Psychological Bulletin
1980, Vol. 88, No. 3, 588-606

Significance Tests and Goodness of Fit in the Analysis of Covariance Structures

P. M. Bentler and Douglas G. Bonett
University of California, Los Angeles

Factor analysis, path analysis, structural equation modeling, and related multivariate statistical methods are based on maximum likelihood or generalized least squares estimation developed for covariance structure models. Large-sample theory provides a chi-square goodness-of-fit test for comparing a model against a general alternative model based on correlated variables. This model comparison is insufficient for model evaluation: In large samples virtually any model tends to be rejected as inadequate, and in small samples various competing models, if evaluated, might be equally acceptable. A general null model based on modified independence among variables is proposed to provide an additional reference point for the statistical and scientific evaluation of covariance structure models. Use of the null model in the context of a procedure that sequentially evaluates the statistical necessity of various sets of parameters places statistical methods in covariance structure analysis into a more complete framework. The concepts of ideal models and pseudo chi-square tests are introduced, and their roles in hypothesis testing are developed. The importance of supplementing statistical evaluation with incremental fit indices associated with the comparison of hierarchical models is also emphasized. Normed and nonnormed fit indices are developed and illustrated.

Covariance structure analysis represents a set of techniques for theory testing with correlational data. The theories that can be tested are those that can be represented as a system of equations that describe the unidirectional and bidirectional influences of several variables on each other. Typically, a covariance structure theory or model is specified via a simultaneous set of structural linear regressions of particular variables on other variables. The method is called *covariance structure analysis* because the implications of the simultaneous regressions are studied primarily at the level of correlations or covariances. The main statistical problems involved are those of estimating the parameters of one or several competing models and evaluating the relative goodness of fit of competing models by significance tests.

The field of covariance structure analysis actually covers a wide range of topics, including confirmatory factor analysis, path analysis, and simultaneous equation and structural equation modeling. An overview of the methods that are involved is beyond the scope of this article, but introductions can be found in Bentler (1978, 1980), Bielby and Hauser (1977), Jöreskog (1978), and Kenny (1979). More advanced topics are covered by Aigner and Goldberger (1977), Bentler (in press), Bentler and Weeks (1979, 1980, in press), Jöreskog and Sörbom (1979), and Jöreskog and Wold (in press). It is not the purpose of this article to survey the voluminous literature that is involved. Rather, we examine the logic of statistical significance testing in covariance structure models. This

This investigation was supported in part by U. S. Public Health Service Research Scientist Development Award K02-DA00017 and U.S. Public Health Service Research Grant DA01070.

We would like to thank M. W. Browne, G. J. Huba, B. McGarvey, D. G. Weeks, J. A. Wingard, and J. A. Woodward, who provided valuable ideas that led to this formulation.

Requests for reprints should be sent to P. M. Bentler, Department of Psychology, University of California, Los Angeles, California 90024.

logic can be understood without a familiarity with the technical literature of the field, and hence the limitations of current practice can be established without much mathematical detail. Some relatively simple suggestions are made to improve current practice. The points made are clarified through the use of examples.

Statistical Basis of Covariance Structure Analysis

The statistical theory involved in covariance structure analysis exists only in rudimentary form. Only large-sample theory has been developed to any extent, and the relevance of this theory to small samples has not been established. Although the statistical theory associated with some covariance structure models based on multinormally distributed variables existed years ago (cf. Anderson & Rubin, 1956; Lawley, 1940), Jöreskog (1967, 1969, 1978; Jöreskog & Sörbom, 1979) must be given credit for establishing that maximum likelihood estimation could be routinely applied to a variety of covariance structure models. Although various researchers were studying specialized statistical problems and searching for estimators that might be easy to implement, Jöreskog showed that complex models could be estimated by maximum likelihood methods based on a standard covariance structure approach. The most general alternative approach to estimation in covariance structure models was developed by Browne (1974). Building on the work of Jöreskog and Goldberger (1972) and Anderson (1973), who had developed generalized least squares estimators for the factor analytic model and for linear covariance structures, Browne showed that a class of generalized least squares estimators could be developed that have many of the same large-sample properties as the maximum likelihood estimators, that is, consistency, normality, and efficiency. He also developed the associated goodness-of-fit tests. Lee (1977) showed that maximum likelihood and generalized least squares estimators are asymptotically equal.

Little is known about the relative robustness of these estimators to violation of assumptions or model misspecification and about their relative small-sample properties (see Boomsa, in press; Geweke & Singleton, 1980; Olsson, 1979). Unlike maximum likelihood estimators, the generalized least squares estimators are not based on the assumption of multivariate normality of variables. Unfortunately, the empirical meaning of loosening the normality assumption is open to question, since simple procedures for evaluating the less restrictive assumption (that fourth-order cumulants of the distribution of the variables are zero) do not appear to be available. Although certain generalized least squares estimators are easier to compute than maximum likelihood estimators, there is some evidence that they may be more biased (Browne, 1974; Jöreskog & Goldberger, 1972). A less restrictive statistical theory based on minimal assumptions about variables is currently being developed (Browne, Note 1).

Overview of Estimation

In covariance structure analysis a sample of multivariate data based on $N = n + 1$ subjects and p variables is summarized in the $(p \times p)$ sample covariance matrix $\mathbf{S}$ based on n degrees of freedom. The elements s_{ij} of $\mathbf{S}$ are the variances of the variables and their covariances. (A covariance is a correlation multiplied by the two standard deviations involved.) It is hypothesized that the corresponding population covariance matrix Σ with elements σ_{ij} is generated by q true though unknown parameters that can be assembled in the $q \times 1$ vector Θ, so that each element of the covariance matrix is a function of the q elements of Θ under a given model. Thus $\sigma_{ij} = f_{(ij)}(\Theta)$ may be said to be the model for the covariance structure, where the function $f_{(ij)}$ describes the particular structure under investigation that relates the q parameters in Θ to σ_{ij}. Although models will also contain parameters that are fixed or treated as known in a given application, these parameters are, by convention, not included in Θ. A more complete covariance structure representation might be $\sigma_{ij} = f_{(ij)}(\Theta, \omega)$, where ω represents a vector of fixed parameters. However, even this would be incomplete because the structural form relating parameters is not given in the representation. (See Bentler &

Weeks, 1979, for the range of structural models involved, which, e.g., include factor analytic and simultaneous equation models.) To illustrate the meaning of $\sigma_{ij} = f_{(ij)}(\Theta)$, in exploratory factor analysis the unknown factor loadings and unique variances are the parameters that are the elements of the vector Θ, whereas the factor analytic equations that relate these parameters to the covariance matrix represent the function $f_{(ij)}$. Specifically, this is given by the sum of cross-products of factor loadings, plus the unique variance if $i = j$. In some models one hypothesizes that the elements of σ_{ij} have a certain relationship to each other, summarized in $f_{(ij)}$, and the parameter vector Θ contains elements such as σ_{gh}. Thus if the model states that $\sigma_{ij} = \sigma_{gh}$, the function $f_{(ij)}$ simply selects σ_{gh} from the remaining elements of Θ. In structural equation models, the variances and covariances of the independent variables, the regression coefficients relating independent to dependent variables, the coefficients for regressions relating dependent variables to each other, and residual variances and covariances would typically be elements of the vector Θ, and $f_{(ij)}$ may represent a complex matrix equation. In general, of course, the particular theory or model that is tested would dictate the form and structure (i.e., $\Sigma = \Lambda\Phi\Lambda' + \Psi$, for a confirmatory factor analytic model) that relates the various parameters to each other and to the data.

Assuming a theory to be correct, if the sample size N were arbitrarily large, $\mathbf{S}$ would converge to Σ, and it would be obvious whether the sample data matrix $\mathbf{S}$ corresponded to a particular hypothesized structure (e.g., $\Lambda\Phi\Lambda' + \Psi$). In data analysis, however, where Σ and Θ are not known and where N is not very large, it is first necessary to estimate the parameters of the model, yielding $\hat{\Theta}$ and $\hat{\Sigma}$ via $\hat{\sigma}_{ij} = f_{(ij)}(\hat{\Theta})$. The closeness of the model-based estimated covariance matrix $\hat{\Sigma}$ to $\mathbf{S}$ not only serves as a criterion to be optimized in estimating the parameters, as is described later, but it represents an index of the validity of the model itself. If $\hat{\Sigma}$ is virtually identical element by element under the model to $\mathbf{S}$, that is, $\hat{\sigma}_{ij} \cong s_{ij}$, the model $f_{(ij)}(\Theta)$ that generates $\hat{\Sigma}$ via $f_{(ij)}(\hat{\Theta})$ is a possible candidate for the structure underlying the population Σ. If even the best estimate of Σ under the model is very different from $\mathbf{S}$, it is unlikely that the hypothesized model accurately mirrors the process that generates the data, thus providing a basis for using sample data to reject a given hypothesized model. These points are shortly made more precise.

Loss Functions and Estimators

The long tradition of least squares estimation in statistics might dictate the use of the unweighted least squares goodness-of-fit function,

$$U(\Theta) = \tfrac{1}{2}\text{trace}(\mathbf{S} - \Sigma)^2,$$

in which the q parameters in Θ are estimated so as to minimize $U(\Theta)$. This loss function represents an unweighted sum of squares of the residual elements $(s_{ij} - \sigma_{ij})$ of the matrix $(\mathbf{S} - \Sigma)$. Of course, if $\hat{\Sigma} = \mathbf{S}$, $U(\hat{\Theta}) = 0$. Unfortunately, a statistical theory associated with $U(\hat{\Theta})$ has not been developed. Thus statistical estimation in covariance structure models hinges on the generalized least squares function,

$$G(\Theta) = \tfrac{1}{2}\text{trace}[(\mathbf{S} - \Sigma)\mathbf{W}]^2,$$

which minimizes the sum of squares of a weighted residual matrix ($\mathbf{W}$ is a weight matrix), and the maximum likelihood method, which minimizes

$$L(\Theta) = \log|\Sigma| + \text{trace}(\mathbf{S}\Sigma^{-1}) - \log|\mathbf{S}| - p.$$

The complicated loss function $L(\Theta)$ is a positive quantity that approaches zero as $\mathbf{S}$ and Σ become identical. Browne (1974) has shown that $L(\Theta)$ and $G(\Theta)$ become equivalent as the residuals approach zero and $\mathbf{W}$ estimates Σ^{-1}.

Let $\tilde{\Theta}$ be the estimator that minimizes $G(\Theta)$, with $\mathbf{W}$ being any weight matrix (such as $\mathbf{S}^{-1}$) that converges in probability to Σ^{-1}, and let $\bar{\Theta}$ be the estimator that minimizes $L(\Theta)$. These generalized least squares and maximum likelihood estimators have the following asymptotic (large-sample) properties: (a) They are consistent, (b) they have minimal Cramer-Rao sampling variance, (c) they are asymptotically equal, and (d) they have a multivariate normal distribution with mean vector Θ and covariance matrix given by the inverse of the Fisher information matrix. (For more details on these properties, see Browne, 1974, and Lee, 1977.) These properties are technically important but are not especially relevant to the concerns of this article.

For our purposes, a final property is more important. This property relates the minimum function values $G(\tilde{\Theta})$ and $L(\bar{\Theta})$ to a chi-square statistic. Let $F(\hat{\Theta})$ be either $G(\tilde{\Theta})$ or $L(\bar{\Theta})$, whichever is relevant. Then

$$v = nF(\hat{\Theta})$$

has an asymptotic chi-square distribution, with degrees of freedom equal to the number of free sample variances and covariances minus the number of parameters estimated, that is, $p(p+1)/2 - q$, typically. This final property provides a statistical basis for evaluating a model.

Logic of the Goodness-of-Fit Test

The statistic v, which is a function of sample size N and of the closeness of $\hat{\Sigma}$ to $\mathbf{S}$, provides a probabilistic basis for evaluating the goodness of fit of a model. The chi-square statistic provides a test of the proposed model $\sigma_{ij} = f_{(ij)}(\Theta)$ against the general alternative that the variables are simply correlated to an arbitrary extent. This alternative model proposes that $\sigma_{ij} = f_{s(ij)}(\Theta_s)$, where Θ_s contains all $p(p+1)/2$ elements of Σ (except for duplicated elements σ_{ji} for σ_{ij}); of course, $f_{s(ij)}$ simply picks σ_{ij} from Θ_s. If v is large compared with the degrees of freedom, one concludes that the model $f_{(ij)}(\Theta)$ does not appropriately mirror the process that generates the data in the population. If the statistic is small compared with degrees of freedom, one concludes that the model provides a plausible representation of the system of influences among variables in the population. After all, in that instance the q parameters in Θ are as good as the $p(p+1)/2$ alternative parameters in Θ_s, relating $\hat{\Sigma}$ to $\mathbf{S}$ under random sampling of subjects from the population, given the sample size and degrees of freedom involved in the comparison.

While the chi-square test provides valuable information about a statistically false model, problems associated with sample size mitigate the value of the information that is obtained. The increase in ability to detect a false model with increasing sample size represents an important aspect of statistical power, but in the context of most applications in which the exactly correct model is almost certainly unknowable, this effect of sample size is a mixed blessing. Since the chi-square variate is a direct function of sample size, the probability of rejecting any model increases as N increases, even when the model is minimally false, that is, when the residual matrix $(\mathbf{S} - \hat{\Sigma})$ contains trivial discrepancies between data s_{ij} and estimated model $\hat{\sigma}_{ij}$. As a consequence, in very large samples virtually all models that one might consider would have to be rejected as statistically untenable. Although the statistical conclusion is reasonable, namely, that the residual matrix may contain additional valuable information that a better model could in principle explain, the matrix $\hat{\Sigma}$ based on $\hat{\Theta}$ may contain virtually all of the information that one may be concerned with in practical circumstances.

There is another problem. In many circumstances one would like to establish that the model provides a plausible representation of the data. In effect, a *non*significant chi-square value is desired, and one attempts to infer the validity of the hypothesis of no difference between model and data. Such logic is well-known in various statistical guises as attempting to prove the null hypothesis. This procedure cannot generally be justified, since the chi-square variate v can be made small by simply reducing sample size. Ignoring the as yet unresolved problem of the applicability of large-sample statistical theory to small samples in covariance structure models, it is apparent that the probability of accepting a model increases as N decreases. Thus, one's favorite model will stand the best chance of being accepted when tested against the data of small samples. Of course, accepting the null hypothesis of no difference between $\mathbf{S}$ and $\hat{\Sigma}$, that is, accepting the hypothesis that s_{ij} is obtained as a random sample from $f_{(ij)}(\Theta)$, seems inappropriate in small samples.

These difficulties can be illustrated by two examples. McGaw and Jöreskog (1971) reported an eight-factor exploratory factor analysis of 21 variables based on $N = 11{,}743$. They obtained $v(70) = 403$. The probability of the associated solution based on the tabled values of the chi-square distribution was less than .01, so that the hypothesis that an eight-factor model generates the variances and covariances of the measured variables in the population has to be rejected with a high degree of confidence. However, in view of the large sample size, it is likely that no factor model with positive degrees of freedom could be found that would fit the data with probability greater than .05; that is, no model could be established that would adequately account for all of the statistically reliable data in this large sample. Nonetheless, the absolute size of the residuals would verify that virtually all the important statistical information had been extracted from the data. The converse problem is illustrated in a study by Bentler and Lee (1979). They studied the intercorrelations of four personality variables measured by peer, teacher, and self-ratings in a sample of 68 children. One of their models for the covariance matrix had four trait factors and three method factors. This solution yielded $v(35) = 43.88$. This value does not exceed critical cutoff values in the chi-square distribution, so that the model was considered to be a plausible representation of the process that generated the data. However, in view of the small sample size, numerous competing models, if evaluated, might similarly be accepted. In fact, the simple competing model of complete independence among variables might also be a plausible representation of the data in the population. To minimize interpretive problems associated with significance testing under varying sample size conditions, the logic of significance testing in covariance structure analysis needs more explication. Hierarchical models are a key ingredient of an appropriate statistical methodology.

Hierarchical Models

Provided that models can be framed so as to be hierarchical or nested, that is, with one model able to be considered as a specialization of another model, both generalized least squares and maximum likelihood estimation provide for a chi-square difference test that evaluates the statistical significance of the parameters that differentiate between two competing models.

Let M_k be any covariance structure model of the form M_k: $\sigma_{ij} = f_{k(ij)}(\Theta_k)$, and let M_ℓ be any alternative model of the form M_ℓ: $\sigma_{ij} = f_{\ell(ij)}(\Theta_\ell)$ under the restriction that the set of matrices $f_{k(ij)}(\Theta_k)$ is a subset of the larger set of matrices $f_{\ell(ij)}(\Theta_\ell)$. These sets of matrices are generated as the parameter vectors Θ_k and Θ_ℓ vary over the whole parameter space of admissible values under the covariance structure functions $f_{k(ij)}$ and $f_{\ell(ij)}$. Henceforth, we let the simpler notation f_k and f_ℓ replace the more accurate notation $f_{k(ij)}$ and $f_{\ell(ij)}$. To achieve clarity and generality, we introduce the concepts of *covariance matrix nesting* and *parameter nesting*. In covariance matrix nesting, the function f_k and f_ℓ need not be identical, and the parameter vector Θ_k does not need to be a specialization of the vector Θ_ℓ, but the set of covariance matrices $f_k(\Theta_k)$ must be a subset of the set of matrices $f_\ell(\Theta_\ell)$. In covariance matrix nesting, there is a set of overidentifying restrictions that takes one model into the more specialized model. Unfortunately, these restrictions may not be easy to describe. To illustrate, let M_k: $\Sigma = \Lambda\Lambda' + \Psi^2$ be the traditional factor analytic model for a Λ of rank $k < p$, and let the alternative be M_ℓ: $\Sigma = \Gamma\Gamma'$, where Σ is simply restricted to being positive definite, with Γ being rank p. Evidently, the matrices Σ under M_k are a subset of the matrices Σ under M_ℓ, but it is not possible to describe the common factor loading λ_{ij} and unique loading Ψ_{ii} parameters under M_k as a simple subset of the parameters γ_{ij} under M_ℓ. As another illustration, let M_k: $\Sigma = \Lambda\Lambda' + \Psi^2$ as before, and let M_ℓ: $\Sigma = \Lambda\Lambda' + \mathbf{D}$, where the rank k of Λ under M_k is less than the rank ℓ of Λ under M_ℓ. With Ψ^2 and $\mathbf{D}$ both taken as diagonal matrices, it is apparent that M_k is a specialization of M_ℓ. However, $f_k \neq f_\ell$, since the parameters Ψ_{ii} and d_{ii} are unequal. In this case, the simple reparameterization $d_{ii} = \Psi_{ii}^2$ for M_k (or M_ℓ) would produce a representation in which $f_k = f_\ell$. We may also take Λ under M_k to be identical to Λ under M_ℓ, with the exception that Λ under M_k contains one or more null columns that are instead considered to represent free parameters under M_ℓ. As a consequence, the example has been transformed to the typical case of nesting by parameter restriction.

In parameter nesting, the functions of f_k and f_ℓ are taken to be identical, and the models to be compared differ only in that the parameter vector Θ_k is a special case of the vector Θ_ℓ, obtained by constraining free parameters to equalities or known constants. The case of more general functional constraints is discussed by Lee and Bentler (in press). The comparison of covariance structure models by chi-square difference tests is typically limited to parameter nesting, in which nesting can be easily verified.

Chi-Square Difference Test

The rationale for chi-square difference tests is most clearly seen in the maximum likelihood method, in which a likelihood ratio test (Anderson, 1958; Neyman & Pearson, 1928; Wilks, 1938) provides the basis for comparing competing models. We assume that M_k is a more restricted model than M_ℓ. It is known that, in general, the function $L(\Theta)$ is related to the logarithm of the likelihood function of the observations via

$$L^*(\Theta) = -nL(\Theta)/2 + c,$$

where c is independent of Θ (e.g., Jöreskog, 1967). Let $L^*(\bar{\Theta}_k)$ be the maximum of $L^*(\Theta)$ under M_k, and let $L^*(\bar{\Theta}_\ell)$ be the maximum of $L^*(\Theta)$ under M_ℓ. Thus

$$L^*(\bar{\Theta}_k) \leq L^*(\bar{\Theta}_\ell),$$

since the maximum under a space of restricted range cannot exceed the maximum under a space of less restricted range. Consequently,

$$\log \lambda = L^*(\bar{\Theta}_k) - L^*(\bar{\Theta}_\ell)$$

is negative, with $0 < \lambda \leq 1$. Under the null hypothesis H_0 of model equivalence, $(-2 \log \lambda)$ is asymptotically distributed as a chi-square variate, with degrees of freedom given by the difference in the number of parameters estimated under M_ℓ and M_k. The null hypothesis H_0 associated with parameter nesting, with Θ_k and Θ_ℓ containing *all* parameters, is

$$H_0\colon \Theta_k = \Theta_\ell;$$

that is, it is a test of the equality of parameters under the two models. Since the free parameters in Θ_k are a subset of the free parameters in Θ_ℓ, the null hypothesis also has several equivalent forms in various applications. For example, if M_k simply specializes M_ℓ by setting certain parameters to a vector κ of known constants—typically taken as the null vector—we can write

$$H_0\colon \Theta_{\ell-k} = \kappa,$$

where $\Theta_{\ell-k}$ represents the free parameters in M_ℓ that are not free in M_k. Alternatively, if M_k introduces equality constraints for parameters that are free under M_ℓ, then the null hypothesis can be framed as

$$H_0\colon \theta_a = \theta_b = \ldots = \theta_c,$$

where θ_a, θ_b, and θ_c are free parameters under M_ℓ but are constrained to be equal under M_k. In general, then, the chi-square variate provides a test on the overidentifying restrictions that differentiate the models. If the chi-square variate is large compared with the degrees of freedom, when evaluated against a critical value in the central chi-square distribution (e.g., at $\alpha = .05$), one can reject the corresponding null hypothesis and accept the alternative hypothesis,

$$H_1\colon \Theta_k \neq \Theta_\ell.$$

If the chi-square variate does not exceed the critical value, the null hypothesis is not rejected (i.e., H_0 is "accepted").

The null hypothesis associated with model comparisons takes on another form under covariance matrix nesting. The null hypothesis is that the covariance matrices generated by the parameter vectors under the structural models are equivalent under M_k and M_ℓ; that is,

$$H_0\colon f_k(\Theta_k) = f_\ell(\Theta_\ell).$$

The alternative hypothesis is that

$$H_1\colon f_k(\Theta_k) \neq f_\ell(\Theta_\ell),$$

and the model comparison provides a test of the overidentifying restrictions, evaluated statistically as described earlier.

The overall goodness-of-fit test described previously can be obtained from the rationale described earlier. Let the substantive model

$$M\colon \sigma_{ij} = f(\Theta)$$

be compared with the alternative model

$$M_s\colon \sigma_{ij} = f_s(\Theta_s),$$

where $\Theta_{s(k)} = \sigma_{ij}$. We can call M_s a completely "saturated" model, that is, a model with $p(p+1)/2$ parameters corresponding to the variances σ_{ii} and covariances σ_{ij} of Σ. Typically, then, the saturated model has as many parameters as covariance data points, leaving zero degrees of freedom. If certain elements σ_{ij} are taken as known, M_s has fewer parameters, of course. In most covariance structure models it can easily be shown via the covariance matrix nesting argument given earlier that M is a more restricted model than M_s, that is, that $f(\Theta)$ is a subset of $f_s(\Theta_s)$. It is well-known that the maximum $L^*(\bar{\Theta}_s)$ is obtained as

$$f_s(\bar{\Theta}_s) = s_{ij},$$

the sample covariance matrix. Evaluating $L^*(\bar{\Theta}_s)$ when $f_s(\bar{\Theta}_s) = s_{ij}$ yields

$$L^*(\bar{\Theta}_s) = c,$$

since $L(\bar{\Theta}_s) = 0$. Thus the fit of the saturated model to data is perfect. The maximum $L^*(\bar{\Theta})$ is obtained at $L(\bar{\Theta})$ so that

$$L^*(\bar{\Theta}) = -nL(\bar{\Theta})/2 + c.$$

It follows that

$$-2 \log \lambda = nL(\bar{\Theta}),$$

which was previously shown to equal v, the asymptotic chi-square variate for goodness of fit. This chi-square variate provides a test of the null hypothesis that the covariance matrices generated under M and M_s are equivalent; that is,

$$H_0: f(\Theta) = f_s(\Theta_s),$$

implying no loss of statistical information in using the more restricted model M. If the null hypothesis is rejected in favor of the alternative hypothesis

$$H_1: f(\Theta) \neq f_s(\Theta_s)$$

because the chi-square variate exceeds an appropriate critical value of the chi-square distribution, the model M cannot be assumed to adequately represent the covariance matrix in the population. Crucial statistical information is missing in M.

In practice, the chi-square tests described earlier are applied under conditions of model misspecification, violation of basic assumptions, and data-based model modifications. Although the theoretical properties of the tests, such as the associated probability levels, are likely to be compromised under extremely unfavorable circumstances, little is known as yet about the robustness of the tests in the range of applications typically encountered in practice (Boomsa, in press; Geweke & Singleton, 1980; Lawley & Maxwell, 1971; Olsson, 1979). Further work clearly needs to be directed at understanding the virtues and limitations of the chi-square procedures.

It seems that no technical development of the rationale for a chi-square difference test based on generalized least squares estimators has been published. We propose two procedures for establishing that the chi-square difference test is applicable to generalized least squares estimators, based on the asymptotic equivalence of the respective chi-square functions and on additive features of the chi-square statistic. Browne (1974) showed that the chi-square statistics $nG(\tilde{\Theta})$ and $nL(\bar{\Theta})$, based on optimized generalized least squares and maximum likelihood functions, converge stochastically. This convergence occurs under both M_k and M_t of two nested models. Consequently, the difference between these converging statistics also converges. Alternatively, it is known (Lancaster, 1969) that a chi-square based on t degrees of freedom can be decomposed into the sum of squares of t independent, unit-normal variates. Thus $nG(\tilde{\Theta})$ can be so decomposed. Now, partition $\tilde{\Theta}$ into $\tilde{\Theta}^r$ and $\tilde{\Theta}^s$, based on r and s $(r + s = t)$ degrees of freedom, and identify $\tilde{\Theta}^r$ with the more restrictive model M_k of two nested models (M_k, M_t) and $\tilde{\Theta}^s$ with the parameters that differentiate the models. Thus if $nG(\tilde{\Theta}^r)$ and $nG(\tilde{\Theta}^s)$ are the chi-square statistics for the restricted and difference sets of parameters, based on a sum of r and s squares of independent unit-normal variates, respectively, then the sum $nG(\tilde{\Theta}^r) + nG(\tilde{\Theta}^s) = nG(\tilde{\Theta})$ is a sum based on t such squares of variates. A more general, technical development is provided by Lee and Bentler (in press).

The chi-square difference test can be used to establish a general hierarchy of tests that is more informative than the test that compares a given model with a saturated model. The value of such nested model comparisons has been thoroughly discussed by Jöreskog (1967, 1969, 1978; Jöreskog & Sörbom, 1979), and this emphasis need not be reiterated here. The comparisons that are made should ideally be dictated by theory, as in the comparison of two competing models of a given phenomenon. In addition to such theoretical comparisons, the comparison of a given model to a saturated model represents an important nested comparison, as pointed out earlier. Equally important, in many circumstances, is the comparison between a given model and a suitably framed null model.

Null Model and General Hierarchical Comparisons

Consider the hierarchically nested covariance structure models M_k, M_ℓ, and M_s, where M_k is the most restricted model and M_s is the saturated model. Since M_s can always be fit to any set of data without error, it does not represent a structural model than can be evaluated or rejected. Rather, it serves as a standard of comparison for M_k and M_ℓ. Although M_k and M_ℓ must of necessity be explicated specifically in the context of a given research design, it may be fruitful to inquire whether a model M_0 more restrictive than M_k can be developed in the general case. If so, it would be possible to recommend a model-testing strategy for covariance structure analysis that minimizes the difficulties associated with sample size discussed earlier. In such a model-testing strategy, M_0, M_k, M_ℓ, and M_s would be evaluated relative to each other using chi-square difference tests. These tests, in general, evaluate the statistical necessity of individual or groups of parameters that differentiate the models involved. In small samples, the parameters that differentiate between M_0 and the least restrictive structural model M_ℓ may not be statistically significant, and the meaning of the $M_\ell - M_s$ comparison may need to be reinterpreted as a result. The $M_k - M_\ell$ comparison has been well explicated in the literature, but it too may need reinterpretation in the rare situation in which the $M_0 - M_\ell$ comparison does not yield a statistically significant effect. In very large samples, a significant statistical effect for the parameters that differentiate M_0 and M_ℓ could indicate that valuable information has been extracted from the data, even if the comparison of M_ℓ with M_s yields the conclusion that additional statistical information remains to be extracted from the data. No general discussion of the role of a null model M_0 in model comparisons appears in the literature. Before describing a general null model, it may be worthwhile to fix ideas involving a covariance structure model via matrix representation.

Model Comparison and Matrix Representations

A complete structural representation for a model of substantive interest in a research problem requires specification of the fixed or known parameters. Thus we might write M_ℓ: $\sigma_{ij} = f(\Theta_\ell, \omega_\ell)$, where the $(q \times 1)$ vector Θ_ℓ represents the q_ℓ unknown parameters of the model, and ω_ℓ represents the vector of known or fixed parameters under a given structural representation. To help make the meaning of M_ℓ clear, consider the general matrix representation of covariance structure models provided by Bentler and Weeks (1979, 1980). These authors claim that all structural models that are linear in variables can be obtained as a specialization of the covariance matrix model

$$\Sigma = \mathbf{G}\mathbf{B}^{-1}\Gamma\Phi\Gamma'\mathbf{B}^{-1\prime}\mathbf{G}',$$

where Φ represents the covariance matrix of all independent variables, Γ represents the regression weights of dependent on independent variables, $(\mathbf{I} - \mathbf{B})$ represents the regression weights of dependent variables on each other, and $\mathbf{G}$ is a known matrix that relates measured variables to all variables in the system. An independent variable is never a dependent variable in any regression equation, whether it is based on measured, latent, or residual variables; all the remaining variables are considered dependent. With this notation, the q_ℓ elements of Θ_ℓ are the unknown and unconstrained elements of $\mathbf{B}$, Γ, and Φ, whereas the elements of ω_ℓ represent the known or fixed parameters in $\mathbf{B}$, Γ, and Φ and all the elements of $\mathbf{G}$. If functional relations are allowed among the free parameters, then Θ_ℓ represents the independent free parameters of the model after consideration of all functional relations. Thus although there may be more than q_ℓ unknown parameters in $\mathbf{B}$, Γ, and Φ, only q_t of these are allowed to freely vary in fitting the model to data (see Lee & Bentler, in press).

The competing general alternative model is M_s: $\sigma_{ij} = f_s(\Theta_s, \omega_s)$, corresponding to a saturated model that specifies q_s arbitrary variances and covariances arranged in the $(q_s \times 1)$ vector Θ_s and the remaining fixed values of these parameters in ω_s, as discussed earlier. Alternatively, ω_s represents the number of independent restrictions placed on the $p(p+1)/2$ elements of the covariance matrix. In the previous matrix notation, M_s would be represented by $\Sigma = \Phi$, a symmetric matrix. The free parameters of Φ are given by Θ_s, and the fixed parameters of $\mathbf{G} = \mathbf{I}$, $\mathbf{B} = \mathbf{I}$, $\Gamma = \mathbf{I}$, and Φ would be given in ω_s. The comparison of M_t to M_s provides an index of goodness of fit of the model M_t, as described before, with d_t degrees of freedom. Since M_s may not be associated with $q_s = p(p+1)/2$ free parameters, it is more accurate to say $d_t = q_s - q_t$. A comparison between M_t and M_s is made implicitly in the chi-square statistic v associated with M_t, which is judged relative to d_t.

If M_k: $\sigma_{ij} = f(\Theta_k, \omega_k)$ is a more restricted intermediate model obtained by parameter nesting, then Θ_k must be a subset of all the parameters in Θ_t, and ω_k must contain all restrictions found in ω_t. If, in the previous matrix notation, M_k is a model such that all regressions among variables have been eliminated, then all elements of matrices $\mathbf{B}$ and Γ that corresponded to free parameters under M_t would be fixed, and only Φ would contain free parameters to be estimated in fitting the model to data. The vector of fixed parameters and constraints ω_k would include all the fixed parameters ω_t of M_t, plus additional parameters that were free under M_t but are fixed (usually at zero) under M_k, as well as constraints imposed on parameters that were free under M_t. The comparison of M_k with M_t will in this example provide an overall test of the value of the free regression parameters of M_t considered simultaneously.

Null Model

We propose that a general null model is given by

$$M_0\text{: } \sigma_{ij} = f_0(\Theta_0, \omega_0),$$

where Θ_0 represents the $(q_0 \times 1)$ vector of free variances of the measured variables and ω_0 represents the vector of known or fixed parameters. In general ω_0 must contain all restrictions found in ω_k plus new restrictions that are generated by moving from M_k to M_0. In most research contexts, the fixed parameters ω_0 are covariances that are presumed to be zero, and M_0 represents the severely restricted model specifying that the variables are mutually independent. In some research contexts, certain variances, covariances, or regression coefficients may be treated as known (possibly nonnull), and the model M_0 specifies that the remaining covariances are zero.

When ω_0 consists only of fixed zero covariances among manifest variables, the model M_0, as compared with M_s, has been previously discussed as a test of complete independence (Larzelere & Mulaik, 1977), but its role in covariance structure analysis generally has been ignored. In exploratory factor analysis, M_0 can serve as a baseline model of no common factors, against which a factor model can be compared (McGaw & Jöreskog, 1971; Tucker & Lewis, 1973). In this case, the unique variances become equated with the variances of the manifest variables, and all common factor loadings are simultaneously set to zero. However, the model M_0 is relevant to confirmatory factor analysis, path analysis, structural equation models, higher order and three-mode factor analytic models, and other structural models that in their most general form implicate a nonzero covariance among several variables. In the context of the matrix representation for Σ, one can usually obtain the null model by setting $\Sigma = \Phi$, with Φ being diagonal except for certain fixed covariances. Thus it is apparent that M_0 is a special case of models M_k or M_t that may be of substantive interest. As one of the examples (later) shows, however, there are cases when it is not a simple matter to verify that M_0 is a special case of M_k, M_t, or M_s. Nonetheless, this is a fundamental requirement of M_0.

Let us assume that M_t corresponds to a model of special interest. A comparison of M_0 with M_t provides a test of whether the restrictions made in going from M_t to M_0 are reasonable. If the chi-square difference test does not yield a value of the chi-square statistic that exceeds typical cutoff points in the chi-square distribution (e.g., at the .05 level of significance), the free parameters that are included under M_t but not under M_0 provide no additional statistical gain in fitting a model to data. In that case, even if M_t provides an adequate model for the data, as evaluated by the goodness-of-fit comparison with M_s, M_t is no more adequate a model than the relatively uninformative model of modified independence. Such a situation is likely to arise only in small samples. Another possible but improbable situation involves a statistically significant difference between M_k and M_t, with the comparable $M_0 - M_t$ difference not being significant; in this case, the null model M_0 also serves to cast doubt on the model M_t, even though it appears to contain statistically valuable parameters when evaluated against M_k. If the chi-square statistic for the $M_0 - M_t$ comparison is large compared with the degrees of freedom, the free parameters that differentiate the models provide a significant increment in fit of model to data. Such a result becomes even more informative if each comparison $M_0 - M_k$ and $M_k - M_t$ is also statistically significant, for then each model comparison demonstrates the value of the parameters that differentiate models. Of course, it is also possible to obtain a significant overall effect of parameters that differentiate M_0 and M_t, although the specific sources of the difference lie in $M_0 - M_k$ and not $M_k - M_t$ (or vice versa). Note, however, that statistical rejection of M_t (compared with that of M_s) does not imply that an acceptable, more restricted model M_k cannot be found. In any case, if any model M_t provides a significant increment in fit over M_0, valuable statistical effects have been localized, even if M_t does not account for all the data, as indicated by the $M_t - M_s$ comparison. It is also important to use a nonstatistical method for evaluating goodness of fit. Incremental fit indices for this purpose are discussed in a subsequent section.

Ideal Models and Pseudo Chi-Square Tests

Existing methods of model comparison are limited to the goodness-of-fit test that compares a given model M_t with the saturated alternative M_s and to the comparison between two existing hierarchical models M_k and M_t. A new, additional model comparison procedure is possible, however. This comparison involves a given model M_t and all conceivable alternative submodels, such as an M_k that has not even been defined. It is based on a general idealized significance test logic that is spelled out in greater detail by Bentler, Bonett, and Browne (Note 2).

Idealized Significance Test Logic

As was done previously, assume that M_t is based on q_t unknown parameters with d_t goodness-of-fit degrees of freedom. Assume that the chi-square goodness-of-fit variate for M_t takes on the value v_t. It follows that any nested model M_k has a chi-square variate $v_k \geq v_t$ and degrees of freedom $d_k > d_t$. The upper bound on degrees of freedom for any model M_k is given by $d_0 - 1$, based on the degrees of freedom d_0 of the null model minus at least one additional parameter. (M_0 is the most restricted model that is entertained.) Now consider an "ideal" model $M_k{}^i$ that fits best and has the largest degrees of freedom: This model will have a chi-square statistic equal to v_t and degrees of freedom $d_0 - 1$. The *pseudo chi-square test* associated with this ideal model $M_k{}^i$, based on v_t and $d_0 - 1$, provides a goodness-of-fit test for the ideal submodel. *If the data are not consistent with the ideal submodel, all conceivable submodels can also be rejected statistically.* There is then no point to actually testing any given submodel M_k, since any such model will have (a) a larger chi-square value (larger than v_t) and/or (b) smaller degrees of freedom (smaller than $d_0 - 1$). Consequently, an actual chi-square test based on any real submodel M_k will also be statistically significant if the pseudo chi-square test is significant. If the ideal submodel cannot be rejected, of course, no conclusion can be drawn about potential actual submodels.

The pseudo chi-square test can be carried out starting with any given model M_t, using as a base the degrees of freedom of any rational, known submodel besides M_0. One of the most informative of such tests is to be found with complete latent variable models (cf. Bentler, in press) in which the dimensionality of the latent variables, including an error of measurement or unique latent variable for each measured variable, is greater than the dimensionality of the measured variables. For example, suppose there are $p = 20$ measured variables, an equal number of unique or error factors, and $t = 5$ common factors. The relation of the p measured variables to the $p + t$ (25) latent variables describes the *measurement submodel* of a given model, and the interrelations of the t common factors describes the *structural submodel* (or simultaneous equation or structural equation submodels) (see Bentler, 1978; Jöreskog & Sörbom, 1978; Wiley, 1973). (In more complex types of models, this distinction is not particularly useful; see Bentler, in press.) Suppose the model with t factors and their regression relations is called M_k. Now consider the more general model that is identical to M_k but replaces all regression and residual relations in M_k with a saturated structural submodel [i.e., the common factor relations are replaced by $t(t + 1)/2$ factor variances and covariances]; call this model M_ℓ. Finally, consider the specialization of M_k that eliminates all structural regressions and factor covariances and leaves only the t factor variances, that is, the null structural submodel M_n. Of course, the null structural model must maintain all structural submodel parameters that are fixed at arbitrary nonzero values in M_k. It is thus apparent that (M_n, M_k, M_ℓ) are hierarchically nested.

Suppose that M_ℓ is based on q_ℓ unknown parameters and d_ℓ goodness-of-fit degrees of freedom. Then M_n has $q_n = q_\ell - t(t - 1)/2$ unknown parameters and $d_n = d_\ell + t(t - 1)/2$ degrees of freedom. If M_ℓ is associated with a chi-square statistic v_ℓ, we can evaluate an ideal submodel $M_k{}^i$ that falls between M_n and M_ℓ. This ideal structural submodel will have a pseudo chi-square statistic v_ℓ, based on d_n degrees of freedom. (Note that v_n does not need to be computed.) Any other model intermediate between M_n and M_ℓ will have a larger chi-square statistic and/or have fewer degrees of freedom. If the data are not consistent with the ideal structural submodel, all conceivable structural submodels having the same measurement submodel can also be rejected statistically. In particular, the proposed model M_k must also not fit statistically, and there is no way the structural submodel can be modified to yield an acceptable overall model. Stated differently, this pseudo chi-square test can determine when a measurement submodel is fundamentally misspecified or flawed. The flaw is fundamental, since no model with the given measurement structure could ever be found to be statistically consistent with the data.

Minimal Model-Differentiating Parameters

The general idealized significance test logic can also provide an ordinal index of the minimum number of parameters that are necessary to statistically differentiate two nested models. Again, let M_ℓ based on d_ℓ degrees of freedom have a chi-square statistic v_ℓ, and let the more restricted model M_k have corresponding values d_k and v_k. Assume that the chi-square difference test $v_{k\ell} = v_k - v_\ell$ based on $d_{k\ell} = d_k - d_\ell$ degrees of freedom is not statistically significant but, for simplicity, that both models M_k and M_ℓ do not differ significantly from M_s, that is, fit the data by a goodness-of-fit test. As an example, M_k may be the model of modified independence, and M_ℓ may be a substantive model of interest. Then one may be interested in knowing the minimal number of parameters required to be added to M_k before a statistically significant improvement in fit is even theoretically possible. Alternatively, one may be interested in other intermediate models from the standpoint of wanting to know the minimum number of parameters that can always be dropped from M_ℓ no matter what the model. These *minimally differentiating* parameter numbers are determined by the same logic.

The ideal differentiating model $M_k{}^i$ that is intermediate between M_k and M_ℓ cannot differ from M_k or M_ℓ by more than $v_{k\ell}$ units in the chi-square distribution. This model may have anywhere from 1 to $d_{k\ell}$ degrees of freedom. Since $v_{k\ell}$ with $d_{k\ell}$ degrees of freedom is not statistically significant, we can step down $d_{k\ell}$ in unit steps until a critical cutoff point in the chi-square distribution (e.g., at $\alpha = .05$) is reached. If $d_{k\ell}{}^*$ represents the degrees of freedom for which $v_{k\ell}$ exceeds the cutoff point, then at least $q_{k\ell} = d_{k\ell} - d_{k\ell}{}^*$

parameters must be added to M_k to yield a new model that has a significant increment in fit over M_k. Stated differently, any number of parameters less than q_{kt} can always be added to M_k without yielding a significant chi-square difference. Viewed from the point of M_t, $d_{kt}^* + 1$ or more parameters dropped from M_t in any specialization toward M_k will always guarantee a nonsignificant decrement from M_t in statistical fit.

If a real model close to the ideal model M_k^i can be specified, it may also yield an acceptable overall goodness-of-fit test if both M_k and M_t do so. However, there would be a significant decrement in fit in going to the more restricted M_k, making the acceptability of M_k questionable. If M_t fits statistically but M_k does not do so, in spite of the nonsignificant chi-square difference in models, then the real model that is close to the ideal model would probably also yield an acceptable overall goodness-of-fit chi-square.

Incremental Fit Indices

Although the process of comparing hierarchical models via tests of significance yields a valuable perspective on the analysis of covariance structures, an index of the amount of information gained in the comparison would provide important additional information about the usefulness of competing models. Such an index of information gained should be independent of sample size and statistical significance test information, though it must reflect the goodness of fit of competing models. Authors of literature on psychological statistics are well aware of the distinction between statistical significance and practical significance; an incremental fit index can provide information about practical significance, in which a statistically significant effect can be evaluated for its practical usefulness in explaining the data. Furthermore, providing that the index is framed in a general way, it would apply to analyses such as those based on least squares procedures that do not as yet have a statistical basis and to analyses in which the probability levels associated with a statistically based procedure may not be appropriate.

A general index of incremental fit appropriate to alternative estimation methods and to an arbitrary covariance structure model, it seems, has not appeared in the literature. Tucker and Lewis (1973) provided an index for evaluating a very specialized covariance structure with a particular estimation method, namely, the exploratory factor analytic model when evaluated by maximum likelihood methods. They proposed the index

$$\rho = (Q_0 - Q_k)/(Q_0 - 1),$$

where $Q_0 = (v/df)_0$ and $Q_k = (v/df)_k$; that is, the quantities Q_0 and Q_k represent ratios of a chi-square variate v to df, evaluated with a given number $(0, \ldots, k)$ of common factors. The index ρ represents an index of increment in fit obtained by using k common factors rather than none. Thus the index compares a null model of independence with a model with k common factors. Unfortunately, ρ is not normed to necessarily lie between zero and one, and its potential relevance to arbitrary covariance structures requires justification. The index is extended here to the general models M_0, M_k, and M_t described earlier for arbitrary covariance structures, where the particular comparison of M_k and M_t is summarized by

$$\rho_{kt} = (Q_k - Q_t)/(Q_0 - 1).$$

This *nonnormed fit index* represents the increment in fit obtained in evaluating any hierarchical step-up comparison of two models (M_k, M_t). Note that although ρ_{kt} is relevant to any hierarchical model comparison, we use the common denominator $(Q_0 - 1)$ in all comparisons, so that the null model M_0 plays a crucial role in the comparison. As applied to (M_0, M_k, M_t), one obtains $\rho_{0t} = \rho_{0k} + \rho_{kt}$, and ρ_{0t} represents an index of the overall fit of M_t in relation to M_0. The ratios Q_k $[= (v/df)_k]$, Q_t, and Q_0 can be based on generalized least squares estimates as well as maximum likelihood estimates. Since it cannot be guaranteed that $Q_k \geq Q_t$, it is possible for ρ_{kt} to be negative.

A slightly more general *normed fit index* that does not require a statistical basis for model fitting is given by

$$\Delta_{kt} = (F_k - F_t)/F_0,$$

where F is any fit function such as $U(\Theta)$, $G(\Theta)$, or $L(\Theta)$, F_0 is the function evaluated under the null model previously proposed, and F_k and F_ℓ correspond to the minimum function values $F(\hat{\Theta})$ for the hierarchically defined step-up models (M_k, M_ℓ). It is apparent that $F_0 \geq F_k \geq F_\ell \geq 0$, so that the index is additive and lies in the interval $0 \leq \Delta_{k\ell} \leq 1$. As applied to ($M_0$, M_k, M_ℓ), $\Delta_{0\ell} = \Delta_{0k} + \Delta_{k\ell}$, and $\Delta_{0\ell}$ represents the overall fit of M_ℓ (again, in relation to M_0).

In some circumstances the most restrictive model M_0 that one might consider would also contain theoretically uninteresting free parameters. For example, in the random regression model, the covariances of the predictors are not considered known, but they are often theoretically uninteresting. In general, the most restrictive, theoretically defensible model should be used in the denominator of the fit indices presented earlier.

The normed and nonnormed fit indices can be used in the comparison of nested or nonnested models within any sample, including possible cross-validation or replication samples (Huba, Woodward, Bentler, & Wingard, Note 3). In addition, the fit indices can be useful in the comparison of a particular model across samples, when the chi-square statistics and the associated probability levels may not be comparable due to unequal sample sizes.

Since the scale of the fit indices is not necessarily easy to interpret (e.g., the indices are not squared multiple correlations), experience will be required to establish values of the indices that are associated with various degrees of meaningfulness of results. In our experience, models with overall fit indices of less than .9 can usually be improved substantially. These indices, and the general hierarchical comparisons described previously, are best understood by examples.

Illustrative Applications

Although the material developed earlier represents a natural extension and clarification of previous work in the area of covariance structure analysis, researchers have not used the proposed null model, fit indices, and pseudo chi-square tests to clarify the role of inference in particular applications. (See Olsson & Bergman, 1977, however, for an application of the Tucker-Lewis, 1973, coefficient.) All the examples are based on maximum likelihood estimation.

As described previously, McGaw and Jöreskog (1971) carried out an eight-factor exploratory factor analysis of 21 variables based on $N = 11{,}743$. They obtained a value of v_1, the chi-square statistic, of 403 with 70 *df*. Thus the eight-factor model M_1 could not be said to account for the data when compared with M_s, the saturated model, since the probability of obtaining data such as the sample covariance matrix is less than .001 if the eight-factor model is true. However, the null model M_0 yields a value of $v_0(210) = 57{,}915$, $p < .001$. Consequently, the improvement in fit obtained by the eight-factor model over the independence model is highly significant, based on a chi-square difference of 57,512 with 140 *df* ($p < .001$). The incremental fit indices yield $\hat{\rho}_{01} = (275.8 - 5.8)/274.8 = .983$ and $\hat{\Delta}_{01} = (57915 - 403)/57915 = .993$, thus verifying that M_1 is a substantial improvement over M_0. These results also demonstrate that the remaining improvement $1 - \hat{\Delta}_{01} = .007$ that might be obtained with a more adequate model is insignificant from a practical viewpoint, even though M_1 represents a statistically significant lack of fit to the data.

The small-sample problem mentioned previously is also amenable to further statistical and incremental fit analyses. Bentler and Lee (1979) studied the intercorrelations of four personality variables measured by peer, teacher, and self-ratings in a sample of 68 children. A model M_1 associated with four trait factors and three method factors yielded $v_1(35) = 43.88$, $p > .05$, thus verifying that M_1 could not be distinguished statistically from M_s. However, Bentler and Lee did not compare their model with the null model M_0. The null model yields $v_0(66) = 400.45$, so that it must be concluded that the 12 variables are not statistically independent. The chi-square difference test comparing M_0 and M_1 yields a chi-square difference of 356.47. With 31 *df*, this difference is highly significant ($p < .001$). Consequently, important additional information has been obtained from the comparison, namely, that the theory underlying M_1 accounts for a statistically significant portion of the data. In the metric

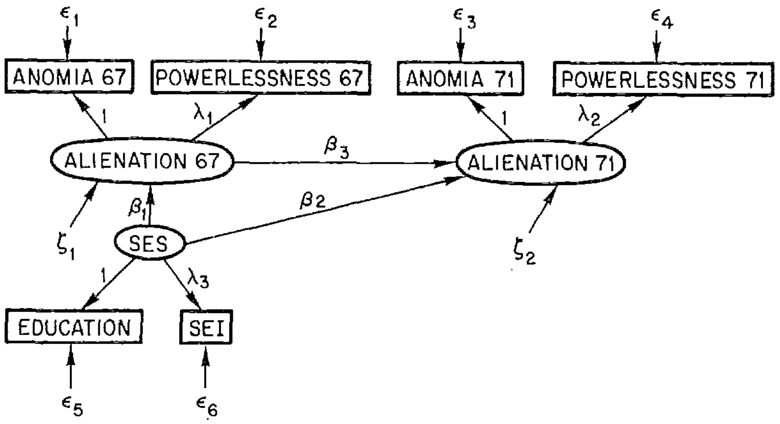

Figure 1. A model with six manifest and three common factor latent variables. (SES = socioeconomic status; SEI = socioeconomic index. The manifest variables are in rectangles. Unidirectional arrows represent regression weights, with λ_g = the regression of manifest on latent variable, β_g = the regression of latent on latent variable, ζ_g = the regression of latent on residual variable, and ϵ_g = the regression of manifest on residual variable.)

of incremental fit indices, one obtains $\hat{\rho}_{01} = .950$ and $\hat{\Delta}_{01} = .890$. Since these increments in fit are associated with a statistically significant difference between M_0 and M_1, the insignificant difference between M_1 and M_s need not be a source of concern due to the small sample size.

An example taken from Wheaton, Muthén, Alwin, and Summers (1977) and described in Jöreskog and Sörbom (1978) and various other sources more completely illustrates the process of model comparison with three models M_0, M_1, and M_2. In this case six manifest variables were taken to be indicators of three latent variables, as shown in Figure 1. The measured variables consisted of anomia and powerlessness measured in 1967 and 1971, as well as education and a socioeconomic index. The latent variables were alienation, both in 1967 and 1971, and a general socioeconomic status factor. The λ_g in the figure represent factor loadings, the ϵ_g represent measurement residuals, the β_g represent regression coefficients for regressions among latent variables, and the ζ_g represent regression residual variables.

The model M_2 considered in Figure 1 is not compatible with the data, $v_2(6) = 71.47$, $p < .001$. M_2 is thus substantially different from M_s, the saturated model, and one might inquire whether a better model could be developed for the data. The authors indeed develop a model that allows correlation between the error terms (ϵ_1, ϵ_3) and (ϵ_2, ϵ_4). Rather than pursue their final model, we might inquire whether the model of Figure 1 has any redeeming features. Although the chi-square goodness-of-fit statistic is very large, M_2 may nonetheless capture a significant aspect of the data. The relevant model tests and comparisons are given in Table 1. The left part of the table shows the chi-square values for the null model M_0 consisting of the six variances associated with ϵ_i only, for a Model M_1 similar to that in Figure 1 but with the three β_g coefficients set to zero and the three λ_g coefficients set to one, and the complete model M_2, as reported previously. M_0, M_1, and M_2 are obviously hierarchical models. As mentioned earlier, a goodness-of-fit chi-square test of a given model represents an implicit model comparison, and hence all model tests in the left part of Table 1 represent implicit comparisons of a given model against M_s. M_s has $v_s(0) = 0$, so that the chi-square of each model is equivalent to the chi-square difference between the model and M_s. The chi-square difference tests and fit indices associated with the three model variants are shown in the right part of Table 1. It is apparent that even though the final model M_2 is inadequate when compared with M_s, the improvement in fit over M_0 that is due to M_1 and the improvement of M_2 over M_1 are

Table 1
Evaluation of Submodels of Figure 1

Model test*			Model comparison*				
Model	χ^2	*df*	Comparison	χ^2	*df*	$\hat{\rho}_{kl}$	$\hat{\Delta}_{kl}$
M_0	2,131.43	15	M_0–M_1	1,408.03	3	.580	.661
M_1	723.40	12	M_1–M_2	651.93	6	.343	.305
M_2	71.47	6	M_0–M_2	2,059.96	9	.923	.967

* $p < .001$ for all chi-square statistics.

both statistically as well as practically significant. Furthermore, if one were not able to produce a better model than that shown in Figure 1, one would at least have the satisfaction of realizing that the remaining increment in fit that might be possible would be only .033 (= 1.0 − .967). Thus although the model in Figure 1 cannot statistically account for the data in this sample of $N = 932$, the importance of the data remaining to be explained, considered in the context of the entire sample data, is in practical terms quite small. In this example, of course, the authors were able to obtain $v_3(4) = 4.73$, $p > .05$, with a model M_3 allowing the correlated errors mentioned earlier (not shown in Figure 1). The corresponding incremental fit indices are $\hat{\rho}_{23} = .076$ and $\hat{\Delta}_{23} = .031$, so that $\hat{\rho}_{03} = .999$ and $\hat{\Delta}_{03} = .998$, indicating virtually perfect fit for the final model.

The model of Figure 1 may also be used to illustrate ideal models and pseudo chi-square tests. It can be shown that M_2 is equivalent to a model with a saturated structural submodel; that is, the coefficients β_i in Figure 1 can be replaced by factor covariances. A model M_n with a null structural submodel would be obtained by setting the three $\beta_i = 0$. Although this model is not identified, it is clear that its goodness-of-fit v_n would exceed 71.47, based on 9 *df*. Consequently, an ideal submodel based on the measurement structure of Figure 1 would have $v_2(9) = 71.47$. The corresponding pseudo chi-square test ($p < .001$) indicates that the measurement submodel contains a fundamental misspecification. This conclusion is consistent with the fact reported previously that a model M_3 required a modification in the measurement model to yield a statistically acceptable fit.

Finally, we turn to an unusual example to illustrate some of the complexities that can arise in verifying that a null model M_0 is a special case of a substantive model of interest. The model is more properly considered econometric rather than psychological in nature, representing Tintner's model of the U. S. meat market. However, the data, model, and analysis are widely known in psychological circles via Jöreskog and Sörbom's (1978) analysis, and the problem involves inference difficulties stemming from sample size. The data represent time-series observations on five variables based on a sample size of 23. The model is given as a statement of two regression equations dealing with consumption (y_1) and price (y_2) as a function of disposable income (x_1), meat processing costs (x_2), and agricultural costs (x_3). It is presumed that the six variances and covariances among x_1, x_2, and x_3 are known, so there remain nine free covariances (actually, deviation sums of squares and cross-products) in the symmetric $5(6)/2 = 15$ element sample matrix S. The variables are related under a theoretical model M_2 via

$$y_1 = \beta_1 y_2 + \gamma_1 x_1 + \zeta_1,$$
$$y_1 = \beta_2 y_2 + \gamma_2 x_2 + \gamma_3 x_3 + \zeta_2,$$

where ζ_1 and ζ_2 are residuals that are allowed to correlate. The parameters of the model are the β and γ coefficients, as well as the variances and covariance of (ζ_1, ζ_2). There are thus eight parameters in M_2, and the fit of M_2 compared with that of M_s yields $v(1) = 2.68$, $p > .05$. Since the model cannot be rejected, one may conclude that the sample data are drawn from a population having the proposed structure. However, the sample size is so small that it may be possible for virtually any model to be accepted.

To strengthen inference, models M_1 and M_0 were developed by us and compared with M_2. M_1 consists of that specialization of M_2 in which all regression coefficients β and γ are fixed as known, that is, a model with three free parameters and 6 *df*. M_0 consists of that further specialization of M_1 that yields modified independence, that is, a model having only the variances of y_1 and y_2 as free parameters with 7 *df*. Models M_0 and M_1 are best understood after further study of the equations under M_2.

The second equation under M_2 can be manipulated to yield

$$y_2 = (1 + \beta_2)y_2 - y_1 + \gamma_2 x_2 + \gamma_3 x_3 + \zeta_2,$$

which expresses y_2 as a function of itself as well as other variables. It is now obvious that simply setting all β and γ coefficients to zero will not yield an equation in which y_2 is expressed as a function of other variables in the system, since y_2 drops out of both equations. Thus under M_1 we specify $\beta_2 = -1$ and $\beta_1 = \gamma_1 = \gamma_2 = \gamma_3 = 0$. It follows that

$$y_1 = \zeta_1 \quad \text{and} \quad y_2 = \zeta_2 - \zeta_1.$$

Consequently, since under M_1 $\text{var}(\zeta_1)$ and $\text{var}(\zeta_2)$ and $\text{cov}(\zeta_1, \zeta_2)$ are free parameters, the effect of these parameters is one of inducing variance and covariance between y_1 and y_2 only. That is, M_1 predicts zero covariance between the *y*s and *x*s. Although M_1 may or may not be substantively interesting, it provides a useful avenue for establishing M_0. Under M_1, $\text{var}(y_1) = \text{var}(\zeta_1)$, $\text{var}(y_2) = \text{var}(\zeta_1) + \text{var}(\zeta_2) - 2 \text{cov}(\zeta_1, \zeta_2)$, and $\text{cov}(y_1, y_2) = \text{cov}(\zeta_1, \zeta_2) - \text{var}(\zeta_1)$. It follows that simply setting $\text{cov}(\zeta_1, \zeta_2) = 0$ as a special case of M_1 does *not* yield M_0, since $\text{cov}(y_1, y_2) = -\text{var}(\zeta_1) \neq 0$. The special case of M_1 that yields M_0 is given by the added constraint $\text{cov}(\zeta_1, \zeta_2) = \text{var}(\zeta_1)$, since then $\text{cov}(y_1, y_2) = 0$. Finally, note that M_0 is a test of *modified* independence, not complete mutual independence among all variables, since the covariances among the *x*s are still fixed at known nonzero values as they are under M_2 and M_1. (Note again that we use variance–covariance labels simply as a shorthand for deviation sums of squares and sums of cross-products in this example.)

The model tests and model comparisons associated with this example are presented in Table 2, using the same format as Table 1. The analysis verifies that although M_2 provides an acceptable representation of the data ($p > .05$), this goodness-of-fit test is not unduly influenced by the small sample size. The parameters of M_2 obviously improve goodness of fit over M_0 both statistically as well as by the incremental indices. However, it is apparent that the statistical information is carried by the regression weights and the variance of the residuals ζ_1 and ζ_2 but not by the constraint $\text{cov}(\zeta_1, \zeta_2) = \text{var}(\zeta_1)$. Finally, the $M_0 - M_1$ comparison yields $\hat{\rho}_{01} = -.159$, which shows how this nonnormed fit index can be negative while the normed fit index $\hat{\Delta}_{01} = .021$ remains positive but small.

Conclusion

Covariance structure analysis promises to become an important technique for comparing competing substantive theories on nonexperimental data. Although the statistical basis of the methodology provides a means for using correlational data in a truly model-testing way, there is a danger in current practice of overemphasizing goodness-of-fit tests and the associated probability levels while ignoring or

Table 2
Evaluation of Meat Market Submodels

Model test			Model comparison				
Model	χ^2	*df*	Comparison	χ^2	*df*	$\hat{\rho}_{kl}$	$\hat{\Delta}_{kl}$
M_0	61.38**	7	M_0–M_1	1.30*	1	−.159	.021
M_1	60.08**	6	M_1–M_2	57.40**	5	.943	.935
M_2	2.68*	1	M_0–M_2	58.70**	6	.784	.956

* $p > .05$. ** $p < .001$.

minimizing the practical importance associated with various model comparisons. As Guttman (1977) observed, "A test of statistical significance is not a test of scientific importance" (p. 92). An overemphasis on probability values is particularly dangerous with large-sample data. We have seen research reports that inappropriately were rejected by reviewers of journal articles, as well as researchers who were unnecessarily dejected by their inability to account for every bit of sample variation, in instances in which the proposed models clearly represented a superior understanding of a phenomenon, compared with competing or inadequately specified theories. Although we would not agree with a recommendation to abandon statistical significance testing (Carver, 1978), the incremental fit indices recommended here should provide important adjunct information in evaluating models. At the very least, these indices should help to create an appropriate perspective on statistical significance testing. Such a perspective has been urged on psychologists for many years in more traditional research and statistical contexts (Bakan, 1966; Hays, 1963; Lykken, 1968; Morrison & Henkel, 1970; Rozeboom, 1960).

The incremental fit indices depend critically on the availability of a suitably framed null model. The model of modified independence, that is, mutual independence among variables subject to the known, fixed covariation of variables that is specified in or generated by a substantive model of interest, seems to meet this requirement. The null model also provides a baseline for statistically evaluating increments in goodness of fit. In most moderately sized samples, substantive models can be expected to provide a statistically significant improvement in fit over the null model, so that the statistical consequences of using the null model are primarily limited to providing probabilistic assurance about information gained in contexts in which even the best substantive model falls short of ideal when compared with the saturated model. Of course, this probabilistic emphasis should not replace the pragmatic and scientific emphasis provided by the incremental fit indices.

The statistical value of the null model seems to be more important in small samples in which a nonsignificant difference between substantive and saturated models may simply reflect lack of power rather than a substantive achievement. In such situations the demonstration of a significant increment in fit due to the substantive model over the null model would provide assurance that the data are adequate to the task of model evaluation. Atypical results may also be observed, however. For example, it is possible that a substantive model cannot be statistically differentiated from either the null or the saturated model. In such cases, it would certainly be inappropriate to conclude that the substantive model provides an unambiguous representation of the data. Although the null model should thus prove to have valuable applications in small samples, it must be remembered that statistical theory in covariance structure analysis has been developed primarily for large samples. Further work is needed to develop small-sample theory. Nonetheless, the use of the null model and model-testing and incremental fit evaluation strategies recommended here will also be relevant under such a theory.

The ideal models and pseudo chi-square tests proposed here are likely to prove of special value in situations in which competing models cannot be completely specified and in which model development is part of a research program. The formulation of ideal models and evaluation of the associated pseudo chi-square statistics can help to locate fundamental specification errors in models. These misspecifications represent important areas in which model modification must be undertaken. In particular, the proposed method allows a researcher to evaluate the adequacy of a measurement model in which all possible latent variable regression structures would be embedded. There may be little point to evaluating a given regression structure if the measurement model is totally inadequate. Of course, when a model is modified empirically rather than theoretically, cross-validation or another method for assuring that the statistical theory is not grossly violated becomes essential (Huba et al., Note 3).

Reference Notes

1. Browne, M. W. *Asymptotically distribution free methods for the analysis of covariance structures.*

Paper presented at the meeting of the Psychometric Society, Iowa City, Iowa, May 1980.

2. Bentler, P. M., Bonett, D. G., & Browne, M. W. *Theory and application of an idealized significance test.* Manuscript in preparation, 1980.
3. Huba, G. J., Woodward, J. A., Bentler, P. M., & Wingard, J. A. *Causal modeling in social/personality psychology: Issues in quality control.* Unpublished manuscript, 1978. (Available from G. J. Huba, Department of Psychology, University of California, Los Angeles, California 90024.)

References

Aigner, D. J., & Goldberger, A. S. (Eds.). *Latent variables in socioeconomic models.* Amsterdam, The Netherlands: North-Holland, 1977.

Anderson, T. W. *An introduction to multivariate statistical analysis.* New York: Wiley, 1958.

Anderson, T. W. Asymptotically efficient estimation of covariance matrices with linear structure. *Annals of Statistics*, 1973, *1*, 135–141.

Anderson, T. W., & Rubin, H. Statistical inference in factor analysis. *Proceedings of the 3rd Berkeley Symposium on Mathematics, Statistics, and Probability*, 1956, *5*, 111–150.

Bakan, D. The test of significance in psychological research. *Psychological Bulletin*, 1966, *66*, 423–437.

Bentler, P. M. The interdependence of theory, methodology, and empirical data: Causal modeling as an approach to construct validation. In D. B. Kandel (Ed.), *Longitudinal research on drug use: Empirical findings and methodological issues.* New York: Wiley, 1978.

Bentler, P. M. Multivariate analysis with latent variables: Causal modeling. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 31). Palo Alto, Calif.: Annual Reviews, 1980.

Bentler, P. M. Linear systems with multiple levels and types of latent variables. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction.* Amsterdam: North-Holland, in press.

Bentler, P. M., & Lee, S. Y. A statistical development of three-mode factor analysis. *British Journal of Mathematical and Statistical Psychology*, 1979, *32*, 87–104.

Bentler, P. M., & Weeks, D. G. Interrelations among models for the analysis of moment structures. *Multivariate Behavioral Research*, 1979, *14*, 169–185.

Bentler, P. M., & Weeks, D. G. Linear structural equations with latent variables. *Psychometrika*, 1980, *45*, 289–308.

Bentler, P. M., & Weeks, D. G. Multivariate analysis with latent variables. In P. R. Krishnaiah & L. Kanal (Eds.), *Handbook of statistics* (Vol. 2). Amsterdam: North-Holland, in press.

Bielby, W. T., & Hauser, R. M. Structural equation models. In A. Inkeles, J. Coleman, & N. Smelser (Eds.), *Annual review of sociology* (Vol. 3). Palo Alto, Calif.: Annual Reviews, 1977.

Boomsa, A. The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction.* Amsterdam: North-Holland, in press.

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

Carver, R. P. The case against statistical significance testing. *Harvard Educational Review*, 1978, *48*, 378–399.

Geweke, J. F., & Singleton, K. J. Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association*, 1980, *75*, 133–137.

Guttman, L. What is not what in statistics. *Statistician*, 1977, *26*, 81–107.

Hays, W. L. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

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

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

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

Jöreskog, K. G., & Goldberger, A. S. Factor analysis by generalized least squares. *Psychometrika*, 1972, *37*, 243–260.

Jöreskog, K. G., & Sörbom, D. *LISREL IV user's guide.* Chicago: National Educational Resources, 1978.

Jöreskog, K. G., & Sörbom, D. *Advances in factor analysis and structural equation models.* Cambridge, Mass.: Abt, 1979.

Jöreskog, K. G., & Wold, H. (Eds.). *Systems under indirect observation: Causality, structure, prediction.* Amsterdam: North-Holland, in press.

Kenny, D. A. *Correlation and causality.* New York: Wiley, 1979.

Lancaster, H. O. *The chi-squared distribution.* New York: Wiley, 1969.

Larzelere, R. E., & Mulaik, S. A. Single-sample tests for many correlations. *Psychological Bulletin*, 1977, *84*, 557–569.

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

Lawley, D. N., & Maxwell, A. E. *Factor analysis as a statistical method.* New York: American Elsevier, 1971.

Lee, S. Y. Some algorithms for covariance structure analysis (Doctoral dissertation, University of California, Los Angeles, 1977). *Dissertation Abstracts International*, 1977, *38*, 480B–481B. (University Microfilms No. 77-17,230)

Lee, S. Y., & Bentler, P. M. Some asymptotic properties of constrained generalized least squares estimation in covariance structure models. *South African Statistical Journal*, in press.

Lykken, D. T. Statistical significance in psychological research. *Psychological Bulletin*, 1968, *70*, 151–159.

McGaw, B., & Jöreskog, K. G. Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. *British Journal of Mathematical and Statistical Psychology*, 1971, *24*, 154–168.

Morrison, D. E., & Henkel, R. E. *The significance test controversy—A reader*. Chicago: Aldine, 1970.

Neyman, J., & Pearson, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 1928, *20A*, 175–240.

Olsson, U. On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research*, 1979, *14*, 485–500.

Olsson, U., & Bergman, L. R. A longitudinal factor model for studying change in ability structure. *Multivariate Behavioral Research*, 1977, *12*, 221–241.

Rozeboom, W. W. The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 1960, *57*, 416–428.

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

Wheaton, B., Muthén, B., Alwin, D. F., & Summers, G. F. Assessing reliability and stability in panel models. In D. R. Heise (Ed.), *Sociological methodology, 1977*. San Francisco: Jossey-Bass, 1977.

Wiley, D. E. The identification problem for structural equation models with unmeasured variables. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences*. New York: Academic Press, 1973.

Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 1938, *9*, 60–62.

Received September 12, 1979 ■