Psychological Methods
2006, Vol. 11, No. 1, 19–35

Copyright 2006 by the American Psychological Association
1082-989X/06/$12.00 DOI: 10.1037/1082-989X.11.1.19

Testing Differences Between Nested Covariance Structure Models: Power Analysis and Null Hypotheses

Robert C. MacCallum
University of North Carolina at Chapel Hill

Michael W. Browne
Ohio State University

Li Cai
University of North Carolina at Chapel Hill

For comparing nested covariance structure models, the standard procedure is the likelihood ratio test of the difference in fit, where the null hypothesis is that the models fit identically in the population. A procedure for determining statistical power of this test is presented where effect size is based on a specified difference in overall fit of the models. A modification of the standard null hypothesis of zero difference in fit is proposed allowing for testing an interval hypothesis that the difference in fit between models is small, rather than zero. These developments are combined yielding a procedure for estimating power of a test of a null hypothesis of small difference in fit versus an alternative hypothesis of larger difference.

Keywords: power analysis, structural equation modeling

When using covariance structure modeling in any of its variety of forms and special cases, researchers routinely encounter the need to evaluate differences between two or more models. Model comparisons are commonly used for purposes such as selection from among competing theoretical models representing different patterns of relationships among latent variables, comparison of models representing different degrees of measurement invariance across groups or across measurement occasions, comparison of alternative latent growth curve models representing different patterns of change over time, or determination of an optimal number of factors to retain in exploratory factor analysis. All of these questions, and many more, are routinely addressed by fitting the alternative models to sample data and then conducting statistical tests of the difference in fit between the models.

By far, the most common statistical method used for such model comparisons is the likelihood ratio test (to be defined in detail below), which tests the null hypothesis that the fit of two models is identical in the population. The present article addresses two issues relevant to the use of this test. The first is statistical power. Power analysis for a test of the difference between two models requires the investigator to define an effect size representing the degree of difference under the alternative hypothesis. One procedure for doing so, based on the work of Satorra and Saris (1985), defines effect size in terms of the values of parameters that differentiate the two models. In this article, we propose an alternative procedure in which effect size is defined in terms of the difference in overall fit of the two models. We discuss differences between these approaches along with circumstances in which each would be useful. A procedure for determining minimum sample size necessary to achieve a desired level of power is also presented.

The second issue addressed in this article involves the specification of the null hypothesis. The null hypothesis of zero difference in model fit, like virtually any point hypothesis, will always be false in empirical studies. Even when the true difference in fit between the two models is very small, with an adequately large sample, this null hypothesis will always be rejected, implying a statistically significant difference and thus favoring the less constrained model. To address these problems, we propose the use of an interval null hypothesis for evaluating model differences where the null hypothesis specifies that the difference in fit between the two models is small, rather than zero. We suggest that such a null hypothesis is of more empirical interest than the traditional null hypothesis of zero difference, and we provide a method for testing such a null hypothesis. Finally, we

Robert C. MacCallum and Li Cai, Department of Psychology, University of North Carolina at Chapel Hill; Michael W. Browne, Department of Psychology, Ohio State University.

Additional materials are on the Web at http://dx.doi.org/10.1037/1082-989X.11.1.19.supp.

Correspondence concerning this article should be addressed to Robert C. MacCallum, Department of Psychology, CB#3270 Davie Hall, University of North Carolina, Chapel Hill, NC 27599-3270. E-mail: rcm@email.unc.edu

combine the two developments just described, allowing for power analysis where the null hypothesis specifies a small difference in model fit and the alternative hypothesis specifies a larger difference. To facilitate implementation of the methods presented in this article, we provide a number of SAS programs described later in the article. These materials are also on the Web at http://dx.doi.org/10.1037/1082-989X.11.1.19.supp.

Much of the work presented here is an extension of developments for power analysis and hypothesis tests for fit of single covariance structure models by MacCallum, Browne, and Sugawara (1996). The current article uses a similar rationale and statistical theory to extend that earlier work to the current context of evaluating differences between models.

Power Analysis for Tests of Differences Between Models

Statistical Theory and Methods

We consider the case in which we wish to assess the difference between two covariance structure models, designated Models A and B, where Model A is nested in Model B. *Nesting* refers to the property that the set of possible covariance matrices generated by Model A is a subset of those generated by Model B. The most common form of nesting involves the free parameters in Model A being a subset of those in Model B. Stated another way, Model A includes all of the same constraints as Model B, plus additional constraints. Let degrees of freedom for these two models be designated d_A and d_B, and let sample size be designated N, with $n = (N - 1)$. Let us assume for the moment that model estimation is to be conducted using maximum likelihood (ML), which minimizes the ML discrepancy function defined as

$$F_{ML} = \ln|\hat{\Sigma}| - \ln|S| + tr(S\hat{\Sigma}^{-1}) - p, \tag{1}$$

where ln is natural logarithm, S is the sample covariance matrix, $\hat{\Sigma}$ is the covariance matrix implied by the model, and p is the number of measured variables. Let sample values of this discrepancy function be designated $\hat{F}$ and population values be designated F^*. A likelihood ratio test can be conducted of the null hypothesis (H_0) that there is no difference in the population discrepancy function values for the two models. That is, $H_0 : (F_A^* - F_B^*) = 0$. Under this H_0, and assuming multivariate normality, the test statistic $T = n(\hat{F}_A - \hat{F}_B)$ will asymptotically follow a χ^2 distribution with degrees of freedom $d_{A-B} = d_A - d_B$. In practice, a finding of a significant test statistic results in rejection of H_0 and a conclusion that there is a statistically significant difference in the fit of the models, with Model B fitting better than Model A. Failure to reject H_0, on the other hand, indicates that the more constrained Model A would be preferred.

Present developments focus on the statistical power of this test. That is, we wished to examine the probability of rejecting the null hypothesis of zero difference in population discrepancy function values when those values in fact differ to some specified degree. Determination of statistical power requires knowledge of the distribution of the test statistic, T, when H_0 is true, as given above, as well as the distribution when H_0 is false. When H_0 is false, an alternative hypothesis (H_1) of the form $H_1 : (F_A^* - F_B^*) = \delta$ must be true, where $\delta > 0$. Under the same assumptions given above, plus one new assumption to be defined shortly, the distribution of T when H_0 is false will be approximately noncentral χ^2 with degrees of freedom $d_{A-B} = (d_A - d_B)$ and noncentrality parameter

$$\lambda = n(F_A^* - F_B^*) = n\delta. \tag{2}$$

Use of the noncentral χ^2 in this context requires an additional assumption called the *population drift* (or parameter drift) *assumption*, which in essence states that neither Model A nor Model B is badly misspecified. For each model, lack of fit due to model misspecification is assumed to be of approximately the same magnitude as lack of fit due to sampling error. For a more formal statement and discussion of the population drift assumption, see Steiger, Shapiro, and Browne (1985, p. 256). Thus, under an alternative hypothesis, $H_1 : (F_A^* - F_B^*) = \delta$, the distribution of the test statistic, T, is defined.

Let us assume for the moment that a value of the noncentrality parameter λ defined in Equation 2 has been specified. (Later in this article, we will discuss in detail ways in which an appropriate value of λ can be established.) Once λ is specified, the computation of statistical power is straightforward. The distribution of the test statistic under H_0 is central χ^2 with degrees of freedom $d_{A-B} = (d_A - d_B)$. From this distribution and given a specified α level, one can determine a critical value of χ^2, designated χ_C^2. The alternative hypothesis is then defined as $H_1 : (F_A^* - F_B^*) = \delta$. The distribution of the test statistic under H_1 is noncentral χ^2 with $d_{A-B} = (d_A - d_B)$ and noncentrality parameter λ. Statistical power is then obtained by computing the area to the right of χ_C^2 under this alternative distribution. Distributions and relevant areas for a typical case are illustrated in Figure 1. The shaded area in the figure shows the probability of rejecting $H_0 : (F_A^* - F_B^*) = 0$ if in fact $H_1 : (F_A^* - F_B^*) = \delta = \lambda/n$ is true.

Specifying a Value of the Noncentrality Parameter

Approach based on parametric misspecification. The procedure just described requires the investigator to provide a value of the noncentrality parameter, λ. Of course, this value is not known in practice. Considering Equation 2, the

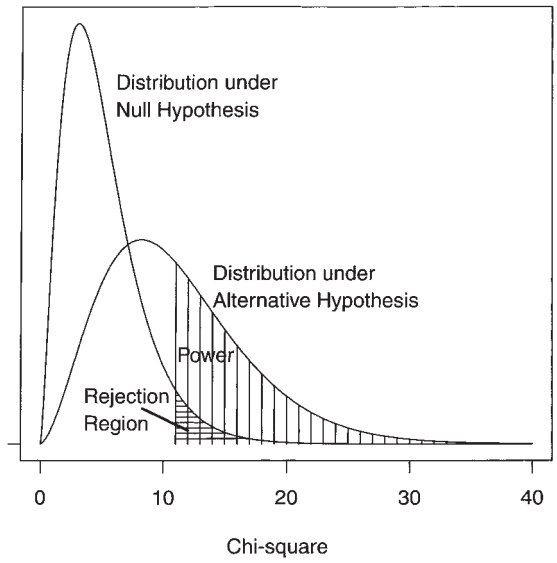

Figure 1. Null and alternative distributions of the test statistic for determining statistical power. The distribution to the left is a central χ^2 distribution representing the distribution of the test statistic under the null hypothesis of no difference in model fit. The distribution to the right is a noncentral χ^2 distribution representing the distribution of the test statistic under the alternative hypothesis of a specified nonzero difference in fit. The critical value of χ^2 for the test of no difference is defined with respect to the central χ^2 distribution, and power is the area under the noncentral χ^2 distribution to the right of that critical value.

investigator will not know the values of F_A^* or F_B^*. A procedure is needed for establishing an appropriate value of λ so that statistical power can be computed.

One option for accomplishing this step can be drawn from the work of Satorra and Saris (1985) on power analysis procedures for tests of fit of single models. Adapting their approach to the case of two competing models, with Model A nested in Model B, the first step is to establish values for all parameters in Model B. This could be done in a variety of ways. One approach would be to fit Model B to sample data and use the resulting estimates as parameter values. Another would be to fit Model A to sample data, use the resulting parameter estimates as parameter values for those parameters in Model B that both models have in common, and then assign numerical values to the additional parameters in Model B. Yet another method would be to assign parameter values of interest to all parameters in Model B. In any event, given such a set of parameter values for Model B, the implied covariance matrix Σ_B can then be computed. Model A is then fit to Σ_B. Note that because Model A is nested in Model B, Model A will probably not fit Σ_B perfectly and will yield a nonzero discrepancy function value designated F_A^*. Note also that Model B will fit Σ_B perfectly, yielding $F_B^* = 0$. The noncentrality parameter for the distribution of the test statistic, T, for the test of $H_0 : (F_A^* - F_B^*) = 0$ is then defined as $\lambda = nF_A^*$. (Note that in general $\lambda = n(F_A^* - F_B^*)$ but that $F_B^* = 0$ in the present context.) This procedure thus provides a value of the noncentrality parameter, λ, and computation of statistical power then proceeds as described earlier and as illustrated in Figure 1. The resulting power indicates the probability of rejecting the null hypothesis $H_0 : (F_A^* - F_B^*) = 0$ if Model B is correct in the population. Satorra and Saris (1985) provided a simple example.

In earlier work, Satorra and Saris (1983) presented a method allowing for the case where both Models A and B could be considered as misspecified in the population. That procedure requires in effect that some less restrictive Model C be considered as correct. From Model C, with all parameter values specified, one then generates Σ_C, and both Models A and B are fit to Σ_C. This process yields F_A^* and F_B^*, neither of which will be zero in all likelihood, and the noncentrality parameter is then obtained as $\lambda = n(F_A^* - F_B^*)$. Power computation then proceeds as indicated above, with the result representing the likelihood of rejecting the null hypothesis, $H_0 : (F_A^* - F_B^*) = 0$, when a less restricted Model C is correct. Compared with the Satorra–Saris (1985) procedure, this procedure requires further information from the user so as to construct Σ_C. For purposes of this article, we will focus on the more well-known Satorra–Saris (1985) procedure, although readers should keep in mind that this earlier alternative exists.

Approach based on overall model fit. In the present article, we propose an alternative approach to establishing an appropriate value of the noncentrality parameter. Whereas the procedure based on the Satorra–Saris (1985) framework defines the difference between the two models in terms of values of parameters that differentiate them, our proposed approach defines the difference in terms of overall model fit.

Considering again Equation 2, the problem is to establish a value of λ. Thus, the problem in effect is how to establish sensible or interesting values of F_A^* and F_B^*. In this context, we define effect size as the value of $\delta = (F_A^* - F_B^*)$, which represents the degree to which the null hypothesis is false. (The noncentrality parameter is then given by $\lambda = n\delta$.) Specifying an effect size, δ, is difficult if one attempts to choose values of F_A^* and F_B^* using the ML discrepancy function defined in Equation 1. The scale of this function is quite difficult to work with directly. Researchers need some rational method for determining useful values of F_A^* and F_B^* that does not require direct use of the function F_{ML}. One viable option is to establish these values by using a measure of goodness of fit that is directly related to the population discrepancy function values for the two models, F_A^* and F_B^*. An obvious candidate is the root-mean-square error of approximation (RMSEA) measure (Browne & Cudeck, 1993; Steiger & Lind, 1980), which will be designated throughout

this article as ε. In the population, RMSEA is defined for Models A and B as $\varepsilon_A = \sqrt{F_A^*/d_A}$ and $\varepsilon_B = \sqrt{F_B^*/d_B}$, respectively. Solving these expressions for values of F^*, we obtain $F_A^* = d_A\varepsilon_A^2$ and $F_B^* = d_B\varepsilon_B^2$. From these relationships, we can define effect size as

$$\delta = (F_A^* - F_B^*) = (d_A\varepsilon_A^2 - d_B\varepsilon_B^2). \quad (3)$$

In turn, the noncentrality parameter can be specified in terms of RMSEA values as follows:

$$\lambda = n(F_A^* - F_B^*) = n(d_A\varepsilon_A^2 - d_B\varepsilon_B^2). \quad (4)$$

This development provides a simple method for establishing a value of λ for the power analysis procedure being proposed here. The investigator may choose RMSEA values for each model in such a way as to represent a difference in model fit that would be desirable to detect. Later in this article, we will consider in detail the issue of how to choose appropriate values of RMSEA, and we will offer principles and guidelines for choosing such values in practice. For now, an investigator may choose a pair of values, ε_A and ε_B, such as .06 and .04, and then compute δ using Equation 3. Once this step is complete, the computation of statistical power proceeds in the fashion described above and illustrated in Figure 1. An SAS program (designated Program A) is provided on the Web at http://dx.doi.org/10.1037/1082-989X.11.1.19.supp.

To illustrate the procedure just described, consider a case where we wish to test the difference between Models A and B, where Model A is nested in Model B, and $d_A = 22$ and $d_B = 20$. Suppose we intend to collect a sample of size $N = 200$, and we will use $\alpha = .05$ as a significance level. To set an arbitrary effect size for purposes of illustration, let us choose $\varepsilon_A = .06$ and $\varepsilon_B = .04$, meaning we wish to determine power for detecting a difference this large or larger in fit, when testing the null hypothesis of no difference in population discrepancy function values. From these values and Equation 3, we obtain $\delta = .047$, which represents the specified effect size. The null and alternative hypotheses are then expressed, respectively, as $H_0 : (F_A^* - F_B^*) = 0$ and $H_1 : (F_A^* - F_B^*) = .047$. Given this scenario and the usual assumptions, the distribution of the test statistic under H_0 would be central χ^2 with $d_{A-B} = 22 - 20 = 2$. The resulting critical value based on this distribution is found to be $\chi_C^2 = 5.99$. The distribution of the test statistic under H_1 is noncentral χ^2 with $d_{A-B} = 2$ and noncentrality parameter $\lambda = n\delta = 9.393$. Finally, as illustrated in Figure 1, we compute the area under this noncentral distribution to the right of $\chi_C^2 = 5.99$. This area is found to be .79. Thus, for Models A and B as described above, if the true difference in fit is represented by $\varepsilon_A = .06$ and $\varepsilon_B = .04$, and if we conduct a test of the null hypothesis of no difference in fit, using $N = 200$ and $\alpha = .05$, the probability of rejecting that null hypothesis is approximately .79.

Of course, it is of great potential interest to consider a closely related question. That is, given Models A and B and a specified effect size, can we determine the sample size necessary to achieve a desired level of power? The capability to address questions of this form would be valuable in research design, where an investigator wishes to assure having adequate power to detect specified differences in fit between models of interest. It is quite feasible to carry out this computation, although an iterative procedure is required. A simple procedure has been developed and is described along with corresponding SAS code (designated Program B) on the Web at http://dx.doi.org/10.1037/1082-989X.11.1.19.supp. To illustrate, to achieve power of .90 in the example given above, a sample of $N = 269$ would be necessary.

Comparison of approaches. We have described two approaches for establishing a value of the noncentrality parameter, which forms the basis of the power computation. Under the parametric misspecification approach based on the work of Satorra and Saris (1985), Models A and B are defined as differing in the sense that the additional constraints in Model A are defined as misspecifications in comparison with Model B, which is defined as correct. Power is then the probability of detecting the cumulative effect of the particular set of parametric misspecifications. In the approach proposed here, which is based on model fit, the difference between Models A and B is defined in terms of overall fit using the RMSEA index. Power is then the probability of detecting the specified difference in fit.

In considering which approach to use in a given situation, one guiding principle is that the investigator should make use of all available relevant information and should exercise the capacity to make educated estimates of unknown quantities necessary for the power analysis. Available information may include prior research on the models in question (or similar models), results from pilot studies, or strong theoretical hypotheses. The parametric misspecification approach may be more useful when research in a particular domain is sufficiently advanced to allow for a focus on questions where the difference between models is defined in terms of values of parameters that differentiate the models, and where sufficient prior information is available to allow the investigator to specify reasonable values for all model parameters. This approach may be especially feasible for smaller models where the number of parameters to be specified is not large or in models such as latent growth curve models where many nonzero parameters have specified fixed values, leaving relatively few free parameters. In any case, if there is sufficient basis for justifiable specification of parameter values in Model B, then the parametric misspecification approach of Satorra and Saris (1985) may be preferred. If not, such as in earlier stages of

research when less information is available, or when specification of parameter values is considered to be too unreliable, then the current approach based on overall model fit may be preferred in that it requires less information from the user. In addition, there are generalizations and extensions of the overall fit approach, to be considered in this article, that are not made easily using the parametric misspecification approach.

Results from the Satorra–Saris (1985) procedure are a direct function of the specified values of the parameters that differentiate Models A and B. As a result, effect size is readily interpreted in terms of units associated with those parameters. For example, if Model B includes a correlation between two latent variables and Model A does not include that parameter, then effect size is a function of the value that the investigator assigns to that parameter (e.g., .40), and power is the probability of rejecting the hypothesis that Models A and B are equivalent with respect to model fit if that parameter takes on the assigned value. This characteristic of the Satorra–Saris (1985) approach lends a kind of precision to interpretation of results in that the results are associated with particular parameters, their assigned values, and their meaningful units.

By comparison, the procedure based on overall model fit requires less information from the investigator but sacrifices direct interpretation in terms of values of model parameters. Effect size is defined in terms of a difference in overall fit, and power is the probability of rejecting the hypothesis of equal fit if the true difference in fit corresponds to the specified RMSEA values. Some awkwardness is introduced in that the units associated with the RMSEA fit measure are not as directly interpretable as units associated with parameter values, making substantive interpretation of effect size less precise in a sense. Nevertheless, as mentioned earlier, the current procedure would have utility when a reasonably correct specification of values for all model parameters is not feasible.

It is important to recognize that, for both approaches, different specifications of the conditions under which power is calculated can produce the same value of the noncentrality parameter and the same estimate of power. In the parametric misspecification approach, different specified parameter values or even different parametric misspecifications would exist that would yield the same final estimate of power. Saris and Satorra (1993) demonstrated the use of isopower contours showing sets of parameter values that would yield the same power. In the overall fit approach proposed in this article, different selected pairs of RMSEA values for Models A and B would yield the same power. In general, users of any of these methods should keep in mind that there is not a one-to-one isomorphism between the outcome of the power analysis and the specified conditions for that analysis.

Alternative Methods

The critical feature of the method for power analysis proposed here is the use of the RMSEA fit measure as the basis for a rational approach to establishing an effect size and in turn a value of the noncentrality parameter. It would be possible to follow the same procedure using a different fit index, although options are limited. The only indexes that could serve the purpose would be those that can be expressed as a function of the population discrepancy function, F^*. The most obvious are the goodness-of-fit index (GFI) and adjusted goodness-of-fit index (AGFI; Jöreskog & Sörbom, 2001). MacCallum and Hong (1997) examined the use of these measures for establishing effect sizes for tests of fit of single models, extending the earlier work of MacCallum et al. (1996). The results of MacCallum and Hong showed serious problems with the use of GFI for this purpose as well as evidence of less systematic behavior of AGFI than of RMSEA. Given these findings, we have not pursued the potential use of GFI and AGFI in the present context of testing differences between models. Raykov and Penev (1998) explored the use of the root discrepancy per restriction (RDR) for power analysis for tests of model differences. The RDR index was proposed by Browne and du Toit (1992) as an initial attempt to adapt the logic of RMSEA to the problem of assessing the difference between two models. However, experience with RDR has indicated that it is difficult to use or interpret. Michael W. Browne, who initially proposed RDR and is the second author of the current article, has abandoned the RDR index and recommends against its use because it is incompatible with the RMSEA. The RDR can indicate a large difference in fit between two models that both have small RMSEA values and consequently both fit closely. In sum, currently the RMSEA appears to be the best available index to serve as a basis for the procedures proposed in this article.

Factors Affecting Statistical Power

We now consider the question of what factors affect the outcome of power computations using procedures presented here. This question is most easily addressed by focusing on factors that influence the noncentrality parameter, λ, for the distribution of the test statistic under H_1. The value of λ exerts a strong influence on resulting power in that it determines the separation of the two distributions (see Figure 1). The more separated the distributions, the greater is power, holding all else constant. However, it should be kept in mind that the value of λ does not completely determine power because the shape (variance) of the two χ^2 distributions in question will also depend on degrees of freedom. That is, the same value of λ will yield a different level of power as degrees of freedom change. In any event, all of the factors that influence λ and, in turn, power, are represented in Equation 4. The influences of these factors are interactive

in nature in that the strength of the effect of one factor will depend on the level of other factors. In the ensuing discussion, we make use of the scenario in the example above, altering specific aspects of the scenario while holding others fixed.

As always, power will increase as N increases. Equation 4 shows a direct relationship between N and λ. To illustrate, the middle curve in Figure 2 shows power as a function of N, holding all other aspects of our original example fixed ($d_A = 22$, $d_B = 20$). This power curve shows the typical pattern, with power asymptotically approaching 1.0 as N becomes large.

Equation 4 also indicates that the noncentrality parameter will be affected by degrees of freedom for Models A and B. In fact, degrees of freedom exert interesting and large influences on power of these tests. We first consider the effect of the magnitude of d_A and d_B, holding their difference fixed (in this case, $d_A - d_B = 2$). To illustrate, recall that in the above example, with $d_A = 22$ and $d_B = 20$, power was found to be .79. Keeping other features of the example fixed but changing d_A and d_B to 42 and 40, respectively, increases power to .97. Changing d_A and d_B to 8 and 6, respectively, reduces power to only .40. Thus, it becomes more difficult to detect a specified difference in overall fit between two models as the degrees of freedom of those models become smaller. The separation among the three curves in Figure 2 illustrates this effect and shows power as a function of N for these three pairs of values of d_A and d_B. This finding is consistent with results presented by MacCallum et al. (1996) in their study of statistical power for tests of fit of single models. In that context, power associated with a test of fit of a given model was found to decrease rapidly as degrees of freedom decreased, meaning it is more difficult to detect a given level of model misfit in models with smaller degrees of freedom. Models with low degrees of freedom arise when the number of parameters being estimated approaches the number of distinct elements in the covariance matrix for the measured variables (plus the number of measured variable means if the model includes a mean structure). Such models impose relatively few constraints on the covariance structure (and mean structure). Path analysis models with small numbers of measured variables typically exhibit low degrees of freedom. Models with larger numbers of variables may exhibit low degrees of freedom if large numbers of free parameters are estimated. Current results extend those of MacCallum et al. by showing that it is difficult to detect a specified difference in overall fit between two models that have relatively few constraints.

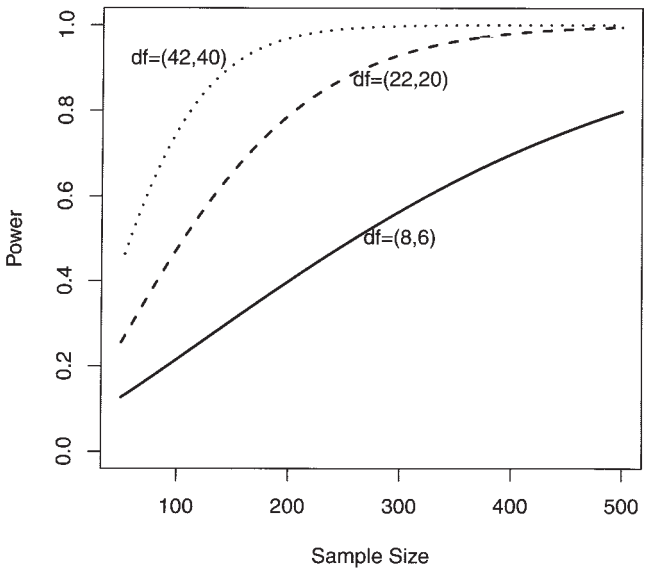

Figure 2. Relationship between power and sample size for selected levels of degrees of freedom (df) for Models A and B. Each curve represents a different level of degrees of freedom for Models A and B and shows the relationship between power (y-axis) and sample size (x-axis). The separation of the curves indicates power increasing as degrees of freedom increase, and the rise in each curve from left to right indicates power increasing as N increases. All computations are based on root-mean-square error of approximation values of .06 and .04 for Models A and B, respectively.

It is also interesting to consider the effect of the difference between d_A and d_B on power. In practice, model comparisons may involve very small or very large differences in model degrees of freedom. For instance, a comparison of two structural equation models having the same measurement structure but differing in only a small number of structural parameters will involve a small difference between d_A and d_B. By contrast, a comparison of two models of measurement invariance, where one involves a large number of equality constraints on parameters across groups or occasions and the other imposes no such constraints, would involve a large difference between d_A and d_B. When the difference between d_A and d_B becomes very large, power will tend to be high. Figure 3 shows power as a function of pairs of values of d_A and d_B holding $N = 200$, $\varepsilon_A = .06$, and $\varepsilon_B = .04$. The figure shows that when both d_A and d_B are high or when the difference is large, power tends to be quite high. When both d_A and d_B are in the middle to lower range, power is lower, especially when the difference between d_A and d_B is small.

We next consider the effect of the specified RMSEA values on power. As would be expected, when the difference between specified ε_A and ε_B values is greater, power will increase. This can be seen by considering the effect of these values on λ in Equation 4. To illustrate, we modify the example above in which $\varepsilon_A = .06$ and $\varepsilon_B = .04$ and power was .79. Keeping $\varepsilon_B = .04$ but changing ε_A to .08 increases power to .99, whereas changing ε_A to .05 reduces power to .47. These changes clearly alter the effect size and in turn

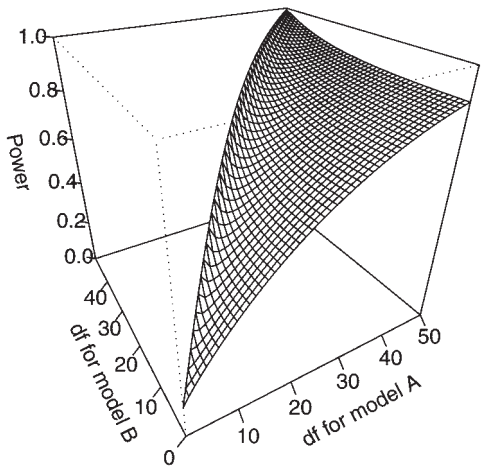

Figure 3. Relationship between power and degrees of freedom (df). The vertical axis represents power, and the two axes on the floor of the plot represent degrees of freedom for Models A and B. For each pair of degrees of freedom, with d_A ranging from 1 to 50 and $d_B < d_A$, the corresponding level of power is computed and plotted, resulting in a surface in the three-dimensional space. The region where the surface is low indicates low power when both models have low degrees of freedom. The high level of the surface along the back of the plot indicates high power when Model A has high degrees of freedom or when the difference in degrees of freedom is large. The curvature in the surface indicates the change in power as degrees of freedom change. All computations are based on root-mean-square error of approximation values of .06 and .04 for Models A and B, respectively, and $N = 200$.

have substantial effects on power. Power is higher when the specified difference in fit between the models is greater.

Holding the difference in RMSEA values fixed, power will also change as the location of the RMSEA values on the scale is shifted. Referring again to Equation 4, larger values of ε_A and ε_B will yield larger values of λ, even when holding the difference between ε_A and ε_B constant. Again modifying the original example, if we keep the difference between ε_A and ε_B fixed at .02 but shift the pair of values along the scale, power will vary. For example, if we set $\varepsilon_A = .10$ and $\varepsilon_B = .08$, power will be .99; if we set $\varepsilon_A = .03$ and $\varepsilon_B = .01$, power decreases to .37. Thus, other things being equal, it is more difficult to detect a difference in fit between models when both models fit well than when both fit poorly. Another way to view this phenomenon is that, in terms of effect size as defined in Equation 3, a given magnitude of difference (e.g., .02) between ε_A and ε_B implies a different effect size depending on the region of the RMSEA scale. Thus, effect size is determined by the pair of RMSEA values chosen and not simply by their difference.

Effects of the choice of RMSEA on power are illustrated in Figure 4. Holding $N = 200$, $d_A = 22$, and $d_B = 20$, the figure shows power as a function of pairs of RMSEA values, with ε_A ranging from .02 to .08, and ε_B ranging from 0 to ($\varepsilon_A - .01$). Results show that when RMSEA values are chosen in the low range with a small difference power is low, but in the higher range or with a larger difference, power becomes higher.

To summarize all of the effects just described, power for a test of the difference between two models will generally decrease as N decreases, as degrees of freedom for each model decrease or the difference becomes smaller, as the specified difference in model fit decreases, and as the specified difference in fit falls in the lower range of RMSEA values. These effects are illustrated in Figures 2–4 with other factors held constant, but it should be kept in mind that these effects will change in degree as levels of other factors change. Our previous example illustrated a set of conditions yielding moderately high statistical power. By contrast, and to illustrate the effects just described, consider an alternative set of conditions where $N = 100$, $d_A = 8$, $d_B = 4$, $\varepsilon_A = .03$, and $\varepsilon_B = .01$. Using $\alpha = .05$, the power of the test of the null hypothesis of no difference in population discrepancy functions between these models would be .09. Clearly the chance of detecting the specified difference in fit between these models would be very small.

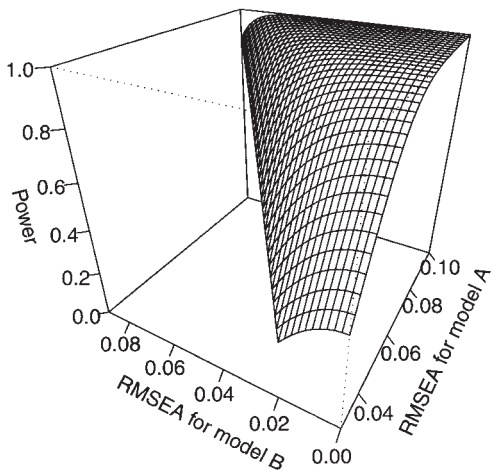

Figure 4. Relationship between power and root-mean-square error of approximation (RMSEA) holding $N = 200$, $d_A = 22$, and $d_B = 20$, where d refers to degree of freedom and A and B refer to Models A and B, respectively. The vertical axis represents power, and the two axes on the floor of the plot represent RMSEA (ε) for Models A and B. For pairs of RMSEA values, with ε_A ranging from .03 to .10, and ε_B ranging from .00 to ($\varepsilon_A - .01$), the corresponding level of power was computed and plotted, resulting in a surface in the three-dimensional space. The region where the surface is low near the front of the figure indicates low power when both models have low RMSEA. The region where the surface is high toward the back of the figure indicates high power when both RMSEA values are large or the difference is large. The curvature in the surface indicates the change in power as RMSEA values change. All computations are based on $d_A = 22$, $d_B = 20$, and $N = 200$.

We extend this last example to again illustrate the computation of minimum N required to achieve a desired level of power. Using SAS Program B in the group of programs provided on the Web, it can be determined that in order to achieve power of .80 in the conditions specified in this example, N would have to be at least 1,756. Although such a prospect would be daunting in most empirical research settings, the critical point is that such information is readily available and can be put to use during the research design phase of substantive studies.

Adaptation to Multisample Models

Some natural applications of the methods just described would involve multisample designs and models, where competing models specify different constraints across groups with respect to parameters of interest. A common such case would involve studies of measurement invariance, where competing models specify invariance, or lack thereof, with respect to parameters such as factor loadings and factor means (Meredith, 1993). Extension of the developments in this article to the multisample case requires consideration of whether or how to modify the definition of RMSEA for that case. Steiger (1998) proposed a modification of this definition wherein RMSEA is expressed as $\varepsilon = \sqrt{G}\sqrt{F^*/df}$, where G is the number of groups. This expression alters the original interpretation of RMSEA as (the square root of) discrepancy per degree of freedom. In the current article, we choose to retain the usual expression for RMSEA as $\varepsilon = \sqrt{F^*/df}$ even in the multisample case and thus also retain the original interpretation. Researchers who wish to apply developments in this article to multisample problems and who also wish to adopt Steiger's (1998) alternative formula would have to modify computations accordingly, as represented in the SAS programs provided on the Web at http://dx.doi.org/10.1037/1082-989X.11.1.19.supp. To facilitate, the SAS programs provided on the Web include an optional statement for implementing Steiger's (1998) adjustment to RMSEA for the multisample case.

Application to Examples From Published Empirical Studies

In this section, we illustrate the methods just developed using results from two published empirical studies that involve testing differences between models. Results show how these power analysis techniques can both shed light on existing findings and also aid in research design. These results are not intended in any manner to reflect negatively on the published studies but merely to illustrate methods and their potential utility. Note also that the Satorra–Saris (1985) method could be applied to these same examples if there were sufficient information available to allow for specification of values of all model parameters.

Sadler and Woody (2003). On pages 84–85, Sadler and Woody examined the issue of measurement equality across genders by using a likelihood ratio test of no difference in fit between two covariance structure models. They were not performing a multisample analysis, so the two models they were comparing simply differed in terms of whether there were equality constraints on the factor loadings across gender-specific latent variables. The model without equality constraints (Model B) yielded $\chi^2(24, N = 112) = 16.63$, and the model with constraints (Model A) yielded $\chi^2(27, N = 112) = 23.73$. These results produced a test of the difference between models with $\chi^2(3, N = 112) = 7.10, p > .05$. On the basis of this nonsignificant result, Sadler and Woody favored Model A and concluded that the measurement was equivalent across gender. Applying power analysis procedures described above, we consider the statistical power of this test with $d_A = 27$, $d_B = 24$, and $d_{A-B} = 3$. If we set $\varepsilon_A = .06$ and $\varepsilon_B = .04$, the noncentrality parameter can be calculated from Equation 4 to be $\lambda = 111[(27)(.06)^2 - (24)(.04)^2] = 6.53$. Using $\alpha = .05$, the critical value of χ^2 for the test of no difference between the two models is $\chi^2_C = 7.81$. On the basis of this set of information, the power of the test of zero difference in overall fit between the models is then calculated as the area under the alternative distribution beyond 7.81, which is found to be 0.57. Thus, Sadler and Woody's finding of a nonsignificant difference between the models might possibly have been the result of modest statistical power for detecting a difference in model fit. In addition, it can be determined that the minimum N needed to achieve power of .80, holding all other aspects of the problem fixed, would be 187.

Manne and Glassman (2000). On page 160, Manne and Glassman compared two nested mediational models. In the first model, every mediating path was freely estimated. This model (Model B) served as a baseline and was found to fit the data very well, $\chi^2(19, N = 191) = 15.98$. In the second model (Model A), Manne and Glassman fixed seven of the paths to zero, resulting in a model with $\chi^2(26, N = 191) = 26.96$. A χ^2 test of the difference between the two models yielded $\chi^2(7, N = 191) = 10.98, p > .05$. On the basis of this finding of a nonsignificant difference between the models, Manne and Glassman adopted Model A, the more constrained model. It is informative to consider the power of the test of the difference between these two models as conducted by the authors. Let $d_A = 26$, $d_B = 19$, and $d_{A-B} = 7$. For purposes of power analysis, let us set $\varepsilon_A = .06$ and $\varepsilon_B = .04$. From Equation 4, the noncentrality parameter is calculated to be $\lambda = 190[(26)(.06)^2 - (19)(.04)^2] = 12.008$. Using $\alpha = .05$, the critical value of the test of no difference between the two models is $\chi^2_C = 14.07$. From this information and proceeding as in the earlier example, we find the power of this test to be .71. Thus, although power is not extremely low, it would generally be considered as inadequate under the conditions

specified in the power analysis. Holding all aspects of the problem constant, it can further be determined that the minimum N necessary to achieve power of .80 would be 228.

Specifying the Null Hypothesis

As noted earlier, in assessing differences between models, it is conventional to test the null hypothesis of no difference in population discrepancy function values, that is, $H_0 : (F_A^* - F_B^*) = 0$. As with any test of a point hypothesis, however, it can be argued that this null hypothesis is essentially never true in practice and is of limited empirical interest. We do not realistically expect that two nested models would ever fit exactly the same in the population. It follows that with sufficiently large N, this null hypothesis would always be rejected. Thus, a finding of a nonsignificant test statistic could be viewed simply as an indication that N was not large enough to obtain a significant test statistic and to reject a null hypothesis that is always false.

The same issue has been discussed often in the context of tests of fit of single models. The conventional likelihood ratio test of fit of a single model provides a test of $H_0 : F^* = 0$ or that the model is exactly correct in the population. Of course, this hypothesis is always false in practice for any parsimonious model and, given sufficiently large N, will always be rejected even for models that fit quite well. This fact has led to the development of a wide array of alternative measures for assessing fit. Within the hypothesis testing framework, Browne and Cudeck (1993) have suggested circumventing this problem by testing a null hypothesis of close fit rather than exact fit. Using RMSEA as a basis for their approach, they suggested testing $H_0 : \varepsilon \leq .05$, meaning that the model fits closely in the population. This null hypothesis may well be true and is certainly of more empirical interest, and a test of this hypothesis is not compromised by having a very large N. Browne and Cudeck provided a specific method for conducting this test.

This approach can be viewed as a direct application of *the good-enough principle*, as presented by Serlin and Lapsley (1985). In response to problems associated with the testing of point hypotheses that are never true, Serlin and Lapsley suggested that the investigator should set a standard for what outcomes would be good enough to support or falsify a theory. This range of outcomes could then serve as the basis for defining an interval null hypothesis rather than a point null hypothesis, where the interval hypothesis defines a region of interest for the relevant parameter(s) that will result in support of or failure to support the theory under scrutiny. A statistical test could then be conducted to assess the likelihood that the parameter of interest falls in the region of interest. This principle can serve as a heuristic providing the basis for an alternative approach to testing hypotheses about model fit. Browne and Cudeck (1993) implemented the good-enough principle by postulating that an RMSEA value of .05 or lower is good enough to indicate close fit of a model. More specifically, in Browne and Cudeck's test of close fit, the outcome provides an exceedance probability (a p value) that yields a decision about $H_0 : \varepsilon \leq .05$. Failure to reject H_0 provides support for close fit (observed fit is good enough), whereas rejection of H_0 implies that fit is probably not close in the population (observed fit is not good enough). The outcome of such a test is of more interest and relevance than is the outcome of a test of exact fit.

We suggest that the good-enough principle can be applied constructively to the present problem of evaluating the difference between two models. Rather than testing the null hypothesis that there is zero difference between population discrepancy function values for two models, we propose testing a null hypothesis that the difference in population discrepancy function values is small. Rejection of such a null hypothesis will imply that the observed difference between the models is too large (not good enough) for us to believe the true difference is small. And failure to reject will imply that the observed difference is small enough (good enough) for us to believe that the true difference is small. Again the good-enough principle serves as a heuristic allowing a restructuring of the null hypothesis, with implementation requiring precise specification of what constitutes a good enough, or small enough, difference in overall fit between models. Such an approach will also alleviate the sample size problem, wherein the test of zero difference will virtually always result in rejection when N is adequately large.

Statistical Theory and Method for Testing for a Small Difference in Fit

Given Models A and B, we propose to test a null hypothesis of the form $H_0 : (F_A^* - F_B^*) \leq \delta^*$, where δ^* is some specified small number. Shortly we shall provide a rational method for choosing a useful value of δ^*. To test a null hypothesis of this form, we use the same test statistic as in the conventional test of zero difference but a different reference distribution. For the test of zero difference, the reference distribution is central χ^2, but for the test of small difference, the reference distribution will be noncentral χ^2 (analogous to the alternative distribution in the power analysis procedure described earlier). Specifically, under $H_0 : (F_A^* - F_B^*) \leq \delta^*$, and under the same assumptions discussed earlier, the test statistic $T = n(\hat{F}_A - \hat{F}_B)$ will approximately follow a noncentral χ^2 distribution with degrees of freedom $d_{A-B} = d_A - d_B$ and noncentrality parameter $\lambda = n(F_A^* - F_B^*) = n\delta^*$. Using this reference distribution, one can determine a critical value of the test statistic and then, following usual procedures, determine whether the

sample value of the test statistic is sufficiently large to reject H_0.

Clearly, the key to using this procedure in practice is to be able to specify an appropriate value of δ*. As before, we suggest the use of RMSEA as a basis for establishing such a value. Following the good-enough principle as a heuristic, one could specify values of ε_A and ε_B so as to represent a small difference in fit between the models (e.g., $\varepsilon_A = .06$ and $\varepsilon_B = .05$). This pair of values of RMSEA then can be used to compute a value of δ*. Using the relationship between RMSEA and the population discrepancy function as noted earlier (i.e., $F_A^* = d_A\varepsilon_A^2$ and $F_B^* = d_B\varepsilon_B^2$), the value of δ* can be computed using the expression

$$\delta^* = (F_A^* - F_B^*) = (d_A\varepsilon_A^2 - d_B\varepsilon_B^2). \quad (5)$$

This step provides the value of δ* for $H_0 : (F_A^* - F_B^*) \leq \delta^*$ and for computation of the noncentrality parameter for the reference distribution according to $\lambda = n(F_A - F_B^*) = n\delta^*$. The hypothesis test then proceeds by usual methods.

To illustrate, consider Models A and B, with Model A nested in Model B, and with $d_A = 22$ and $d_B = 20$. Suppose we have $N = 200$ and we wish to carry out a test of a null hypothesis of a small difference between Model A and Model B. We choose $\varepsilon_A = .06$ and $\varepsilon_B = .05$ to represent this small difference. Substituting these values into Equation 5 yields $\delta^* = [(22)(.06)^2 - (20)(.05)^2] = .0292$. The null hypothesis for this test of small difference can then be expressed as $H_0 : (F_A^* - F_B^*) \leq .0292$. The reference distribution for this test is then noncentral χ^2 with degrees of freedom $d_{A-B} = d_A - d_B = 22 - 20 = 2$ and noncentrality parameter $\lambda = n\delta^* = (199)(.0292) = 5.8108$. For this reference distribution, using $\alpha = .05$, the critical value of the test statistic is found to be 17.75. A sample value of the test statistic, T, that exceeds this critical value leads to rejection of this null hypothesis of a small difference in fit, whereas a test statistic smaller than 17.75 indicates that the null hypothesis of this small difference (or smaller) is not rejected. SAS code (designated Program C) for carrying out the necessary computations is provided on the Web at http://dx.doi.org/10.1037/1082-989X.11.1.19.supp.

It is important to note the potential difference between the outcome of a test of a null hypothesis of a small difference and a test of a null hypothesis of zero difference. In the present example, the conventional test of $H_0 : (F_A^* - F_B^*) = 0$ would be conducted using a central χ^2 distribution as a reference distribution, with $d_{A-B} = 2$. Using $\alpha = .05$, the resulting critical value would be 5.99. Thus, for this example, the critical values for the tests of zero difference versus small difference are quite discrepant (5.99 vs. 17.75). Any sample value of the test statistic that lay between these values would result in rejection of the null hypothesis of zero difference but failure to reject the null hypothesis of the specified small difference (defined by $\varepsilon_A = .06$ and $\varepsilon_B = .05$). We suggest that such outcomes may not be unusual in practice when differences between models are relatively small and N is large. Furthermore, the potential occurrence of such outcomes emphasizes the value of this alternative approach to specifying H_0. Rejection of the null hypothesis of zero difference is seen in a substantially different light if a null hypothesis of a small difference is not rejected.

Application to Examples From Published Studies

In this section, we present results of testing a null hypothesis of a small difference in fit in two published studies. The outcomes show the potential utility of this approach in evaluating differences between models. Again, results are merely illustrative and in no way reflect on the quality of the published studies.

Joireman, Anderson, and Strathman (2003). This study used a path analysis model for predicting verbal aggression. On page 1295, Joireman et al. examined the difference between their initial model and the final model. The final model (Model B) differed from the initial model (Model A) by freeing two path coefficients. Model A yielded $\chi^2(12, N = 154) = 31.08$, and Model B yielded $\chi^2(10, N = 154) = 11.95$. The test of no difference between the models resulted in $\chi^2(2, N = 154) = 19.13$, $p < .05$. Rejection of the null hypothesis of zero difference between the models implies rejection of the constraint placed on the two path coefficients in Model B. It is of interest to consider a test of the null hypothesis of a small difference in fit between these models. Following the procedures described above, if we set $\varepsilon_A = .06$ and $\varepsilon_B = .05$, we can use Equation 5 to compute $\delta^* = [(12)(.06)^2 - (10)(.05)^2] = .0182$, and we can then define the null hypothesis as $H_0 : (F_A^* - F_B^*) \leq .0182$. The reference distribution for the test of this null hypothesis is noncentral χ^2 with $d_{A-B} = 2$ and noncentrality parameter $\lambda = (153)(.0182) = 2.7846$. Using this reference distribution and $\alpha = .05$, the critical value of the test statistic is $\chi_C^2 = 12.40$. The observed value of the test statistic, given above, is $\chi^2 = 19.13$, yielding a significance level of $p < .05$, indicating clear rejection of the null hypothesis of small difference. Thus, the test of small difference also leads to Joireman et al.'s decision to reject the initial model.

Dunkley, Zuroff, and Blankstein (2003). On page 241, Dunkley et al. investigated the difference between a fully mediated structural model, in which the direct path between two variables is constrained to zero, and a partially mediated model, with the direct path freely estimated. The partially mediated model (Model B) yielded $\chi^2(337, N = 163) = 584.88$, and the fully mediated model (Model A) yielded $\chi^2(338, N = 163) = 588.72$. The test of no difference between the models resulted in $\chi^2(1, N = 163) = 3.84$, $p < .05$, implying a statistically significant difference and rejection of the constraint in question. Even though the χ^2 was

significant, the authors adopted the fully mediated model primarily because of the fact that the two models were almost indistinguishable in terms of their ability to account for variance in the outcome variable. In addition, though AIC (Akaike information criterion) favored the partially mediated model, BIC (Bayesian information criterion) actually selected the fully mediated model. Furthermore, the partially mediated model was not informative theoretically. It is interesting that different types of information indicated different conclusions. The use of a test of a small difference between the models is informative in this case. Setting $\varepsilon_A = .06$ and $\varepsilon_B = .05$, and conducting the test in the fashion described above, we obtain a critical value of $\chi^2_C = 88.96$ and $p > .99$. The null hypothesis of a small difference is highly plausible, providing further support for Dunkley et al.'s decision to accept the more constrained model.

Power Analysis Using the Null Hypothesis of Small Difference

To this point, this article has presented two methodological developments. The first is a method for computing statistical power for a test of zero difference between population discrepancy function values for Models A and B when the true difference between the models is represented by specified RMSEA values, ε_A and ε_B. The second development is a method for testing a null hypothesis that the difference between Models A and B is small, where a small difference is defined again by a pair of RMSEA values for the models. It is straightforward to combine these two developments so as to allow for a more general power analysis procedure in which the null hypothesis specifies a small difference in fit and the alternative hypothesis specifies a larger difference, with those differences defined in terms of specified RMSEA values for the models.

The procedure would be implemented in the following fashion. The null hypothesis would be of the form described in the previous section, representing a small difference in fit: $H_0 : (F^*_A - F^*_B) \leq \delta^*_0$, where δ^*_0 is defined as in Equation 5 under H_0. The alternative hypothesis specifies that the difference $(F^*_A - F^*_B)$ takes on some specific value that is greater than δ^*_0; that is, $H_1 : (F^*_A - F^*_B) = \delta^*_1$, where $\delta^*_1 > \delta^*_0$. As in the various procedures described earlier in this article, we suggest establishing useful values of δ^*_0 and δ^*_1 by selecting RMSEA values and obtaining δ^*_0 and δ^*_1 using Equation 5. To establish a value for δ^*_0, one would select a pair of RMSEA values, denoted now as ε_{0A} and ε_{0B}, to represent the small difference in fit that defines H_0. The value of δ^*_0 is then computed using Equation 5. To establish a value for δ^*_1, one chooses RMSEA values ε_{1A} and ε_{1B} to represent a larger difference in fit under H_1 and then computes δ^*_1 using Equation 5. At that point, the investigator has defined a null hypothesis representing a small difference in fit and an alternative hypothesis specifying a larger difference in fit. Power analysis can then be used to determine the probability of rejecting the specified H_0 when H_1 is in fact true. The procedure follows the same general template as the method described earlier, with the simple modification that now both of the relevant distributions (as shown in Figure 1) are noncentral χ^2 distributions, whereas earlier (when H_0 represented exact fit), the distribution under H_0 was a central χ^2 distribution. Specifically, under the current procedure and under the same assumptions discussed earlier, the distribution of the test statistic under H_0 will be noncentral χ^2 with $d_{A-B} = d_A - d_B$ and noncentrality parameter $\lambda_0 = n\delta^*_0$. The distribution under H_1 will be noncentral χ^2 with the same d_{A-B} and noncentrality parameter $\lambda_1 = n\delta^*_1$. Given a specified level of α, a critical value of χ^2 is then established using the null distribution, and power is computed as the area under the alternative distribution to the right of that critical value, just as shown in Figure 1 earlier. We provide SAS code (designated Program D) for carrying out the necessary computations on the Web at http://dx.doi.org/10.1037/1082-989X.11.1.19.supp.

To illustrate, let us extend the same example used above for showing a test of a null hypothesis of a small difference in fit. In that example, $d_A = 22$, $d_B = 20$, $N = 200$, and the small difference in fit was specified using RMSEA values of .06 and .05, now denoted as $\varepsilon_{0A} = .06$ and $\varepsilon_{0B} = .05$. Using Equation 5, this scenario yielded a value of $\delta^*_0 = .0292$ and a null hypothesis expressed as $H_0 : (F^*_A - F^*_B) \leq .0292$. Under this H_0, the distribution of the test statistic is noncentral χ^2 with $d_{A-B} = 2$ and noncentrality parameter $\lambda_0 = n\delta^*_0 = (199)(.0292) = 5.8108$. From this distribution and using $\alpha = .05$, the critical value of χ^2 is 17.7474. For purposes of power analysis, one must next define an alternative hypothesis specifying a larger difference in fit than that represented under H_0. To illustrate, we choose RMSEA values of $\varepsilon_{1A} = .08$ and $\varepsilon_{1B} = .05$. The power analysis procedure then determines the probability of rejecting the H_0 above, on the basis of $\varepsilon_{0A} = .06$ and $\varepsilon_{0B} = .05$, if the true difference in fit is represented by $\varepsilon_{1A} = .08$ and $\varepsilon_{1B} = .05$. Using these RMSEA values, the value of δ^*_1 could be determined from Equation 5 to be .0908, implying the alternative hypothesis $H_1 : (F^*_A - F^*_B) = .0908$. Under this H_1, the distribution of the test statistic is noncentral χ^2 with $d_{A-B} = 2$ and noncentrality parameter $\lambda_1 = n\delta^*_1 = (199)(.0908) = 18.0692$. The two distributions are now defined and power can be computed to be .56. That is, if we test $H_0 : (F^*_A - F^*_B) \leq .0292$ when $H_1 : (F^*_A - F^*_B) = .0908$ is actually true, and we use $N = 200$ and $\alpha = .05$, the probability that we will reject H_0 is approximately .56. Of course, to understand this scenario, one must keep in mind that H_0 is established to represent a small difference in RMSEA values ($\varepsilon_{0A} = .06$ and $\varepsilon_{0B} = .05$) and H_1 is established to represent a larger difference ($\varepsilon_{1A} = .08$ and $\varepsilon_{1B} = .05$). If this larger difference holds in truth and we test

H_0 representing the smaller difference, power will be approximately .56.

This small example illustrates the potential utility of this combination of the two methods developed earlier in this article: power analysis for testing differences in model fit, combined with null hypotheses specifying a small difference rather than zero difference. Thus, the methods presented here allow for conducting power analysis using any combination of null and alternative hypotheses based on pairs of RMSEA values representing differences in model fit. This development helps the investigator to escape from the limitation of using the null hypothesis of zero difference in fit, which is essentially always false in practice.

Issues Inherent in Use of These Methods

Choice of RMSEA Values

A primary factor influencing the outcome of the power analysis procedures described in this article is the investigator's choice of RMSEA values for the models being compared. This choice plays a major role in determining the value of the noncentrality parameter(s) according to Equation 4 and in turn the level of power. Therefore, it is important that this choice be made with as much care as possible. In considering this problem, it should be recognized that the impact of this choice will also depend on the levels of other factors that influence power; specifically, sample size and the degrees of freedom of the two models being compared. Depending on the levels of those factors, the choice of RMSEA values may have relatively little effect on power or may have a great effect. Thus, we examine this issue in two steps: (a) We describe conditions under which the choice of RMSEA values will have small or large impact. (b) When the impact is likely to be large, we offer some guiding principles for making this choice.

Conditions where choice has little impact. When N is extremely large, the choice of RMSEA values, within reasonable bounds, will have little impact on power, which will tend to be consistently high regardless of other factors. In addition, when the difference between d_A and d_B is large, the choice of RMSEA values will again make little difference. Likewise, the choice will make little difference when d_A and d_B are both high even if $d_A - d_B$ is small. Power will tend to be quite high in such cases unless the investigator chooses ε_A and ε_B to have an extremely small difference, which is probably not often a useful approach. These three phenomena are illustrated simultaneously in Figure 5, which shows extremely high power when N is large, both d_A and d_B are substantial, and $d_A - d_B$ is substantial. Holding $N = 500$, $d_A = 60$, and $d_B = 50$, the figure shows power as a function of pairs of values of ε_A and ε_B, with ε_A ranging from .03 to .10 and ε_B ranging from .00 to ($\varepsilon_A - .01$). Power is always high under these conditions regardless of the choice of RMSEA values in this range.

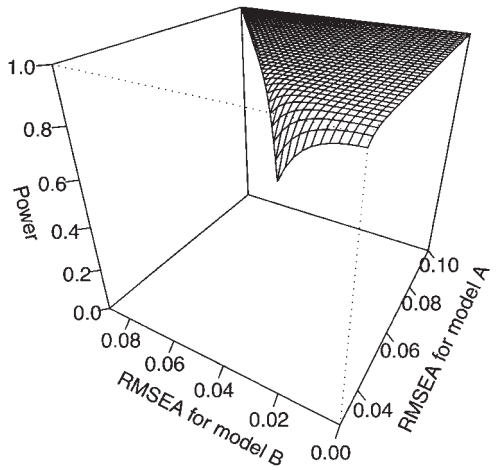

Figure 5. Relationship between power and root-mean-square error of approximation (RMSEA) for large N and degrees of freedom. The vertical axis represents power, and the two axes on the floor of the plot represent RMSEA (ε) for Models A and B. For pairs of RMSEA values, with ε_A ranging from .03 to .10, and ε_B ranging from .00 to ($\varepsilon_A - .01$), the corresponding level of power was computed and plotted, resulting in a surface in the three-dimensional space. For each such computation, N was fixed at 500, and degrees of freedom were $d_A = 60$ and $d_B = 50$, where A and B refer to Models A and B, respectively. The uniformly high level of the surface indicates a high level of power for any pair of RMSEA values in the specified range but with some decline when both RMSEA values are small and the difference is small.

Under other sets of conditions, power will be consistently low regardless of the choice of RMSEA values within reasonable bounds. When N is small, power will tend to be low. And when d_A and d_B are both small, power will also be consistently low assuming N is not extremely large. These phenomena are illustrated simultaneously in Figure 6. Holding $N = 100$, $d_A = 5$, and $d_B = 4$, the figure shows power as a function of pairs of values of ε_A and ε_B, with ε_A ranging from .03 to .10 and ε_B ranging from .00 to ($\varepsilon_A - .01$). Power is always inadequate under these conditions regardless of the choice of RMSEA values in this range.

Recommendations when choice has greater impact. The investigator's choice of RMSEA values will matter most when N is moderate and d_A and d_B are also moderate (probably less than 50 or so, and greater than 5 or so, as rough guidelines), with the difference not being very large. Such a set of conditions was illustrated earlier in Figure 4 where it can be seen that power varies substantially with the choice of RMSEA values. Although this region of design factors may seem limited, it probably includes quite a wide range of substantive applications of covariance structure modeling. The question then becomes what values of RMSEA are appropriate in those situations where the choice will have a substantial impact on the outcome of a power com-

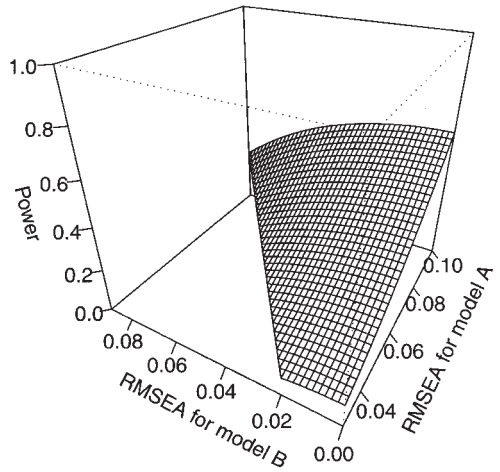

Figure 6. Relationship between power and root-mean-square error of approximation (RMSEA) for small N and degrees of freedom. The vertical axis represents power, and the two axes on the floor of the plot represent RMSEA (ε) for Models A and B. For pairs of RMSEA values, with ε_A ranging from .03 to .10, and ε_B ranging from .00 to (ε_A – .01), the corresponding level of power was computed and plotted, resulting in a surface in the three-dimensional space. For each such computation, N was fixed at 100 and degrees of freedom were $d_A = 5$ and $d_B = 4$, where A and B refer to Models A and B, respectively. The uniformly low level of the surface indicates a low level of power for any pair of RMSEA values in the specified range, reaching moderate levels only when the difference in RMSEA values is very large.

putation. From this perspective, we consider some general principles. First, it is probably sensible to avoid choosing RMSEA values in either the very low range of the RMSEA scale or the very high range. Regarding the low range, very small differences between ε_A and ε_B in the low range (e.g., .02 vs. .01) would be difficult to detect and would be of limited practical significance. Regarding the high range (e.g., .15 vs. .12), it would be of limited scientific value to know that power was high for detecting a difference in fit between two very poor models. Thus, it is probably most sensible to consider pairs of RMSEA values in the midrange of the scale, roughly .03 to .10. Within this range, it should be kept in mind that power will tend to be low when the difference between specified values of ε_A and ε_B is very small. We suggest choosing pairs of values with a moderate to large difference (e.g., at least .01–.02).

In addition to relying on such principles, researchers should also consider other sources of information to aid in the choice of RMSEA values. If any prior information is available regarding overall model quality, it should be taken into account, with lower values of RMSEA selected when more evidence is available that the models can be expected to fit reasonably well. With respect to the difference in specified RMSEA values, users might consider the number of parameters that differentiate the models along with their likely importance. If Models A and B are differentiated by only a single parameter, it is probably reasonable to choose RMSEA values that are not highly discrepant, such as .06 and .05. On the other hand, if the models are differentiated by a large number of parameters or constraints, then a larger difference might be specified, such as .08 versus .05.

It is often the case that little useful information is available or the investigator has little confidence in specification of such values. In such situations, we recommend that power (or minimum N) be computed for multiple pairs of RMSEA values so as to provide information about a range of conditions. One simple strategy for establishing baselines would be to compute power for two pairs of RMSEA values, one representing a small difference in the low range (e.g., $\varepsilon_A = .05$ and $\varepsilon_B = .04$) and another representing a larger difference in the higher range (e.g., $\varepsilon_A = .10$ and $\varepsilon_B = .07$). For the first pair, if power is found to be high, then it can be safely assumed that power will be high for other pairs of values in the recommended range; this is a best case situation. For the second pair, if power is found to be low, then power will also be low for other pairs of values in the recommended range; this would be a worst case situation.

In addition to these baseline computations, we suggest computing power, or minimum N, for a larger set of pairs of ε_A and ε_B in the recommended range. For example, given $N = 200$, $d_A = 22$, $d_B = 20$, and $\alpha = .05$, one could compute power for all pairs of RMSEA values letting ε_A vary from .04 to .10 in steps of .01, and ε_B vary from .03 to (ε_A – .01) in steps of .01. An SAS program (designated Program E and available on the Web at http://dx.doi.org/10.1037/1082-989X.11.1.19.supp) allows for such computations, and Table 1 shows power values for the scenario just described. Such results provide the investigator with information about the likelihood of detecting differences of various degrees in overall fit. Alternatively, one could com-

Table 1
Power for Selected Range and Pairs of Root-Mean-Square Error of Approximation (RMSEA) Values

	ε_B						
ε_A	.03	.04	.05	.06	.07	.08	.09
.04	.36						
.05	.68	.47					
.06	.89	.79	.57				
.07	.97	.95	.87	.66			
.08	1.00	.99	.97	.92	.75		
.09	1.00	1.00	1.00	.99	.96	.82	
.10	1.00	1.00	1.00	1.00	1.00	.98	.87

Note. Power computed by setting $N = 200$, $d_A = 22$, $d_B = 20$, and $\alpha = .05$, where d refers to degrees of freedom and A and B refer to Models A and B, respectively. ε_A and ε_B represent RMSEA values for Models A and B, respectively.

pute the minimum N required to achieve a desired level of power, such as .80, across a range of paired RMSEA values. A separate SAS program (designated Program F and available on the Web at http://dx.doi.org/10.1037/1082-989X.11.1.19.supp) provides for such computations. Table 2 shows results of such computations for the same scenario, providing an investigator with information about the level of sample size needed to have a strong likelihood for detecting differences in fit of various degrees.

It should be kept in mind that investigators should not rely exclusively on commonly used guidelines for interpretation of values of RMSEA. For example, Browne and Cudeck (1993) and Steiger (1995) suggested that values less than .05 indicate close fit, values in the range of .08 indicate fair fit, and values above .10 indicate poor fit. These guidelines are of some help in choosing RMSEA values for power analysis but are not fail-safe. Evidence is accumulating that sample values of RMSEA indicating close fit may not be independent of other factors. For example, Hu and Bentler (1999) observed that rejection rates for correct models using fixed RMSEA cutoffs decreased (improved) as N increased. Curran, Bollen, Chen, Paxton, and Kirby (2003) have shown that when a model is correct in the population, the mean value of the estimated RMSEA decreases with increasing sample size, indicating sensitivity to sample size. Results indicate that a .05 threshold applied to sample values of RMSEA and applied across levels of N will not produce consistently valid inferences about fit in the population. However, this phenomenon is not directly relevant in the present context in that the values of ε_A and ε_B specified for power analysis are population values and thus not impacted by sampling error.

Effects of degrees of freedom on the conventional guidelines for interpretation of RMSEA would be relevant in the present context. There is some evidence that perhaps guidelines should be adjusted downward when degrees of freedom are high. Curran et al. (2003) found that the same parametric misspecification in models with very different degrees of freedom produced smaller RMSEA values when degrees of freedom were larger.

Table 2
Minimum Sample Size for Selected Range and Pairs of Root-Mean-Square Error of Approximation (RMSEA) Values

	ε_B						
ε_A	.03	.04	.05	.06	.07	.08	.09
.04	561						
.05	261	420					
.06	159	205	331				
.07	108	128	168	270			
.08	79	89	107	141	227		
.09	61	67	76	92	121	193	
.10	49	52	58	66	80	106	167

Note. Sample size computed by setting desired power at .80, $d_A = 22$, $d_B = 20$, and $\alpha = .05$, where d refers to degrees of freedom and A and B refer to Models A and B, respectively. ε_A and ε_B represent RMSEA values for Models A and B, respectively.

Using observed RMSEA values for power analysis. A tempting alternative approach to selection of RMSEA values for power analysis would be to use observed values. That is, results may be available from fitting Models A and B to sample data, thus providing point estimates of ε_A and ε_B. Those point estimates could be used to compute what is called *observed power* in the statistical literature (e.g., Hoenig & Heisey, 2001). We do not encourage this approach. Observed power is a direct function of the observed p value (significance level) from the test of the null hypothesis in the sample from which the RMSEA values were obtained. If p is small, power will be high and vice versa. No new information is provided by observed power (Hoenig & Heisey, 2001). We do not recommend the use of observed RMSEA values for power analysis, and we suggest that more useful information can be obtained by choosing ranges of RMSEA values as described earlier.

Summary of recommendations. It is probably most reasonable to choose RMSEA values in the midrange of the scale, with a small to moderate difference. It is also advisable to conduct power analyses using several different pairs of RMSEA values rather than just a single pair. A more extended approach, which we encourage, involves computation of power (or minimum sample size) across a range of paired values of RMSEA. This procedure can be implemented easily using SAS programs described above (Programs E and F). Results provide the investigator with clear information about the magnitude of difference in fit that one can expect to detect or the level of sample size needed to have reasonable likelihood of detecting differences in model fit of various sizes.

Issues in Statistical Theory

Violations of assumptions. All of the methodological developments presented in this article rely on well-known statistical distribution theory and its inherent assumptions. Specifically, we make extensive use of the assumption that the test statistic $T = n(\hat{F}_A - \hat{F}_B)$ follows a central χ^2 distribution when $(F_A^* - F_B^*) = 0$, and a noncentral χ^2 distribution when $(F_A^* - F_B^*) \neq 0$. These distributional forms will hold asymptotically under certain theoretical assumptions mentioned earlier. As always, however, such assumptions never hold exactly in the real world. Although it is to some extent an open question as to how strongly violations of assumptions will invalidate results of the methods we have presented, there are some things that can be said in this regard.

Consider the usual assumption of multivariate normality. Theoretical work (Amemiya & Anderson, 1990; Browne & Shapiro, 1988) has shown that the asymptotic central or

noncentral χ^2 distribution of the likelihood ratio test statistic can remain valid when the latent variables have nonnormal distributions in a wide range of covariance structure models. Because nonnormal latent variables result in nonnormal measured variables, the results carry over to measured variables with some nonnormal distributions. For the exploratory factor analysis model, in particular, the requirements for asymptotic robustness are simple and plausible. The asymptotic distribution of the test statistic remains valid if all latent variables assumed to be uncorrelated in the model are in fact independently distributed, even if they have arbitrary nonnormal distributions. The robustness requirements become a little more complicated when certain restrictions are imposed on latent variable covariances. These findings that independence rather than normality is critical for the asymptotic χ^2 distribution of the test statistic imply that at least for large samples there will often be a strong degree of robustness with respect to violations of multivariate normality.

With respect to effects of violations when sample size is not large, there exists some relevant simulation research. For correctly specified models, a number of simulation studies (e.g., Chou, Bentler, & Satorra, 1991; Curran, West, & Finch, 1996; Hu, Bentler, & Kano, 1992; Muthén & Kaplan, 1992) have shown that as nonnormality becomes more severe, some inflation of T occurs compared with its expectation from a central χ^2 distribution. Of relevance in the present context is whether this same trend would extend to approximate models and noncentral χ^2 distributions. The degree to which this would occur is unknown currently. To the extent that the same phenomenon holds for approximate models and noncentral χ^2 distributions, the impact would be that the empirical likelihood of rejecting a model would exceed the theoretical power. Correspondingly, calculated sample sizes necessary to achieve a desired level of power might overestimate the sample size actually needed to achieve the desired likelihood of rejection in practice. The validation and degree of such effects remains a topic for further study.

A second important assumption is the population drift assumption described earlier. Briefly, this assumption requires that neither Model A nor Model B be badly specified and that the lack of fit due to model misspecification is of approximately the same magnitude as the lack of fit due to sampling error. Clearly, this assumption cannot hold in practice because it implies that the degree of model misspecification in the population changes (gets smaller) as N increases; that is, the population is not fixed, hence the name *population drift.* In practice, the critical requirement is that neither of the models being compared should be very badly specified. If they are, then the use of the noncentral χ^2 distribution would be unjustified and results may be subject to considerable error. The degree of this effect under various levels of model misspecification is a complex question that is open to study, although there is evidence that some effect is present (Curran, Bollen, Paxton, Kirby, & Chen, 2002; Yuan, 2005). Results of simulation studies suggest that the noncentral χ^2 approximation holds up well in terms of its mean and variance for moderately misspecified models and moderate sample size ($N \geq 200$). However, for severely misspecified models, especially when N is relatively small, the variance of T may become too large. If such a trend is consistent, estimates of power would tend to be too low, and estimates of minimum N would tend to be too high under such conditions.

The asymptotic nature of the reliance on χ^2 distributions must also be kept in mind. That is, even if other assumptions hold, the distribution of T will approach χ^2 closely only as N becomes large. Results of the use of methods presented in this article may be subject to more error when N is small. Simulation studies (e.g., Curran et al., 1996) have shown that for correctly specified models, the value of T tends to be inflated relative to its expectation in a central χ^2 distribution when N is small, although the degree of inflation is relatively small at $N = 100$. But again, it is an open question as to the degree to which such a trend extends to approximate models and noncentral χ^2 distributions. To the extent that it does, the impact would be for power and sample size computations to be at least slightly conservative for smaller levels of N, meaning observed rejection rates may exceed theoretical power, and minimum sample size levels may be overestimated. In any event, we believe that the procedures proposed in the present article should provide approximate and usable, although certainly not exact results, for levels of $N = 100$ or more.

Of course, it must be kept in mind that none of the factors just described operates in isolation across the range of empirical studies. The effect of violation of one assumption may well depend on the degree of violation of another assumption. For example, results of Curran et al. (2002) indicate that the effect of the degree of misspecification on the quality of the noncentral χ^2 approximation is reduced as sample size increases. In the context of approximate models and noncentral χ^2 distributions, further study of these issues will be valuable.

Estimation method. Another statistical issue inherent in the methods we propose here involves the estimation method that is used. At the beginning of our presentation of proposed methods, we stated that we were assuming the use of ML estimation as defined by the discrepancy function in Equation 1. Undoubtedly, ML is by far the most commonly used method, but the procedures we propose could be used under other estimation methods, including generalized least squares and asymptotically distribution free (Browne, 1984). The essential feature of the estimation method is that it provide a test statistic that asymptotically follows a χ^2 distribution, as discussed earlier in this section. The meth-

ods proposed here can be used under any such estimation method.

Relevance of statistical issues to other methods. The statistical issues just discussed are relevant to a wide range of inferential methods commonly used in structural equation modeling. These include the power analysis methods presented in the current article as well as those of MacCallum et al. (1996) and of Satorra and Saris (1985). There are no differences among these methods with regard to reliance on the statistical theory discussed in the current article. Furthermore, some other commonly used inferential methods are likewise dependent on this same statistical theory. Such methods include the computation of confidence intervals for RMSEA and for the population discrepancy function and the test of close fit (Browne & Cudeck, 1993). The use of noncentral χ^2 distributions underlies many modern approaches to model evaluation, and the methods for power analysis proposed in the current article represent additional techniques built on this same foundation.

Conclusion

We have proposed an approach to power analysis and determination of minimum sample size for tests of differences between nested covariance structure models, where effect size is defined in terms of a difference in overall model fit. This approach differs from methods developed by Satorra and Saris (1985) where effect size is represented in terms of values of parameters that differentiate the models. The current procedure may be primarily useful when the investigator has relatively little precise information about the models being studied. The procedure requires little information from the user—only specification of RMSEA values to represent a hypothesized difference in overall fit between the models of interest. When considerably more information is available or can be estimated with confidence, in particular values of all model parameters, then it may be preferable to use the methods based on parametric misspecification as proposed by Satorra and Saris (1985). Those methods have advantages with respect to interpretation of effect size in terms of units associated with model parameters rather than more ambiguous units associated with measures of model fit as well as interpretation of power with respect to detection of particular parametric misspecifications. By comparison, the procedures proposed in this article have the advantage of requiring far less information from the investigator and also allowing for straightforward extensions to incorporate tests of small differences between models.

Regardless of which approach is used, there will be uncertainty and some subjectivity in implementation of these procedures and in interpretation of results. This is the normal state of affairs in power analysis in that results are always conditional on specification or estimation of unknown quantities by the investigator. Estimates of statistical power or minimum N should always be viewed as rough approximations, conditional on the specification of unknown quantities and dependent on the validity of statistical assumptions. Nevertheless, these estimates of power or minimum N can be highly useful in planning and evaluation of research design. In the present context, we believe the procedures presented in this article should provide useful estimates of power or minimum N that can aid in design or evaluation of empirical studies involving comparison of nested covariance structure models.

References

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

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

Browne, M. W., & 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., & du Toit, S. H. C. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research, 27,* 269–300.

Browne, M. W., & Shapiro, A. (1988). Robustness of normal theory methods in the analysis of linear latent variable models. *British Journal of Mathematical and Statistical Psychology, 41,* 193–208.

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44,* 347–357.

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. B. (2003). Finite sampling properties of the point estimates and confidence intervals of the RMSEA. *Sociological Methods and Research, 32,* 208–252.

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research, 37,* 1–36.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16–29.

Dunkley, D. M., Zuroff, D. C., & Blankstein, K. R. (2003). Self-critical perfectionism and daily affect: Dispositional and situational influences on stress and coping. *Journal of Personality and Social Psychology, 84,* 234–252.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55,* 19–24.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351–362.

Joireman, J., Anderson, J., & Strathman, A. (2003). The aggression paradox: Understanding links among aggression, sensation seeking, and the consideration of future consequences. *Journal of Personality and Social Psychology, 84,* 1287–1302.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8 user's reference guide* [Computer software manual]. Lincolnwood, IL: Scientific Software.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

MacCallum, R. C., & Hong, S. (1997). Power analysis in covariance structure modeling using GFI and AGFI. *Multivariate Behavioral Research, 32,* 193–210.

Manne, S., & Glassman, M. (2000). Perceived control, coping efficacy, and avoidance coping as mediators between spouses' unsupportive behaviors and cancer patients' psychological distress. *Health Psychology, 19,* 155–164.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45,* 19–30.

Raykov, T., & Penev, S. (1998). Nested covariance structure models: Noncentrality and power of restriction test. *Structural Equation Modeling, 5,* 229–246.

Sadler, P., & Woody, E. (2003). Is who you are who you're talking to? Interpersonal style and complementarity in mixed-sex interactions. *Journal of Personality and Social Psychology, 84,* 80–96.

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.

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. P. Middendorp, B. Niemoller, & W. E. Saris (Eds.), *Sociometric research 1982* (pp. 127–190). Amsterdam: Sociometric Research Foundation.

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

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83.

Steiger, J. H. (1995). Structural equation modeling (SEPATH). In *Statistica 5* (Vol. 3, pp. 3539–3688) [Computer software manual]. Tulsa, OK: Statsoft.

Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling, 5,* 411–419.

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

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

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40,* 115–148.

Received September 8, 2004
Revision received December 5, 2005
Accepted December 5, 2005 ■