1 and the quadratic fit function of Equation 2 are minimized through iterative algorithms (cf. Bentler, 1986b).

The specific GLS method of estimation is specified by the value of $\mathbf{U}$ in Equation 2. Specifying $\mathbf{U}$ as $\mathbf{I}$ implies that minimizing F is the minimization of the sum of squared residuals, that is, ordinary, or "unweighted," least squares estimation. Alternately, when it is updated as a function of the most recent parameter estimates obtained at each iteration during the estimation process, $\mathbf{U}$ can be chosen so that minimizing Equation 2 is asymptotically equivalent to minimizing the likelihood fit function of Equation 1 (Browne, 1974; Lee & Jennrich, 1979).

Other choices of $\mathbf{U}$ result in estimation procedures that do not assume multivariate normality. The most general procedure, provided by Browne (1984), yields asymptotically distribution-free (ADF) "best" generalized least squares estimates, with corresponding statistical tests that are "asymptotically insensitive to the distribution of the observations" (p. 62). These estimators are provided by the EQS program and the LISREL 7 program. The EQS program refers to these ADF GLS estimators as *arbitrary distribution theory generalized least squares* (AGLS; Bentler, 1985), whereas the LISREL 7 program refers to them as *weighted least squares* (WLS; Joreskog & Sorbom, 1987).

The value of $\mathbf{U}$ for ADF estimation is noteworthy in at least two respects. First, the elements of $\mathbf{U}$ involve not only the second-order product moments about the respective means (variances and covariances) of the observed variables but also the fourth-order product moments about the respective means. Therefore, as seen from Equation 2, although covariances are still being fitted by the estimation process, as in traditional maximum likelihood estimation, $\mathbf{U}$ now becomes the asymptotic covariance matrix of the sample variances and covariances. Second, in ML or GLS estimation under multivariate normal theory, Equation 2 simplifies to a more computationally tractable expression, such as in Equation 1. By contrast, in ADF estimation, one must employ the full $\mathbf{U}$ matrix. For example, when there are only 20 observed variables, $\mathbf{U}$ has 22,155 unique elements (Browne, 1984). Thus, the computational requirements of ADF estimation can quickly surpass the capability of present computers as the number of observed variables becomes moderately large.

To address this problem of computational infeasibility when the number of variables is moderately large, both EQS and LISREL 7 use approximations of the full ADF method. Bentler and Dijkstra (1985) developed what they called *linearized estimators,* which involve a single iteration beginning from appropriate initial estimates, such as those provided by normal theory ML. This linearized (L) estimation procedure is referred to as *LAGLS* in EQS. The approximation approach implemented in LISREL 7 (Joreskog & Sorbom, 1987) uses an option for ignoring the off-diagonal elements in $\mathbf{U}$, providing what are called *diagonally weighted least squares* (DWLS) estimates.

Bentler (1985) also implemented in the EQS program an estimation approach that assumes a somewhat more general underlying distribution than the multivariate normal assumed for ML estimation: elliptical estimation. The multivariate normal distribution assumes that each variable has zero skewness (third-order moments) and zero kurtosis (fourth-order moments). The multivariate elliptical distribution is a generalization of the multivariate normal in that the variables may share a common, nonzero kurtosis parameter (Bentler, 1983; Beran, 1979; Browne, 1984). As with the multivariate normal, iso-density contours are ellipsoids, but they may reflect more platykurtic or leptokurtic distributions, depending on the magnitude and direction of the kurtosis parameter. The elliptical distribution with regard to Equation 2 is a generalization of the multi-

variate normal and, thus, provides more flexibility in the types of data analyzed. Another advantage of this distribution is that the fourth-order moments can be expressed as a function of the second-order moments with only the addition of a single kurtosis parameter, greatly simplifying the structure of **U**.

Bentler (1983) and Mooijaart and Bentler (1985) have outlined an estimation procedure even more ambitious than any of those presently implemented in EQS or LISREL 7. This procedure, called *asymptotically distribution-free reweighted least squares* (ARLS), generalizes on Browne's (1984) ADF method. In an ADF method (or AGLS in EQS notation), **U** is defined as a constant before the minimization of Equation 2 begins. By contrast, in ARLS, **U** is updated at each iteration of the minimization algorithm. This updating is based on Bentler's (1983) expression of higher order moment structures, specified as a function of the current estimates of the model parameters, thereby representing a generalization of presently estimated second-order moment structures.

In addition to the relaxation of multivariate normality, recent developments in estimation procedures have addressed at least two other issues. One problem is that when the data are standardized, the covariances are not rescaled by known constants but by data-dependent values (i.e., standard deviations) that will randomly vary across samples. Because of this, when the observed variable covariances are expressed as correlations, the asymptotic standard errors and overall chi-square goodness-of-fit tests are not correct without adjustments to the estimation procedure (Bentler & Lee, 1983). A companion program to LISREL 7, PRELIS (Joreskog & Sorbom, 1987), can provide such adjustments. A second problem is the use of product–moment correlations when the observed variables cannot be regarded as continuous (cf. Babakus, Ferguson, & Joreskog, 1987). PRELIS also can account for this potential shortcoming of current usage by calculating the correct polychoric and polyserial coefficients (Muthen, 1984) and then adjusting the estimation procedure accordingly.

In summary, these new estimation methods represent important theoretical advances. The degree, however, to which estimation methods that do not assume multivariate normality will supplant normal theory estimation methods in practice has yet to be determined. Many data sets may be adequately characterized by the multivariate normal, much as the univariate normal often adequately describes univariate distributions of data. And, as Bentler (1983) noted, referring to the weight matrix **U**, "an estimated optimal weight matrix should be adjusted to reflect the strongest assumptions about the variables that may be possible" (p. 504). Related to this, the limited number of existing Monte Carlo investigations of normal theory ML estimators applied to nonnormal data (Browne, 1984; Harlow, 1985; Tanaka, 1984) has provided support for the robustness of ML estimation for the recovery of parameter estimates, though their associated standard errors may be biased. Because assessments of the multivariate normality assumption now can be readily made by using the EQS and PRELIS programs, a researcher can make an informed choice on estimation methods in practice, weighing the trade-offs between the reasonableness of an underlying normal theory assumption and the limitations of arbitrary theory methods (e.g., constraints on model size and the need for larger sample sizes, which we discuss later in the next section).

Confirmatory Measurement Models

A confirmatory factor analysis model, or confirmatory measurement model, specifies the posited relations of the observed variables to the underlying constructs, with the constructs allowed to intercorrelate freely. Using the LISREL program notation, this model can be given directly from Joreskog and Sorbom (1984, pp. I.9–10) as

$$\mathbf{x} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{3}$$

where $\mathbf{x}$ is a vector of q observed measures, $\boldsymbol{\xi}$ is a vector of n underlying factors such that $n < q$, $\mathbf{\Lambda}$ is a $q \times n$ matrix of pattern coefficients or factor loadings relating the observed measures to the underlying construct factors, and $\boldsymbol{\delta}$ is a vector of q variables that represents random measurement error and measure specificity. It is assumed for this model that $E(\boldsymbol{\xi}\boldsymbol{\delta}') = \mathbf{0}$. The variance–covariance matrix for $\mathbf{x}$, defined as $\mathbf{\Sigma}$, is

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Theta}_\delta, \tag{4}$$

where $\mathbf{\Phi}$ is the $n \times n$ covariance matrix of $\boldsymbol{\xi}$ and $\mathbf{\Theta}_\delta$ is the diagonal $q \times q$ covariance matrix of $\boldsymbol{\delta}$.

Need for Unidimensional Measurement

Achieving unidimensional measurement (cf. Anderson & Gerbing, 1982; Hunter & Gerbing, 1982) is a crucial undertaking in theory testing and development. A necessary condition for assigning meaning to estimated constructs is that the measures that are posited as alternate indicators of each construct must be acceptably unidimensional. That is, each set of alternate indicators has only one underlying trait or construct in common (Hattie, 1985; McDonald, 1981). Two criteria, each representing necessary conditions, are used in assessing unidimensionality: internal consistency and external consistency.

The internal consistency criterion can be presented in the following fundamental equation (Hart & Spearman, 1913, p. 58; Spearman, 1914, p. 107):

$$\frac{\rho_{ac}}{\rho_{ad}} = \frac{\rho_{bc}}{\rho_{bd}}, \tag{5}$$

where a, b, c, and d are measures of the same construct, ξ. This equality should hold to within sampling error (Spearman & Holzinger, 1924), and at least four measures of a construct are needed for an assessment. A related equation is the product rule for internal consistency:

$$\rho_{ab} = \rho_{a\xi}\rho_{b\xi}, \tag{6}$$

where a and b are measures of some construct, ξ.

The external consistency criterion can be given by a redefinition of Equation 3, where (a) a, b and c are alternate indicators of a given construct and d is redefined as an indicator of another construct or (b) both c and d are redefined as alternate indicators of another construct. A related equation is the product rule for external consistency:

$$\rho_{ad} = \rho_{a\xi}\rho_{\xi\xi^*}\rho_{\xi^*d}, \quad (7)$$

where a is any indicator of construct ξ and d is any indicator of another construct, ξ^*. Note that when $\xi = \xi^*$, Equation 7 reduces to Equation 6; that is, internal consistency represents a special case of external consistency. Because it often occurs in practice that there are less than four indicators of a construct, external consistency then becomes the sole criterion for assessing unidimensionality. The product rules for internal and external consistency, which are used in confirmatory factor analysis, can be used to generate a predicted covariance matrix for any specified model and set of parameter estimates.

In building measurement models, multiple-indicator measurement models (Anderson & Gerbing, 1982; Hunter & Gerbing, 1982) are preferred because they allow the most unambiguous assignment of meaning to the estimated constructs. The reason for this is that with multiple-indicator measurement models, each estimated construct is defined by at least two measures, and each measure is intended as an estimate of only one construct. Unidimensional measures of this type have been referred to as *congeneric measurements* (Joreskog, 1971). By contrast, measurement models that contain correlated measurement errors or that have indicators that load on more than one estimated construct do not represent unidimensional construct measurement (Gerbing & Anderson, 1984). As a result, assignment of meaning to such estimated constructs can be problematic (cf. Bagozzi, 1983; Fornell, 1983; Gerbing & Anderson, 1984).

Some dissent, however, exists about the application of the confirmatory factor analysis model for assessing unidimensionality. Cattell (1973, 1978) has argued that individual measures or items, like real-life behaviors, tend to be factorially complex. "In other words, to show that a given matrix is rank one is not to prove that the items are measuring a pure unitary trait factor in common: it may be a mixture of unitary traits" (Cattell, 1973, p. 382). According to Cattell (1973), although these items are unidimensional with respect to each other, they simply may represent a "bloated specific" in the context of the true (source trait) factor space. That is, the items represent a "psychological concept of something that is behaviorally very narrow" (Cattell, 1973, p. 359).

We agree with Cattell (1973, 1978) that estimated first-order factors may not correspond to the constructs of interest (cf. Gerbing & Anderson, 1984). The measurement approach that we have advocated is not, however, necessarily inconsistent with Cattell's (1973, 1978) approach. The two approaches can become compatible when the level of analysis shifts from the individual items to a corresponding set of composites defined by these items. Further analyses of these composites could then be undertaken to isolate the constructs of interest, which would be conceptualized as higher order factors (Gerbing & Anderson, 1984). One possibility is a second-order confirmatory factor analysis as outlined by, for example, Joreskog (1971) or Weeks (1980). Another possibility is to interpret the resulting composites within an existing "reference factor system," such as the 16 personality dimensions provided by Cattell (1973) for the personality domain.

Specification

Setting the metric of the factors. For identification of the measurement model, one must set the metric (variances) of the factors. A preferred way of doing this is to fix the diagonal of the phi matrix at 1.0, giving all factors unit variances, rather than to arbitrarily fix the pattern coefficient for one indicator of each factor at 1.0 (Gerbing & Hunter, 1982). Setting the metric in this way allows a researcher to test the significance of each pattern coefficient, which is of interest, rather than to forgo this and test whether the factor variances are significantly different from zero, which typically is not of interest.

Single indicators. Although having multiple indicators for each construct is strongly advocated, sometimes in practice only a single indicator of some construct is available. And, as most often is the case, this indicator seems unlikely to perfectly estimate the construct (i.e., has no random measurement error or measure-specificity component). The question then becomes "At what values should the theta-delta and lambda parameters be set?" To answer this, ideally, a researcher would like to have an independent estimate for the error variance of the single indicator, perhaps drawn from prior research, but often this is not available.

In the absence of an independent estimate, the choice of values becomes arbitrary. In the past, a conservative value for θ_δ, such as $.1s_x^2$, has been chosen, and its associated λ has been set at $.95s_x$ (e.g., Sorbom & Joreskog, 1982). Another conservative alternative to consider is to set θ_δ for the single indicator at the smallest value found for the other, estimated error variances ($\hat{\Theta}_\delta$). Although this value is still arbitrary, it has the advantage of being based on information specific to the given research context. That is, this indicator shares a respondent sample and survey instrument with the other indicators.

Sample size needed. Because full-information estimation methods depend on large-sample properties, a natural concern is the sample size needed to obtain meaningful parameter estimates. In a recent Monte Carlo study, Anderson and Gerbing (1984) and Gerbing and Anderson (1985) have investigated ML estimation for a number of sample sizes and a variety of confirmatory factor models in which the normal theory assumption was fully met. The results of this study were that although the bias in parameter estimates is of no practical significance for sample sizes as low as 50, for a given sample, the deviations of the parameter estimates from their respective population values can be quite large. Whereas this does not present a problem in statistical inference, because the standard errors computed by the LISREL program are adjusted accordingly, a sample size of 150 or more typically will be needed to obtain parameter estimates that have standard errors small enough to be of practical use.

Related to this, two problems in the estimation of the measurement model that are more likely to occur with small sample sizes are nonconvergence and improper solutions. (We discuss potential causes of these problems within the Respecification subsection.) Solutions are nonconvergent when an estimation method's computational algorithm, within a set number of iterations, is unable to arrive at values that meet prescribed, termination criteria (cf. Joreskog, 1966, 1967). Solutions are improper when the values for one or more parameter estimates

are not feasible, such as negative variance estimates (cf. Dillon, Kumar, & Mulani, 1987; Gerbing & Anderson, 1987; van Driel, 1978). Anderson and Gerbing (1984) found that a sample size of 150 will usually be sufficient to obtain a converged and proper solution for models with three or more indicators per factor. Measurement models in which factors are defined by only two indicators per factor can be problematic, however, so larger samples may be needed to obtain a converged and proper solution.

Unfortunately, a practical limitation of estimation methods that require information from higher order moments (e.g., ADF) is that they correspondingly require larger sample sizes. The issue is not simply that larger samples are needed to provide more stable estimates for consistent estimators. Perhaps of greater concern, the statistical properties of full-information estimators are asymptotic; that is, they have proven to be true only for large samples. Thus, a critical task is to establish guidelines regarding minimum sample sizes for which the asymptotic properties of these more general, arbitrary distribution theory estimators can be reasonably approximated.

Presently, such guidelines on minimum sample sizes have not been determined. Initial studies by Tanaka (1984) and Harlow (1985) suggest that a sample size of at least 400 or 500 is needed. Furthermore, consider Browne's (1984) comments regarding the choice of the best generalized least squares (BGLS) estimators:

> We note that the term "best" is used in a very restricted sense with respect to a specific asymptotic property which possibly may not carry over to finite samples. It is possible that other estimators may have other properties which render them superior to BGLS estimators for practical applications of samples of moderate size. (p. 68)

Related to these comments, in a small, preliminary Monte Carlo study, Browne (1984) found the BGLS estimates provided by the asymptotic distribution-free procedure to have "unacceptable bias" (p. 81) for some of the parameters with a sample size of 500.

Assessment of Fit

After estimating a measurement model, given a converged and proper solution, a researcher would assess how well the specified model accounted for the data with one or more overall goodness-of-fit indices. The LISREL program provides the probability value associated with the chi-square likelihood ratio test, the goodness-of-fit index, and the root-mean-square residual (cf. Joreskog & Sorbom, 1984, pp. I.38–42). Anderson and Gerbing (1984) gave estimates of the expected values of these indices, and their 5th- or 95th-percentile values, for a variety of confirmatory factor models and sample sizes. The chi-square probability value and the normed and nonnormed fit indices (Bentler & Bonett, 1980) are obtained from the EQS program (Bentler, 1985, p. 94).[1]

Convergent validity can be assessed from the measurement model by determining whether each indicator's estimated pattern coefficient on its posited underlying construct factor is significant (greater than twice its standard error). Discriminant validity can be assessed for two estimated constructs by constraining the estimated correlation parameter ($\hat{\phi}_{ij}$) between them to 1.0 and then performing a chi-square difference test on the values obtained for the constrained and unconstrained models (Joreskog, 1971). "A significantly lower χ^2 value for the model in which the trait correlations are not constrained to unity would indicate that the traits are not perfectly correlated and that discriminant validity is achieved" (Bagozzi & Phillips, 1982, p. 476). Although this is a necessary condition for demonstrating discriminant validity, the practical significance of this difference will depend on the research setting. This test should be performed for one pair of factors at a time, rather than as a simultaneous test of all pairs of interest.[2] The reason for this is that a nonsignificant value for one pair of factors can be obfuscated by being tested with several pairs that have significant values. A complementary assessment of discriminant validity is to determine whether the confidence interval (±two standard errors) around the correlation estimate between the two factors includes 1.0.

Respecification

Because the emphasis of this article is on structural equation modeling in practice, we recognize that most often some respecification of the measurement model will be required. It must be stressed, however, that respecification decisions should not be based on statistical considerations alone but rather in conjunction with theory and content considerations. Consideration of theory and content both greatly reduces the number of alternate models to investigate (cf. Young, 1977) and reduces the possibility of taking advantage of sampling error to attain goodness of fit.

Sometimes, the first respecification necessary is in response to nonconvergence or an improper solution. Nonconvergence can occur because of a fundamentally incongruent pattern of sample covariances that is caused either by sampling error in conjunction with a properly specified model or by a misspecification. Relying on content, one can obtain convergence for the model by respecifying one or more problematic indicators to different constructs or by excluding them from further analysis.

Considering improper solutions, van Driel (1978) presented three potential causes: sampling variations in conjunction with true parameter values close to zero, a fundamentally misspecified model, and indefiniteness (underidentification) of the model. Van Driel showed that it is possible to distinguish which

---

[1] The normed fit index (Bentler & Bonett, 1980) can also be calculated by using the LISREL program. This is accomplished by specifying each indicator as a separate factor and then fixing lambda as an identity matrix, theta delta as a null matrix, and phi as a diagonal matrix with freely estimated variances. Using the obtained chi-square value for this overall null model ($\chi^2_\varnothing$), in conjunction with the chi-square value (χ^2_m) from the measurement model, one can calculate the normed fit index value as $(\chi^2_\varnothing - \chi^2_m)/\chi^2_\varnothing$.

[2] When a number of chi-square difference tests are performed for assessments of discriminant validity, the significance level for each test should be adjusted to maintain the "true" overall significance level for the family of tests (cf. Finn, 1974). This adjustment can be given as $\alpha_o = 1 - (1 - \alpha_i)^t$, where α_o is the overall significance level, typically set at .05; α_i is the significance level that should be used for each individual hypothesis test of discriminant validity; and t is the number of tests performed.

of these causes is the likely one by examining the confidence interval constructed around the negative estimate. When positive values fall within this confidence interval and the size of the interval is comparable to that for proper estimates, the likely cause of the improper estimate is sampling error. Building on this work, Gerbing and Anderson (1987) recently found that for improper estimates due to sampling error, respecifying the model with the problematic parameter fixed at zero has no appreciable effect on the parameter estimates of other factors or on the overall goodness-of-fit indices. Alternately, this parameter can be fixed at some arbitrarily small, positive number (e.g., .005) to preserve the confirmatory factor model (cf. Bentler, 1976).

Given a converged and proper solution but unacceptable overall fit, there are four basic ways to respecify indicators that have not "worked out as planned": Relate the indicator to a different factor, delete the indicator from the model, relate the indicator to multiple factors, or use correlated measurement errors. The first two ways preserve the potential to have unidimensional measurement and are preferred because of this, whereas the last two ways do not, thereby obfuscating the meaning of the estimated underlying constructs. The use of correlated measurement errors can be justified only when they are specified a priori. As an example, correlated measurement errors may be expected in longitudinal research when the same indicators are measured at multiple points in time. By contrast, correlated measurement errors should not be used as respecifications because they take advantage of chance, at a cost of only a single degree of freedom, with a consequent loss of interpretability and theoretical meaningfulness (Bagozzi, 1983; Fornell, 1983). Gerbing and Anderson (1984) demonstrated how the uncritical use of correlated measurement errors for respecification, although improving goodness of fit, can mask a true underlying structure.

In our experience, the patterning of the residuals has been the most useful for locating the source of misspecification in multiple-indicator measurement models. The LISREL program provides normalized residuals (Joreskog & Sorbom, 1984, p. I.42), whereas the EQS program (Bentler, 1985, pp. 92–93) provides standardized residuals. Although Bentler and Dijkstra (1985) recently pointed out that the normalized residuals may not be strictly interpretable as standard normal variates (i.e., normalized residuals greater than 1.96 in magnitude may not be strictly interpretable as statistically significant), nonetheless, the pattern of large normalized residuals (e.g., greater than 2 in magnitude) is still informative for respecification. For example, an indicator assigned to the wrong factor will likely have a pattern of large negative normalized residuals with the other indicators of the factor to which it was assigned (representing overfitting), and when another factor on which it should belong exists, an obverse pattern of large positive residuals will be observed with the indicators of this factor (representing underfitting). As another example, indicators that are multidimensional tend to have large normalized residuals (the result of either underfitting or overfitting) with indicators of more than one factor, which often represents the only large normalized residual for each of these other indicators.

Useful adjuncts to the pattern of residuals are similarity (or proportionality) coefficients (Anderson & Gerbing, 1982; Hunter, 1973) and multiple-groups analysis (cf. Anderson & Gerbing, 1982; Nunnally, 1978), each of which can readily be computed with the ITAN program (Gerbing & Hunter, 1987). A similarity coefficient, v_{ij}, for any two indicators, x_i and x_j, can be defined for a set of q indicators as

$$v_{ij} = \frac{\sum_{k=1}^{q} r_{x_i x_k} r_{x_j x_k}}{(\sum_{k=1}^{q} r_{x_i x_k}^2)^{1/2} (\sum_{k=1}^{q} r_{x_j x_k}^2)^{1/2}}. \quad (8)$$

The value of this index ranges from -1.0 to $+1.0$, with values greater in magnitude indicating greater internal and external consistency for the two indicators. Thus, similarity coefficients are useful because they efficiently summarize the internal and external consistency of the indicators with one another. Alternate indicators of the same underlying factor, therefore, should have similarity coefficients that are typically .8 or greater.

Multiple-groups analysis is a confirmatory estimation method that is complementary to full-information estimation of multiple-indicator measurement models. With multiple-groups analysis, each construct factor is defined as simply the unit-weighted sum of its posited indicators. The factor loadings are simply the correlation of each indicator with the composite (construct factor), and the factor correlations are obtained by correlating the composites. Communalities are computed within each group of indicators by iteration. By using communalities, the resultant indicator–factor and factor–factor correlations are corrected for attenuation due to measurement error. Because multiple-groups analysis estimates are computed from only those covariances of the variables in the equation on which the estimates are based, these estimates more clearly localize misspecification, making it easier to detect (Anderson & Gerbing, 1982). For example, if an indicator is specified as being related to the wrong factor, then the multiple-groups analysis shows this by producing a higher factor loading for this indicator on the correct factor. Full-information methods, by contrast, draw on all indicator covariances to produce estimates that minimize the fit function (Joreskog, 1978).

In summary, a researcher should use these sources of information about respecification in an integrative manner, along with content considerations, in making decisions about respecification. In practice, the measurement model may sometimes be judged to provide acceptable fit even though the chi-square value is still statistically significant. This judgment should be supported by the values of the normed fit index and the other fit indices, particularly the root-mean-square residual index in conjunction with the number of large normalized or standardized residuals (and the absolute values of the largest ones).

One-Step Versus Two-Step Modeling Approaches

The primary contention of this article is that much is to be gained from separate estimation and respecification of the measurement model prior to the simultaneous estimation of the measurement and structural submodels. In putting forth a specific two-step approach, we use the concepts of nested models, pseudo chi-square tests, and sequential chi-square difference tests (SCDTs) and draw on some recent work from quantitative

psychology (Steiger, Shapiro, & Browne, 1985). These tests enable a separate assessment of the adequacy of the substantive model of interest, apart from that of the measurement model. We first present the structural model and discuss the concept of interpretational confounding (Burt, 1973, 1976).

A confirmatory structural model that specifies the posited causal relations of the estimated constructs to one another can be given directly from Joreskog and Sorbom (1984, p. I.5). This model can be expressed as

$$\eta = \mathbf{B}\eta + \Gamma\xi + \zeta, \tag{9}$$

where η is a vector of m endogenous constructs, ξ is a vector of n exogenous constructs, $\mathbf{B}$ is an $m \times m$ matrix of coefficients representing the effects of the endogenous constructs on one another, Γ is an $m \times n$ matrix of coefficients representing the effects of the exogenous constructs on the endogenous constructs, and ζ is a vector of m residuals (errors in equations and random disturbance terms).

The definitional distinction between endogenous and exogenous constructs is simply that endogenous constructs have their causal antecedents specified within the model under consideration, whereas the causes of exogenous constructs are outside the model and not of present interest. Note that this distinction was not germane in the specification of confirmatory measurement models, given in Equation 3. Because of this, all observed measures were denoted simply as $\mathbf{x}$. In contrast, when structural models are specified, only observed measures of exogenous constructs are denoted as $\mathbf{x}$, whereas observed measures of endogenous constructs are denoted as $\mathbf{y}$. Separate measurement submodels are specified for $\mathbf{x}$ and $\mathbf{y}$ (cf. Joreskog & Sorbom, 1984, pp. I.5–6), which then are simultaneously estimated with the structural submodel.

In the presence of misspecification, the usual situation in practice, a one-step approach in which the measurement and structural submodels are estimated simultaneously will suffer from *interpretational confounding* (cf. Burt, 1973, 1976). Interpretational confounding "occurs as the assignment of empirical meaning to an unobserved variable which is other than the meaning assigned to it by an individual a priori to estimating unknown parameters" (Burt, 1976, p. 4). Furthermore, this empirically defined meaning may change considerably, depending on the specification of free and constrained parameters for the structural submodel. Interpretational confounding is reflected by marked changes in the estimates of the pattern coefficients when alternate structural models are estimated.

The potential for interpretational confounding is minimized by prior separate estimation of the measurement model because no constraints are placed on the structural parameters that relate the estimated constructs to one another. Given acceptable unidimensional measurement, the pattern coefficients from the measurement model should change only trivially, if at all, when the measurement submodel and alternate structural submodels are simultaneously estimated. With a one-step approach, the presence of interpretational confounding may not be detected, resulting in fit being maximized at the expense of meaningful interpretability of the constructs.

Recommended Two-Step Modeling Approach

For assessing the structural model under a two-step approach, we recommend estimating a series of five nested structural models. A model, M_2, is said to be *nested within* another model, M_1, when its set of freely estimated parameters is a subset of those estimated in M_1, and this can be denoted as $M_2 < M_1$. That is, one or more parameters that are freely estimated in M_1 are constrained in M_2. Typically, these parameters are fixed at zero, although equality constraints may be imposed so that two or more parameters are constrained to have the same value.

A *saturated* structural submodel (cf. Bentler & Bonett, 1980), M_s, can be defined as one in which all parameters (i.e., unidirectional paths) relating the constructs to one another are estimated. Note that this model is formally equivalent to a confirmatory measurement model. Obversely, a *null* structural submodel, M_n, can be defined as one in which all parameters relating the constructs to one another are fixed at zero (i.e., there are no posited relations of the constructs to one another).

A third structural submodel, M_t, represents the researcher's *theoretical or substantive model of interest.* Finally, the structural submodels M_c and M_u represent, respectively, the *"next most likely" constrained and unconstrained alternatives* from a theoretical perspective to the substantive model of interest. That is, in M_c, one or more parameters estimated in M_t are constrained, whereas in M_u, one or more parameters constrained in M_t are estimated. Given their definitions, this set of five structural submodels is nested in a sequence such that $M_n < M_c < M_t < M_u < M_s$.

Under a two-step approach, a researcher could first assess whether any structural model that would have acceptable goodness of fit existed. This would be accomplished with a pseudo chi-square test (Bentler & Bonett, 1980), in which a pseudo chi-square statistic is constructed from the chi-square value for M_s (the smallest value possible for any structural model) with the degrees of freedom for M_n (the largest number of degrees of freedom for any structural model). Note that M_s and M_n need not be estimated, because M_s is equivalent to the final measurement model, and only the associated degrees of freedom for M_n are needed. If this pseudo chi-square statistic is significant, then no structural model would give acceptable fit, because it would have a chi-square value greater than or equal to the value for M_s with fewer degrees of freedom than for M_n. Significance, then, would suggest a fundamental misspecification of the measurement model needs to be remedied, rather than a need to estimate additional structural models. A researcher using a one-step approach would not know this.

Sequential chi-square difference tests (SCDTs). Continuing with the two-step approach, a researcher would next estimate M_c, M_t and M_u, obtaining a likelihood ratio chi-square statistic value for each. These sequential chi-square tests (SCTs) provide successive fit information, although these tests are not independent. A preferred approach is to employ these test statistic values and their respective degrees of freedom in a set of SCDTs (cf. Steiger et al., 1985), each of which is framed as a null hypothesis of no significant difference between two nested structural models (denoted as $M_2 - M_1 = 0$). The difference between chi-square statistic values for nested models is itself asymptoti-

cally distributed as chi-square, with degrees of freedom equal to the difference in degrees of freedom for the two models. In a recent development, Steiger et al. (1985) proved analytically that these sequential chi-square difference tests are asymptotically independent.[3,4]

What this means is that to maintain asymptotically independent tests, a researcher would first use the SCDT comparison of $M_u - M_s$ to assess the reasonableness of the structural constraints imposed by M_u on the estimated construct covariances. If the null hypothesis associated with this test was upheld, the SCDT comparison of $M_t - M_u$ would be made. If the null hypothesis was upheld for this test, a researcher would then proceed to $M_c - M_t$. Each test assesses whether there is a significant difference in explanation of the estimated construct covariances given by the two structural models. For each SCDT in which the associated null hypothesis was upheld, the more constrained model of the two would be tentatively accepted.

In practice, though, when this sequence of tests indicates that M_t or M_c should be accepted, a researcher would also like to know whether or not M_t or M_c also provides acceptable explanation of the construct covariances. That is, a researcher would like to know if the null hypothesis associated with the SCDT comparison of $M_t - M_s$ or of $M_c - M_s$ also is upheld. Note that finding that $M_t - M_u$ and $M_u - M_s$ are each not significant does not necessarily mean that $M_t - M_s$ will not be significant. Conversely, finding that $M_t - M_u$ is significant, when $M_u - M_s$ is not significant, does not necessarily mean that $M_t - M_s$ also will be significant. A similar situation holds for $M_c - M_t$ in relation to $M_c - M_s$. Therefore, to provide a greater understanding of the acceptability of a given structural model, a researcher would perform these additional SCDT comparisons in conjunction with the earlier sequence of tests.

Fortunately, the SCDT comparisons of $M_t - M_s$ and $M_c - M_s$ are each asymptotically independent of the chi-square test of M_s, which represents the baseline model. Because the SCDT value and associated degrees of freedom for $M_t - M_s$ are simply the respective sums of those for $M_t - M_u$ and $M_u - M_s$, note that it will not, however, be independent from these tests. In a similar way, the SCDT comparison of $M_c - M_s$ will not be independent from the earlier sequence of three SCDT comparisons (or from $M_t - M_s$). Nevertheless, the additional SCDT comparisons of $M_t - M_s$ and $M_c - M_s$ can be usefully interspersed with the earlier sequence of SCDT comparisons to provide a decision-tree framework that enables a better understanding of which, if any, of the three alternative theoretical models should be accepted. We present one decision-tree framework for this set of SCDT comparisons in Figure 1.

A decision-tree framework. As can be seen from Figure 1, under this decision-tree framework, a researcher would first perform the SCDT of $M_t - M_s$. This SCDT provides an asymptotically independent assessment of the theoretical model's explanation of the relations of the estimated constructs to one another. In other words, one can make an asymptotically independent test of nomological validity. Note that because $M_t - M_s$ is asymptotically independent of M_s, a researcher can build a measurement model that has the best fit from a content and statistical standpoint, where respecification may have been employed to accomplish this, and still provide a statistical assessment of the adequacy of the theoretical model of interest.

Before continuing with this decision tree, we should mention another comparative strength of the two-step approach. Not only does the SCDT comparison of $M_t - M_s$ provide an assessment of fit for the substantive model of interest to the estimated construct covariances, but it also requires the researcher to consider the strength of explanation of this theoretical model over that of a confirmatory measurement model. Comparing the degrees of freedom associated with this SCDT with the total number available, $[(m + n)(m + n - 1)]/2$, indicates this inferential strength. That is, the ability to make any causal inferences about construct relations from correlational data depends directly on the available degrees of freedom. Thus, for example, a researcher who specifies a substantive model in which each construct is related by direct causal paths to all others would realize from this test the inability to make any causal inferences. This is because no degrees of freedom would exist for the SCDT; the theoretical "causal" model is indistinguishable from a confirmatory measurement model, and any causal interpretation should be carefully avoided. To the extent, however, that a "considerable" proportion of possible direct causal paths are specified as zero and there is acceptable fit, one can advance qualified causal interpretations.

The SCDT comparison of $M_c - M_t$ provides further understanding of the explanatory ability afforded by the theoretical model of interest and, irrespective of the outcome of the $M_t - M_s$ comparison, would be considered next. Bagozzi (1984) recently noted the need to consider rival hypotheses in theory construction and stressed that whenever possible, these rival explanations should be tested within the same study. Apart from this but again stressing the need to assess alternative models, MacCallum (1986) concluded from his research on specification searches that "investigators should not interpret a nonsignificant chi-square as a signal to stop a specification search" (p. 118). SCDTs are particularly well-suited for accomplishing these comparisons between alternative theoretical models.

Consider first the upper branch of the decision tree in Figure 1, that is, the null hypothesis that $M_t - M_s = 0$ is not rejected. Given this, when both the $M_c - M_t$ and the $M_c - M_s$ comparisons also are not significant, M_c would be accepted because it is the most parsimonious structural model of the three hypothesized, theoretical alternatives and because it provides adequate explanation of the estimated construct covariances. When ei-

[3] Steiger et al. (1985) developed these analytic results within the context of exploratory maximum likelihood factor analysis, in which the question of interest is the number of factors that best represents a given covariance matrix. However, their derivations were developed for a general discrepancy function, of which the fit function used in confirmatory analyses of covariance structures (cf. Browne, 1984; Joreskog, 1978) is a special case. Their results even extend to situations in which the null hypothesis need not be true. In such situations, the SCDTs will still be asymptotically independent but asymptotically distributed as noncentral chi-square variates.

[4] A recent development in the EQS program (Bentler, 1986a) is the provision of Wald tests and Lagrange multiplier tests (cf. Buse, 1982), each of which is asymptotically equivalent to chi-square difference tests. This allows a researcher, within a single computer run, to obtain overall goodness-of-fit information that is asymptotically equivalent to what would be obtained from separate SCDT comparisons of M_c and M_u with the specified model, M_t.

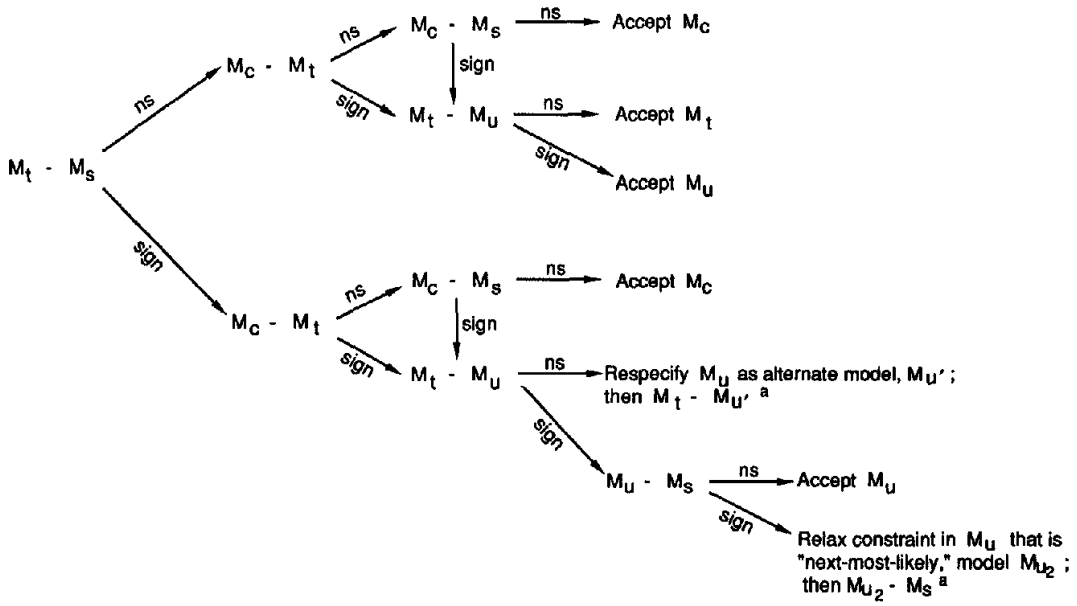

Figure 1. A decision-tree framework for the set of sequential chi-square difference tests (SCDTs). M_t = theoretical model of interest; M_s = measurement (or "saturated") model; M_c and M_u = next most likely constrained and unconstrained structural models, respectively; ns and sign indicate that the null hypothesis for each SCDT is not or is rejected, respectively, at the specified probability level (e.g., .05).
[a]The modeling approach shifts from being confirmatory to being increasingly exploratory.

ther $M_c - M_t$ or $M_c - M_s$ is significant, the $M_t - M_u$ comparison would be assessed next. If this SCDT is not significant, it indicates that relaxing the next most likely constraint or constraints from a theoretical perspective in M_t does not significantly add to its explanation of the construct covariances, and with parsimony preferred when given no difference in explanation, M_t would be accepted. Obversely, a significant result would indicate that the additional estimated parameter or parameters incrementally contribute to the explanation given by M_t and would lead to the acceptance of M_u. Note that because of the additive property of chi-square values and their associated degrees of freedom, one need not perform the SCDT of $M_u - M_s$, which must be nonsignificant given the earlier pattern of SCDT results.

Consider now the lower branch of the decision tree, that is, the null hypothesis that $M_t - M_s = 0$ is rejected. As with the upper branch, when both the $M_c - M_t$ and $M_c - M_s$ comparisons are not significant, a researcher would accept M_c. The explanation in this situation, however, would be that one or more parameters that were being estimated in M_t were superfluous in that they were not significantly contributing to the explanation of the construct covariances but were "costing" their associated degrees of freedom. Constraining these irrelevant parameters in M_c gains their associated degrees of freedom, with no appreciable loss of fit. As a result, although $M_t - M_s$ was significant, $M_c - M_s$, which has essentially the same SCDT value, is not because of these additional degrees of freedom.

When either the $M_c - M_t$ or $M_c - M_s$ comparison is significant, the SCDT of $M_t - M_u$ is considered next. Given that $M_t - M_s$ has already been found significant, a nonsignificant value for $M_t - M_u$ would indicate the need to respecify M_u as some alternative structural model, $M_{u'}$, such that $M_t < M_{u'}$. Put differently, in this situation, a researcher needs to reconsider the set of parameters from M_t that were freed in M_u and pursue an alternative theoretical tack in specifying $M_{u'}$. Then, a researcher would perform the SCDT of $M_t - M_{u'}$, and the modeling approach would shift from being confirmatory to being increasingly exploratory. In practice, a researcher might at this point also constrain to zero, or "trim" any parameters from M_t that have nonsignificant estimates ($M_{t'}$). A respecification search for $M_{u'}$ would continue until both a significant value of $M_t - M_{u'}$ and a nonsignificant value of $M_{u'} - M_s$ are obtained or until there are no further constrained parameters that would be theoretically meaningful to relax.

Before moving on to consider the $M_u - M_s$ branch, we should note that even though $M_t - M_s$ is significant and $M_t - M_u$ is not, it is possible to obtain a SCDT value for $M_u - M_s$ that is not significant. Although this may not occur often in practice, a researcher would still not accept M_u in this situation. The rationale underlying this is that, given that $M_t - M_s$ is significant, there must be one or more constrained parameters in M_t that, when allowed to be unconstrained (as $M_{u'}$), would provide a significant increment in the explanation of the estimated construct covariances over M_t; that is, the SCDT of $M_t - M_{u'}$ would be significant. Given this and that $M_u - M_s$ was not significant, $M_{u'} - M_s$ must also not be significant. Therefore, $M_{u'}$ would provide a significant increment in explanation over M_t and would provide adequate explanation of the estimated construct covariances.

The final SCDT comparison of $M_u - M_s$ is performed when $M_t - M_u$ is significant (as is $M_t - M_s$). When this SCDT value is significant, a researcher would accept M_u. The next most likely

unconstrained theoretical alternative, though less parsimonious than M_t, is required for acceptable explanation of the estimated construct covariances. Finally, when $M_u - M_s$ is significant, a researcher would relax one or more parameters in M_u that is "next most likely" from a theoretical perspective, yielding a model M_{u_2}, such that $M_u < M_{u_2}$. Then, a researcher would perform the SCDT of $M_{u_2} - M_s$, and as with $M_t - M_{u'}$, the modeling approach would shift from being confirmatory to being increasingly exploratory. A respecification search for M_{u_2} would continue until a nonsignificant value of $M_{u_2} - M_s$ is obtained or until no further constrained parameters are theoretically meaningful to relax. Note that the critical distinction between M_{u_2} and $M_{u'}$ is that with M_{u_2}, the respecification search continues along the same theoretical direction, whereas with $M_{u'}$, the respecification search calls for a change in theoretical tack. This is reflected by the fact that $M_u < M_{u_2}$, whereas M_u will not be nested within $M_{u'}$.

We should mention two further advantages of this two-step approach over a one-step approach. Paths that are specified as absent and are then supported by an SCDT also provide theoretical information, and this should not be overlooked. A two-step approach focuses attention on the trade-off between goodness of fit and strength of causal inference that is implicit in a one-step approach. Adding more paths will likely improve goodness of fit, but it correspondingly compromises the ability to make meaningful, causal inferences about the relations of the constructs to one another. As a final comparative advantage, separate assessments of the measurement model and the structural model preclude having good fit of one model compensate for (and potentially mask) poor fit of the other, which can occur with a one-step approach.

Additional Considerations in Structural Model Interpretation

Practical versus statistical significance. To this point, we have considered significance only from the perspective of formal, statistical tests. As has been noted by Bentler and Bonett (1980) and others (e.g., Joreskog, 1974), however, the value of the chi-square likelihood ratio statistic is directly dependent on sample size. Because of this, with large sample sizes, significant values can be obtained even though there are only trivial discrepancies between a model and the data. Similarly, with large sample sizes, a significant value for an SCDT may be obtained even when there is only a trivial difference between two nested structural models' explanations of the estimated construct covariances. Therefore, an indication of goodness of fit from a practical standpoint, such as that provided by the normed fit index (Δ) of Bentler and Bonett, is useful in conjunction with formal statistical tests. The normed fit index, which ranges from 0 to 1, can be thought of as the percentage of observed-measure covariation explained by a given measurement or structural model (compared with an overall, null model [$M_\varnothing$] that solely accounts for the observed-measure variances).

Under the two-step approach, a normed fit index value would be calculated in conjunction with each SCDT. As an example, Δ_{ts} would provide supplementary information on the practical decrement in fit of the theoretical model of interest from that of the measurement model, expressed as a percentage difference in covariation explained. Put differently, this value would indicate the practical loss of explanatory ability that resulted from constraining to zero the paths that were hypothesized as such in the substantive, structural model.

Depending on the research setting, a researcher may place greater emphasis on the normed fit index values than on the SCDT values in making decisions about which of the alternative structural models to accept. For example, a researcher may decide to accept M_t over M_u on the basis of a practically insignificant Δ_{tu}, even though the SCDT of $M_t - M_u$ indicates a statistically significant difference between the two models. That is, from a practical standpoint, the more parsimonious M_t provides adequate explanation.

Finally, $\Delta_{t\varnothing}$ would indicate the overall percentage of observed-measure covariation explained by the structural and measurement submodels.

Considerations in drawing causal inferences. Causal inferences made from structural equation models must be consistent with established principles of scientific inference (cf. Cliff, 1983). First, models are never confirmed by data; rather, they gain support by failing to be disconfirmed. Although a given model has acceptable goodness of fit, other models that would have equal fit may exist, particularly when relatively few paths relating the constructs to one another have been specified as absent. Second, temporal order is not an infallible guide to causal relations. An example that Cliff noted is that although a father's occupation preceded his child's performance on an intelligence test and the two are correlated, this does not mean that the father's occupation "caused" the child's intelligence.

Third, in what is known as the *nominalistic fallacy,* naming something does not necessarily mean that one understands it. An inherent gap in meaning exists between an observed variable (indicator) and its corresponding, underlying construct because of (a) *invalidity*—the observed variable measures, at least partly, something other than what was intended—and (b) *unreliability*—the values of the observed variable are partly due to random measurement error. Finally, although use of the two-step approach preserves the ability to make some inferences, respecification typically limits the ability to infer causal relations.

Ideally, a researcher would want to split a sample, using one half to develop a model and the other half to validate the solution obtained from the first half. (For a discussion of cross-validation for covariance structures, see Cudeck & Browne's, 1983, article.) However, because large samples are needed to attain the desirable asymptotic properties of full-information ML or GLS estimators, in practice the ability to split a sample most often will be precluded. Application of these principles will have the effect that, in most research situations, only qualified statements of causal inference can be justified.

Conclusion

We have attempted to provide some guidance for substantive researchers regarding the construction and evaluation of structural equation models in practice. Gaining a working understanding of these relatively new, confirmatory methods can be facilitated by the suggested guidelines. The primary contribution of this article is to present a comprehensive, two-step mod-

eling approach that draws on past research and experience, as well as some recent analytic developments. We have also offered guidance regarding the specification, assessment, and respecification of confirmatory measurement models.

As we have advocated, there is much to be gained from a two-step approach, compared with a one-step approach, to the model-building process. A two-step approach has a number of comparative strengths that allow meaningful inferences to be made. First, it allows tests of the significance for all pattern coefficients. Second, the two-step approach allows an assessment of whether any structural model would give acceptable fit. Third, one can make an asymptotically independent test of the substantive or theoretical model of interest. Related to this, because a measurement model serves as the baseline model in SCDTs, the significance of the fit for it is asymptotically independent from the SCDTs of interest. As a result, respecification can be made to achieve acceptable unidimensional construct measurement. Finally, the two-step approach provides a particularly useful framework for formal comparisons of the substantive model of interest with next most likely theoretical alternatives.

Structural equation modeling, properly employed, offers great potential for theory development and construct validation in psychology and the social sciences. If substantive researchers employ the two-step approach recommended in this article and remain cognizant of the basic principles of scientific inference that we have reviewed, the potential of these confirmatory methods can be better realized in practice.

References

Anderson, J. C., & Gerbing, D. W. (1982). Some methods for respecifying measurement models to obtain unidimensional construct measurement. *Journal of Marketing Research, 19,* 453–460.

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49,* 155–173.

Babakus, E., Ferguson, C. E., Jr., & Joreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 24,* 222–228.

Bagozzi, R. P. (1983). Issues in the application of covariance structure analysis: A further comment. *Journal of Consumer Research, 9,* 449–450.

Bagozzi, R. P. (1984). A prospectus for theory construction in marketing. *Journal of Marketing, 48,* 11–29.

Bagozzi, R. P., & Phillips, L. W. (1982). Representing and testing organizational theories: A holistic construal. *Administrative Science Quarterly, 27,* 459–489.

Bentler, P. M. (1976). Multistructural statistical models applied to factor analysis. *Multivariate Behavioral Research, 11,* 3–25.

Bentler, P. M. (1978). The interdependence of theory, methodology, and empirical data: Causal modeling as an approach to construct validation. In D. B. Kandel (Ed.), *Longitudinal drug research* (pp. 267–302). New York: Wiley.

Bentler, P. M. (1983). Some contributions to efficient statistics in structural models: Specification and estimation of moment structures. *Psychometrika, 48,* 493–517.

Bentler, P. M. (1985). *Theory and implementation of EQS: A structural equations program.* Los Angeles: BMDP Statistical Software.

Bentler, P. M. (1986a). *Lagrange multiplier and Wald tests for EQS and EQS/PC.* Los Angeles: BMDP Statistical Software.

Bentler, P. M. (1986b). Structural modeling and Psychometrika: An historical perspective on growth and achievements. *Psychometrika, 51,* 35–51.

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

Bentler, P. M., & Dijkstra, T. (1985). Efficient estimation via linearization in structural models. In P. R. Krishnaiah (Ed.), *Multivariate analysis—VI* (pp. 9–42). Amsterdam: Elsevier.

Bentler, P. M., & Lee, S. Y. (1983). Covariance structures under polynomial constraints: Applications to correlation and alpha-type structural models. *Journal of Educational Statistics, 8,* 207–222, 315–317.

Beran, R. (1979). Testing for ellipsoidal symmetry of a multivariate density. *The Annals of Statistics, 7,* 150–162.

Bock, R. D., & Bargmann, R. E. (1966). Analysis of covariance structures. *Psychometrika, 31,* 507–534.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal, 8,* 1–24.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge, England: 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.

Burt, R. S. (1973). Confirmatory factor-analytic structures and the theory construction process. *Sociological Methods and Research, 2,* 131–187.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods and Research, 5,* 3–52.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician, 36,* 153–157.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist, 15,* 546–553.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cattell, R. B. (1973). *Personality and mood by questionnaire.* San Francisco: Jossey-Bass.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum Press.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research, 18,* 115–126.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

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

Dijkstra, T. (1983). Some comments on maximum likelihood and partial least squares methods. *Journal of Econometrics, 22,* 67–90.

Dillon, W. R., Kumar, A., & Mulani, N. (1987). Offending estimates in covariance structure analysis: Comments on the causes of and solutions to Heywood cases. *Psychological Bulletin, 101,* 126–135.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37,* 36–48.

Finn, J. D. (1974). *A general model for multivariate analysis.* New York: Holt, Rinehart & Winston.

Fornell, C. (1983). Issues in the application of covariance structure analysis: A comment. *Journal of Consumer Research, 9,* 443–448.

Fornell, C., & Bookstein, F. L. (1982). Two structural equation models:

LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research, 19,* 440–452.

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research, 11,* 572–580.

Gerbing, D. W., & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. *Multivariate Behavioral Research, 20,* 255–271.

Gerbing, D. W., & Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika, 52,* 99–111.

Gerbing, D. W., & Hunter, J. E. (1982). The metric of the latent variables in the LISREL-IV analysis. *Educational and Psychological Measurement, 42,* 423–427.

Gerbing, D. W., & Hunter, J. E. (1987). *ITAN: A statistical package for ITem ANalysis including multiple groups confirmatory factor analysis.* Portland, OR: Portland State University, Department of Management.

Harlow, L. L. (1985). Behavior of some elliptical theory estimators with nonnormal data in a covariance structures framework: A Monte Carlo study. *Dissertation Abstracts International, 46,* 2495B.

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

Hart, B., & Spearman, C. (1913). General ability, its existence and nature. *British Journal of Psychology, 5,* 51–84.

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

Hunter, J. E. (1973). Methods of reordering the correlation matrix to facilitate visual inspection and preliminary cluster analysis. *Journal of Educational Measurement, 10,* 51–61.

Hunter, J. E., & Gerbing, D. W. (1982). Unidimensional measurement, second-order factor analysis, and causal models. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 4, pp. 267–299). Greenwich, CT: JAI Press.

Joreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika, 31,* 165–178.

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

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

Joreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol 2, pp. 1–56). San Francisco: Freeman.

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

Joreskog, K. G., & Sorbom, D. (1984). *LISREL VI: Analysis of linear structural relationships by the method of maximum likelihood.* Chicago: National Educational Resources.

Joreskog, K. G., & Sorbom, D. (1987, March). *New developments in LISREL.* Paper presented at the National Symposium on Methodological Issues in Causal Modeling, University of Alabama, Tuscaloosa.

Joreskog, K. G., & Wold, H. (1982). The ML and PLS techniques for modeling with latent variables: Historical and comparative aspects. In K. G. Joreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction. Part I* (pp. 263–270). Amsterdam: North-Holland.

Kmenta, J. (1971). *Elements of econometrics.* New York: MacMillan.

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.* New York: American Elsevier.

Lee, S. Y., & Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika, 44,* 99–113.

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

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

McDonald, R. P., & Mulaik, S. A. (1979). Determinacy of common factors: A nontechnical review. *Psychological Bulletin, 86,* 297–308.

Mooijaart, A., & Bentler, P. M. (1985). The weight matrix in asymptotic distribution-free methods. *British Journal of Mathematical and Statistical Psychology, 38,* 190–196.

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

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Sorbom, D., & Joreskog, K. G. (1982). The use of structural equation models in evaluation research. In C. Fornell (Ed.), *A second generation of multivariate analysis* (Vol. 2, pp. 381–418). New York: Praeger.

Spearman, C. (1914). Theory of two factors. *Psychological Review, 21,* 105–115.

Spearman, C., & Holzinger, K. (1924). The sampling error in the theory of two factors. *British Journal of Psychology, 15,* 17–19.

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's: Some interesting parallels. *Psychometrika, 44,* 157–166.

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

Tanaka, J. C. (1984). Some results on the estimation of covariance structure models. *Dissertation Abstracts International, 45,* 924B.

van Driel, O. P. (1978). On various causes of improper solutions of maximum likelihood factor analysis. *Psychometrika, 43,* 225–243.

Weeks, D. G. (1980). A second-order longitudinal model of ability structure. *Multivariate Behavioral Research, 15,* 353–365.

Wold, H. (1982). Soft modeling: The basic design and some extensions. In K. G. Joreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction. Part II* (pp. 1–54). Amsterdam: North-Holland.

Young, J. W. (1977). The function of theory in a dilemma of path analysis. *Journal of Applied Psychology, 62,* 108–110.

Received September 26, 1986
Revision received August 15, 1987
Accepted August 25, 1987 ■