5

STATISTICAL POWER IN COVARIANCE STRUCTURE MODELS

Ross L. Matsueda

UNIVERSITY OF WISCONSIN, MADISON

William T. Bielby

UNIVERSITY OF CALIFORNIA, SANTA BARBARA

During the preparation of this chapter, Ross L. Matsueda received support from the National Institute of Justice, U.S. Department of Justice (82-IJ-CX-0060), and the Graduate School of the University of Wisconsin–Madison. William T. Bielby was supported by a grant from the National Science Foundation to the Center for Advanced Study in the Behavioral Sciences (NSF-BNS-76-22443). Opinions expressed in this chapter are those of the authors and do not necessarily represent the official position or policies of the funding agencies. An earlier version of this chapter was presented at the August 1984 meetings of the American Sociological Association, San Antonio. The authors are grateful to Michael E. Sobel and several anonymous reviewers for helpful comments on that version.

Advances in covariance structure analysis have made structural equation modeling an important methodological tool in the social sciences. These models attempt to specify, estimate, and test causal relationships underlying observable variables. (See Goldberger 1972; Bielby and Hauser 1977; Jöreskog and Sörbom 1982; Bentler 1980, for overviews.) Typically the parameters of such models are estimated by the method of maximum likelihood, and hypotheses are evaluated with the likelihood-ratio χ^2 test. As many researchers have recently noted, however, hypothesis testing and model fitting using the likelihood ratio test can be obscured for two reasons. First, the test statistic may be sensitive to violations of the following assumptions: (1) that the observable variables are distributed multivariate normal; (2) that the observations are independent; and (3) that the sample size is large enough to capitalize on asymptotic properties. Second, hypothesis testing can be obscured because the test statistic is sensitive to the size of the sample. This problem is readily apparent when evaluating the overall goodness of fit of a model's overidentifying restrictions. In large samples, even models with relatively trivial departures from hypothesized restrictions are likely to be rejected. Conversely, in small samples, even models with large departures are likely to be accepted.

Several strategies for addressing these two issues have been proposed. Jöreskog (1979) and others suggest comparing the χ^2 statistic to degrees of freedom; Jöreskog and Sörbom (1984) propose several goodness-of-fit indexes that measure the relative amount of covariation jointly explained by the model; and Bentler and Bonett (1980) follow the lead of Tucker and Lewis (1973) in proposing fit indexes based on comparisons to baseline models. These strategies may be useful for evaluating hypotheses when departures from the assumptions of normality, independence, and large samples are so extreme that they seriously threaten conventional test results.[1] On the other hand, each of these proposals represents an ad hoc strategy for offsetting the influence of sample size on the likelihood ratio statistic. This problem is more appropriately conceived as an issue of

[1] When these assumptions are violated, however, the optimal properties of maximum-likelihood estimation (best asymptotic normal) are suspect. Thus any use of the LISREL approach to estimate or test parameters is drawn into question. Little is known of the robustness of either maximum-likelihood estimation or likelihood ratio tests to violations of these assumptions. For preliminary work, see Beardon, Sharma, and Teel (1982), Boomsma (1982, 1983), Fuller and Hemmerle (1966), Geweke and Singleton (1980), Lawley and Swanson (1954), and Olsson (1979). Furthermore, research on more robust procedures has yet to provide definitive results (see Browne, 1984).

statistical power and more directly addressed with procedures for protecting against type II error (the error of retaining a false null hypothesis). Furthermore, the issue of power is relevant to all conventional statistical tests of parameters in structural equation modeling. This implies that the power of tests of both overall goodness of fit and specific parameters can be treated within a single unified framework.

In this chapter, we draw upon principles of classical inference to examine statistical power for covariance structure models. Focusing on maximum-likelihood estimation and likelihood ratio tests of Jöreskog's LISREL approach (Jöreskog and Sörbom 1982, 1984), we proceed in five steps. First, we review estimation and testing in covariance structure models. Second, we review ad hoc strategies for dealing with statistical power in covariance structure models. Third, we present an approximate power function of the likelihood ratio test and show how power can be calculated routinely. Fourth, we discuss the implications of power calculations for testing hypotheses in structural equation models. While our recommendations are sometimes difficult to implement in complex models, they do provide a formal conceptualization of the problem and distinguish between formal hypothesis testing and informal model fitting. Fifth, we investigate the influence of certain parametric structures on the power of a test and discuss ways of increasing power.

COVARIANCE STRUCTURE MODELS

The LISREL Model. In Jöreskog's LISREL model for the analysis of covariance structures, random vectors of n latent independent variables, $\boldsymbol{\xi}' = (\xi_1, \xi_2, \ldots, \xi_n)$, and m latent dependent variables, $\boldsymbol{\eta}' = (\eta_1, \eta_2, \ldots, \eta_m)$, are linearly related by the following system of equations:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{1}$$

where $\mathbf{B}(m \times m)$ and $\boldsymbol{\Gamma}(m \times n)$ are matrices of coefficients, $\boldsymbol{\zeta}(m \times 1)$ is a vector of random disturbances, elements of $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, and $\boldsymbol{\xi}\boldsymbol{\zeta}'$ have zero expectations, and $\mathbf{I} - \mathbf{B}$ is nonsingular. The vectors of observed variables, $\mathbf{x}' = (x_1, x_2, \ldots, x_q)$ and $\mathbf{y}' = (y_1, y_2, \ldots, y_p)$, are related to latent variables by the following measurement equations:

$$\mathbf{y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{2}$$

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

where $\boldsymbol{\Lambda}_y(p \times m)$ and $\boldsymbol{\Lambda}_x(q \times n)$ are matrices of coefficients and $\boldsymbol{\varepsilon}(p \times 1)$ and $\boldsymbol{\delta}(q \times 1)$ are vectors of random measurement errors. Elements of $\boldsymbol{\varepsilon}$

and $\boldsymbol{\delta}$ have zero expectations and are uncorrelated both with one another and with elements of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. To simplify matters, the observable variables are expressed as deviations from their means, so that $E(\mathbf{x}) = \mathbf{0}$ and $E(\mathbf{y}) = \mathbf{0}$. The covariance matrices (with dimensions) for $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{\delta}$ are respectively: $\boldsymbol{\Phi}(n \times n) = E(\boldsymbol{\xi}\boldsymbol{\xi}')$, $\boldsymbol{\Psi}(m \times m) = E(\boldsymbol{\zeta}\boldsymbol{\zeta}')$, $\boldsymbol{\theta}_\varepsilon(p \times p) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$, and $\boldsymbol{\theta}_\delta(q \times q) = E(\boldsymbol{\delta}\boldsymbol{\delta}')$, where $\boldsymbol{\psi}$ and $\boldsymbol{\Phi}$ are positive definite. Then Equations (1) through (3) can be manipulated so that the $(p+q) \times (p+q)$ nonsingular covariance matrix $\boldsymbol{\Sigma}$ for the vector of observables $\mathbf{z} = (\mathbf{y}', \mathbf{x}')'$ is expressed as a function of the following structural parameters (Jöreskog and Sörbom 1984, I.8):

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})(\mathbf{I} - \mathbf{B}')^{-1}\boldsymbol{\Lambda}_y' + \boldsymbol{\theta}_\varepsilon & \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Lambda}_x' \\ \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I} - \mathbf{B}')^{-1}\boldsymbol{\Lambda}_y' & \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}_x' + \boldsymbol{\theta}_\delta \end{bmatrix} \tag{4}$$

Maximum-Likelihood Estimation. If the joint distribution of observed variables is multivariate normal, then maximizing the likelihood function is equivalent to choosing values of unconstrained parameters that minimize

$$F = \log|\boldsymbol{\Sigma}| - \log|\mathbf{S}| + \operatorname{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - (p+q) \tag{5}$$

where $\mathbf{S}$ is the sample covariance matrix for $\mathbf{z} = (\mathbf{x}', \mathbf{y}')'$. The maximum-likelihood estimator $\hat{\boldsymbol{\theta}}$ for the vector of t free parameters $\boldsymbol{\theta}$ is minimum-variance asymptotic normal. The $t \times t$ asymptotic covariance matrix of parameter estimates $\mathbf{V}$ is a function of the inverse of Fisher's information matrix:

$$\mathbf{V} = \left(\frac{2}{N}\right)\left[E\left(\frac{\partial^2 F}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right)\right]^{-1} \tag{6}$$

where N is the sample size. The asymptotic standard errors are the square roots of the diagonal elements of $\mathbf{V}$.[2]

The Likelihood Ratio Test. Specific hypotheses in nested models can be tested using Neyman–Pearson's likelihood ratio method. Let $\boldsymbol{\theta}_t = (\boldsymbol{\theta}_r', \boldsymbol{\theta}_s')'$ be the partitioned vector of $r + s = t$ parameters, where elements in $\boldsymbol{\theta}_t$ correspond to parameters in the matrices on the right-hand side of

[2] Note that $\mathbf{V}$ refers to the population value of the variance-covariance matrix of the maximum-likelihood estimator. When applying the estimation procedure to data in practice, one typically uses the Fletcher–Powell iteration method (or some variant) to obtain a maximum-likelihood estimate both of the parameters and of $\hat{\mathbf{V}}$, which is an estimate of $\mathbf{V}$. (See Gruvaeus and Jöreskog 1970.)

Equation (4). The structure corresponding to the null hypothesis, H_0: $\boldsymbol{\theta}_r = \boldsymbol{\theta}_{r0}$, is nested within a less restrictive structure corresponding to the alternative hypothesis, H_1: $\boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r0}$. Given nested parameterizations, the structure representing the null hypothesis can be tested against the structure representing the alternative hypothesis. Let F_H denote the minimized value of Equation (5) under the null hypothesis, and let F_A denote the value under the less restricted alternative. Then -2 times the log-likelihood ratio is

$$\nu = N\left[F_H - F_A\right] \tag{7}$$

which is asymptotically distributed χ^2 with r degrees of freedom under the null hypothesis.

We can show that Equation (7) assesses how well the null hypothesis fits the data. If the null hypothesis is true—that is, if $\boldsymbol{\theta}_r = \boldsymbol{\theta}_{r0}$ in the population—then the r corresponding overidentifying restrictions on the population covariance matrix $\boldsymbol{\Sigma}$ must also be satisfied (Jöreskog 1973, 1977). Let $\hat{\boldsymbol{\Sigma}}_H$ denote the covariance matrix implied by estimates for the model corresponding to the null hypothesis, where $\boldsymbol{\theta}_r = \boldsymbol{\theta}_{r0}$ is assumed. It estimates $\boldsymbol{\Sigma}$ subject to the r constraints. Furthermore, let $\hat{\boldsymbol{\Sigma}}_A$ denote the covariance matrix implied by estimates for the less restrictive model corresponding to the alternative hypothesis, where $\boldsymbol{\theta}_r$ is assumed to be unconstrained ($\boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r0}$). It also estimates $\boldsymbol{\Sigma}$, but the elements of $\hat{\boldsymbol{\Sigma}}_A$ need not satisfy the r restrictions. It follows that the test statistic in Equation (7) can be reformulated as

$$\nu = N\left[\log|\hat{\boldsymbol{\Sigma}}_H|/|\hat{\boldsymbol{\Sigma}}_A|\right] + \mathrm{tr}\left[\mathbf{S}\left(\hat{\boldsymbol{\Sigma}}_H^{-1} - \hat{\boldsymbol{\Sigma}}_A^{-1}\right)\right] \tag{8}$$

The nonnegative expression in Equation (8) is zero—denoting a perfect fit of the null hypothesis—only when $\hat{\boldsymbol{\Sigma}}_H$ equals $\hat{\boldsymbol{\Sigma}}_A$. Thus for a given sample size N, the likelihood ratio test statistic is proportional to a scalar criterion for assessing the degree to which matrix $\hat{\boldsymbol{\Sigma}}_H$ departs from $\hat{\boldsymbol{\Sigma}}_A$. We can state this another way. Rewrite Equation (8) as $\nu = Nd(\hat{\boldsymbol{\Sigma}}_H, \hat{\boldsymbol{\Sigma}}_A)$, where d denotes a "discrepancy function" for the two implied covariance matrices (Browne 1984, p. 7). Note that for a given N, larger values of the statistic correspond to larger discrepancies between $\hat{\boldsymbol{\Sigma}}_H$ and $\hat{\boldsymbol{\Sigma}}_A$. Conversely, for a given discrepancy, larger values of N correspond to larger values of the χ^2 statistic. Classic inferential procedures for computing the probability of type I and type II error link measures of fit based on sample data to the issue of substantive interest: the degree to which $\boldsymbol{\theta}_r$ departs from $\boldsymbol{\theta}_{r0}$ in the population. In contrast, most alternative approaches to fit focus only on the discrepancy between $\hat{\boldsymbol{\Sigma}}_H$ and $\hat{\boldsymbol{\Sigma}}_A$.

The Overall Goodness-of-Fit Statistic. The overall goodness-of-fit statistic is a special case of the likelihood ratio test statistic. Any overidentified model implies restrictions on the population covariance matrix, which can be assessed by testing the model incorporating the null hypothesis against a just-identified alternative. In this case, $\boldsymbol{\theta}_t = (\boldsymbol{\theta}_r', \boldsymbol{\theta}_s')'$ has $t = (1/2)(p+q)(p+q+1)$ elements; $\boldsymbol{\theta}_s$ refers to the s unconstrained parameters of the hypothesized model; and $\boldsymbol{\theta}_r$ refers to the r parameters that differentiate the hypothesized model from the just-identified alternative model. The null hypothesis, then, is $\boldsymbol{\theta}_r = \mathbf{0}$. A just-identified model reproduces the sample moments exactly, so that $\hat{\boldsymbol{\Sigma}}_A = \mathbf{S}$. Therefore, for the overall goodness-of-fit test, Equation (8) reduces to the following:

$$\nu = N\left[\log\left(|\hat{\boldsymbol{\Sigma}}_H|/|\mathbf{S}|\right) + \operatorname{tr}\left(\mathbf{S}\hat{\boldsymbol{\Sigma}}_H^{-1}\right) - (p+q)\right] \tag{9}$$

This statistic assesses the goodness of fit of the hypothesized model by indexing the discrepancy between the sample covariance matrix $\mathbf{S}$ and the covariance matrix implied by the hypothesized model $\hat{\boldsymbol{\Sigma}}_H$.

In practice, researchers rarely articulate $\boldsymbol{\theta}_r$—the parameters that differentiate the hypothesized model from the just-identified alternative—especially in complex models when r is large. Indeed, in some instances more than one parameterization of $\boldsymbol{\theta}_r$ is substantively plausible. Nevertheless, the test statistic evaluates the null hypothesis that a specific (but possibly unarticulated) set of r restrictions on the observable moments is satisfied in the population covariance matrix. As in the general case, the test statistic, which indexes departures from the overidentifying restrictions *in the sample*, depends not only on corresponding discrepancies in the population but also on sampling variability. Again classic methods of statistical inference allow one to disentangle these influences.

MODEL FITTING AND HYPOTHESIS TESTING: ALTERNATIVE STRATEGIES

Null Models, Incremental Fit, and Hypothesis Testing. Because a given discrepancy between $\hat{\boldsymbol{\Sigma}}_H$ and $\hat{\boldsymbol{\Sigma}}_A$ yields different values of the likelihood ratio test statistic in samples of different sizes, several researchers have proposed alternative indexes of fit. Jöreskog (1979), for example, suggests that when analyses are in part exploratory, the χ^2 statistic ν should be compared to degrees of freedom r. When ν/r is large, residuals, normalized residuals, and modification indexes are inspected and restrictions in the model are relaxed; if this respecification results in a large drop in ν relative to the number of parameters added, the improvement is presumed real and not due to chance. In effect, this strategy relaxes the formal use of significance tests. Moreover, strictly speaking, it does not offset the influence of sample size, since the ratio of ν to r is just as dependent on sample size as ν alone (Hoelter 1983, p. 330).

Jöreskog and Sörbom (1984) propose a more direct way of offsetting the influence of sample size on criteria of fit using the adjusted goodness-of-fit index (AGFI):

$$\mathrm{AGFI} = 1 - [(p+q)(p+q+1)/2r]\left[-\mathrm{tr}(\hat{\Sigma}_H^{-1}\mathbf{S} - \mathbf{I})^2/\mathrm{tr}(\hat{\Sigma}_H^{-1}\mathbf{S})^2\right] \tag{10}$$

The index is not an explicit function of sample size; like the fitting function of Equation (5), it measures fit in an intuitive way. The distribution of this statistic is unknown, however. Consequently, the index is more useful in exploratory studies, where one is attempting informally to fit various models to data, and less useful in confirmatory studies where one is attempting formally to test a priori hypotheses. For the latter, a formal procedure is needed.

Bentler and Bonett (1980) present a formal strategy for evaluating the overall fit of a hypothesized model and for testing hypotheses about specific parameters. To eliminate the effect of sample size, they propose that comparisons of nested models be assessed relative to the fit of "the most restrictive, theoretically defensible model"—a baseline model (p. 600).[3] Their "incremental fit index" is

$$\Delta_{HA} = \frac{F_H - F_A}{F_B} \tag{11}$$

where F_A, F_H, and F_B are the minimized values of the fitting function (Equation 5 in the maximum-likelihood context) for models corresponding to the alternative hypothesis, the null hypothesis, and the baseline model, respectively. Their statistic depends on the discrepancies between estimated covariance matrices implied by the hypothesized and alternative models, but it is not dependent on sample size. Moreover since $0 \leq (F_H - F_A) \leq F_B$, the incremental fit index Δ_{HA} is bounded by 0 and 1. The index assesses the discrepancy between $\hat{\Sigma}_H$ and $\hat{\Sigma}_A$ as a proportion of the discrepancy

[3] We use the term *baseline* rather than null for two reasons. First, the null model usually refers to a model in which observable variables are mutually independent, whereas the baseline model refers to the general case of the most restrictive model that remains substantively justified (Sobel and Bohrnstedt 1985). Second, we wish to avoid confusion between the baseline model of mutual independent observables and the hypothesized model corresponding to the null hypothesis.

between $\hat{\Sigma}_B$ and **S**, the observable sample moment matrix. The *overall* fit of the hypothesized model would be assessed relative to the baseline model by the statistic $\Delta_{BH} = (F_B - F_H)/F_B = 1 - (F_H/F_B)$, whereas the conventional χ^2 test for the parameters differentiating the two models is $\nu = N(F_B/F_H)$. Similarly, the overall test for the less restrictive alternative relative to the baseline is $\Delta_{BA} = 1 - (F_A/F_B)$, where $\Delta_{BA} = \Delta_{BH} + \Delta_{HA}$. Thus in going from the most restrictive baseline model to the least restrictive alternative, any improvement in fit can be attributed to two additive components: one component Δ_{BH} due to parameters differentiating the hypothesized model from the baseline model and an incremental component Δ_{HA} due to parameters differentiating the hypothesized model from the alternative model. According to Bentler and Bonett (1980, p. 591), if Δ_{BH} is large relative to Δ_{BA}, then $\hat{\Sigma}_H$ (the covariance matrix implied by the estimated hypothesized model) may "contain virtually all the information that one may be concerned with in practical circumstances," even if a conventional statistical test rejects the restrictions implied by the hypothesized model.

In proposing an index of fit independent of sample size, Bentler and Bonett have also effectively changed the focus of comparison: The hypothesized model is now compared to a more restrictive baseline model rather than to a less restrictive alternative. Thus they are claiming that correct inferences can be made about a subset of parameters of the hypothesized model—even if that model is specified incorrectly. Their procedure, however, can yield misleading results (see Sobel and Bohrnstedt 1985).

Testing a hypothesized model against a baseline model can be illuminating, but it is not a *substitute* for the test of the hypothesized model's overidentifying restrictions. As Bentler and Bonett note (1980, p. 595), the tests correspond to different hypotheses, and the decision to perform one or both tests should be based on substantive considerations. But depending on the structure of the data, the test against the baseline model can yield misleading results when the test of the hypothesized model's overidentifying restrictions is ignored.

We can illustrate this point with a simple example. Consider the following competing specifications for a simple regression model:

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{12a}$$

$$y = \beta_1 x_1 \qquad\quad + \varepsilon \tag{12b}$$

$$y = \qquad\qquad\quad + \varepsilon \tag{12c}$$

where Equation (12a) corresponds to the alternative hypothesis, Equation (12b) corresponds to the null hypothesis, and Equation (12c) corresponds to the baseline model. Assume that in the true specification (12a) β_2 is very large, β_1 is very small, and x_1 and x_2 are highly correlated. Assume further that we draw a very large sample so that the estimated standard errors are small. Because of omitted variable bias, the estimate for β_1 can be very large under the hypothesized model (12b); consequently, that model will reproduce the observed moments much more closely than will the baseline model (12c). In going from the hypothesized model (12b) to the alternative model (12c), the *incremental* improvement in fit might be very small; but if we interpret this as supporting the hypothesized model (as Bentler and Bonett suggest), we would erroneously reject the true structure (12a) underlying the observable moments.

Suppose, for example, the sample moments are $s_{x_1x_1} = s_{x_2x_2} = s_{yy} = 1.0$, $s_{x_1x_2} = 0.90$, $s_{x_1y} = 0.81$, and $s_{x_2y} = 0.90$. In this case, all three models reproduce the exogenous moments exactly. The baseline model (12c) fails to reproduce s_{x_1y} and s_{x_2y}, instead implying values of zero for both $\hat{\sigma}_{x_1y}$ and $\hat{\sigma}_{x_2y}$. The hypothesized model (12b) reproduces s_{x_1y} exactly and implies a value of 0.729 for $\hat{\sigma}_{x_2y}$. The alternative model (12a) reproduces all sample moments exactly. The incremental fit for the hypothesized model (12b) against the baseline model (12c) is 0.643. Thus, following an incremental-fit procedure, we would conclude that most of the potential for reproducing the 3 × 3 sample covariance matrix is accomplished by going from the baseline model (12c) to the seriously misspecified hypothesized model (12b). Under the hypothesized model, coefficient β_1 is 0.81. Since the hypothesized model reproduces s_{x_2y} much better than the baseline model, parameter β_1 appears to incorporate meaningful information. But under the true alternative specification (12a), the estimate of β_1 is actually zero.[4]

Criteria for assessing fit derived from comparing a hypothesized model to a baseline model are misleading because any nested comparison requires an accurate specification of the less restrictive model. Equations (12a), (12b), and (12c) are nested in parameters and imply nested restric-

[4] The conventional likelihood-ratio χ^2 test would also yield misleading results if applied only to the test of the hypothesized model against the baseline model. Imposing the constraint $\beta_1 = 0$ on the hypothesized model would yield a large increase in ν, leading us to reject the baseline model. The overall χ^2 test statistic for the hypothesized model would also be large, however, leading us to reject the hypothesized model as well. The Bentler and Bonett procedure suggests that we ignore the latter statistic as Δ_{BH} approaches 1.

tions on observable population moments. The alternative model (12a) implies no restrictions; imposing $\beta_2 = 0$ yields the hypothesized model (12b) with a single restriction:

$$\sigma_{yx_1}\sigma_{x_1x_2} = \sigma_{yx_2}\sigma_{x_1x_1} \tag{13}$$

Assuming that this constraint holds, imposing the additional constraint, $\beta_1 = 0$, on the hypothesized model (12b) yields the baseline model (12c) with an additional restriction:

$$\sigma_{yx_1}\sigma_{x_2x_2} = \sigma_{yx_2}\sigma_{x_1x_2} \tag{14}$$

Bentler and Bonett (1980, pp. 596–597) imply that a comparison of the hypothesized model (12b) with the baseline model (12c) necessarily tests whether the second restriction holds in the population. They argue further that such a test can determine when "valuable statistical effects have been localized"—even when the hypothesized model (12b) fails to fit the data. Neither assertion, however, is necessarily true. In our example, comparing the hypothesized and baseline models tests the restriction described by Equation (14) only when Equation (13) holds.[5] We have demonstrated that the incorrectly specified hypothesized model fits much better than the baseline model even though Equation (14) was satisfied *exactly* in the sample.[6]

In summary, the incremental fit procedure suggested by Bentler and Bonett provides useful results only when one can assume that the less restrictive model is not seriously misspecified. In effect, however, this requires assuming away the problem their strategy was intended to solve. In a large sample, the overall fit of the less restrictive model may be poor even when the parameters differentiating it from the correct model are substantively trivial. Conversely, in a small sample, a poorly specified model may still yield an acceptable overall fit. The problem is best addressed by calculating how sensitive test statistics are to both meaningful and trivial departures from hypothesized parameter constraints. Otherwise the indiscriminate use of incremental fit as a modeling strategy (Jöreskog

[5] Unless x_1 and x_2 are perfectly correlated, the two restrictions hold simultaneously only when $\sigma_{yx_1} = \sigma_{yx_2} = 0$.

[6] Of course, when the alternative, hypothesized, and baseline models are as simple as the three in our example, we would compare the hypothesized model to both the less restrictive and the more restrictive specifications. But with a more complex structure (for example, where y, x_1, and x_2 each represent vectors of many variables), the indiscriminant application of incremental fit indexes could easily support seriously misspecified models.

1981, pp. 75–76; Rorer and Widiger 1983, p. 453) may come at the expense of carefully considering plausible, less restrictive alternative models.

Finally, statistical power is no less a consideration when using baseline model test statistics (or adjusted goodness-of-fit indexes) that are independent of sample size. These statistics index the extent to which sample moments satisfy a set of overidentifying restrictions. But when sampling variability is large—due, for example, to a small sample—we should expect large departures from restrictions even when they hold in the population.

Critical Sample Size. As we have noted, the likelihood-ratio χ^2 statistic ν can be written as the product of sample size and a scalar discrepancy function denoting the difference between covariance matrices implied by the null and alternative hypotheses: $\nu = Nd(\hat{\Sigma}_H, \hat{\Sigma}_A)$. The expression shows, for example, that in a sample of 5,000 the discrepancy corresponding to a type I error rate of 0.05 would be on the average 10 times smaller than the discrepancy corresponding to the same error rate in a sample of 500. Suppose we test a null hypothesis with a very large sample, and ν is statistically significant at conventional levels ($p < 0.05$). It might be informative to know whether the discrepancy between $\hat{\Sigma}_H$ and $\hat{\Sigma}_A$ is small enough to retain the null hypothesis with an even smaller sample size. Hoelter (1983) has proposed an index of critical sample size (CN) showing—for a given discrepancy, $d(\hat{\Sigma}_H, \hat{\Sigma}_A)$, and a given level of type I error—the sample size below which the null hypothesis could not be rejected. For the likelihood-ratio χ^2 statistic with r degrees of freedom,

$$\mathrm{CN} = \frac{\chi^2_\alpha(r)}{\nu/N} \tag{15}$$

where $\chi^2_\alpha(r)$ is the critical value of the χ^2 distribution with r degrees of freedom and type I error rate of α. For $r < 100$, the critical value can be determined from statistical tables of the χ^2 distribution (Beyer 1968, pp. 296–298). For larger values of r, the numerator of Equation (15) is approximately $(1/2)\,[z_\alpha + (2r-1)^{1/2}]^2$, where z_α is the critical value for the standard normal distribution.[7]

[7] Equation (15) differs from the index proposed by Hoelter (1983, p. 331):

$$\mathrm{CN} = \frac{\left[Z_\alpha + (2r-1)^{1/2}\right]}{2\nu/(N-1)} + 1$$

If we know, for example, that models retained in samples of 200 or more observations typically have small discrepancies between $\hat{\Sigma}_H$ and $\hat{\Sigma}_A$, then we may safely conclude that so long as CN < 200, departures from the null hypothesis detected with large samples are substantively trivial (Hoelter 1983, p. 331). The index can also suggest when a model has too little statistical power. If we retain a null hypothesis and CN is substantially larger than 200, then there are likely to be substantively large but statistically nonsignificant discrepancies between $\hat{\Sigma}_H$ and $\hat{\Sigma}_A$.

Although the CN index addresses the issue of power more directly than the incremental fit strategy, Hoelter's procedure poses an indirect method of protecting against a desired level of type II error, since the critical sample size is chosen on informal grounds. Moreover, like incremental fit procedures, this strategy avoids explicit consideration of alternative, less restrictive parameterizations.

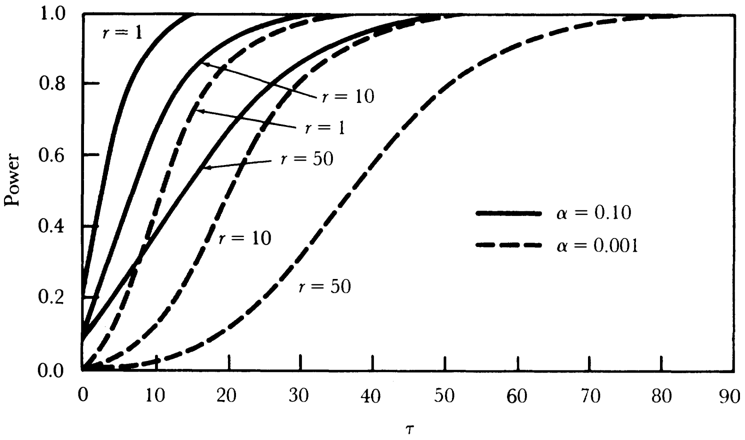
In sum, then, when applied with extreme care the strategies that Jöreskog and Sörbom (1984), Bentler and Bonett (1980), and Hoelter (1983) propose for fitting models can be useful. These strategies, however, are based on indexes with unknown distributional properties. Furthermore, they are all ad hoc methods for offsetting the influence of sample size on the likelihood ratio test statistic. Formally, this influence is more directly addressed by protecting against a desired level of type II error. Thus the problem should be conceptualized as an issue of statistical power. Moreover, classic tenets of statistical inference also address influences on the power of tests other than sample size. Now we shall show how conventional power analyses can provide a more thorough approach for testing structural equation models within the context of multivariate normality.

AN APPROXIMATE NONCENTRAL DISTRIBUTION FOR THE LIKELIHOOD-RATIO χ^2 STATISTIC

Kendall and Stuart (1979, pp. 246–247) show that the likelihood ratio test statistic in Equation (7), testing H_0: $\boldsymbol{\theta}_r = \boldsymbol{\theta}_{r0}$ against H_1: $\boldsymbol{\theta}_r \neq$

Hoelter (1983) uses the normal approximation regardless of r, even though it is biased toward zero when r is small. Moreover, his index is based on a slightly different expression for the χ^2 statistic: $\nu = (N-1)(F_H - F_A)$. Hoelter's index and Equation (15) give nearly identical results for $r \geq 5$, and the difference between N and $N-1$ is inconsequential in samples large enough to justify using statistics with asymptotic properties.

FIGURE 1. Power function of the noncentral χ^2 distribution by type I error rate (α) and degrees of freedom (r).



$\boldsymbol{\theta}_{r0}$, is asymptotically equal to

$$\nu = (\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_{r0})'\mathbf{V}_r^{-1}(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_{r0}) \tag{16}$$

where $\mathbf{V}_r$, an r-dimensional submatrix of $\mathbf{V}$ from Equation (6), is the asymptotic covariance matrix for $\hat{\boldsymbol{\theta}}_r$. The quadratic form in Equation (16) is distributed as the noncentral χ^2 distribution with r degrees of freedom and noncentrality parameter

$$\tau = (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r0})'\mathbf{V}_r^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r0}) \tag{17}$$

This result allows us to calculate the power of the likelihood ratio test in four steps.[8] First, the parametric structure representing the null hypothesis is specified. Second, the structure of the alternative hypothesis is specified. This step, ignored in both conventional testing of covariance structure models as well as in the recommendations reviewed above, is crucial since power cannot be calculated without first specifying the values of the population parameters. The third step is to compute the information matrix of the alternative model. This is accomplished by applying maximum-likelihood estimation to the moments implied by the alternative model and then inverting the r-dimensional submatrix of $\mathbf{V}$ corresponding to $\boldsymbol{\theta}_r$. (Note that we are computing population values of $\mathbf{V}$, not estimating

[8] In a recently published paper, Satorra and Saris (1985) also apply this result to covariance structure models.

V.) Fourth, the noncentrality parameter is calculated by using Equation (17), and power is obtained from power tables for the noncentral χ^2 distribution (see Haynam, Govindarajulu, and Leone 1970).

For a given r and type I error rate α, the power $(1 - \beta)$ to detect θ—that is, the ability to reject the null hypothesis given certain values of the parameters in question—is a monotonically increasing function of the noncentrality parameter τ (see Figure 1). The noncentrality parameter is a quadratic form that weights departures from the null hypothesis in the population $(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r0})$ by the amount of sampling variability $\mathbf{V}_r^{-1}$ for estimates of those departures. Equations (5) and (6) show that τ is proportional to sample size and is also a function of the parametric structure of the population covariance matrix $\boldsymbol{\Sigma}$.

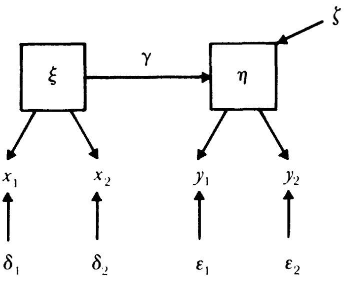
For the test that a single coefficient is zero (H_0: $\theta = 0$), the noncentrality parameter is $\tau = \theta^2/\text{var}(\hat{\theta})$, the ratio of the squared population value of the parameter to the sampling variance of its estimator.[9] For $r = 1$ the estimator $\hat{\theta}$ has an asymptotic normal distribution under the alternative hypothesis (H_1: $\theta \neq 0$), so power can be computed either by referring τ to the noncentral χ^2 tables or by referring $\tau^{1/2}$ to the cumulative normal distribution (Kendall and Stuart, 1979, p. 249).

POWER, SAMPLE SIZE, AND PARAMETRIC STRUCTURE

In multiple-indicator models, relationships among unobservable constructs can be obtained from relationships among observable indicators measured with error. For these models, the power of detecting structural coefficients among unobservables is influenced by measurement parameters that can often be manipulated by the researcher. In most nonexperimental research, covariation among exogenous variables and structural disturbances—parametric influences on power in the general linear model—are properties of the underlying structural process and consequently are beyond the control of the researcher. Power can be manipulated only by controlling the size of the sample and the rate of type I error (Bielby and

[9] Belsley (1982) proposes a signal-to-noise index for testing multicollinearity in the general linear model that is the ratio of the parameter (signal) to its sampling standard error (noise). This ratio is equivalent to the noncentrality parameter of the Wald (1943) test, and the corresponding test statistic is distributed noncentral *F*. Note that the power function of the likelihood ratio test statistic, as we have applied it to covariance structure models, can be used as a diagnostic test for extreme multicollinearity among exogenous latent factors.

FIGURE 2. A multiple-indicator model.



Structural equations:

$$\eta = \gamma\xi + \zeta$$
$$x_1 = \xi + \delta_1$$
$$x_2 = \lambda_{x_2}\xi + \delta_2$$
$$y_1 = \eta + \varepsilon_1$$
$$y_2 = \lambda_{y_2}\eta + \varepsilon_2$$

Parameters:

$\gamma = 1/3$ $\sigma_{\xi\xi} = 9$ $\sigma_{\zeta\zeta} = 12$

$\lambda_{x_2} = 1$ $\lambda_{y_2} = 1$

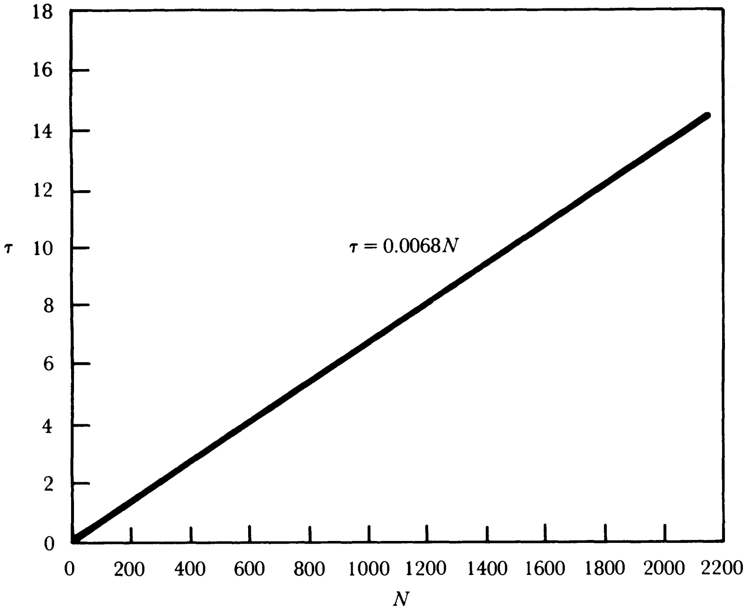
$\sigma_{\delta_1\delta_1} = \sigma_{\delta_2\delta_2} = 27$ $\sigma_{\varepsilon_1\varepsilon_1} = \sigma_{\varepsilon_2\varepsilon_2} = 27$

$N = 1200$

Kluegal 1977). But in multiple-indicator models, power can also be influenced by modifying the structure of the measurement model. In this section, we focus on a two-factor multiple-indicator model (see Figure 2) and analyze power as a function of the size of the sample, the reliability of the indicators, and the number of indicators.

When the population parameters have the values listed in Figure 2 and the sample size is 1,200, the noncentrality parameter for the null hypothesis $\gamma = 0$ has a value of 8.112. Figure 1 shows that the null hypothesis will be rejected with a probability (power) of about 0.90 when the type I error rate (α) is 0.10 and with a probability of approximately 0.35 when $\alpha = 0.001$. Thus, protecting against type I error more conservatively would reduce the likelihood of detecting $\gamma = 0.333$ (a standardized effect of about 0.28) to just over one chance in three. Increasing sample

FIGURE 3.



size, improving measurement reliability, and obtaining additional indicators are alternative ways of increasing the power to detect $\gamma = 0$. We shall examine each in turn.

Sample Size. Figure 3 shows the proportional relationship between τ and sample size, given the parameter values in Figure 2. According to Figure 1, to detect $\gamma = 0.333$ with a probability (power) of 0.90 when the type I error rate is 0.001 requires a value of $\tau = 20.9$ or, according to Figure 3, a sample of about 3,100 cases ($20.9/0.0068 = 3{,}074$). If we relax our level of protection against type II error to $\beta = 0.20$ (power of 0.80), we need only 2,500 cases to detect $\gamma = 0.333$; at $\beta = 0.30$, we need just 2,140 cases.

Reliability. In principle the relationship between τ and the reliability of indicators of ξ and η can be expressed analytically. Specifically, Equation (5) would be twice differentiated with respect to the nine free parameters to obtain $\mathbf{V}^{-1}$ (Jöreskog, 1973, pp. 107–110). The diagonal element corresponding to γ would give the noncentrality parameter as a function of the model's parameters, including $\sigma_{\varepsilon_i\varepsilon_i}$ and $\sigma_{\delta_i\delta_i}$. In practice, such an exercise becomes prohibitively complex. Therefore we determined the relationship through a simulation by computing values of τ for different combinations of $\sigma_{\varepsilon_i\varepsilon_i}$ and $\sigma_{\delta_i\delta_i}$, graphing these values, and inferring patterns from the results.

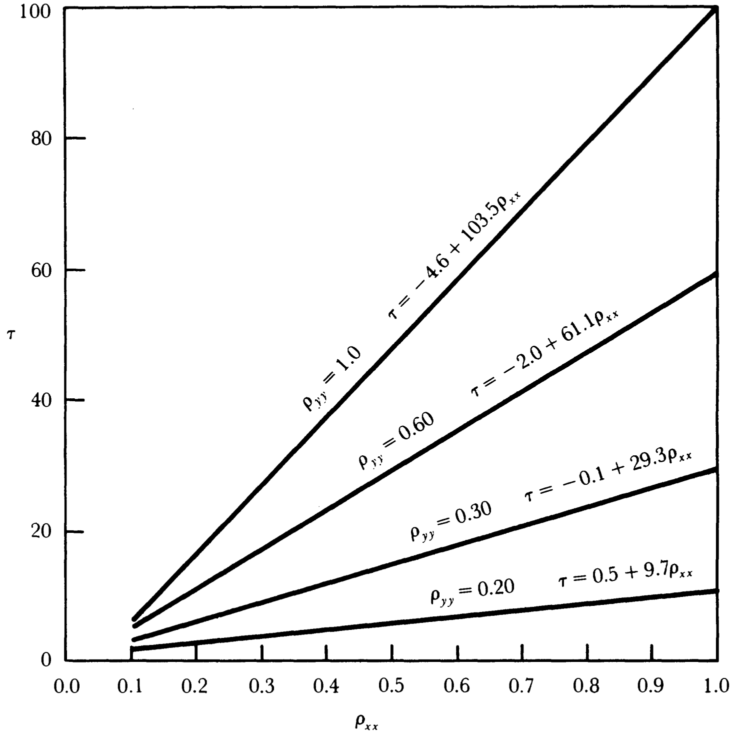
For the model depicted in Figure 2, measurement error variances of 27 correspond to reliabilities of 0.250 and 0.325 for the x and y indicators, respectively, where $\rho_{xx} = 1 - (\sigma_{\delta_i\delta_i}/\sigma_{\xi\xi})$ and $\rho_{yy} = 1 - (\sigma_{\varepsilon_i\varepsilon_i}/\sigma_{\eta\eta})$. We computed τ for 70 combinations of $\sigma_{\delta_i\delta_i}$ and $\sigma_{\varepsilon_i\varepsilon_i}$, with values of each ranging from 0 to 99. The corresponding reliabilities ranged from 0.11 to 1.0 for the x indicators and from 0.12 to 1.0 for the y indicators. We discovered that within that range the relationship between τ and reliability is described almost perfectly ($R^2 = 0.9997$) by the following equation:

$$\tau = 1.8 - 2.5\rho_{xx} - 6.4\rho_{yy} + 106.0\rho_{xx}\rho_{yy} \tag{18}$$

For any given application, the specific values of these coefficients depend on values of the other parameters as well as the size of the sample. Nevertheless, this equation does show that the reliability of the x's affects power contingent on the reliability of the y's and vice versa (Fornell and Larcker 1981).

Figures 4 and 5 illustrate this relationship more clearly. When η is measured perfectly, increasing the reliability of the two measures of ξ by 0.10 increases τ by more than 10. But the slopes in Figure 4 decrease dramatically with ρ_{yy}. Thus when η is measured with reliability 0.30, changes in the reliability of the x's have a negligible impact on power. Figure 5 shows a similar relationship between τ and ρ_{yy} for selected values of ρ_{xx}. In our original example, the noncentrality parameter would have to increase from 8.1 to 20.9 to obtain power of 0.90 with $\alpha = 0.001$. This would require increasing the reliability of the two indicators of the latent exogenous variable from 0.25 to 0.66 (leaving $\rho_{yy} = 0.325$). Similarly, without changing ρ_{xx}, unreliability in the indicators of the latent endogenous variable would have to be all but eliminated (increasing ρ_{yy} from 0.325 to 0.98) in order to have the same effect. In contrast, increasing the reliability of *both* x's and y's to just 0.50 increases τ to 23.8, providing more than enough power (greater than 0.9) to detect $\gamma = 0.333$. That increase in power is comparable to increasing the sample size nearly threefold (from 1,200 to more than 3,500). For this example, if it takes fewer resources to double reliability than it takes to nearly triple sample size, then improving the quality of measurement is more cost-effective than adding cases.

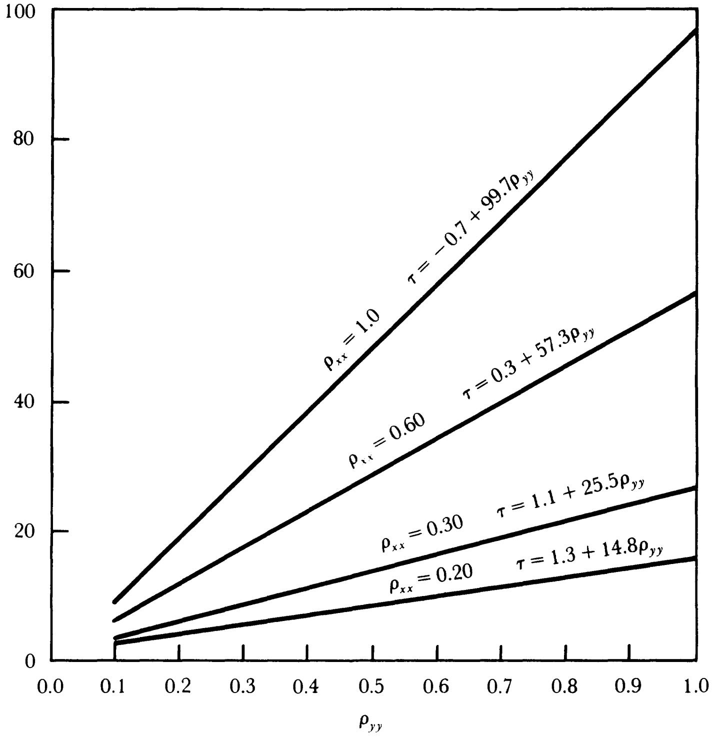
FIGURE 4. Noncentrality parameter as a function of exogenous indicator reliability at selected levels of endogenous indicator reliability.



We have constructed our example so that τ equals 100 when all variables are measured perfectly. Consequently, we can interpret Figures 4 and 5 and Equation (18) as showing the proportionate reduction in effective sample size due to imperfect measurement. For example, measuring η perfectly but reducing the reliability of ξ's indicators from 1.0 to 0.60 is equivalent to reducing the number of cases by 42 percent (from 100 to 58). Furthermore, measuring both η and ξ with a reliability of 0.30 instead of 1.0 reduces the effective sample size by 81 percent. Again, calculations such as these can show when improving the accuracy of measurement is cost-effective relative to increasing the size of the sample. (See Cleary and Linn 1969; Cleary, Linn, and Walster 1970.)

Number of Indicators of the Endogenous Latent Variable. The relationship between power and the number of indicators of the endogenous latent

FIGURE 5. Noncentrality parameter as a function of endogenous indicator reliability at selected levels of exogenous indicator reliability.



variable can be established analytically for a specific class of models. Suppose ξ is measured without error and J indicators of η are obtained so that the structural equations are

$$\eta_i = \gamma\xi_i + \zeta_i \tag{19}$$

$$y_{ij} = \eta_i + \varepsilon_{ij} \tag{20}$$

where $i = 1, \ldots, N$ indexes individuals and $j = 1, \ldots, J$ indexes indicators of the latent endogenous variable. The indicators of η are assumed to have parallel measurement properties, so that $\lambda_{y_j} = 1$ and $\sigma_{\varepsilon_j\varepsilon_j} = \sigma_{\varepsilon\varepsilon}$ for all j.[10]

[10] Following the terminology of Lord and Novick (1968), the indicators are parallel but not "essentially parallel." The latter requires that the indicators' means be constrained to be equal as well. Although this is possible within a LISREL framework (see Sörbom 1981), we have not parameterized the means in the models considered here.

Substituting Equation (19) into (20) gives

$$y_{ij} = \gamma\xi_i + \zeta_i + \varepsilon_{ij} \tag{21}$$

Equation (21) is equivalent to a variance components model for panel data where j indexes indicators instead of time periods. (See Maddala 1971; Judge, Griffiths, Hill, and Lee 1980.) The structural disturbance ζ_i is a random component varying across persons, whereas ε_{ij} is a random component uncorrelated with ζ_i and varying across both persons and indicators. There is no indicator-specific component, since the y_{ij} are deviated from indicator-specific means. Averaging Equation (21) across the j indicators gives

$$\bar{y}_{i.} = \gamma\xi_i + (\zeta_i + \bar{\varepsilon}_{i.}) \tag{22}$$

where the composite disturbance is independently distributed with zero mean and variance $\sigma_{\zeta\zeta} + (1/J)\sigma_{\varepsilon\varepsilon}$. The maximum-likelihood estimate of γ can be obtained by applying ordinary least squares (OLS) to Equation (22).[11] The noncentrality parameter for the hypothesis $\gamma = 0$ is

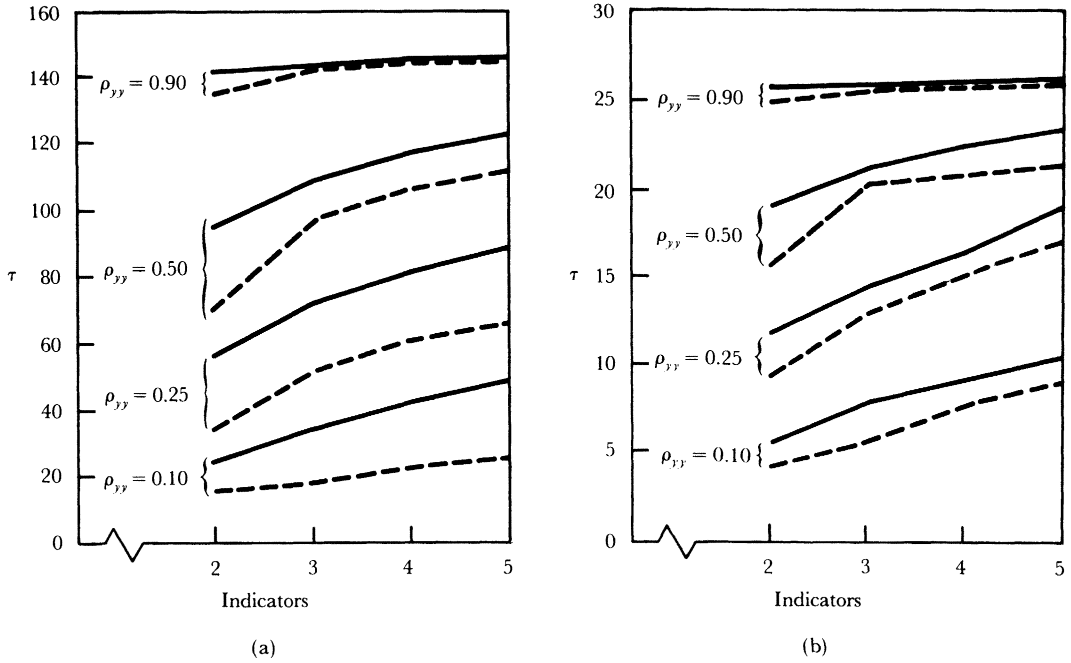
$$\tau = \frac{\gamma^2}{\operatorname{var}(\hat{\gamma})} = \frac{N\hat{\gamma}\sigma_{\xi\xi}}{\sigma_{\zeta\zeta} + (1/J)\sigma_{\varepsilon\varepsilon}} \tag{23}$$

When the structural equation is deterministic, then $\sigma_{\zeta\zeta} = 0$ and the noncentrality parameter is proportional to J. In this case, adding another indicator is equivalent to adding N independent observations. As $\sigma_{\varepsilon\varepsilon}$ approaches zero, however, additional indicators become redundant and do not increase the power to detect γ. Finally, whenever $\sigma_{\zeta\zeta}$ and $\sigma_{\varepsilon\varepsilon}$ are both positive, adding indicators increases τ at a decreasing rate.

We confirmed these results with a simulation that is summarized in Figure 6. The solid lines in Figure 6(a) graph the noncentrality parameter for departures from the null hypothesis $\gamma = 0$ against the number of

[11] In terms of the variance components analogy, the ML estimate is equivalent to the "between" estimator, since for the ith individual, ξ_i does not vary across j. If ξ_i is replaced by a vector of exogenous variables, Equations (19) and (20) constitute a multiple-indicator, multiple-cause (MIMIC) model (Jöreskog and Goldberger 1975). Equation (22) shows that a MIMIC model with parallel measures of the endogenous variable is equivalent to a single-indicator regression model for the additive composite, $\bar{y}_{i.} = (1/J)\Sigma_{j=1}^{J} y_{ij}$. Of course the latter formulation provides neither separate estimates for $\sigma_{\zeta\zeta}$ and $\sigma_{\varepsilon\varepsilon}$ nor a test of the overidentifying restrictions implied by the MIMIC specification. In the context of classic test theory, Equation (22) also shows that doubling the number of parallel indicators is equivalent to improving reliability by doubling the length of tests comprised of parallel items.

FIGURE 6. Noncentrality parameter as a function of number of endogenous indicators at selected levels of indicator reliability. (a) Multiple y's, ξ_1 Measured Without Error (b) Multiple y's, two x's with $\rho_{xx} = 0.25$



indicators and their reliability. Values of $\sigma_{\varepsilon\varepsilon}$ equal to 1, 9, 27, and 90 correspond to reliabilities of 0.9, 0.5, 0.25, and 0.1, respectively, where $\gamma = 1/3$, $N = 1{,}200$, $\sigma_{\xi\xi} = 9$, and $\sigma_{\zeta\zeta} = 8$. At $\rho_{yy} = 0.10$, increasing the number of indicators from two to five doubles τ (from 24.7 to 49.5), thus having an effect equivalent to doubling the sample size. But when reliability is 0.50, going from two to five indicators increases τ by only 28 percent. At $\rho_{yy} = 0.90$, additional indicators are largely redundant; adding three more indicators increases the noncentrality parameter by just 4 percent. Clearly, adding indicators, increasing reliability, and increasing sample size are alternative ways of increasing power. Choosing a strategy to attain a desired protection against type II error should be determined through a cost-benefit analysis of the alternatives. For example, adding indicators can be a cost-effective way to improve the ability to detect departures from $\gamma = 0$ even when reliability is high, providing that the other two strategies are comparatively expensive.[12]

The impact of adding indicators of the endogenous latent variable is somewhat different when they are not in equivalent metrics. Replacing Equation (21) by

$$y_{ij} = \lambda_{y_j}\eta_i + \varepsilon_{ij} \tag{24}$$

precludes averaging across indicators unless $\lambda_{y_1} = \lambda_{y_2} = \cdots = \lambda_{y_j}$; therefore the variance components analogy expressed in Equation (21) breaks down.[13] The dotted lines in Figure 6(a) show τ as a function of sample size when λ_{y_j} is a free parameter to be estimated for $j = 2, \ldots, J$. These simulation results are based on the identical parameter values as our model with parallel measures of η: Each λ_{y_j} equals 1.0 in the population, but they are now free parameters for $j > 1$. Although the results for parallel indicators can be either derived analytically or simulated empirically, we have only empirical results for the case of unconstrained λ_{y_j}.

[12] Evaluating alternatives strictly on the basis of "tau per dollar" is inappropriate if one strategy threatens the viability of the specification more than the others. Adding cases is unlikely to change the specification in most applications, while the risk of misspecification could be greater with the other strategies. If reliability is increased in a way that also introduces undetected error correlations, for example, then estimates of error variances will be biased downward and statistical power will be overstated.

[13] In contrast, when items are tau-equivalent but not parallel—that is, when $\lambda_{y_1} = \lambda_{y_2} = \cdots = \lambda_{y_J}$ and measurement error variances differ across indicators—Equation (21) still holds with λ_y absorbed in the metrics of ξ and ζ. The specification can be expressed as a "pooling" model for NJ observations with heteroscedastic item-specific disturbances.

Comparing solid and dotted lines in Figure 6(a) reveals several patterns. First, freeing the constraints on λ_{y_j} reduces the probability of detecting departures from the hypothesis $\gamma = 0$: The dotted line lies below the solid line at every level of ρ_{yy}. Second, the proportionate reduction in τ due to freeing λ_{y_j} decreases as the reliability of indicators of the endogenous latent variable increases. At $\rho_{yy} = 0.10$, freeing the constraints on λ_{y_j} has roughly the same effect as reducing the sample size by one-half, while at $\rho_{yy} = 0.90$ the reduction is quite small. Third, when the λ_{y_j} are unconstrained, adding a third indicator can have a particularly large impact on power, and the gain appears largest at moderate levels of reliability. At $\rho_{yy} = 0.50$, for example, adding a third indicator increases τ by nearly 40% (from 70.5 to 97.8), while adding fourth and fifth indicators increases the noncentrality parameter by 8 percent and 5 percent, respectively. From a different perspective, the pattern suggests that the potential gain in power from imposing equality constraints on λ_{y_j} is greatest when there are just two indicators.[14]

Figure 6(b) shows what happens when the exogenous variable is also measured by multiple fallible indicators. Here we computed values of the noncentrality parameter for models in which measurement parameters for indicators of the exogenous latent variable were not constrained to be equal.[15] The calculations for Figure 6(b) are comparable to those for Figure 6(a) except that ξ is measured with two indicators, x_1 and x_2, with a reliability of 0.25 ($\sigma_{\delta_1\delta_1} = \sigma_{\delta_2\delta_2} = 27$). The pattern of returns to additional indicators is similar in Figures 6(a) and 6(b), but noncentrality parameters are roughly one-fifth as large in the latter figure. That is, the major impact of unreliability in the x's is to reduce the noncentrality parameter (as implied by Equation 18 for the case of two x's and two y's) regardless of either the number or the reliability of the y's. Closer

[14] It is not uncommon for researchers to estimate a model with unconstrained λ_{y_j}, test for equality of the λ_{y_j}, and then reestimate the model subject to the equality constraint. Because of the sequential estimation procedure, estimates of γ and estimates of sampling variability produced under the constrained model will be biased. Similarly, the noncentrality parameters reported in Figure 6 for models with parallel indicators apply only when constraints on λ_{y_j} are imposed a priori. See Judge, Griffiths, Hill, and Lee (1980, pp. 61–67) for a discussion of this issue in the context of the general linear model.

[15] That is, distinct parameters were specified for λ_{x_1}, λ_{x_2}, $\sigma_{\delta_1\delta_1}$, and $\sigma_{\delta_2\delta_2}$, even though $\lambda_{x_1} = \lambda_{x_2}$ and $\sigma_{\delta_1\delta_1} = \sigma_{\delta_2\delta_2}$ in the population model that generated the data. We chose this strategy because the assumption of parallel measures is seldom imposed a priori in sociological applications of covariance structure models.

examination of the two figures reveals that the returns to additional indicators are somewhat flatter in Figure 6(b); thus the proportionate reduction in τ due to unreliability in the x's increases slightly with the number of indicators.[16] Moreover, the relative impact of unreliability in the x's on the proportionate reduction in τ is greatest when it matters least: when ρ_{yy} is close to 1. For models with λ_{y_j} unconstrained, for example, values of τ in Figure 6(b) are roughly 18% as large as corresponding parameters in Figure 6(a) when $\rho_{yy} = 0.90$, 26% as large when $\rho_{yy} = 0.25$, and about 33% as large when $\rho_{yy} = 0.10$.

Number of Indicators of the Exogenous Latent Variable. Figures 7(a) and 7(b) show how power increases with the number of indicators of the exogenous latent variable for a similar set of models. Solid lines correspond to models with equality constraints on λ_{x_j} and $\sigma_{\delta_j\delta_j}$; dotted lines apply to models where the x's are not assumed to be parallel. In Figure 7(a), the endogenous latent variable is measured by two indicators having perfect reliability; in Figure 7(b), it is measured by two indicators having reliability 0.25. In every instance, the relationship between τ and the number of x's parallels the relationship between τ and the number of y's.[17] For a given value of ρ_{xx}, adding x's increases τ at a decreasing rate (Figure 7a); for different values of ρ_{xx}, the proportionate increase in the noncentrality parameter is greatest when reliability is low (Figure 7b). Power is sacrificed when λ_{x_j} and $\sigma_{\delta_j\delta_j}$ are unconstrained—especially when $j = 2$. (Compare solid and dotted lines in Figure 7.) Finally, the effect of introducing unreliability in the y's parallels what we found previously: Noncentrality parameters are attenuated, but the patterns are largely unaltered.

[16] There is one exception in our example: When endogenous indicators are very unreliable and not assumed to have parallel measurement properties, the proportionate reduction in γ *decreases* with the number of indicators. That is, for unconstrained measurement parameters, the pattern described in the text is nonmonotonic in ρ_{yy}, reversing somewhere between $\rho = 0.25$ and $\rho = 0.10$.

[17] The variance components analogy that produced analytic results for the case of endogenous indicators does not work here. Solving out the unobservable ξ_i yields

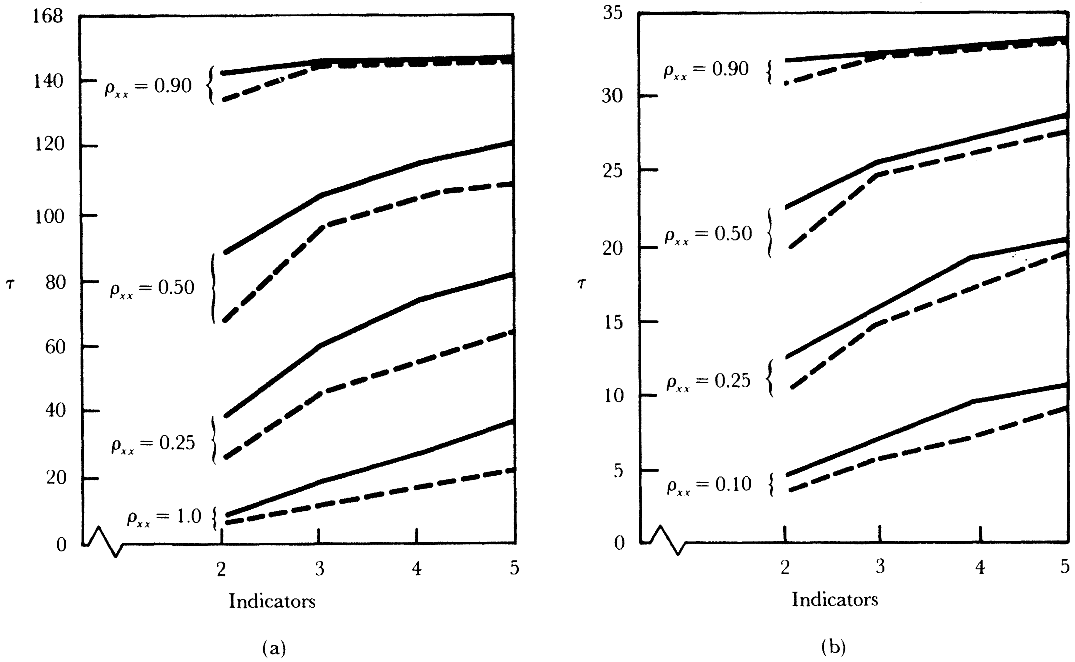
$$\eta_i = \gamma x_{ij} + \zeta_i - \gamma\delta_{ij}$$

and averaging over J exogenous indicators gives

$$\eta_i = \gamma\bar{x}_{i.} + (\zeta_i - \gamma\bar{\delta}_{i.})$$

Unlike Equation (22), the composite disturbance $(\zeta_i - \gamma\bar{\delta}_{i.})$ covaries with the observable exogenous variable, $\bar{x}_{i.}$, since $\text{cov}(x_{ij}, \delta_{ij}) \neq 0$. Further, the composite disturbance variance, $\sigma_{\zeta\zeta} + (\gamma^2/J)\sigma_{\delta\delta}$, is a function of γ. Consequently, the model averaged over indicators cannot be recast as one meeting OLS assumptions.

FIGURE 7. Noncentrality parameter as a function of number of exogenous indicators at selected levels of indicator reliability. (a) Multiple x's, η Measured Without Error (b) Multiple x's, two y's with $\rho_{yy} = 0.25$



Summary: Power and Parametric Structure. We have demonstrated the effects of sample size, measurement properties, and number of indicators on statistical power with a very simple example: a test on a single structural coefficient in a two-construct multiple-indicator model. The overall pattern of findings, consisting of five results, generalizes to models with more than two unobservable constructs. First, the noncentrality parameter is proportional to the size of the sample under any parameterization. Second, the ability to detect structural relationships among unobservables is in general an interactive multiplicative function of reliabilities of all indicators. More precisely, the partial derivative of τ with respect to the reliabilities of one construct's indicators are a function of the reliabilities of other constructs' indicators. Therefore, increasing the reliability of indicators of one construct has a greater effect when other constructs are measured precisely. Third, the ability to detect structural relationships among unobservables is increased by adding indicators, but additional indicators are largely redundant when reliability is high. Fourth, power is increased by assuming equal metrics across indicators (constraining λ_{y_j} to be equal when appropriate), especially when reliability is low. Fifth, the ability to detect elements of $\boldsymbol{\Gamma}$ depends on the moments among elements of $\boldsymbol{\xi}$ in a way that parallels results for the general linear model. For example, the power to detect effects of the kth exogenous unobservable ξ_k increases with the variance of ξ_k and the degree to which ξ_k is independent of other elements of $\boldsymbol{\xi}$.

More important than the generalizations we can make about a limited class of models is the generality of the *procedure* we have followed. The quadratic form of Equation (17) can be computed routinely for hypotheses about any set of parameters in models estimated by maximum-likelihood methods. Thus the concept of statistical power—the ability to detect meaningful effects—is tractable though rarely invoked in applications of covariance structure models.

AN APPLICATION TO THE OVERALL GOODNESS-OF-FIT STATISTIC

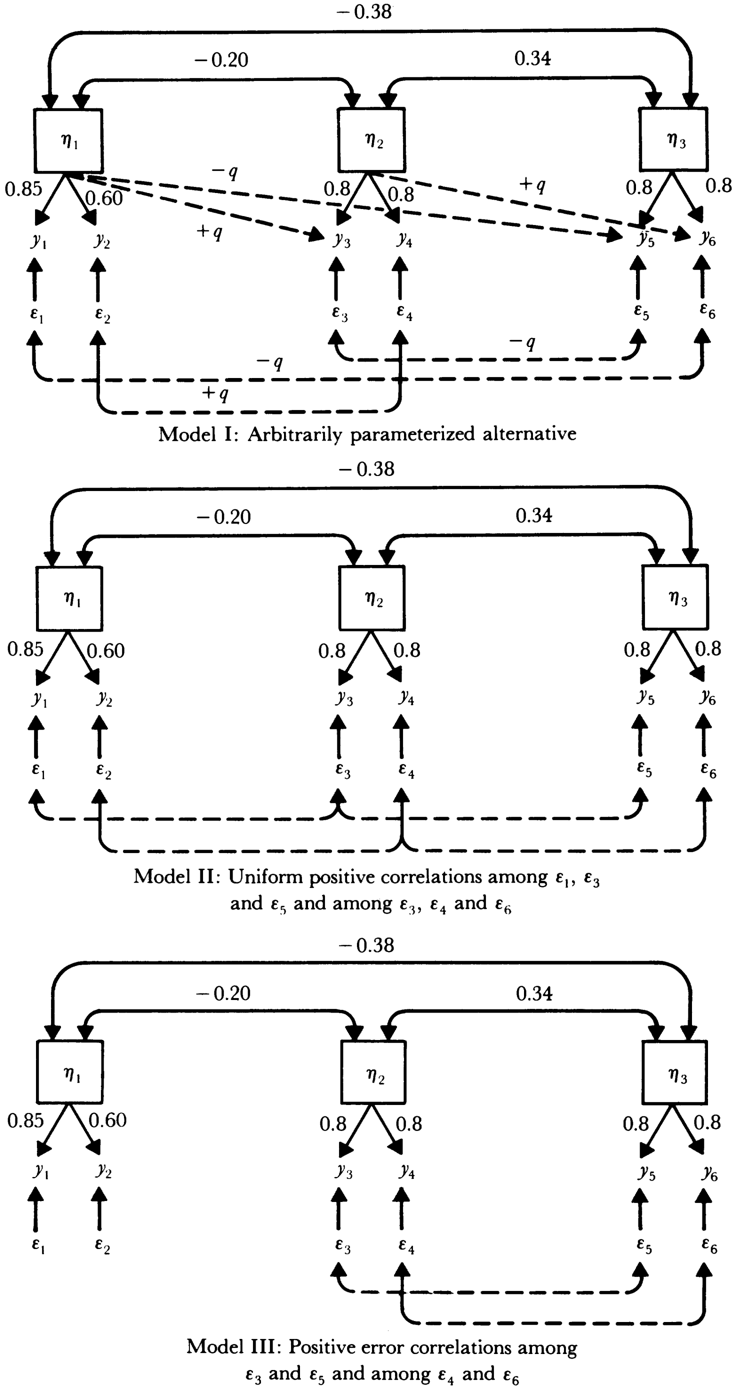
In general, computing the power of statistical tests for covariance structure models requires an explicit specification of parameter values for the less restrictive model corresponding to the alternative hypothesis. While this principle holds equally for computing the power of the overall goodness-of-fit statistic, parameterizing the just-identified alternative model can be very difficult in practice. There are at least three reasons why this problem can become intractable, especially for complex hypothesized models with many overidentifying restrictions. First, there may be more than one just-identified parameterization corresponding to plausible alternative hypotheses. In our computations, we have discovered that Equation (17) applied to different parameterizations produces slightly different results, even when those parameterizations imply identical moments among observable variables. Second, when the number of variables and the number of degrees of freedom are large, computing $\mathbf{V}_r^{-1}$ can be prohibitively expensive. Third, the most plausible alternative parameterization may not be identified from moments among observables. While the moments implied by such a model can be determined, $\mathbf{V}_r^{-1}$ (and therefore τ) cannot be computed.

In this section we present a procedure for calculating power that does not require computing $\mathbf{V}_r^{-1}$ for just-identified alternative models. The procedure can be applied to alternative specifications even when the parameters of the alternative are not identified from moments among observables. We apply this procedure to an example taken from the literature and then use the example to illustrate the implications of power analyses for testing a model's overall goodness of fit.

Our example is derived from Wheaton, Muthén, Alwin, and Summers' (1977) model of the stability of alienation. The hypothesized model has two observable indicators for each of three unobservable constructs. Following Bentler and Bonett (1980) and Sobel and Bohrnstedt (1985), our example is an abbreviated version of Wheaton and colleagues' model, where η_1 is socioeconomic status in year t, η_2 is alienation in year $t+1$, and η_3 is alienation in year $t+5$. Specifically, we are evaluating the power of the test of the null hypothesis that the overidentifying restrictions implied by the hypothesized model hold in the population, given that unconstrained parameters have the values reported in Figure 8. One might, for example, undertake such power computations before replicating the stability of alienation study of Wheaton and colleagues (1977).

Three versions of this model appear in Figure 8. The numbers on the solid lines denote values of standardized coefficients for the hypothesized model and correspond closely to estimates obtained by Wheaton and colleagues. Dashed lines in the three panels of Figure 8 correspond to parameterizations of three different alternatives to the hypothesized model. Model I depicts an arbitrarily parameterized alternative. To obtain this alternative, six parameters presumed to be of no substantive interest were

FIGURE 8. Three parameterizations of departures from a maintained three-construct, six-indicator model with uncorrelated errors.



Model I: Arbitrarily parameterized alternative

Model II: Uniform positive correlations among ε_1, ε_3 and ε_5 and among ε_3, ε_4 and ε_6

Model III: Positive error correlations among ε_3 and ε_5 and among ε_4 and ε_6

added to the hypothesized model and randomly assigned positive and negative signs. Testing the hypothesized model against this alternative allows us to address two issues. On the one hand, we want to discover whether the overall goodness-of-fit statistic is overly sensitive to small and substantively unimportant departures from the hypothesized model. Specifically, if q is the absolute value of standardized coefficients differentiating the two models, we want to avoid a situation in which the power to reject the null hypothesis is large when the magnitude of q is small. On the other hand, we do want to detect departures from restrictions implied by a seriously misspecificed hypothesized model. That is, we want a high probability that the goodness-of-fit test will reject the null hypothesis when q is large. Thus even though the parameters are substantively uninteresting, failure to detect them when they are large in the population can cause biases in estimates of other substantively meaningful parameters.

In model II of Figure 8, we assumed that the parameters differentiating the hypothesized and alternative models are of substantive interest. Specifically, we assume that error terms ε_1, ε_3, and ε_5 are positively correlated, as are ε_2, ε_4, and ε_6. Assuming that each of the six error correlations takes on the same value q, we determine how large q must be before the correlations are likely to be detected by the overall test statistic.

In model III, the models are differentiated by just two parameters: correlations between ε_3 and ε_5 and between ε_4 and ε_6. We compute the ability of the goodness-of-fit statistic to detect departures from restrictions implied by the hypothesized model when four of the six restrictions hold in the population. Again we assume that both error correlations take on the same value q, and we compute power as a monotonically increasing function of q.

Models I and II of Figure 8 are underidentified without equality constraints or other restrictions on the parameters.[18] Consequently, we must modify the procedure we followed when computing the power to detect γ in the model of Figure 2. First, as before, we generate hypothetical population moments among observables implied by specific values of the alternative model's parameters. Second, we use the LISREL program (Jöreskog and Sörbom, 1984) to fit the *hypothesized* model to the popula-

[18] In general, an alternative model need not be identified. Indeed, a case could be made that in most quantitative social science research, moments among observable variables are generated by underidentified models. By posing a plausible underidentified alternative model with specific parameter values, we generate departures from restrictions implied by an overidentified hypothesized model. Power calculations determine the probability of rejecting the restrictions implied by the hypothesized model, given those departures. Obviously we are not computing the probability of "detecting" underidentified parameters.

tion moments generated by the *alternative* model. Fitting the alternative model to obtain $\mathbf{V}_r^{-1}$ (as we did in the previous examples) is impossible, since $\mathbf{V}_r^{-1}$ does not exist for an underidentified model. Therefore, instead of computing τ from Equation (17) we adopt the procedure of Satorra and Saris (1985) to approximate what τ would be under a just-identified alternative.[19] (For alternative models that are identified, we typically find that the two procedures for computing τ yield results within 5 to 10 percent of one another.) We then proceed as before, computing power by referring τ to power tables for the noncentral χ^2 distribution.

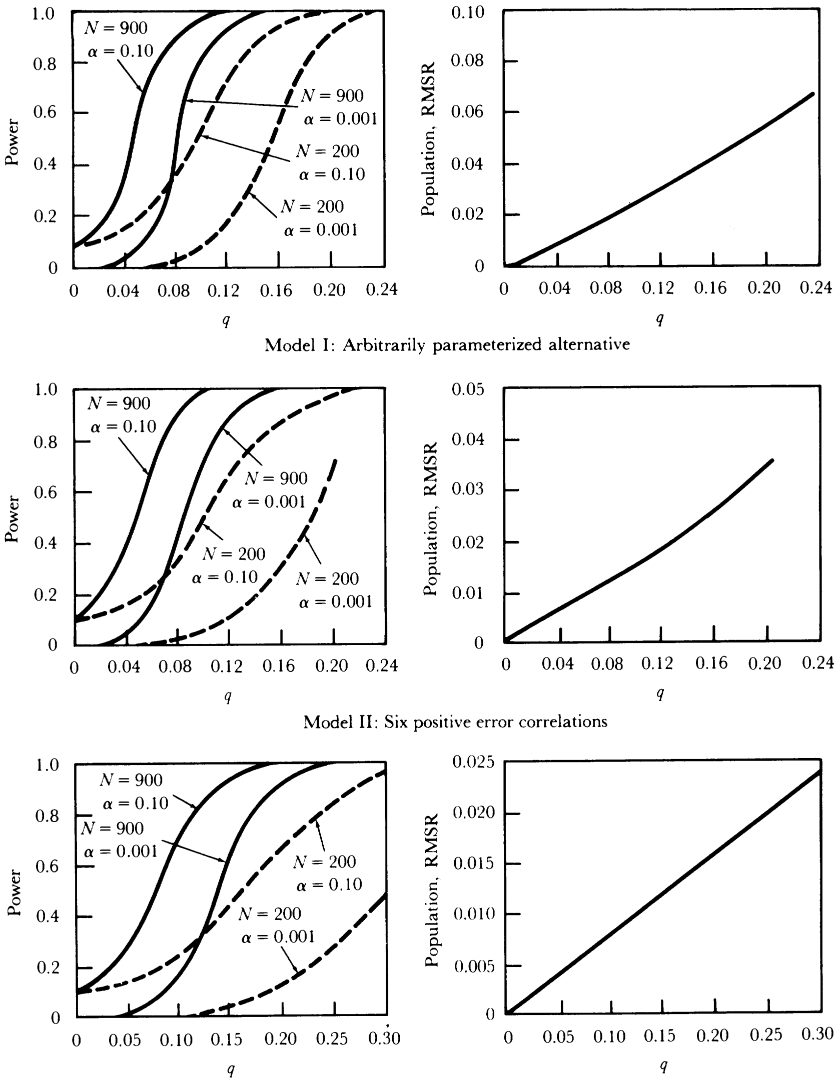
The top panel of Figure 9 presents results for the arbitrarily parameterized alternative. The power of the overall goodness-of-fit statistic is computed as a function of q under four conditions: sample sizes of 900 and 200 and type I error rates (α) of 0.10 and 0.001.[20] The right-hand plots in Figure 9 graph the population root mean square residual (RMSR; Jöreskog and Sörbom 1984, p. 41), a summary measure of departures from restrictions on the population covariance matrix implied by the hypothesized model, against q, the size of the standardized coefficients corresponding to those departures from restrictions. In this example, the typical residual correlation (as indexed by RMSR) is almost directly proportional to q, increasing nonlinearly at a slightly increasing rate.

As depicted in the top panel of Figure 9, power is a monotonically increasing function of the magnitude of q, the size of the sample, and the probability of type I error (α). This graph shows how to obtain a desired amount of power by manipulating the size of the sample and the level of protection against type I error. Suppose, for example, that a value of $q = 0.08$ or smaller (RMSR = 0.02) represents a substantively trivial departure from the hypothesized model, while a value of 0.12 or larger (RMSR = 0.03) represents a meaningful departure. Suppose further that a level of protection against type II error of at least 0.10 (power of 0.90) is desired. One would therefore want a decision rule that rejects the null

[19] See Satorra and Saris (1985) for a derivation of the use of the χ^2 statistic, produced by the LISREL program after estimating the hypothesized model from moments implied by the alternative model, as an approximation of τ, the noncentrality parameter. They also report Monte Carlo results of this approximation for a simple recursive model and find it to be highly accurate even with small sample sizes.

[20] When testing the null hypothesis that a single coefficient θ is zero, the noncentrality parameter is equal to $\theta^2/V(\hat{\theta})$, the squared parameter under the alternative model divided by its sampling variance. We have no algebraic results for the relationship between q and τ when testing a composite hypothesis on two or more parameters, but for the examples reported here the relationship between $\tau^{1/2}$ and q is nearly linear.

FIGURE 9. Power and population root mean square residual correlation (RMSR) as a function of departures (q) from the maintained model under three different parameterizations of alternative models.



Model I: Arbitrarily parameterized alternative

Model II: Six positive error correlations

Model III: Two positive error correlations hypothesis when $q > 0.12$ in the population and retains the null when $q < 0.08$. With a sample of 200, a value of $q = 0.12$ will not be detected at conventional levels of type I error: At $\alpha = 0.001$, only q greater than 0.18 can be detected; at $\alpha = 0.01$, only q greater than 0.14 can be detected. Clearly, given that the sample size is fixed and a power of 0.90 or more is crucial, one must relax the level of type I error above 0.01 to detect $q = 0.12$.

If, however, one can increase the size of the sample to 900, a more stringent level of protection against type I error is possible. In fact, when $\alpha = 0.10$ there is an overabundance of power: The null hypothesis is almost certain to be rejected when $q = 0.12$ (power exceeds 0.99) and also when $q = 0.08$ (power exceeds 0.95). Increasing the protection against type I error to 0.001 would yield an adequate decision rule, since power still exceeds 0.99 when q is 0.12 or larger but is only about 0.50 when q is 0.08 or smaller.

By performing power analyses like these at the design stages of research, the researcher can evaluate alternative steps for increasing power. For example, the alternatives of increasing the size of the sample or lowering the level of protection against type I error can be compared to the possibility of adding further or more reliable indicators.

The middle panel of Figure 9 presents comparable calculations for detecting six positive error correlations. The results are similar to those of the top panel except that power is uniformly lower at given levels of q. Thus it is more difficult to detect departures from the hypothesized model due to the substantively grounded positive error correlations than to detect departures generated by the arbitrarily parameterized alternative.[21] Consequently, it is possible for a researcher to be in the unfortunate situation of simultaneously having too much power to detect arbitrarily and substantively trivial departures from a hypothesized model and not enough power to detect substantively important departures. For example, a statistical decision rule designed to detect positive error correlations of 0.10 may also have a high probability of rejecting the null hypothesis when unim-

[21] This can also be seen by comparing RMSR as a function of q in the top and middle panels of Figure 9. At given levels of q, the model with six error correlations generates smaller departures from the overidentifying restrictions implied by the hypothesized model. Of course, two alternatives that imply the same departures are equally likely to be detected. But in substantively motivated, confirmatory applications we are often interested in detecting specific parameters, not just departures from implied restrictions on population moments.

portant parameters are smaller than 0.10. Confronted with such calculations, the researcher is forced to evaluate whether small departures generated by arbitrary models (like Model I of Figure 9) can be ruled out on a priori grounds. If not, the researcher might be forced to conclude that social theory and research have not developed sufficiently to apply covariance structure models to the deductive research problem at hand.

The bottom panel of Figure 9 shows the power of the goodness-of-fit test for the alternative model with two error correlations.[22] Since four of the six overidentifying restrictions hold in the population, the two error correlations must be relatively large before their probability of detection is high.[23] For example, with a sample of 200 and a type I error rate of 0.10, the two error correlations must exceed 0.25 to obtain power of 0.90. Even with a sample of 900, q must exceed 0.12 to obtain a type II error rate of less than 0.10.

Of course, if one knew beforehand that the only possible departures from the hypothesized model were error correlations between ε_3 and ε_5 and between ε_4 and ε_6, one would evaluate the hypothesized model with a more powerful 2-degree-of-freedom test of the two overidentifying restrictions, rather than the 6-degree-of-freedom overall goodness-of-fit test. But one does not always know whether departures from the hypothesized model are distributed uniformly across all (or most) restrictions, whether they are concentrated in one or two restrictions, or whether they involve substantively important or unimportant parameters. Computations like those reported in Figure 9 are particularly important for complex models with many variables and many overidentifying restrictions because they force the researcher to differentiate important parameters from trivial ones.

Consider, for example, the following situation: Out of dozens of restrictions, only a few involve substantively interesting parameters that might be nontrivial in magnitude; power calculations show that the overall goodness-of-fit statistic is particularly responsive to very small departures

[22] Since the alternative model is identified, the noncentrality parameter can be computed directly from Equation (17) as well as by the procedure used for the other two alternative models. For this model, the latter procedure overstates the noncentrality parameter, especially when q is large. At $q = 0.10$, it is 2 percent larger than the value computed from Equation (17); at $q = 0.20$, it is 6 percent larger; and at $q = 0.30$, it is 14 percent larger.

[23] The RMSR scale in the bottom panel in Figure 9 is somewhat misleading, since all but two of the population correlations are reproduced exactly by the hypothesized model. For a given q, the two correlations that are not reproduced differ from the corresponding population correlations by $0.18q$, while the population RMSR is $0.08q$ in the bottom panel of Figure 9.

from the uninteresting restrictions; and there are strong reasons to assume that departures from the uninteresting restrictions are uniformly small in magnitude. When each of these conditions holds, one may be justified in ignoring a statistically significant overall goodness-of-fit test and concentrating instead on a nested comparison of the parameters of substantive interest.

In principle it should be possible to demonstrate—*before* estimating a baseline model on sample data—that each of the three conditions holds. The procedure we have applied to a relatively simple model is applicable to virtually any model estimated by maximum-likelihood methods. In practice, however, rationalizations for ignoring a significant goodness-of-fit statistic are invoked *after* estimating a poor-fitting model from sample data. That is, rules of thumb such as ratios of χ^2 to degrees of freedom are invoked only when the data fail to conform to hypotheses. Clearly such practices undermine the application of principles of statistical inference to covariance structure models.

CONCLUSIONS

Discussions of statistical power in covariance structure models are usually cloaked under the guise of the influence of sample size on measures of goodness of fit. Since the likelihood ratio χ^2 test statistic is a function of sample size as well as the degree to which restrictions implied by the null hypothesis are satisfied by sample moments, it is often argued that one should ignore that statistic in large samples and resort to alternative indexes of fit. We have offered three reasons for using classic principles of statistical inference rather than various ad hoc procedures for offsetting the influence of sample size on statistical tests.

First, computing the power of statistical tests requires researchers to formulate explicitly an alternative hypothesis, which forces them to decide what constitutes meaningful values of parameters differentiating hypothesized and alternative models. Second, while most ad hoc procedures address the problem of "too much" power to detect trivial departures from implied restrictions in large samples, classic procedures also confront the issue of "too little" power to detect meaningful departures in relatively small samples. Ad hoc indexes of fit that are not explicit functions of sample size fail to consider entirely the impact of sampling variability. Specifically, they do not recognize that departures from overidentifying restrictions *in the sample* are larger on average in smaller samples. Classic procedures simultaneously consider the influence of sample size, sampling variability, and true departures from restrictions in the population on measures of fit, whereas ad hoc procedures deal only with the influence of sample size. Third, the likelihood of rejecting a hypothesized model depends on the entire parametric structure of the true population model. The particular moments among exogenous constructs, the size of disturbance variances, the magnitude of coefficients of structural equations, and the reliability and number of indicators all influence the probability of detecting departures from a hypothesized model. Classic procedures allow us to disentangle the effects of each on statistical power; ad hoc procedures deal only with sample size.

Our exposition suggests several avenues for further research, some more manageable than others. The most straightforward is to extend our results for the hypothesis on a vector of r parameters, $\boldsymbol{\theta}_r = \boldsymbol{\theta}_{r0}$, to the general linear hypothesis of g linear restrictions on r parameters. We can denote this hypothesis as $\mathbf{H}\boldsymbol{\theta}_r = \mathbf{c}$, where $\mathbf{H}$ is a $g \times r$ matrix of coefficients for the g contrasts among the r parameters and $\mathbf{c}$ is a $g \times 1$ vector of constants. Our results generalize by replacing $(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r0})$ with $(\mathbf{H}\boldsymbol{\theta}_r - \mathbf{c})$ and $\mathbf{V}_r$ with $\mathbf{V}_g$ (the matrix of sampling covariances among contrasts in Equation 17).

More complicated statistical issues involve results for small samples and procedures for random variables that depart from a multivariate normal distribution. Somewhat paradoxically, the procedures we propose for addressing statistical power, a finite-sample problem, are based on the large-sample properties of maximum-likelihood estimators (Kendall and Stuart 1979, pp. 246–247). There appears to be little (if any) literature, however, on the finite-sample, nonnull distribution of the likelihood-ratio χ^2 statistic.[24] Furthermore, while there are a variety of estimation methods for covariance structure models that do not require the assumption of multivariate normality (Bentler 1983; Browne 1984), little work has been done on the distribution of appropriate test statistics under either null or alternative hypotheses.

Finally, we need to incorporate the literature on sequential testing of nested hypotheses into our methods for testing covariance structure models. In estimating such models, researchers typically *test* a set of restrictions and then reestimate the model subject to the restrictions that pass the test. For example, we might first test for tau-equivalence (equality

[24] Thus we are assuming that the asymptotic result of the noncentral χ^2 distribution is reasonably approximated in finite samples. A simple way of examining this assumption entails using a Monte Carlo simulation to assess the nonnull distribution of the likelihood ratio statistic in finite samples.

of lambda coefficients) across measures of a latent variable and then reestimate the model subject to the corresponding equality constraints. The resulting gain in statistical power and efficiency, however, is partly illusory since the true distribution of final parameter estimates is some mixture of the sampling distribution of final parameter estimates with and without the restrictions imposed. The work on "pretest estimators" (Judge, Griffiths, Hill, and Lee 1980, pp. 61–67) treats the problem within the context of the general linear hypothesis for linear models, but to date these ideas have not been applied to covariance structure models.

In closing, we advocate extending the frontiers of covariance structure research by developing and applying classic statistical methods. Used with discretion, incremental fit indexes (Bentler and Bonett, 1980), adjusted goodness-of-fit measures (Jöreskog and Sörbom 1984), critical sample size measures (Hoelter 1983), and the like can be valuable practical tools. But by deflecting attention away from basic principles of classic inference and by deemphasizing the careful specifications of a priori theoretical structural equation models, the indiscriminant use of such procedures is transforming a powerful modeling and testing method into the data-dredging, exploratory approach it was meant to supplant.

REFERENCES

BEARDON, WILLIAM O., SHARMA SUBASH, AND JESSE E. TEEL. 1982. "Sample Size Effects on Chi Square and Other Statistics Used in Evaluating Causal Models." *Journal of Marketing Research* 19:425–30.

BELSLEY, DAVID A. 1982. "Assessing the Presence of Harmful Collinearity and Other Forms of Weak Data Through a Test for Signal-to-Noise." *Journal of Econometrics* 20:211–53.

BENTLER, PETER M. 1980. "Multivariate Analysis with Latent Variables: Causal Modeling." *Annual Review of Psychology* 31:419–56.

————————. 1983. "Some Contributions to Efficient Statistics in Structural Models: Specification and Estimation of Moment Structures." *Psychometrika* 48:493–517.

BENTLER, PETER M. AND DOUGLAS G. BONETT. 1980. "Significance Tests and Goodness of Fit in the Analysis of Covariance Structures." *Psychological Bulletin* 88:588–606.

BEYER, WILLIAM H. 1968. *Handbook of Tables for Probability and Statistics* (2nd ed.). Cleveland: Chemical Rubber Company.

BIELBY, WILLIAM T. AND ROBERT M. HAUSER. 1977. "Structural Equation Models." Pp. 137–62 in *Annual Review of Sociology*. Volume 3, edited by A. Inkeles, J. Gleman and N. Smelzer. Palo Alto, CA: Annual Reviews.

BIELBY, WILLIAM T. AND JAMES R. KLUEGAL. 1977. "Simultaneous Statistical Inference and Statistical Power in Survey Research Applications of the General Linear Model." Pp. 283–312 in *Sociological Methodology 1977*, edited by D. R. Heise. San Francisco: Jossey-Bass.

BOOMSMA, ANNE. 1982. "The Robustness of LISREL Against Small Sample Sizes in Factor Analysis Models." Pp. 149–173 in *Systems Under Indirect Observation: Causality, Structure, Prediction*, edited by K. G. Jöreskog and H. Wold. Amsterdam: North Holland.

_______________. 1983. "On the Robustness of LISREL (Maximum Likelihood Estimation) Against Small Sample Size and Non-Normality." Unpublished PhD dissertation, University of Groningen, Groningen.

BROWNE, MICHAEL W. 1984. "Asymptotically Distribution Free Methods for the Analysis of Covariance Structures." *British Journal of Mathematics and Statistical Psychology* 37:62–83.

CLEARY, T. ANNE AND ROBERT L. LINN. 1969. "Error of Measurement and the Power of a Statistical Test." *British Journal of Mathematical and Statistical Psychology* 22:49–95.

CLEARY, T. ANNE, ROBERT L. LINN, AND G. WILLIAM WALSTER. 1970. "Effect of Reliability and Validity on Power of Statistical Tests." Pp. 130–50 in *Sociological Methodology 1970*, edited by E. F. Borgatta. San Francisco: Jossey-Bass.

FORNELL, CLAES AND DAVID F. LARCKER. 1981. "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error." *Journal of Marketing Research* 18:39–50.

FULLER, E. L. JR. AND W. J. HEMMERLE. 1966. "Robustness of the Maximum Likelihood Estimation Procedure in Factor Analysis." *Psychometrika* 31:255–66.

GEWEKE, JOHN F. AND KENNETH J. SINGLETON. 1980. "Interpreting the Likelihood Ratio Statistic in Factor Models When Sample Size is Small." *Journal of the American Statistical Association* 75:133–37.

GOLDBERGER, ARTHUR S. 1972. "Structural Equation Methods in the Social Sciences." *Econometrica* 40:979–1001.

GRUVAEUS, G. T. AND KARL G. JÖRESKOG. 1970. *A Computer Program for Minimizing a Function of Several Variables*. Research Bulletin 70–14. Princeton: Educational Testing Service.

HAYNAN, G. E., A. GOVINDARAJULU, AND F. C. LEONE. 1970. "Tables of the Cumulative Non-Central Chi Square Distribution." Pp. 1–78 in *Selected Tables in Mathematical Statistics*. Volume 1, edited by H. L. Harter and D. B. Owen. Providence: American Mathematical Society.

HOELTER, JON W. 1983. "The Analysis of Covariance Structures: Goodness-of-Fit Indices." *Sociological Methods and Research* 11:325–44.

JÖRESKOG, KARL G. 1973. "A General Method for Estimating a Linear Structural Equation System." Pp. 85–112 in *Structural Equation Models in the Social Sciences*, edited by A. S. Goldberger and O. D. Duncan. New York: Seminar Press.

______________. 1977. "Structural Equation Models in the Social Sciences: Specification, Estimation and Testing." Pp. 265–87 in *Applications of Statistics*, edited by P. R. Krishnaiah. Amsterdam: North Holland.

______________. 1979. "Analyzing Psychological Data by Structural Analysis of Covariance Matrices." Pp. 45–100 in *Advances in Factor Analysis and Structural Equation Models*, edited by K. G. Jöreskog and D. Sörbom. Cambridge, MA: Abt.

______________. 1981. "Analysis of Covariance Structures." *Scandinavian Journal of Statistics* 8:65–92.

JÖRESKOG, KARL G. AND ARTHUR S. GOLDBERGER. 1975. "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable." *Journal of the American Statistical Association* 10:631–39.

JÖRESKOG, KARL G. AND DAG SÖRBOM. 1982. "Recent Developments in Structural Equation Modeling." *Journal of Marketing Research* 19:400–16.

______________. 1984. *LISREL VI: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*. Mooresville, IN: Scientific Software, Inc.

JUDGE, GEORGE D., WILLIAM E. GRIFFITHS, R. CARTER HILL, AND TSOUNG-CHAO LEE. 1980. *The Theory and Practice of Econometrics*. New York: Wiley.

KENDALL, SIR MAURICE AND ALAN STUART. 1979. *The Advanced Theory of Statistics*. Vol. 3: *Inference and Relationship*. London: Griffin.

LAWLEY, D. N. AND Z. SWANSON. 1954. "Tests of Significance in Factor Analysis of Artificial Data." *British Journal of Statistical Psychology* 7:75–9.

LORD, FREDERICK AND MELVIN R. NOVICK. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

MADDALA, G. S. 1971. "The Use of Variance Components in Pooling Cross Section and Time Series Data." *Econometrica* 39:341–58.

MCGAW, BARRY AND KARL G. JÖRESKOG. 1971. "Factorial Invariance of Ability Measures in Groups Differing in Intelligence and Socioeconomic Status." *British Journal of Mathematical and Statistical Psychology* 24:154–68.

OLSSON, ULF. 1979. "On the Robustness of Factor Analysis Against Crude Classification of Observations." *Multivariate Behavioral Research* 14:485–500.

RORER, LEONARD G. AND THOMAS A. WIDIGER. 1983. "Personality Structure and Assessment." *Annual Review of Psychology* 34:431–63.

SARIS, WILLEM E., W. M. DE PIJPER AND P. ZEGWAART. 1979. "Detection of Specification Errors in Linear Structural Equation Models." Pp. 151–71 in *Sociological Methodology 1979*, edited by K. F. Schuessler. San Francisco: Jossey-Bass.

SATORRA, ALBERT AND WILLEM E. SARIS. 1985. "Power of the Likelihood Ratio Test in Covariance Structure Analysis." *Psychometrika* 50:83–90.

SOBEL, MICHAEL E. AND GEORGE W. BOHRNSTEDT. 1985. "The Use of Null Models in Evaluating the Fit of Covariance Structure Models." Pp. 152–78 in *Sociological Methodology 1985*, edited by N. B. Tuma. San Francisco: Jossey-Bass.

SÖRBOM, DAG G. 1975. "Detection of Correlated Errors in Longitudinal Data." *British Journal of Mathematical and Statistical Psychology* 28:138–51.

———————. 1981. "Structural Equation Models with Structured Means." Pp. 183–95 in *Systems Under Indirect Observation: Causality, Structure, Prediction*, edited by K. G. Jöreskog and H. Wold. Amsterdam: North Holland.

TUCKER, LEDYARD AND CHARLES LEWIS. 1973. "A Reliability Coefficient for Maximum Likelihood Factor Analysis." *Psychometrika* 38:1–10.

WALD, ABRAHAM. 1983. "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large." *Transactions of the American Mathematical Society* 54:426–82.

WHEATON, BLAIR, BENGTH MUTHÉN, DUANE ALWIN, AND GENE F. SUMMERS. 1977. "Assessing Reliability and Stability in Panel Models." Pp. 84–136 in *Sociological Methodology 1977*, edited by D. R. Heise. San Francisco: Jossey-Bass.