Psychological Methods
1996, Vol. 1, No. 2, 130–149

1082-989X/96/$3.00

Power Analysis and Determination of Sample Size for Covariance Structure Modeling

Robert C. MacCallum, Michael W. Browne, and Hazuki M. Sugawara
Ohio State University

A framework for hypothesis testing and power analysis in the assessment of fit of covariance structure models is presented. We emphasize the value of confidence intervals for fit indices, and we stress the relationship of confidence intervals to a framework for hypothesis testing. The approach allows for testing null hypotheses of not-good fit, reversing the role of the null hypothesis in conventional tests of model fit, so that a significant result provides strong support for good fit. The approach also allows for direct estimation of power, where effect size is defined in terms of a null and alternative value of the root-mean-square error of approximation fit index proposed by J. H. Steiger and J. M. Lind (1980). It is also feasible to determine minimum sample size required to achieve a given level of power for any test of fit in this framework. Computer programs and examples are provided for power analyses and calculation of minimum sample sizes.

A major aspect of the application of covariance structure modeling (CSM) in empirical research is the assessment of goodness of fit of an hypothesized model to sample data. There is considerable literature on the assessment of goodness of fit of such models, providing a wide array of fit indices along with information about their behavior (e.g., Bentler & Bonett, 1980; Browne & Cudeck, 1993; Marsh, Balla, & McDonald, 1988; Mulaik et al., 1989). Empirical applications of CSM typically evaluate fit using two approaches: (a) the conventional likelihood ratio χ^2 test of the hypothesis that the specified model holds exactly in the population; and (b) a variety of descriptive measures of fit of the model to the sample data. In this article we focus on an inferential approach to assessment of fit involving a particular measure of model fit, root-mean-square error of approximation (RMSEA; Steiger & Lind, 1980). Knowledge of distributional properties of this fit index allows for the construction of confidence intervals (CIs) and the formulation and testing of point or interval hypotheses, as shall be shown. We strongly urge the use of CIs for fit measures, and we use the hypothesis-testing framework for RMSEA as a vehicle for defining a procedure for statistical power analysis and determination of minimum sample size for CSM. These developments provide for the estimation of power for any test of model fit framed in terms of RMSEA, as well as the determination of necessary sample size to achieve a given level of power for any such test.

Of relevance throughout this article is the fact that a point estimate of fit is imprecise to some degree when considered as an estimate of model fit in the population. For some fit indices, such as RMSEA, whose distributional properties are known, the degree of imprecision can be captured in a CI. On the basis of the CI, one can say with a certain level of confidence that the given interval contains the true value of the fit index for that model in the population. Alternatively, one can take into account the imprecision in the sample estimate of fit by testing an hypothesis about the population value of the fit index. There is a simple

Robert C. MacCallum, Michael W. Browne, and Hazuki M. Sugawara, Department of Psychology, Ohio State University.

Portions of the material presented in this article are based on work conducted by Hazuki M. Sugawara in her doctoral dissertation, completed in the Psychology Department at Ohio State University in 1994.

Correspondence concerning this article should be addressed to Robert C. MacCallum, Department of Psychology, 1885 Neil Avenue, Ohio State University, Columbus, Ohio 43210.

relationship between a CI for a population value of a fit index and a test of an hypothesis about that population value. An appropriate CI implies the outcome of an hypothesis test. For instance, suppose we formed a 90% CI for some arbitrary fit index ρ. Suppose we also wished to test the null hypothesis that ρ was equal to some specific value ρ_0, using $\alpha = .05$. The outcome of the test is implied by the CI. If the CI contains the value ρ_0, then the hypothesis is not rejected; but if the CI does not contain ρ_0, then the hypothesis is rejected. Of course, the CI provides more information than the hypothesis test because the interval estimate indicates the degree of precision of the sample value of the index. In this article, although we frame many of our developments regarding power analysis for CSM in terms of hypothesis tests, we often make use of the tie between such tests and the more informative CIs.

When testing an hypothesis about model fit, it is of course highly desirable to (a) test a meaningful, relevant hypothesis and (b) draw the correct conclusion about that hypothesis. There is little point in conducting an hypothesis test when the hypothesis being tested is not empirically interesting and for which the outcome is not informative in our efforts to evaluate a model. Once an appropriate hypothesis is defined, it is important to know the likelihood of drawing the correct conclusion when the hypothesis test is conducted. Incorrect conclusions could lead the investigator far astray in the process of model development and evaluation. For example, if the model is truly a good model in terms of its level of fit in the population, we wish to avoid concluding that the model is a bad one. Alternatively, if the model is truly a bad one, we wish to avoid concluding that it is a good one. However, such invalid conclusions can certainly occur, as they always can in an hypothesis-testing context. For instance, if a model fits badly in the population and we test the null hypothesis that the model fits well, the correct outcome is rejection of the null hypothesis. Failure to reject constitutes a Type II decision error. Such errors can occur because our sample measures of fit are imprecise indicators of fit in the population. In the case just described, we may fail to reject the false null hypothesis if we happen to draw a sample wherein the model fits well or if our sample size is not sufficiently large to provide a precise estimate of goodness of fit.

In the present article we have three objectives. First, we define appropriate hypotheses to test regarding fit of covariance structure models, along with methods for testing them. We also show how appropriate CIs imply the outcome of such tests and provide additional information. Second, we provide procedures for conducting power analyses for these hypothesis tests, thus providing a mechanism for determining the likelihood of drawing the correct conclusion about a false null hypothesis regarding model fit. Although there does exist some literature on power analysis for covariance structure models (e.g., Saris & Satorra, 1993; Satorra & Saris, 1985), which are discussed later in this article, our approach is simpler than existing methods and is more easily applied. Finally, we provide procedures for determining minimum sample size necessary to achieve a desired level of power for testing hypotheses about fit of covariance structure models. We anticipate that these procedures will be useful in the design of studies using CSM.

Model Estimation and Assessment of Fit

Discrepancy Functions and Parameter Estimation

Given p manifest variables (MVs), let Σ_0 represent the $p \times p$ population covariance matrix. A covariance structure model represents the elements of Σ_0 as functions of model parameters. Let γ be a vector of order q containing the q parameters of a specified model. Then the model could be represented as

$$\Sigma_0 = \Sigma(\gamma), \tag{1}$$

where $\Sigma(\gamma)$ is a matrix-valued function that specifies the functional relationship between the population covariances and the model parameters. Many models belong to the class represented by Equation 1, including structural equation models with latent variables, factor analysis, path analysis, simultaneous equation models, and others.

In practice a specified model is fitted to a $p \times p$ sample covariance matrix, $\mathbf{S}$. For any selected vector of parameter estimates, $\hat{\gamma}$, the model specified in Equation 1 can be used to obtain a reconstructed or implied covariance matrix, $\hat{\Sigma}$:

$$\hat{\Sigma} = \Sigma(\hat{\gamma}) \tag{2}$$

The objective in parameter estimation is to find $\hat{\gamma}$ so that the resulting $\hat{\Sigma}$ is as similar as possible to $\mathbf{S}$. The difference between $\hat{\Sigma}$ and $\mathbf{S}$ is measured by a discrepancy function, $F(\mathbf{S}, \hat{\Sigma})$, which takes on a value of zero only when $\mathbf{S} = \hat{\Sigma}$ and otherwise is positive, increasing as the difference between $\mathbf{S}$ and $\hat{\Sigma}$ increases. A number of different discrepancy functions have been defined. The most commonly used such function is the normal-theory maximum likelihood (ML) function, defined as

$$F_{ML} = \ln |\hat{\Sigma}| - \ln |\mathbf{S}| + \text{Tr}\,(\mathbf{S}\hat{\Sigma}^{-1}) - p. \quad (3)$$

Developments in this article are not dependent on the use of the ML discrepancy function for parameter estimation but could be used with other discrepancy functions such as generalized least squares (GLS) or asymptotically distribution free (ADF; see Bollen, 1989, for a discussion of discrepancy functions). All that is required for the use of developments in this article is that a discrepancy function be used that provides an asymptotic χ^2 fit statistic, discussed shortly, and that the distributional assumptions underlying the selected discrepancy function be adequately satisfied. Given the selection of an appropriate discrepancy function, parameter estimation is then carried out by determining the vector of parameter estimates, $\hat{\gamma}$, that produces a $\hat{\Sigma}$ that in turn yields the minimum value of the discrepancy function. That minimum value is a sample statistic that will be designated $\hat{F}$. The magnitude of $\hat{F}$ reflects the degree of lack of fit of the model to the sample data.

Testing Hypotheses About Model Fit

A variety of methods and statistics have been proposed for evaluating the relative magnitude of $\hat{F}$ so as to achieve an assessment of model fit. A common procedure is the conventional likelihood ratio (LR) test. Let γ_0 represent the vector of unknown population parameter values. We can define a null hypothesis H_0: $\Sigma_0 = \Sigma(\gamma_0)$, representing the hypothesis that the specified model holds exactly in the population. This null hypothesis can be tested using the test statistic $(N - 1)\hat{F}$. If the distributional assumptions underlying the discrepancy function being used are adequately satisfied, and if N is sufficiently large, then $(N - 1)\hat{F}$ will be approximately distributed as χ^2_d with degrees of freedom $d = p(p + 1)/2 - q$, where q is the number of distinct parameters to be estimated. For a selected α level, one can determine a critical value of χ^2_d. Let that valued be designed χ^2_c. If the observed value of the test statistic exceeds χ^2_c, then H_0 is rejected; if not, H_0 is not rejected.

The result of this test is reported in virtually every application of CSM. In interpreting the outcome of this test, it is important to recognize that the hypothesis being tested is the hypothesis of exact fit, that is, that the specified model is exactly correct in the population and that any lack of fit in the sample arises only from sampling error. This is a stringent null hypothesis, one that in practical applications is always false to some degree for any overidentified model. In the process of specifying and estimating covariance structure models, the best one can hope for realistically is that a model provides a close approximation to real-world relationships and effects. These models can not be expected to fit the real world exactly. However, even if a model is a good one in terms of representing a fairly close approximation to the real world, the test of exact fit will result in rejection of the model if N is adequately large. Because large samples are necessary in CSM so as to obtain precise parameter estimates as well as to satisfy asymptotic distributional approximations, samples will often be large enough to lead to rejection of good models via the test of exact fit. Thus, we believe that the test of exact fit is not particularly useful in practice because the hypothesis being tested is implausible and is not empirically interesting and because the test will result in rejection of good models when N is large. If one is going to test hypotheses about model fit, it is necessary to test realistic hypotheses so as to obtain useful information.

We next consider a line of development in assessment of model fit that provides a capability for establishing CIs for some fit measures and for testing hypotheses other than that of exact fit. These developments began with work by Steiger and Lind (1980) and have been extended in recent work by Browne and Cudeck (1993). We briefly consider here some basic elements discussed in more detail by Browne and Cudeck but that have their origins in the seminal work of Steiger and Lind.

Prior to presenting this information, it is useful to review some basic background material. The procedures proposed by Browne and Cudeck (1993) and Steiger and Lind (1980), and assump-

tions underlying those procedures, make use of noncentral χ^2 distributions. Basic properties of such distributions are reviewed here for the benefit of readers not familiar with them. Given d normal variates $z_1, z_2, \ldots z_d$, with unit variances and zero means, Σz_j^2 follows a central χ^2 distribution with d degrees of freedom and expected value $E(\chi_d^2) = d$. Given d normal variates $x_1, x_2, \ldots x_d$, with unit variances and nonzero means $\mu_1, \mu_2, \ldots \mu_d$, then Σx_j^2 follows a noncentral χ^2 distribution. Such a distribution has two parameters: degrees of freedom, d, and a noncentrality parameter, $\lambda = \Sigma \mu_j^2$. The expected value is given by $E(\chi_{d,\lambda}^2) = d + \lambda$. Thus, the noncentrality parameter shifts the expected value of the distribution to the right of that of the corresponding central χ^2. In the ensuing material we make extensive use of noncentral χ^2 distributions.

Returning to the context of model evaluation, suppose that the population covariance matrix $\mathbf{\Sigma}_0$ were known and that a model of interest were fit to $\mathbf{\Sigma}_0$. We define F_0 as the resulting value of the discrepancy function reflecting lack of fit of the model in the population. If $F_0 = 0$, meaning that exact fit holds in the population, then, as noted earlier, $(N - 1)\hat{F}$ follows approximately a central χ^2 distribution with d degrees of freedom. However, when the model does not hold in the population, which will be the normal case in empirical applications, F_0 will have some unknown nonzero value. It is desirable to obtain an estimate of F_0. When $F_0 \neq 0$, then $(N - 1)\hat{F}$ will be distributed approximately as noncentral $\chi_{d,\lambda}^2$, where the noncentrality parameter $\lambda = (N - 1)F_0$, under the additional assumption that lack of fit of the model in the population is of approximately the same magnitude as lack of fit arising due to sampling error.

All of the developments presented in the remainder of this article make use of the noncentral χ^2 distribution as an approximation for the distribution of $(N - 1)\hat{F}$. Steiger, Shapiro, and Browne (1985) provided the theoretical basis for the use of the noncentral χ^2 distribution in the context considered in this article. This approximation will be satisfactory under the same conditions in which the widely used test of exact fit is appropriate, with the additional assumption just mentioned that lack of fit due to model error and sampling error are of approximately the same magnitude. Thus, assumptions about the population must be adequately satisfied. The nature of these distributional assumptions depends on the discrepancy function being used. For instance, for ML estimation, one nominally assumes multivariate normality in the population; for ADF estimation, no rigid assumption about the population distribution needs to be made. Regarding the assumption about the relative magnitude of model error and sampling error, Steiger et al. (1985) and Browne and Cudeck (1993) stated that the noncentral χ^2 approximation will be adequate as long as distributional assumptions are satisfied adequately, sample size is not too small, and F_0 is not too large. If this condition is violated severely, which would involve the case of a poor model fitted to data from a large sample, then results of model fitting would clearly show the model to be a poor one, and the adequacy of the noncentral χ^2 approximation would be irrelevant. There has been no large-scale Monte Carlo investigation of the adequacy of this approximation under violations of assumptions, although a limited study by Satorra, Saris, and de Pijper (1991) showed the approximation to work fairly well under conditions in which N was small and model misspecification was not too severe. An extensive investigation of these issues is beyond the scope of this article. In any case, these distributional assumptions are necessary to perform any power analysis in the context of CSM and are no more constraining than the assumptions required for other aspects of model fitting and testing.

As discussed by Browne and Cudeck (1993), if $(N - 1)\hat{F}$ has a noncentral χ^2 distribution, then the sample discrepancy function value $\hat{F}$ is a biased estimator of F_0, with expected value given by

$$E(\hat{F}) = F_0 + d/(N - 1). \tag{4}$$

Thus, a less biased estimator of F_0 can be obtained by

$$\hat{F}_0 = \hat{F} - d/(N - 1). \tag{5}$$

If Equation 5 yields a negative value, then $\hat{F}_0$ is defined as 0.

The notion of the population discrepancy function value F_0 and its estimator $\hat{F}_0$ forms the basis of a measure of fit first proposed by Steiger and Lind (1980) and now usually referred to as RMSEA. The definition of RMSEA is based on the property that the minimum value of the discrepancy function is equal to, or closely approxi-

mated by, a sum of d squared discrepancy terms, where the discrepancy terms represent systematic lack of fit of the model. On the basis of this property, the RMS measure of model discrepancy in the population, to be designated ε in this article, can be defined as

$$\varepsilon = \sqrt{F_0/d}. \quad (6)$$

This index indicates discrepancy per degree of freedom and is thus sensitive to the number of model parameters. Given two models with equal fit in the population (i.e., equal values of F_0) but different degrees of freedom, this index will yield a smaller (better) value for the model with more degrees of freedom (fewer parameters). As defined in Equation 6, ε is a population measure that is a function of the unknown value of F_0. An estimate of ε can be obtained by substituting the estimate of F_0 from Equation 5 into Equation 6, yielding

$$\hat{\varepsilon} = \sqrt{\hat{F}_0/d}. \quad (7)$$

Steiger (1989), Browne and Mels (1990), and Browne and Cudeck (1993) offered guidelines for interpretation of the value of ε. By analyzing many sets of empirical data and evaluating the behavior of ε in relation to previous conclusions about model fit, Steiger (1989) and Browne and Mels (1990) arrived independently at the recommendation that values of ε less than 0.05 be considered as indicative of close fit. Browne and Cudeck provided a number of empirical examples to support this guideline, wherein values of ε less than 0.05 yielded conclusions about model fit consistent with previous analyses of the same data sets. Browne and Cudeck also suggested that values in the range of 0.05 to 0.08 indicate fair fit and that values above 0.10 indicate poor fit. We consider values in the range of 0.08 to 0.10 to indicate mediocre fit. Clearly these guidelines are intended as aids for interpretation of a value that lies on a continuous scale and not as absolute thresholds.

A useful feature of this RMSEA fit index is that it is possible to compute CIs for the population value of ε (Steiger & Lind, 1980). Details are presented in Browne and Cudeck (1993). The RMSEA value and its associated CI are now available in many standard CSM computer programs, including LISREL 8 (Jöreskog & Sörbom, 1993), CALIS (SAS Institute, Inc., 1992), RAMONA (Browne, Mels, & Coward, 1994), AMOS (Arbuckle, 1994), SePath (Steiger, 1994) and others. For users without access to recent versions of these programs, a program called FITMOD is available for computing RMSEA and its CI, as well as other information about model fit.[1] We recommend that such CIs be used in practice. The calculation and interpretation of a point estimate of an index of model fit does not take into account the imprecision in the estimate, potentially misleading a researcher. The associated CI provides information about precision of the estimate and can greatly assist the researcher in drawing appropriate conclusions about model quality. For instance, if a model were to yield a low value of $\hat{\varepsilon}$ but a wide CI, the investigator could recognize that there may be substantial imprecision in $\hat{\varepsilon}$, in which case one cannot determine accurately the degree of fit in the population. A very narrow CI, on the other hand, would lend support to the interpretation of the observed value of $\hat{\varepsilon}$ as a precise indicator of fit in the population. It can be shown that the width of the resulting CIs is greatly influenced by both N and d. If both are small, then CIs for ε will be quite wide. If d is small, then a very large N is needed to obtain a reasonably narrow CI. On the other hand, if d is very large, which would occur in studies with a large number of measured variables and models with relatively few parameters, then rather narrow CIs for ε are obtained even with a moderate N.

Given that CIs for ε can be determined, it is also feasible to frame hypothesis tests in terms of this index of model fit. Recall that, given adequate approximation of assumptions, when $F_0 \neq 0$, then $(N - 1)\hat{F}$ will be distributed approximately as noncentral $\chi^2_{d,\lambda}$, where the noncentrality parameter $\lambda = (N - 1)F_0$. Note that the value of the noncentrality parameter is a function of lack of fit of the model in the population. Making use of the fact from Equation 5 that $F_0 = d\varepsilon^2$, we can define the noncentrality parameter in terms of the RMSEA index, ε:

$$\lambda = (N - 1)d\varepsilon^2. \quad (8)$$

This development can be used to reframe proce-

[1] Information on how to obtain the computer program FITMOD can be obtained by writing to Michael W. Browne, Department of Psychology, 1885 Neil Avenue, Ohio State University, Columbus, Ohio 43210.

dures for testing hypotheses about model fit. The conventional LR test of exact fit can be redefined as a test of H_0: $\varepsilon = 0$, that is, perfect fit of the model in the population. Under this null hypothesis, the test statistic $(N - 1)\hat{F}$ would follow a χ^2 distribution with d degrees of freedom and noncentrality parameter $\lambda = 0$ from Equation 8. Thus, the test statistic is evaluated using a central χ^2_d distribution, as described earlier. Given the problems discussed earlier regarding this test, it is useful to consider defining and testing other hypotheses about fit. In the present context, this is quite straightforward. For example, Browne and Cudeck (1993) suggested testing a null hypothesis of close fit, defined as H_0: $\varepsilon \leq .05$. This hypothesis is more realistic than the hypothesis of exact fit and can be tested quite easily. Under this null hypothesis, and given sufficiently large N and adequate approximation of assumptions mentioned earlier, the test statistic $(N - 1)\hat{F}$ would follow a noncentral $\chi^2_{d,\lambda}$ distribution, with $\lambda = (N - 1)d(0.05)^2$. Therefore, for a given α level, the significance of the test statistic is evaluated by comparing it to a critical χ^2_c, where χ^2_c cuts off an area of α in the upper tail of the distribution of $\chi^2_{d,\lambda}$. In comparison with the test of exact fit, the same test statistic is used, but the value of χ^2_c will be greater because critical values in the noncentral distribution of $\chi^2_{d,\lambda}$ are shifted to the right of corresponding values in the central distribution of χ^2_d, as is shown in Figure 1. As a result, an obtained value of $(N - 1)\hat{F}$ that leads to rejection of the hypothesis of exact fit might well lead to failure to reject the hypothesis of close fit. Such an outcome is not at all unusual for good models when N is large; see the examples in Browne and Cudeck (1993).

If one wishes also to conduct a formal hypothesis test about the value of ε, such tests are straightforward using the noncentral χ^2 distribution as described earlier. Although it would be possible to test many such hypotheses, we believe the hypothesis of close fit (H_0: $\varepsilon \leq 0.05$) is a sensible alternative to the hypothesis of exact fit (H_0: $\varepsilon = 0$). Although it must be acknowledged that the value of 0.05 is somewhat arbitrary, that value is supported by independent sources in the literature, as noted earlier (Browne & Mels, 1990; Steiger, 1989). Furthermore, the test of close fit has been recognized by authors of some major CSM software packages and incorporated into their programs (e.g., LISREL8: Jöreskog & Sörbom, 1993; CALIS: SAS Institute, Inc., 1992; and RAMONA: Browne et al., 1994). Most important, testing a hypothesis of close fit using some reasonable definition of close fit (e.g., H_0: $\varepsilon \leq 0.05$) is clearly more sensible and substantively interesting than testing a hypothesis of exact fit. If future research suggests that some value other than 0.05 is preferable for such a test, it is trivial to adapt the methodology accordingly.

The developments presented to this point regarding the RMSEA fit index and associated CIs as well as the test of close fit are discussed in more detail by Browne and Cudeck (1993). We now wish to consider further developments based on this approach. There exists a problem associated with the role of the null hypothesis in both the test of exact fit and the test of close fit. In both cases, assuming one is seeking to support a model under study, one wishes to garner support for the null hypothesis. If the null hypothesis is rejected, we conclude that the observed data are highly inconsistent with the hypothesis of exact or close fit, whichever is being tested. The model is not supported. If the null hypothesis is not rejected, we conclude that the data are not sufficiently inconsistent with the null hypothesis for us to reject that hypothesis. This latter outcome does not imply clear support for the model but rather the absence of strong evidence against it. It is difficult to argue for support of a model using the tests of exact or close fit.

The current framework for testing hypotheses about ε offers a mechanism for addressing this problem by reversing the role of the null hypothesis. Consider defining the null hypothesis as representing a lack of close fit in the population, thereby creating a situation in which we hope to reject that hypothesis. For example, we could define H_0: $\varepsilon \geq 0.05$, meaning that the fit of the model in the population is not close. Rejection of this hypothesis would support the conclusion that the fit of the model in the population is close, that is, support for the alternative that $\varepsilon < 0.05$. Testing this null hypothesis is straightforward. Under H_0: $\varepsilon \geq 0.05$, and given sufficiently large N and adequate approximation to assumptions, the test statistic $(N - 1)\hat{F}$ would be distributed as noncentral $\chi^2_{d,\lambda}$, where $\lambda = (N - 1)d(0.05)^2$. One would now conduct a one-tail test using the lower tail of the distribution, because a sufficiently low value of $(N - 1)\hat{F}$ would result in rejection of H_0. Thus,

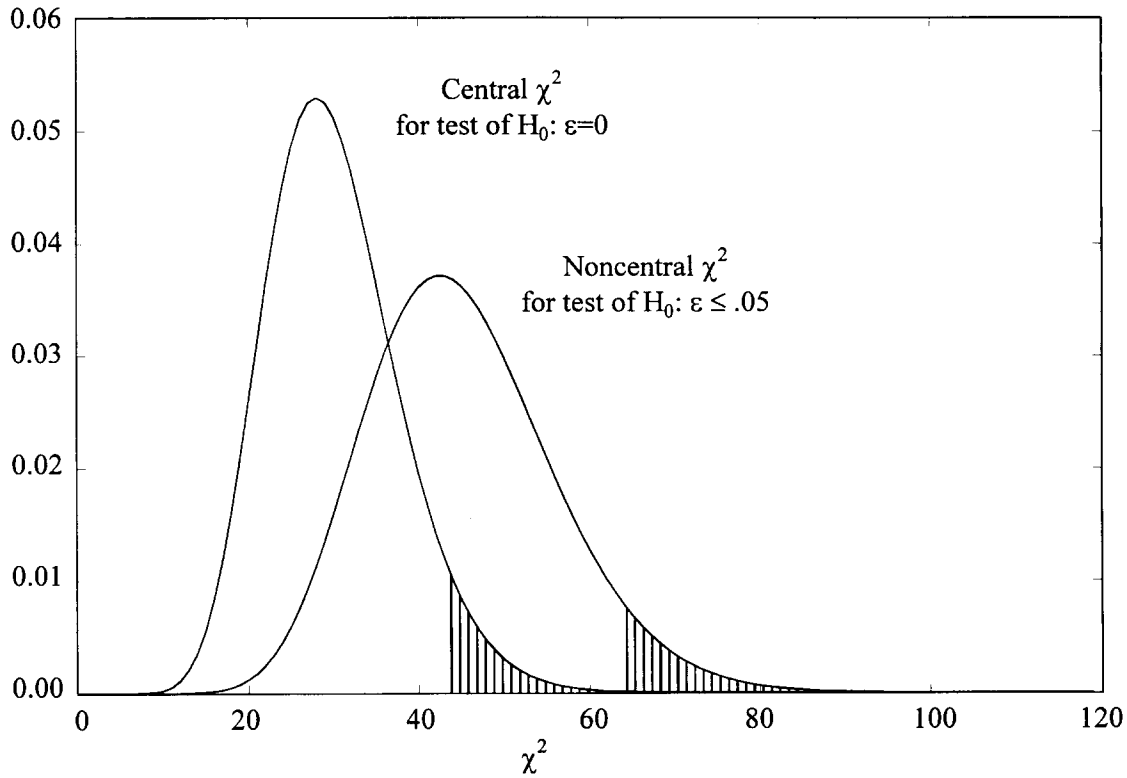

Figure 1. Illustration of difference in critical values between central and noncentral χ^2 distributions.

given α, the critical χ_c^2 would cut off an area of α in the lower tail of the distribution of $\chi_{d,\lambda}^2$, and H_0 would be rejected if $(N - 1)\hat{F} < \chi_c^2$. This case is illustrated in Figure 2. We refer to the test just described as a *test of not-close fit.*

The test of not-close fit provides for more appropriate roles for the null and alternative hypotheses in the context of model evaluation. When specifying and evaluating a model, our research hypothesis would normally be that the model provides a good approximation to the real-world phenomena under study. As is often pointed out in introductory treatments of hypothesis testing (e.g., Champion, 1981), the research hypothesis is most appropriately represented by the alternative hypothesis, so that rejection of the null hypothesis implies support for the research hypothesis. If the research hypothesis corresponds to the null hypothesis, then it becomes very difficult to support the research hypothesis, as is the case in usual tests of model fit in CSM.

To summarize the direct relationship between a CI for ε and the tests of close and not-close fit discussed in this section, consider Table 1. This table shows how a $100(1 - 2\alpha)\%$ CI for ε implies the outcome of tests of close and not-close fit using significance level α. In addition, it is useful to recognize that a CI with a lower bound of 0 will result in failure to reject the hypothesis of exact fit. It is clear that CIs provide more information than is yielded by an hypothesis test. The interval estimate of ε indicates the degree of imprecision in this estimate of fit. This information is not reflected nearly as clearly in an hypothesis test. Thus, we strongly encourage the use of CIs in their own right as well as for purposes of inferring results of hypothesis tests.

Examples of Tests of Model Fit for Empirical Studies

Browne and Cudeck (1993) presented results of tests of exact and close fit for several data sets; we extend their analyses to include the test of not-close fit proposed here. They reported results of a series of factor analyses of data from McGaw

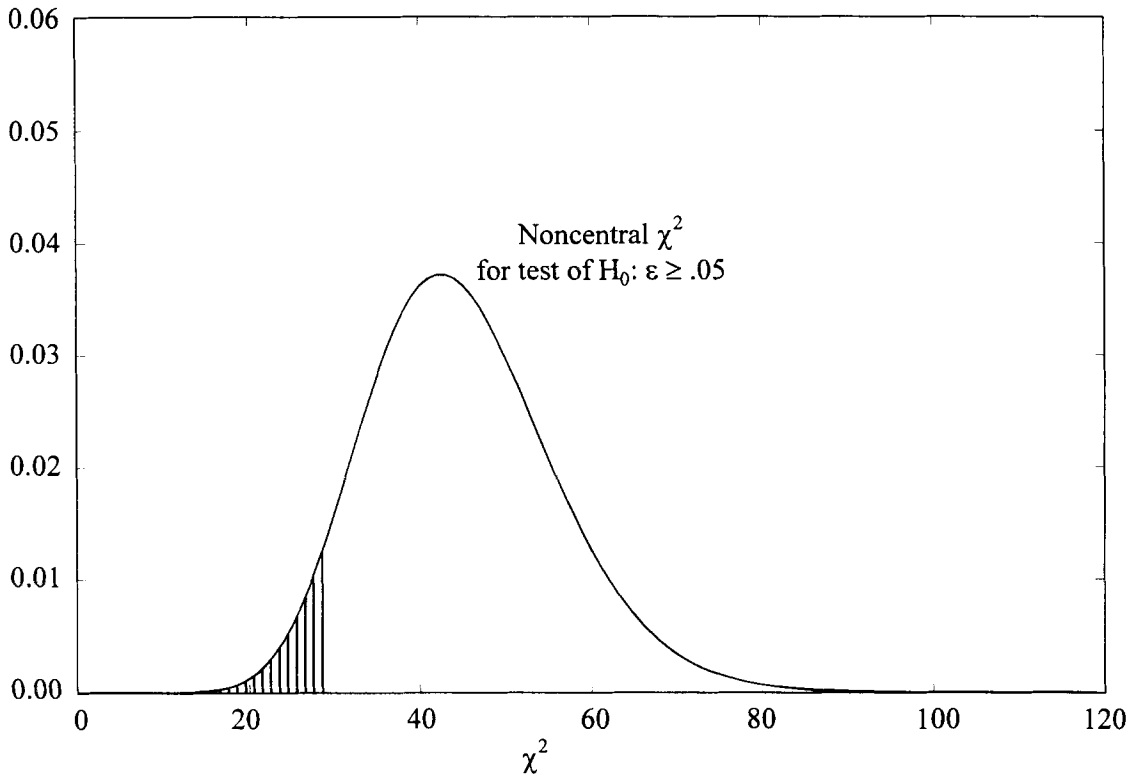

Figure 2. Illustration of critical value of noncentral χ^2 distribution for testing hypothesis of not-close fit.

and Jöreskog (1971); the data consist of measures on 21 ability tests for a sample of 11,743 individuals. In this example, fit measures are used to evaluate fit of the common factor model with a specified number of factors. For instance, a four-factor model yields a 90% CI for ε with bounds of 0.046 and 0.048. The hypothesis of exact fit is rejected at the .05 significance level ($\chi^2 = 3{,}548$, $d = 132$) because the lower bound of the interval is greater than zero, meaning that exact fit in the population is highly implausible. This CI also implies the outcomes of the tests of close and not-close fit, as is summarized in Table 1. The hypothesis of close fit is not rejected at the .05 significance level because the entire interval lies below .05, meaning that close fit is not implausible. A stronger conclusion is implied by the test of not-close fit. Because the upper bound of the CI is less than 0.05, the hypothesis of not-close fit (H_0: $\varepsilon \geq 0.05$) is rejected at the .05 significance level, meaning that not-close fit is highly implausible and providing strong support for close fit of the four-factor model. This

Table 1
Relationship Between Confidence Intervals and Hypothesis Tests

Nature of confidence interval[a]	Reject close fit?	Reject not-close fit?
Entire confidence interval below 0.05	No	Yes
Confidence interval straddles 0.05	No	No
Entire confidence interval above 0.05	Yes	No

[a] This table assumes that close fit is defined as $\varepsilon \leq 0.05$. If hypotheses are constructed on the basis of some other value, ε_0, then that value becomes the reference point for relating confidence intervals to hypothesis tests.

final result provides the clearest support for the model and will tend to occur when a model fits very well and N and d are large, resulting in a low value of $\hat{\varepsilon}$ and a narrow CI.

It is interesting to contrast these results with results obtained by extending another example presented by Browne and Cudeck (1993). In factor analyses on a battery of 24 intelligence tests in a sample of 86 individuals (data from Naglieri & Jensen, 1987), a five-factor model yielded a 90% CI for ε with bounds of 0.034 and 0.081. Because the lower bound of this interval is greater than 0, the hypothesis of exact fit is rejected ($\chi^2 = 215.74$, $d = 166$, $p < 0.05$). Because the CI includes the value of 0.05, neither the hypothesis of close fit nor that of not-close fit is rejected, as is indicated in Table 1. Thus, neither close fit nor not-close fit are ruled out. In this case, it would be a mistake to infer close fit based on failure to reject the hypothesis of close fit. A rigorous inference of close fit would require rejection of the hypothesis of not-close fit, which is not achieved in this case. The wide CI shows that both close fit and not-close fit are plausible.

Power Analysis for Tests of Fit

In the previous section we described a framework for testing various hypotheses about model fit, where the null hypothesis indicates the degree of fit in terms of the ε index. When using these tests for model evaluation, it is important to have adequate power for detecting when an hypothesis about model fit is false. Power analyses for the tests described in the previous section can be conducted fairly easily.

In general, H_0 specifies an hypothesized value of ε; let that value be designated as ε_0. If H_0 is false, then the actual value of ε is some value that is not consistent with H_0; let that value be designated as ε_a. The value of ε_a represents the degree of lack of fit of the specified model in the population. In power analysis terminology, the difference between ε_0 and ε_a reflects the effect size, conceptualized as the degree to which H_0 is incorrect. We emphasize that we are not defining a numerical index of effect size; specifically, the arithmetic difference between ε_0 and ε_a is not a numerical index of effect size because power is not a simple function of this difference. That is, if we define $\delta = \varepsilon_0 - \varepsilon_a$, power is not the same for all choices of ε_0 and ε_a that yield the same δ. Rather, power depends on the particular values of ε_0 and ε_a that are chosen. This same phenomenon occurs in power analysis for other types of hypothesis tests. For instance, Cohen (1988, pp. 110–113, 180–182) described this phenomenon in the contexts of testing differences between correlation coefficients and differences between proportions. In those situations it is conventional to define a numerical measure of effect size as a function of transformed values of correlations or proportions. In the present context it is not clear that a similar approach is viable or necessary. For current purposes, we define effect size in terms of a pair of values, ε_0 and ε_a, and the power analysis methods we present operate on the selected pair of values.

In selecting a pair of values, ε_0 and ε_a, there is an unavoidable element of arbitrariness. Any power analysis requires a specification of effect size, which is unknown in practice (if it were known, then no hypothesis test would be necessary). Cohen (1988) routinely suggested somewhat arbitrary guidelines for designation of small, medium, and large effect sizes for various hypothesis tests. In the present context, we choose values of ε_0 and ε_a on the basis of accepted guidelines for interpretation of ε, as presented earlier. However, we emphasize that the methodology we present for power analysis is not tied to any particular values of ε_0 and ε_a. The method is general and can be used for any pair of such values. However, we believe the values we use here to illustrate the method are reasonable choices that could be useful in empirical studies.

In testing the null hypothesis of close fit (H_0: $\varepsilon \leq 0.05$), ε_0 takes on a value of 0.05. (In general for tests of interval null hypotheses such as those used in the tests of close fit and not-close fit, ε_0 would be defined as the most extreme value of ε in the specified interval.) The value of ε_a must then be specified as some value greater than 0.05, representing the degree to which the model is considered to be incorrect in the population. Although this value is unknown in practice, appropriate values of ε_a can be specified for purposes of power estimation. For instance, ε_a could reasonably be specified as 0.08. One then has framed the following question: If the true value of ε is 0.08 and we test H_0: $\varepsilon \leq 0.05$, what is the power of the test? In other words, if the fit of the model is actually mediocre, and we test the hypothesis that fit is

close, what is the likelihood of rejecting the null hypothesis? For the test of not-close fit, ε_0 also takes on a value of 0.05. In this case, ε_a should be defined as some value less than 0.05, so that H_0: $\varepsilon \geq 0.05$ is false. We suggest setting $\varepsilon_a = 0.01$, representing the case of an extremely good model. Power analysis of this case then addresses the following question: If model fit is actually extremely good, and we test the hypothesis that fit is not close, what is the likelihood of rejecting the null hypothesis? Although these recommendations for values of ε_0 and ε_a are somewhat arbitrary, they are no less arbitrary than many other guidelines used in statistical analysis (e.g., $\alpha = .05$), and we believe that they define interesting, meaningful questions for power analyses. Other investigators are, of course, free to use and study other possible selections of ε_0 and ε_a. We caution investigators to choose meaningful values of ε_0 and ε_a and to recognize that if these two values are specified as very close together, resulting power estimates will generally be quite low. Regardless of the values selected, it is significant to note that the specification of ε_a does not require any statement on the part of the investigator as to how the model is misspecified; rather, ε_a indicates only the degree of lack of fit in the population.

On the basis of the values of ε_0 and ε_a, we can define two overlapping noncentral χ^2 distributions. The first, representing the distribution used to test H_0, is the distribution of χ^2_{d,λ_0}, where $\lambda_0 = (N-1)\,d\varepsilon_0^2$. Under H_0, the test statistic $(N-1)\hat{F}$ follows this distribution. For a given level of α, a critical value χ_c^2 is determined that cuts off an area of α in the upper or lower tail of the distribution of χ^2_{d,λ_0} (depending on whether H_0 represents close or not-close fit, respectively). If $(N-1)\hat{F}$ is more extreme than χ_c^2 in the appropriate tail, then H_0 is rejected. If H_0 is false in reality, and the true value of ε is ε_a, then the test statistic is actually an observation from a different noncentral χ^2 distribution. Specifically, we define the distribution of χ^2_{d,λ_a} as the true distribution of the test statistic, given ε_a, where $\lambda_a = (N-1)d\varepsilon_a^2$. Give χ_c^2, the critical value of the test statistic defined under H_0, the power of the test is then defined as the area under the true distribution of the test statistic, beyond χ_c^2 in the appropriate direction. That is, if $\varepsilon_0 < \varepsilon_a$, which would represent the case of a test of good fit (e.g., exact or close), then power π is given by

$$\pi = Pr(\chi^2_{d,\lambda_a} \geq \chi_c^2). \tag{9}$$

On the other hand, if $\varepsilon_0 > \varepsilon_a$, as in the case of a test of not-close fit, power is given by

$$\pi = Pr(\chi^2_{d,\lambda_a} \leq \chi_c^2). \tag{10}$$

The former case is shown in Figure 3, in which the null hypothesis of close fit is rejected if a sufficiently large value of the test statistic is obtained; the latter case is shown in Figure 4, in which the null hypothesis of not-close fit is rejected if a sufficiently small value of the test statistic is obtained.

We have applied this procedure to several cases of interest, defined according to values of ε_0 and ε_a. We first consider power for the test of H_0: $\varepsilon \leq 0.05$, when true model fit is mediocre; that is, $\varepsilon_a = 0.08$. Following the procedure just described, the distribution of the test statistic under H_0 is noncentral χ^2_{d,λ_0}, where $\lambda_0 = (N-1)d(0.05)^2$. For a given α, the critical value χ_c^2 would cut off an area of α in the upper tail of that distribution, as in Figure 3. The distribution of the test statistic under the alternative that $\varepsilon_a = 0.08$ is noncentral χ^2_{d,λ_a}, where $\lambda_a = (N-1)d(0.08)^2$. Power is then given by Equation 9, as is illustrated in Figure 3. Resulting power estimates indicate the probability of rejecting the hypothesis of close fit when true model fit is mediocre. For a given α level, these estimates are dependent only on d and N. For $\alpha = .05$, Table 2 shows power values for selected levels of d and N in rows labeled *close fit*. For example, with $d = 40$ and $N = 200$, the probability of rejecting H_0: $\varepsilon \leq 0.05$ is approximately .69 if $\varepsilon_a = 0.08$. Inspection of power estimates in Table 2 for the test of close fit indicate that power is consistently low when d is small even when N is relatively large. For studies with moderate to large d, reasonable power is achieved with moderate sample sizes, and very high power is achieved with large samples. For instance, with $d = 100$, power is well above 0.90 if N is 200 or more.

A second case of interest for power analysis involves the hypothesis of not-close fit described earlier. In this case, $\varepsilon_0 = 0.05$ and, as mentioned earlier, we recommend setting $\varepsilon_a = 0.01$. Then the distribution of the test statistic under H_0 is noncentral χ^2_{d,λ_0}, where $\lambda_0 = (N-1)d(0.05)^2$. The critical value χ_c^2 cuts off an area of α in the lower tail of this distribution, as in Figure 4. The distribution of the test statistic under the alternative that $\varepsilon_a = 0.05$ is noncentral χ^2_{d,λ_a}, where $\lambda_a = (N-1)d(0.01)^2$, and power is given by Equation

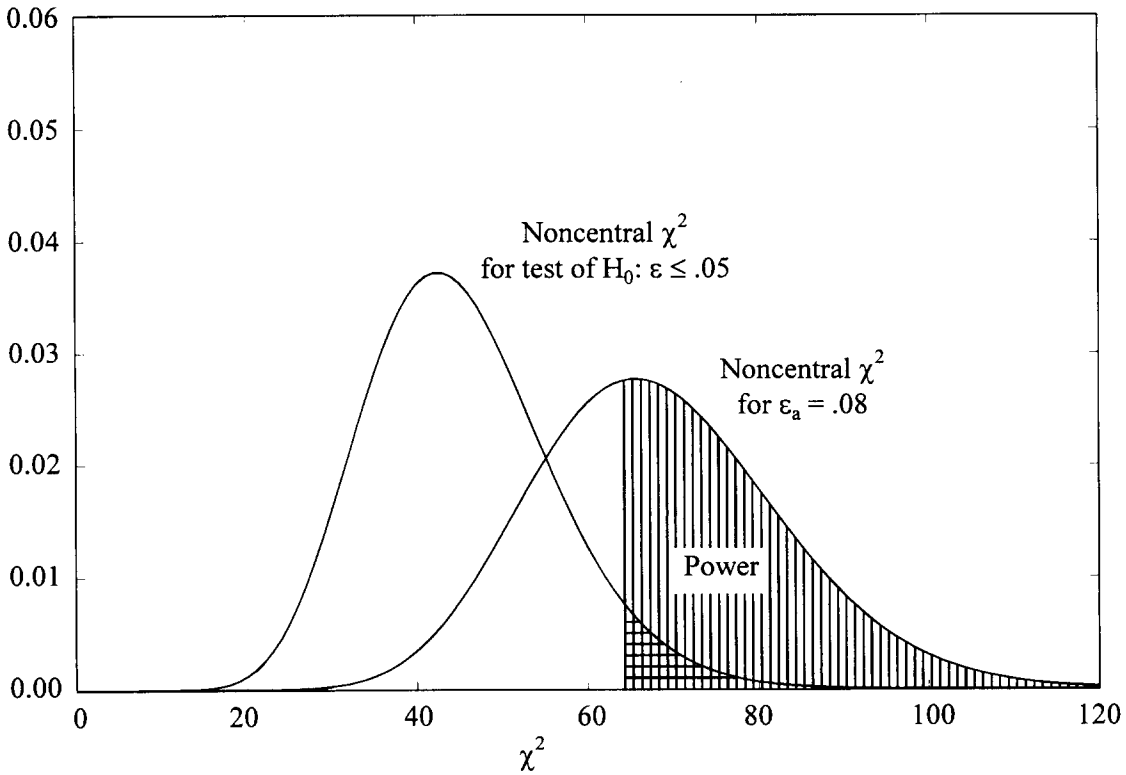

Figure 3. Illustration of power for test of close fit.

10 and illustrated in Figure 4. For selected values of d and N, power values for this condition are shown in Table 2 in the rows labeled *not-close* fit. These power values are a bit smaller than those for the test of close fit when d and N are not large. This finding is a result of the effective difference in effect size represented in the two cases considered here. For the test of close fit, the effect size is represented by the pair of values $\varepsilon_0 = 0.05$ and $\varepsilon_a = 0.08$; for the test of not-close fit, the effect size is reflected by the pair of values $\varepsilon_0 = 0.05$ and $\varepsilon_a = 0.01$. Although the arithmetic difference is larger in the latter case, power analysis results in Table 2 show the latter case to represent an effectively smaller effect size than does the former. That is, except when d and N are both quite large, one would have a better chance of detecting the former effect than the latter.

A third case of interest in power estimation for model tests involves investigation of the test of exact fit when true fit is close. Consider the power of the test of H_0: $\varepsilon = 0$ when the true fit of the model is $\varepsilon_a = 0.05$. Under H_0, the distribution of the test statistic is central χ^2_d, and the critical χ^2_c cuts off an area of α in the upper tail of that distribution. The distribution of the test statistic under the alternative is noncentral χ^2_{d,λ_a}, where $\lambda_a = (N - 1)d(0.05)^2$, and power is given by Equation 9. Graphically, this case corresponds to Figure 3, except that the null distribution in the present case is central rather than noncentral χ^2. Power values for this case indicate the probability for rejecting the hypothesis of exact fit when true fit is close. This phenomenon is often considered to represent a serious problem inherent in the test of exact fit. Table 2 shows power values for this case for selected levels of d and N. Again, values are of roughly the same magnitude as for the other tests considered, with power becoming quite high as d and N increase. One might be tempted to draw a conclusion that it is desirable to have low d and N when testing exact fit, so as to have low probability of rejecting a good model. However, under such conditions power is low for both of the other tests considered also. For instance, for $d = 15$, $N = 100$, $\alpha = .05$, and $\varepsilon_a = 0.05$, power

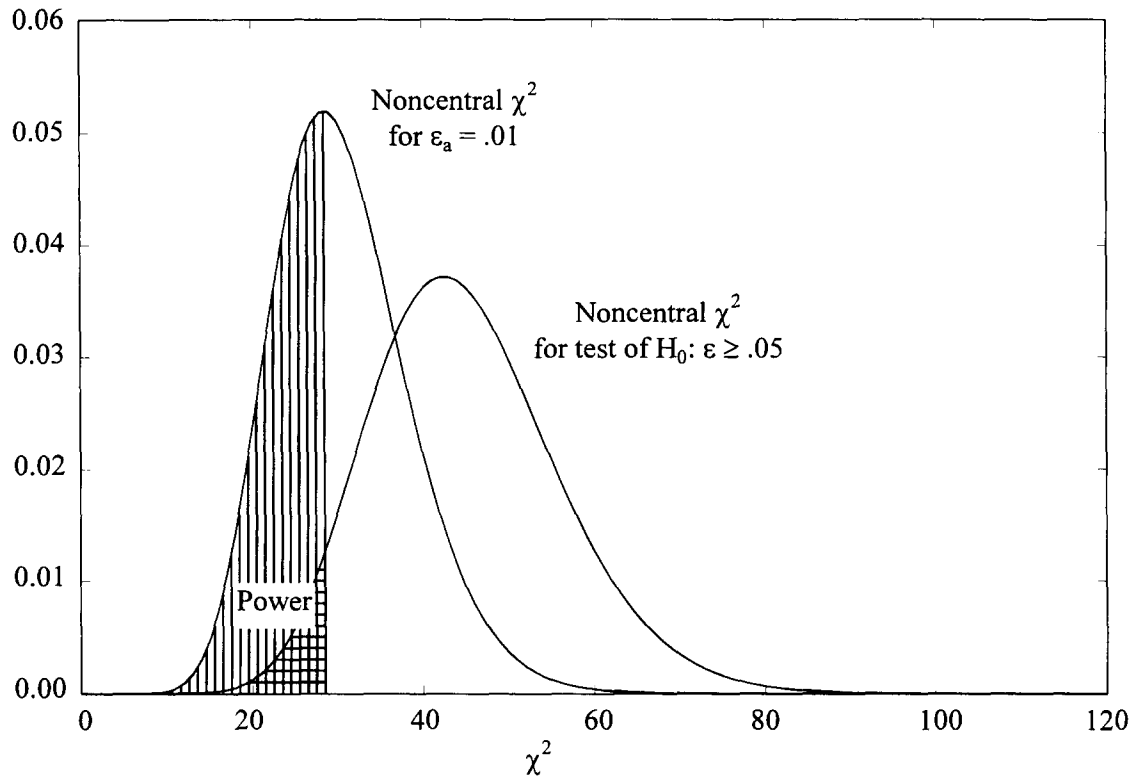

Figure 4. Illustration of power for test of not-close fit.

for the test of exact fit is approximately 0.17, but power for the test of not-close fit is only 0.13. It would be difficult under these conditions to reject either exact fit or not-close fit. The problem here, as discussed earlier, is that under such conditions the confidence interval for ε is quite wide, leaving the investigator with imprecise information about model fit in the population.

For all conditions depicted in Table 2, power increases as d or N increases. This phenomenon can be understood by referring to Equation 8 and Figures 3 and 4. Power is a function of the separation of the distributions in Figures 3 and 4, which is a function of the difference between the noncentrality parameters for the two distributions, λ_0 and λ_a where $\lambda_0 = (N-1)d\varepsilon_0^2$, and $\lambda_a = (N-1)d\varepsilon_a^2$. Clearly, the difference between λ_0 and λ_a is a function of d, N, ε_0, and ε_a. Holding any three of these terms constant, any change in the fourth term that produces a greater difference between λ_0 and λ_a will increase power. Thus, power increases with larger N and with ε_a more discrepant from a fixed ε_0. Furthermore, for fixed N, ε_0, and ε_a, power is greater in models with higher d. That is, a given effect size defined in terms of ε_0 and ε_a is more easily detected when d is higher.

The power estimates computed by the method we have presented can also be interpreted with reference to CIs for ε. Given α, d, N, ε_0, and ε_a, the resulting power can be interpreted as the probability that if $\varepsilon = \varepsilon_a$, the CI will not include ε_0. For example, for $\alpha = .05$, $d = 40$, $N = 200$, $\varepsilon_0 = 0.05$, and $\varepsilon_a = 0.08$, power from Table 2 is 0.69. As explained earlier, this is the probability of rejecting the hypothesis of close fit under these conditions. It is also the probability that a 90% CI for ε will not include the value .05. As is shown in Table 1, the latter event implies the former.

Computer Programs for Power Calculations

The computations involved in these analyses can be carried out easily following the methods described in conjunction with Equations 9 and 10. The Appendix provides a short SAS program for computing power given specified values of α, d, N, ε_0, and ε_a. Note especially that the program allows the user to specify values of ε_0 and ε_a. Thus,

Table 2
Power Estimates for Selected Levels of Degrees of Freedom (df) and Sample Size

df	and test	Sample size 100	200	300	400	500
5	Close	0.127	0.199	0.269	0.335	0.397
	Not close	0.081	0.124	0.181	0.248	0.324
	Exact	0.112	0.188	0.273	0.362	0.449
10	Close	0.169	0.294	0.413	0.520	0.612
	Not close	0.105	0.191	0.304	0.429	0.555
	Exact	0.141	0.266	0.406	0.541	0.661
15	Close	0.206	0.378	0.533	0.661	0.760
	Not close	0.127	0.254	0.414	0.578	0.720
	Exact	0.167	0.336	0.516	0.675	0.797
20	Close	0.241	0.454	0.633	0.766	0.855
	Not close	0.148	0.314	0.513	0.695	0.830
	Exact	0.192	0.400	0.609	0.773	0.882
30	Close	0.307	0.585	0.780	0.893	0.951
	Not close	0.187	0.424	0.673	0.850	0.943
	Exact	0.237	0.512	0.750	0.894	0.962
40	Close	0.368	0.688	0.872	0.954	0.985
	Not close	0.224	0.523	0.788	0.930	0.982
	Exact	0.279	0.606	0.843	0.952	0.988
50	Close	0.424	0.769	0.928	0.981	0.995
	Not close	0.261	0.608	0.866	0.969	0.995
	Exact	0.319	0.684	0.903	0.979	0.997
60	Close	0.477	0.831	0.960	0.992	0.999
	Not close	0.296	0.681	0.917	0.987	0.999
	Exact	0.356	0.748	0.941	0.991	0.999
70	Close	0.525	0.877	0.978	0.997	1.000
	Not close	0.330	0.743	0.949	0.994	1.000
	Exact	0.393	0.801	0.965	0.996	1.000
80	Close	0.570	0.911	0.988	0.999	1.000
	Not close	0.363	0.794	0.970	0.998	1.000
	Exact	0.427	0.843	0.979	0.998	1.000
90	Close	0.612	0.937	0.994	1.000	1.000
	Not close	0.395	0.836	0.982	0.999	1.000
	Exact	0.460	0.877	0.988	0.999	1.000
100	Close	0.650	0.955	0.997	1.000	1.000
	Not close	0.426	0.870	0.990	1.000	1.000
	Exact	0.491	0.904	0.993	1.000	1.000

Note. All power estimates are based on $\alpha = .05$. For the test of close fit, $\varepsilon_0 = 0.05$ and $\varepsilon_a = 0.08$, where ε_0 is the null value of the root-mean-square error of approximation (RMSEA) and ε_a is the alternative value of RMSEA. For the test of not-close fit, $\varepsilon_0 = 0.05$ and $\varepsilon_a = 0.01$. For the test of exact fit, $\varepsilon_0 = 0.00$ and $\varepsilon_a = 0.05$.

if the user wishes to study or use values of these quantities different from those we have suggested (e.g., to define a different criterion for close fit other than $\varepsilon \leq 0.05$), he or she is free to do so. However, we urge users to carefully consider justification for selected values of these quantities.

Examples of Power Calculations for Empirical Studies

We illustrate the power analyses described above by using data from five published applications of CSM. These include the two factor analytic studies mentioned earlier and presented by Browne and Cudeck (1993), as well as three additional studies wherein a particular model was supported (Fredricks & Dossett, 1983; Meyer & Gellatly, 1988; Vance & Colella, 1990). Table 3 shows d and N for a supported model from each study, along with power estimates for tests of close and not-close fit. For the McGaw and Jöreskog (1971) data wherein d and N are both very large, power is essentially 1.0 for both tests. The Naglieri and Jensen (1987) and Fredricks and Dossett (1983) examples show moderate power for both tests. The former data set is characterized by very high d, but rather low N, with the latter having moderate levels of both d and N. The final two data sets in Table 3 show that when d and N are low, power is extremely low for both tests. Such circumstances are problematic in practice, resulting in a very low likelihood of rejecting any sensible hypothesis about fit. In those circumstances, statements of support of models must be considered highly suspect.

Determination of Necessary Sample Size

An important issue in research design involves the determination of sample size necessary to achieve adequate power to carry out planned hypothesis tests. In the present context of testing hypotheses about model fit, it would be desirable to be able to determine necessary N to have adequate power for detecting when such hypotheses are false. In the previous section we provided procedures for power calculation given α, d, N, ε_0, and ε_a. We now consider the closely related problem of determining N, given α, d, ε_0, ε_a, and the desired level of power, π_d. A solution to this problem could be of value in research design by providing investigators with a mechanism for determining necessary N for model testing in CSM studies, thereby avoiding waste and low-power investigations.

This problem can be solved easily in the present framework for hypothesis testing. The solution is

Table 3
Power Estimates and Minimum Sample Sizes for Selected Empirical Studies

Source of data	d	N	Power		Minimum N	
			Close	Not close	Close	Not close
McGaw & Jöreskog (1971)	132	11,743	>0.999	>0.999	110	152
Naglieri & Jensen (1987)	166	86	0.747	0.502	95	133
Fredricks & Dossett (1983)	34	236	0.712	0.566	285	340
Meyer & Gellatly (1988)	8	56	0.107	0.073	954	875
Vance & Colella (1990)	5	90	0.120	0.077	1,463	1,216

Note. For all analyses, $\alpha = .05$. For the test of close fit, $\varepsilon_0 = 0.05$ and $\varepsilon_a = 0.08$, where ε_0 is the null value of the root-mean-square error of approximation (RMSEA) and ε_a is the alternative value of RMSEA. For the test of not-close fit, $\varepsilon_0 = 0.05$ and $\varepsilon_a = 0.01$.

not a direct one, however. If we define N_{min} as the minimum value of N being sought, it is not possible to calculate N_{min} directly from the other relevant factors. Rather, it is necessary to conduct a systematic search for the appropriate value of N_{min}. We use a simple procedure of interval-halving. In this procedure, upper and lower bounds are determined to contain the value of N_{min}, and that interval is successively cut in half in a systematic manner until a very close approximation to the desired N_{min} is found. Details of the procedure, along with a SAS program, are provided in the Appendix. Although it would be possible to use more computationally sophisticated procedures that would arrive at a solution more quickly, we have found the interval-halving procedure to work effectively and quickly, usually in just a few seconds on a PC. Furthermore, this procedure is easy to explain and allows us to provide a simple SAS program to interested users. As with the SAS program for power calculation, note that the user is free to choose values of ε_0 and ε_a. This flexibility, however, must not be abused. Users have the responsibility for justifying their choice of these values.

Consider the application of this procedure in the case in which one plans to test H_0: $\varepsilon \leq 0.05$ when $\varepsilon_a = 0.08$, using $\alpha = 0.05$ and a desired power $\pi_d = 0.80$. Given these conditions, N_{min} depends only on degrees of freedom d. Table 4 shows minimum levels of N for this case for selected levels of d from 2 to 100. For example, for $d = 40$, $N_{min} = 252$ to assure power of at least 0.80 for rejecting the hypothesis of close fit if $\varepsilon_a = 0.08$. Also shown in Table 4 are minimum levels of N for the test of not-close fit, H_0: $\varepsilon \geq 0.05$ when $\varepsilon_a = 0.01$. Once again, this information can be interpreted equivalently in terms of CIs for ε. Given ε_0, ε_a, d, α, and desired power, N_{min} can be interpreted as the minimum sample size required to have the desired probability (power) for the appropriate CI to not include ε_0. As N increases, a CI for ε becomes narrower, thus reducing the likelihood of it including ε_0, which, as is indicated in Table 1, implies rejection of the null hypothesis.

Inspection of Table 4 reveals several interesting phenomena. Most obvious is the strong association between d and N_{min}. When d is small, a very large N is needed to achieve adequate power for these model tests. Studies with small d arise when the number of measured variables is small, when the specified model has a relatively large number of parameters, or both. In such cases, as is seen in Table 2, power is low for almost any sensible hypothesis test; Table 4 indicates that reasonable levels of power cannot be obtained without a very large N.

The relevant phenomenon in such cases involves the relationship of the width of the CI for ε to the levels of d and N. When d is small, these CIs will be very wide unless N is extremely large. Thus, $\hat{\varepsilon}$ is subject to considerable imprecision. To achieve adequate precision and in turn adequate power for the recommended hypothesis tests when d is small, N must exceed the levels shown in Table 4. Given these results, we discourage attempts to evaluate models with low d unless N is extremely large. In conjunction with this view, we discourage the introduction of substantial numbers of parameters into models so as to improve their fit. Such procedures have been shown to be susceptible to capitalization on chance (MacCallum, 1986; MacCallum, Roznowski, & Necowitz, 1992). Furthermore, it is now clear that the resulting reduction

Table 4
Minimum Sample Size to Achieve Power of 0.80 for Selected Levels of Degrees of Freedom (df)

df	Minimum *N* for test of close fit	Minimum *N* for test of not-close fit
2	3,488	2,382
4	1,807	1,426
6	1,238	1,069
8	954	875
10	782	750
12	666	663
14	585	598
16	522	547
18	472	508
20	435	474
25	363	411
30	314	366
35	279	333
40	252	307
45	231	286
50	214	268
55	200	253
60	187	240
65	177	229
70	168	219
75	161	210
80	154	202
85	147	195
90	142	189
95	136	183
100	132	178

Note. For all analyses, $\alpha = .05$. For the test of close fit, $\varepsilon_0 = 0.05$ and $\varepsilon_a = 0.08$, where ε_0 is the null value of the root-mean-square error of approximation (RMSEA) and ε_a is the alternative value of RMSEA. For the test of not-close fit, $\varepsilon_0 = 0.05$ and $\varepsilon_a = 0.01$.

in d causes substantial reduction in power of model tests.

Let us next focus on levels of necessary N as d becomes larger. As is indicated in Table 4, adequate power for the recommended tests can be achieved with relatively moderate levels of N when d is not small. For instance, with $d = 100$, a power of 0.80 for the test of close fit (in comparison with the alternative that $\varepsilon_a = 0.08$) is achieved with $N = 132$. Again, such results reflect the behavior of CIs for ε. With large d, relatively narrow CIs are obtained with only moderate N. This phenomenon has important implications for tests of model fit using hypotheses about ε. For instance, using the test of close fit, if d is large and actual fit is mediocre or worse, one does not need a very large sample to have a high probability of rejecting the false null hypothesis. Consider a specific example to illustrate this point. Suppose one has $p = 30$ manifest variables, in which case there would be $p(p + 1)/2 = 465$ distinct elements in the $p \times p$ covariance matrix. If we tested the null model that the measured variables are uncorrelated, the model would have $q = 30$ parameters (variances of the manifest variables), resulting in $d = p(p + 1)/2 - q = 435$. For the test of close fit, in comparison with the alternative that $\varepsilon_a = 0.08$, we would find $N_{min} = 53$ for power of 0.80. That is, we would not need a large sample to reject the hypothesis that a model specifying uncorrelated measured variables holds closely in the population. In general, our results indicate that if d is high, adequately powerful tests of fit can be carried out on models with moderate N.

This finding must be applied cautiously in practice. Some applications of CSM may involve models with extremely large d. For instance, factor analytic studies of test items can result in models with $d > 2000$ when the number of items is as high as 70 or more. For a model with $d = 2000$, a power of 0.80 for the test of close fit (in comparison with the alternative that $\varepsilon_a = 0.08$) can be achieved with $N_{min} = 23$ according to the procedures we have described. Such a statement is not meaningful in practice for at least two reasons. First, one must have $N \geq p$ to conduct parameter estimation using the common ML method. Second, and more important, our framework for power analysis is based on asymptotic distribution theory, which holds only with sufficiently large N. The noncentral χ^2 distributions on which power and sample size calculations are based probably do not hold their form well as N becomes small, resulting in inaccurate estimates of power and minimum N. Therefore, results that indicate a small value of N_{min} should be treated with caution. Finally, it must also be recognized that we are considering determination of N_{min} only for the purpose of model testing. The magnitude of N affects other aspects of CSM results, and an N that is adequate for one purpose might not be adequate for other purposes. For example, whereas a moderate N might be adequate for achieving a specified level of power for a test of overall fit, the same level of N may not necessarily be adequate for obtaining precise parameter estimates.

Table 5
Minimum Sample Sizes for Test of Exact Fit for Selected Levels of Degrees of Freedom (df) and Power

df	Minimum *N* for power = 0.80	Minimum *N* for power = 0.50
2	1,926	994
4	1,194	644
6	910	502
8	754	422
10	651	369
12	579	332
14	525	304
16	483	280
18	449	262
20	421	247
25	368	218
30	329	196
35	300	180
40	277	167
45	258	157
50	243	148
55	230	140
60	218	134
65	209	128
70	200	123
75	193	119
80	186	115
85	179	111
90	174	108
95	168	105
100	164	102

Note. The $\alpha = .05$, $\varepsilon_0 = 0.0$, and $\varepsilon_a = 0.05$, where ε_0 is the null value of the root-mean-square error of approximation (RMSEA) and ε_a is the alternative value of RMSEA.

An additional phenomenon of interest shown by the results in Table 4 is that the N_{min} values for the two cases cross over as d increases. At low values of d, N_{min} for the test of close fit is larger than N_{min} for the test of not-close fit. For $d > 14$, the relationship is reversed. This phenomenon is attributable to the interactive effect of effect size and d on power. The effect size represented by the test of close fit ($\varepsilon_0 = 0.05$ and $\varepsilon_a = 0.08$) is an effectively larger effect size than that for the test of not-close fit ($\varepsilon_0 = 0.05$ and $\varepsilon_a = 0.01$) at higher levels of d but is effectively smaller at lower levels of d.

Let us finally consider determination of N_{min} for a third case of interest, the test of exact fit when $\varepsilon_a = 0.05$. Using $\alpha = .05$, Table 5 shows values of N_{min} for selected levels of d, for two levels of desired power, 0.80 and 0.50. These results provide explicit information about the commonly recognized problem with the test of exact fit. Levels of N_{min} for power of 0.80 reflect sample sizes that would result in a high likelihood of rejecting the hypothesis of exact fit when true fit is close. Corresponding levels of N_{min} for power of 0.50 reflect sample sizes that would result in a better than 50% chance of the same outcome. For instance, with $d = 50$ and $N \geq 243$, the likelihood of rejecting the hypothesis of exact fit would be at least .80, even though the true fit is close. Under the same conditions, power would be greater than 0.50 with $N \geq 148$. As d increases, the levels of N that produce such outcomes become much smaller. These results provide a clear basis for recommending against the use of the test of exact fit for evaluating covariance structure models. Our results show clearly that use of this test would routinely result in rejection of close-fitting models in studies with moderate to large sample sizes. Furthermore, it is possible to specify and test hypotheses about model fit that are much more empirically relevant and realistic, as has been described earlier in this article.

For the five empirical studies discussed earlier in this article, Table 3 shows values of N_{min} for achieving power of 0.80 for the tests of close fit and not-close fit. These results are consistent with the phenomena discussed earlier in this section. Most important is the fact that rigorous evaluation of fit for models with low d, such as those studied by Meyer and Gellatly (1988) and Vance and Colella (1990), requires extremely large N. Such models are not rare in the literature. Our results indicate that model evaluation in such cases is highly problematic and probably should not be undertaken unless very large samples are available.

Comparison to Other Methods for Power Analysis in CSM

As mentioned earlier, there exists previous literature on power analysis in CSM. Satorra and Saris (1983, 1985; Saris & Satorra, 1993) have proposed a number of techniques for evaluating power of the test of exact fit for a specific model. The methods presented in this earlier work are based on the same assumptions and distributional approxima-

tions as the methods proposed in this article. The major difference between our approach and that of Satorra and Saris involves the manner in which effect size is established. In our procedure, effect size is defined in terms of a pair of values, ε_0 and ε_a, where the latter defines the lack of fit of the specified model in the population. These values are used to determine values of noncentrality parameters for the noncentral χ^2 distributions that are used in turn to determine power. Satorra and Saris used a different approach to reach this end requiring the specification of two models. Given a model under study, they defined a specific alternative model that is different from the original model in that it includes additional parameters; the alternative model is treated as the true model in the population. The effect size is then a function of the difference between the original model and the true model. In their earlier procedures (Satorra & Saris, 1983, 1985), it was necessary for the user to completely specify the alternative model, including parameter values. In later procedures (Saris & Satorra, 1993), it is not necessary to specify parameter values, but effect size is still associated with changes in specific model parameters. For all of these methods, using the difference between the model under study and the alternative model, several methods exist (see Bollen, 1989; Saris & Satorra, 1993) for estimating the noncentrality parameter for the distribution of the test statistic under the alternative model. Once that value is obtained, the actual power calculation is carried out by the same procedures we use (see Equation 9).

Our approach to establishing effect size has several beneficial consequences that differentiate it from the approach of Satorra and Saris (1983, 1985; Saris & Satorra, 1993). First, our procedure is not model-specific with regard to either the model under study or the alternative model. The only feature of the model under study that is relevant under our approach is d. Of course, if one wished to evaluate power of the test of exact fit for a given model versus a specific alternative, then the Satorra and Saris procedures would be useful. Second, our procedure allows for power analysis for tests of fit other than the test of exact fit. In this article we have discussed tests of close and not-close fit, along with associated power analyses. Finally, it is quite simple in our framework to determine minimum sample size required to achieve a desired level of power.

Generalizations of Proposed Procedure

There are at least two ways in which the procedure proposed in this article could be generalized. One would be to use the same procedures for power analysis and determination of sample size using a different index of fit. Our approach uses the RMSEA index (ε). The critical features of the approach involve the capability for specifying sensible null and alternative values of ε and to define the noncentrality parameter values for the relevant χ^2 distributions as a function of ε, as in Equation 8. Those distributions then form the basis for power and sample size calculations. The same procedure could be used with a different fit index as the basis for the hypotheses, as long as one could specify noncentrality parameter values as a function of that index. One possible candidate for such a procedure is the goodness-of-fit index called GFI, reported by LISREL (Jöreskog & Sörbom, 1993). Steiger (1989, p. 84) and Maiti and Mukherjee (1990) showed that GFI can be represented as a simple function of $\hat{F}$, the sample discrepancy function value. Given this finding, one could express the noncentrality parameter as a function of the population GFI and proceed with hypothesis tests and power analyses in the same way as we have done, but by basing hypotheses on GFI rather than on RMSEA. We leave this matter for further investigation.

A second generalization involves the potential use of our approach in contexts other than CSM. There are a variety of other contexts involving model estimation and testing that use discrepancy functions and yield an asymptotic χ^2 test of fit. These other contexts involve different types of data structures and models than those used in CSM. Log-linear modeling is a commonly used procedure in this category. For such techniques, it may be quite appropriate to consider tests of hypotheses other than that of exact fit and to conduct power analyses for such tests. The current framework may well be applicable in such contexts, as well as in CSM.

Summary

We have stressed the value of CIs for fit indices in CSM and the relationship of CIs to a simple

framework for testing hypotheses about model fit. The framework allows for the specification and testing of sensible, empirically interesting hypotheses, including null hypotheses of close fit or not-close fit. The capability for testing a null hypothesis of not-close fit eliminates the problem in which the researcher is in the untenable position of seeking to support a null hypothesis of good fit. We have also provided procedures and computer programs for power analysis and determination of minimum levels of sample size that can be used in conjunction with this hypothesis testing framework. These procedures can be applied easily in practice, and we have included simple SAS programs for such applications in the Appendix.

References

Arbuckle, J. L. (1994). AMOS: Analysis of moment structures. *Psychometrika, 59*, 135–137.

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

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Browne, M. W., & Mels, G. (1990). *RAMONA user's guide.* Unpublished report, Department of Psychology, Ohio State University.

Browne, M. W., Mels, G., & Coward, M. (1994). *Path analysis: RAMONA: SYSTAT for DOS: Advanced applications* (Version 6, pp. 167–224). Evanston, IL: SYSTAT.

Champion, D. J. (1981). *Basic statistics for social research* (2nd ed.). New York: Macmillan.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Fredricks, A. J., & Dossett, D. L. (1983). Attitude-behavior relations: A comparison of the Fishbein-Ajzen and the Bentler-Speckart models. *Journal of Personality and Social Psychology, 45*, 501–512.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Hillsdale, NJ: Erlbaum.

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

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490–504.

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

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis. *Psychological Bulletin, 103*, 391–410.

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

Meyer, J. P., & Gellatly, I. R. (1988). Perceived performance norm as a mediator in the effect of assigned goal on personal goal and task performance. *Journal of Applied Psychology, 73*, 410–420.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*, 430–445.

Naglieri, J. A., & Jensen, A. R. (1987). Comparison of black–white differences on the WISC-R and K-ABC: Spearman's hypothesis. *Intelligence, 11*, 21–43.

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181–204). Newbury Park, CA: Sage.

SAS Institute, Inc. (1992). *The CALIS procedure extended user's guide.* Cary, NC: Author.

Satorra, A., & Saris, W. E. (1983). The accuracy of a procedure for calculating the power of the likelihood ratio test as used within the LISREL framework. In C. O. Middendorp (Ed.), *Sociometric research 1982* (pp. 129–190). Amsterdam, The Netherlands: Sociometric Research Foundation.

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

Satorra, A., Saris, W. E., & de Pijper, W. M. (1991). A comparison of several approximations to the power function of the likelihood ratio test in covariance structure analysis. *Statistica Neerlandica, 45*, 173–185.

Steiger, J. H. (1989). *Causal modeling: A supplementary module for SYSTAT and SYGRAPH.* Evanston, IL: SYSTAT.

Steiger, J. H. (1994). *Structural equation modeling with SePath: Technical documentation.* Tulsa, OK: STATSOFT.

Steiger, J. H., & Lind, J. M. (1980, June). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

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

Vance, R. J., & Colella, A. (1990). Effects of two types of feedback on goal acceptance and personal goals. *Journal of Applied Psychology, 75,* 68–76.

Appendix

SAS Programs for Calculating Power and Minimum Sample Size

Power Analysis

Following is an SAS program for computing power of tests of fit on the basis of root-mean-square error of approximation (RMSEA). The user inputs the null and alternative values of RMSEA (ε_0 and ε_a), the α level, degrees of freedom, and sample size. The program computes and prints the power estimate for the specified conditions.

```
title "power estimation for csm";
data one ;
alpha=.05 ; *significance level ;
rmsea0=.05 ; *null hyp value ;
rmseaa=.08 ; *alt hyp value ;
d=50 ; *degrees of freedom ;
n=200 ; *sample size ;
ncp0=(n-1)*d*rmsea0**2 ;
ncpa=(n-1)*d*rmseaa**2 ;
  if rmsea0<rmseaa then do ;
  cval=cinv(1-alpha,d,ncp0) ;
  power=1-probchi(cval,d,ncpa);
  end ;
  if rmsea0>rmseaa then do ;
  cval=cinv(alpha,d,ncp0) ;
  power=probchi(cval,d,ncpa) ;
  end;
  output ;
proc print data=one ; var rmsea0 rmseaa alpha d n power ; run ;
```

Determination of Minimum Sample Size

We first discuss the interval-halving procedure used to determine the minimum value of N required to achieve a given level of power. Given α, d, ε_0, ε_a, and π_d, we begin by setting $N = 100$ and computing actual power, π_a, by methods described in the section on power analysis. We then increase N as necessary by increments of 100 until $\pi_a > \pi_d$. Let the resulting value of N be called the first trial value, N_{t_1}. We then know that the desired minimum value of N, called $N_{\min}$, lies between N_{t_1} and $(N_{t_1} - 100)$. We define a new trial value N_{t_2} as the midpoint of that interval and recompute π_a. We then compare π_a to π_d to determine whether N_{t_2} is too high or too low. If $\pi_a > \pi_d$, then N_{t_2} is still too high, in which case we set the new N_{t_3} as the midpoint of the interval between N_{t_2} and $(N_{t_1} - 100)$. On the other hand, if $\pi_a < \pi_d$, then N_{t_2} is too low, and we set the new N_{t_3} as the midpoint of the interval between N_{t_2} and N_{t_1}. This process is repeated, setting each new trial value of N as the midpoint of the appropriate interval above or below the current trial value, until the difference between π_a and π_d is less than some small threshold, such as 0.001. The resulting value of N can then be rounded up to obtain $N_{\min}$.

Below is an SAS program that follows the procedure just described for computing minimum sample size for tests of fit on the basis of the RMSEA index. The user inputs the null and alternative values of RMSEA (ε_0 and ε_a), the α level, degrees of freedom, and desired level of power. The program computes and prints the minimum necessary sample size to achieve the desired power.

```
title "Computation of min sample size for test of fit";
data one ;
rmsea0=.05 ; *null hyp rmsea ;
rmseaa=.08 ; *alt hyp rmsea ;
```

```
d=20 ; *degrees of freedom ;
alpha=.05 ; *alpha level ;
powd=.80 ; *desired power ;
*initialize values ;
powa=0.0 ;
n = 0 ;
*begin loop for finding initial level of n ;
do until (powa>powd) ;
n + 100 ;
ncp0=(n-1)*d*rmsea0**2 ;
ncpa=(n-1)*d*rmseaa**2 ;
*compute power ;
if rmsea0>rmseaa then do ;
cval = cinv(alpha,d,ncp0) ;
powa = probchi(cval,d,ncpa) ;
end ;
if rmsea0<rmseaa then do ;
cval = cinv(1-alpha,d,ncp0) ;
powa = 1-probchi(cval,d,ncpa) ;
end ;
end ;
* begin loop for interval halving ;
dir=-1 ;
newn=n ;
intv=200 ;
powdiff=powa-powd ;
do until (powdiff<.001) ;
intv=intv*.5 ;
newn + dir*intv*.5 ;
*compute new power ;
ncp0=(newn-1)*d*rmsea0**2 ;
ncpa=(newn-1)*d*rmseaa**2 ;
*compute power ;
if rmsea0>rmseaa then do ;
cval = cinv(alpha,d,ncp0) ;
powa = probchi(cval,d,ncpa) ;
end ;
if rmsea0<rmseaa then do ;
cval = cinv(1-alpha,d,ncp0) ;
powa = 1-probchi(cval,d,ncpa);
end ;
powdiff=abs(powa-powd) ;
if powa<powd then dir=1; else dir=-1 ;
end ;
minn=newn ;
output ;
proc print data=one;
var rmsea0 rmseaa powd alpha d minn powa ; run ;
```

Received August 15, 1994
Revision received June 12, 1995
Accepted June 15, 1995 ■