CHAPTER THREE

Alternatives to Traditional Model Comparison Strategies for Covariance Structure Models

Kristopher J. Preacher
University of Kansas

Li Cai
Robert C. MacCallum
University of North Carolina at Chapel Hill

In this chapter we discuss two related issues relevant to traditional methods of comparing alternative covariance structure models (CSM) in the context of ecological research. Use of the traditional test of parametrically nested models in applications of CSM (the χ^2 difference or likelihood ratio [LR] test) suffers from several limitations, as discussed by numerous methodologists (MacCallum, Browne, & Cai, 2005). Our primary objection is that the traditional approach to comparing models is predicated on the assumption that it is possible for two models to have identical fit in the population. We argue instead that any method of model comparison which assumes that a point hypothesis of equal fit can hold exactly in the population (e.g., the LR test) is fundamentally flawed. We discuss two alternative approaches to the LR test which avoid the necessity of hypothesizing that two models share identical fit in the population. One approach concerns framing the hypothesis of interest differently, which naturally leads to questions of how to assess statistical power and appropriate sample size. The other approach concerns a radical realignment of how researchers approach model evaluation, avoiding traditional null hypothesis testing altogether in favor of identifying the model that maximizes generalizability.

Power presents a recurrent problem to those familiar with null hypothesis significance testing (NHST). How large should a sample be in order to have

adequate probability of rejecting a false null hypothesis? What is the probability of rejecting a false null if our sample is of size N? These questions present special challenges in the context of CSM because the relative status of null and alternative hypotheses are interchanged from their familiar positions — the null hypothesis in CSM represents the theory under scrutiny, and power is framed in terms of the sample size necessary to reject a false model. Traditional goodness-of-fit tests deal with the null hypothesis under which the model fits exactly in the population (exact fit test). Point hypotheses tested by the exact fit test are likely never true in practice, so how should power be conceptualized? We present an alternative strategy extending earlier work on power for tests of close fit (rather than exact fit) of single models to tests of *small difference* (rather than no difference) in comparisons of nested models. The null hypothesis in a test of small difference states that the model fits nearly as well, but not the same, as a less constrained model.

Another alternative to traditional methods of model assessment is to avoid the hypothesis-testing framework altogether, instead adopting a model selection approach that uses comparative replicability as the criterion for selecting a model as superior to its rivals (Weakliem, 2004). Specifically, we argue that the evaluation of models against arbitrary benchmarks of fit gets the researcher nowhere — only in the context of model comparison can science advance meaningfully (Burnham & Anderson, 2004). Maximizing generalizability involves ranking competing models against one another in terms of their ability to fit present and future data. Adopting this model selection strategy, however, necessitates proper quantification of *model complexity* — the average ability of a model to fit any given data. Most model fit indices include an adjustment for complexity that is a simple function of the number of free model parameters. We argue that this adjustment is insufficient; the average ability of a model to fit data is not completely governed by the number of parameters. Consequently, we present and illustrate the use of a new information-theoretic selection criterion that quantifies complexity in a more appropriate manner. This, in turn, permits the adoption of an appropriate model selection strategy that avoids pitfalls associated with LR tests.

We begin by providing a review of the traditional representation of the covariance structure model (with mean structure), with an emphasis on its application to multiple groups. We then describe advantages granted by adopting a model comparison perspective in CSM. One way around the problems with traditional approaches is to change the hypothesis under scrutiny to a more realistic one. In describing this alternative approach, we describe an approach to power analysis in CSM involving an extension of recently introduced methods to nested model scenarios. Following our discussion of power, we further explore the potential value of adopting a model selection approach that avoids hypoth-

esis testing — and thus most problems associated with LR tests—altogether. In the process, we introduce the topic of model complexity, suggesting and illustrating the use of a new selection criterion that permits appropriate model comparison even for nonnested models.

COVARIANCE STRUCTURE MODELING

Covariance structure modeling (CSM) is an application of the general linear model combining aspects of factor analysis and path analysis. In CSM, the model expresses a pattern of relationships among a collection of observed (manifest) and unobserved (latent) variables. These relationships are expressed as free parameters representing path coefficients, variances, and covariances, as well as other parameters constrained to specific, theory-implied values or to functions of other parameters. For simplicity, we restrict attention to the *all-y* model (LISREL Submodel 3B; Jöreskog & Sörbom, 1996), which involves only four parameter matrices, although the points we discuss later apply more broadly.

Model Specification

Model specification in CSM involves a data model, representing the relationship between manifest indicators and latent variables, as well as mean and covariance structures implied by the data model. The data model can be specified as:

$$\mathbf{y} = \Lambda_y \eta + \varepsilon \tag{1}$$

where $\mathbf{y}$ denotes a vector of response scores, Λ_y denotes a matrix of factor loadings regressing the p items on the m latent variables in the vector η, and ε denotes a vector of unique factors. The covariance structure obtained by taking the expectation of the square of (1) is:

$$\mathbf{\Sigma}_{yy}(\theta) = \Lambda_y \Psi \Lambda_y' + \Theta_{\varepsilon\varepsilon}, \tag{2}$$

where $\Sigma_{yy}(\theta)$ denotes the population covariance matrix of $\mathbf{y}$, with parameters (θ) in Λ_y, Ψ, and $\Theta_{\varepsilon\varepsilon}$, The covariance matrix of the latent variables is denoted Ψ, and $\Theta_{\varepsilon\varepsilon}$ denotes the (usually diagonal) covariance matrix of the unique factors.

A mean structure may also be derived by taking the expectation of (1):

$$\mu_y = \Lambda_y \alpha \tag{3}$$

where μ_y is a vector of population means of measured variables and α is a vector of latent means.

Ecological modeling often involves comparison of key model parameters across two or more groups hypothesized to differ in some important way. Extending these models to the multiple group case is straightforward. For example, Equations 2 and 3 may be extended for multiple groups as:

$$\boldsymbol{\Sigma}_{yy}^{(\mathrm{g})}(\theta) = \Lambda_y^{(\mathrm{g})}\Psi^{(\mathrm{g})}\Lambda_y^{(\mathrm{g})} + \Theta_{\varepsilon\varepsilon}^{(\mathrm{g})}, \tag{4}$$

$$\mu_y^{(\mathrm{g})} = \Lambda_y^{(\mathrm{g})}\alpha^{(\mathrm{g})}, \tag{5}$$

where the addition of a superscripted "g" denotes group membership. Equality constraints may be placed on corresponding parameters across groups.

Free parameters are estimated by employing one of a number of discrepancy minimization techniques, most often maximum likelihood (ML) or weighted least squares (WLS). The value assumed by the discrepancy function at convergence can be used to gauge the model's degree of fit to data. For example, the ML discrepancy function is:

$$F_{ML}(\mathbf{S}, \boldsymbol{\Sigma}) = \ln|\boldsymbol{\Sigma}| - \ln|\mathbf{S}| + tr[\mathbf{S}\boldsymbol{\Sigma}^{-1}] - p \tag{6}$$

When the model is "correct" and if N is large enough, $(\mathrm{N}-1)\hat{F}_{ML}$ is distributed as χ^2 with $df = p(p+1)/2 - q$, where q is the effective number of free parameters. The χ^2 statistic can be used to determine if the degree of model misfit is within chance levels, and serves as the basis for a variety of model fit indices and selection criteria.

The Importance of CSM to Ecological Research

There are several advantages associated with CSM that make it especially appropriate for addressing hypotheses in the context of ecological models. First, CSM permits the specification and testing of complex causal and correlational hypotheses. Sets of hypotheses can be tested simultaneously by constraining model parameters to particular values, or equal to one another within or across multiple groups or occasions of measurement, in ways consistent with theoretical predictions. Second, by permitting several measured variables to serve as indicators of unobserved latent variables, CSM separates meaningful variance from variance specific to items, allowing researchers to test structural hypotheses relating constructs that are not directly observed. Third, CSM is appropriate for testing correlational or causal hypotheses using either (or both) experimental or observational data. One of the central ideas behind ecological modeling is that there is much knowledge to be gained by collecting data observed in context

that would be difficult or impossible to learn under artificial conditions. Finally, CSM is a flexible modeling approach that can easily accommodate many novel modeling problems.

The Importance of Adopting a Model Comparison Perspective

In practice, CSMs are typically evaluated against benchmark criteria of good fit. Based on how well a model fits data relative to these criteria, the model is usually said to fit well or poorly in an absolute sense. The reasoning underlying this strategy of gauging a model's potential usefulness is predicated on an approach to science termed *falsificationism* (e.g., Popper, 1959), which holds that evidence accumulates for theories when their predictions are subjected to, and pass, realistic "risky" tests. If a model passes such a test under conditions where it would be expected to fail if false (i.e., if it shows good fit), evidence accumulates in favor of the theory whose predictions the model represents. If it fails, the model is either rejected or modified, with implications for the revision or abandonment of the theory. Ideally, a model is subjected to repeated risky tests to give a better idea of its long-term performance, but replication is unfortunately rare in the social sciences.

An alternative philosophical perspective maintains that the evaluation of models in isolation tells us very little, and that the fit of a model to a particular data set is nearly uninformative. Rather, science progresses more rapidly if competing theories are compared to one another in terms of their abilities to fit existing data and, as we will discuss, their abilities to fit *future* data arising from the same latent process (Lakatos, 1970; MacCallum, 2003). This approach is sometimes termed *strong inference* (Platt, 1964), and involves model comparison as a signature feature. We know from the outset that no model can be literally true in all of its particulars, unless one is extraordinarily lucky or possesses divinely inspired theory-designing skills. But it stands to reason that, given a set of alternative models, one of those models probably represents the objectively true data-generating process better than other models do. It is the researcher's task to identify this model and use it as the best working hypothesis until an even more appropriate model is identified (which, by design, inevitably happens). Every time a model is selected as the optimal one from a pool of rivals, evidence accumulates in its favor. This process of rejecting alternative explanations and modifying and re-testing models against new data continues *ad infinitum*, permitting scientists to constantly update their best working hypotheses about the unobserved processes underlying human behavior.

Because no model is literally true, there is an obvious logical problem in testing the null hypothesis that a model fits data perfectly in the population. Yet, this is precisely the hypothesis tested by the popular LR test of model

fit. Moreover, most fit indices require the researcher to choose arbitrary values to represent benchmarks of good fit. A model comparison approach goes far in avoiding these problems, although it cannot avoid them altogether. Most damning, it is possible to assert *apriori* that the hypothesis tested with the χ^2 statistic — that a model fits exactly in the population or that two models share exactly the same fit — is false in virtually every setting (Bentler & Bonett, 1980; Tucker & Lewis, 1973). A model selection approach avoids the pitfalls inherent in hypothesis testing by avoiding such tests altogether.

In addition to adhering more closely to scientific ideals and circumventing logical problems inherent in testing isolated models, the practice of model comparison avoids some problems associated with confirmation bias. Confirmation bias reflects the tendency for scientists unconsciously to increase the odds of supporting a preferred hypothesis (Greenwald, Pratkanis, Leippe, & Baumgardner, 1986). Regardless of why or how much the deck is stacked in favor of the researcher's preferred model in terms of absolute fit, one model is virtually guaranteed to outperform its rivals. Model comparison does not entirely eliminate confirmation bias, but it certainly has the potential to improve the researcher's objectivity.

In the foregoing we have explained that the popular LR test is fundamentally flawed in that the hypothesis it tests is rarely or never true in practice; thus, persistent and frequent use of the LR test is of questionable utility. We have also explained that adopting a model selection approach, in which at least two theory-inspired models are compared, has potentially greater scientific potential. In the following two broad sections, we outline some practical solutions to logical problems imposed by use of the traditional LR tests of model fit in ecological research. The first suggested approach emphasizes the utility of avoiding the hypothesis that two models have identical fit in favor of a hypothesis that the difference is within tolerable limits. This approach recognizes that no model can realistically fit perfectly in the population, and points out that shifting the focus to a less stringent hypothesis is more logical, yet has consequences for statistical power and identifying the necessary sample size. We describe and discuss methods that can be used to address these problems. The second section focuses more closely on the model selection perspective just outlined, emphasizing that model fit is overrated as a criterion for the success or usefulness of a theory. Rather, more attention should be paid to a model's ability to cross-validate, or generalize, relative to competing models. Special attention is devoted to a new model selection criterion that considers aspects of model complexity beyond simply the number of free parameters.

POWER ANALYSES FOR TESTS OF DIFFERENCE BETWEEN MODELS

Researchers often conduct tests of the difference between competing models. Such difference tests are commonly performed, for example, when one is interested in determining the level of factorial invariance characterizing a scale administered to two samples from different populations (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). In principle, this strategy involves specifying at least two models, with one model nested within the other, and the test of difference draws upon the general theory of LR tests to construct decision rules. To formalize, suppose that we are given two models, Model A nested in Model B, with degrees of freedom d_A and d_B for A and B, respectively. We assume $d_A > d_B$, and we denote the *population* ML discrepancy function values for the two models as F_A and F_B. When the two models are fitted to the sample covariance matrix, the sample discrepancy function values are minimized, and we denote them as $\hat{F}_A$ and $\hat{F}_B$. The difference between the two sample discrepancy function values, when scaled by a factor of $(N - 1)$, is commonly referred to as the *chi-square difference*, or $\Delta\chi^2$. In this chapter, we denote this well-known likelihood ratio (LR) test statistic for the difference between models as:

$$T = (N-1)(\hat{F}_A - \hat{F}_B). \tag{7}$$

The No-Difference Hypothesis

In applications, the most frequently encountered test of difference involves the specification of the null hypothesis $H_0 : (F_A - F_B) = 0$, i.e., the two models yield the same population discrepancy function values, against the general alternative of no restrictions, using T as the test statistic. We refer to this test as the test of *no-difference*. Under the null hypothesis, the asymptotic distribution of T is that of a central χ^2 variable with $d = (d_A - d_B)$ degrees of freedom. By fixing an α-level, a critical value c can be obtained from a table of the reference chi-square distribution such that

$$\alpha = 1 - G(c; d), \tag{8}$$

where $G(c;d)$ is the cumulative distribution function (CDF) of a central χ^2 random variable with d degrees of freedom, and $1 - G(c;d)$ gives the tail-area probability of this χ^2 variable to the right of c. If T exceeds c, the null hypothesis is rejected. In practice, rejecting the null hypothesis leads to the conclusion that there is a statistically significant difference in the fit of the two models, with B fitting better than A.

Contemplation of the null hypothesis tested in the no-difference test reveals that it is actually rather uninteresting from a substantive perspective, for the very same reason why the null hypothesis in the χ^2 test of goodness of fit for a single model is like a "straw man." We do not expect that two nested models would ever yield exactly the same discrepancy function values in the population, and whether or not the null hypothesis is rejected is primarily a function of the sample size (Bentler & Bonett, 1980; Tucker & Lewis, 1973). Therefore, the usual no-difference test, if applied blindly, can have serious consequences for model selection. That being said, a pressing issue in applications of the no-difference test is the lack of a simple procedure to perform power analysis, so that researchers can have at least some sense of the power of the test given the size of the existing sample, or can plan ahead in study designs to ensure that N is large enough to achieve an adequate level of power to detect the difference.

Power Analysis for the No-Difference Test

Conducting power analysis requires knowledge of the distribution of the test statistic under the alternative hypothesis. The power analysis procedure outlined here is an extension of the results in MacCallum, Browne, and Sugawara (1996), and hence follows their general principle. To begin, we state a well-known distributional result given in Steiger, Shapiro, and Browne (1985). When H_0 is false, an alternative hypothesis of the form $H_1 : (F_A - F_B) = \delta$ must be true, where $\delta > 0$. Under the assumption of *population drift*, the distribution of T under H_1 is approximately noncentral χ^2 with d degrees of freedom and noncentrality parameter

$$\lambda = (N-1)(F_A - F_B) = (N-1)\delta. \tag{9}$$

The *population drift* assumption basically stipulates that neither Model A nor B be badly misspecified (for details, see Steiger et al., 1985, p. 256).

Given both the null and alternative distributions of the test statistic T, computation of power of the no-difference test requires specification of the noncentrality parameter, δ, but this proves difficult if one attempts to somehow compute it directly using the ML discrepancy function defined in Equation 6, because the scale of the maximum Wishart likelihood is not directly interpretable. We need a sensible way to establish δ, preferably in terms of a measure of effect size that is easily interpretable and on a more standardized scale. We note that the specification of the noncentrality parameter is a common theme in all power analysis procedures, and once it is specified, computation of power becomes straightforward.

One viable option is to establish δ by using a measure of goodness of fit that is a function of the population discrepancy function value. More specifically,

we propose using the RMSEA measure (Browne & Cudeck, 1993; Steiger & Lind, 1980) because the scale of this measure is somewhat better understood in comparison with alternative measures such as the GFI and AGFI (Jöreskog & Sörbom, 1996), especially when it is applied in the context of power analysis for covariance structure models (see, e.g., MacCallum & Hong, 1997). There already exist guidelines for interpreting this measure (Browne & Cudeck, 1993) and it has been studied in large-scale simulations (e.g., Curran, Bollen, Paxton, Kirby, & Chen, 2002; Hu & Bentler, 1999). Although there are no rigid decision rules regarding the interpretation of RMSEA values, it is relatively common in applications to view RMSEA values in the range of .05 or lower as indicating close fit, values in the range of .07 – .08 as fair fit, and values greater than .10 as poor fit. Note that there is also simulation evidence that these cutoff values may change as model characteristics change (Curran et al., 2002). It has also been shown that the magnitude of error variances may impact RMSEA values (Browne, MacCallum, Kim, Andersen, & Glaser, 2002). Although not infallible, we feel that in general RMSEA serves the purpose of specifying the noncentrality parameter reasonably well.

We designate the two population RMSEA values as $\varepsilon_A = \sqrt{F_A/d_A}$ and $\varepsilon_B = \sqrt{F_B/d_B}$, for Model A and Model B, respectively. By simple algebra, we find that δ can be expressed in terms of the pair of RMSEA values as

$$\delta = (F_A - F_B) = (d_A\varepsilon_A^2 - d_B\varepsilon_B^2). \tag{10}$$

Therefore, the researcher may simply choose RMSEA values for each model in such a way as to represent the smallest difference in model fit that would be desirable to detect. Then δ and λ can be computed immediately from Equations 9 and 10. Note that one would normally choose $\varepsilon_A > \varepsilon_B$ because Model A is more constrained than model B and would tend to have poorer fit. Once this pair of RMSEA values is chosen, and λ determined, the distribution of T under the alternative hypothesis is completely specified, and computation of power becomes routine. Let π be the power of the test under consideration, then

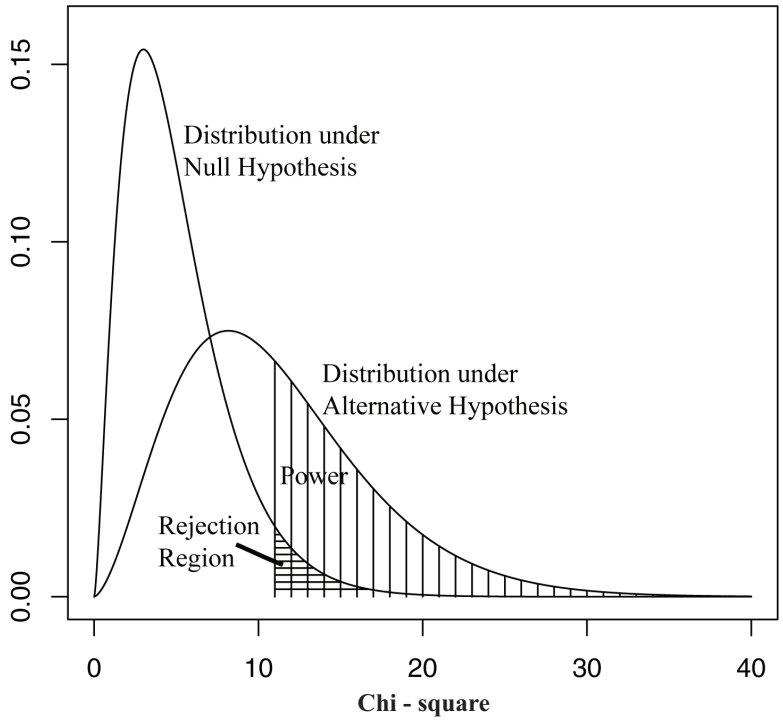
$$\pi = 1 - G(c; d, \lambda), \tag{11}$$

where $G(c;d,\lambda)$ is the cumulative distribution function of a noncentral χ^2 variable with d degrees of freedom and noncentrality parameter λ, and $1 - G(c;d,\lambda)$ gives the tail-area probability of this noncentral χ^2 variable to the right of the critical value c. Distributions and relevant areas for a typical case are illustrated in Figure 3.1. A SAS program for the computation of power is provided in MacCallum et al. (2005).

As an illustration, we apply the foregoing procedure to an empirical study. Shorey, Snyder, Yang, and Lewin (2003) compared a series of structural equation models using the no-difference test (pp. 700-701, esp. Table 2), and we

demonstrate how power can be computed for the test of the difference between what they called Model 1 and Model 5, where Model 5 is nested in Model 1.

FIGURE 3.1
Null and alternative distributions of the test statistic for determining statistical Power.



Adapting their numbering system, their Model 1 corresponds to Model B in our notation, and their Model 5 is our Model A. Their sample size was $N = 196$. To compute power, we need to specify a pair of RMSEA values. Some general guidelines for choosing the RMSEA values can be found in MacCallum et al. (2006). There is no doubt that a better choice can be made by incorporating substantive knowledge, but here we simply choose $\varepsilon_A = .06$ and $\varepsilon_B = .04$ for illustrative purposes. Using these values, we obtain $\delta = .2144$ from Equation 10, so the noncentrality parameter is $\lambda = (N - 1)d = 41.808$ using Equation 9.

From Equation 11, we compute the tail-area probability to the right of c under the noncentral χ^2 distribution with 4 degrees of freedom and noncentrality λ, and this probability is equal to .99. Thus, for Models A and B as described earlier, 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 $N = 196$ and $\alpha = .05$, the probability of rejecting that null hypothesis is approximately .99.

The result of the test conducted by Shorey et al. (2003) can be summarized as follows. With a sample size of $N = 196$, Model A yielded a minimum discrepancy function χ^2 of 191.64 with $d = 104$ degrees of freedom, and for Model B, $\chi^2 = 190.11$ with $d = 100$ degrees of freedom. Consequently, $T = 1.53$. Referring T to the CDF of a central χ^2 distribution with $d = 4$ degrees of freedom, we find the p-value for the null hypothesis, $\mathrm{H}_0 : (F_A - F_B) = 0$, to be 0.82, so there is not enough evidence to reject the null. Because the probability of rejecting the null hypothesis of no difference was about .99, the authors' finding of a nonsignificant difference between the two models in question cannot be attributed to low statistical power (at least under the conditions of the power analysis just presented).

A related question is the determination of the sample size N necessary to achieve a desired level of power given Models A and B and a specified effect size. The capability to address questions of this form would be valuable in research design. A simple procedure has been developed and is described along with corresponding SAS code in MacCallum et al. (2005).

It is also of interest to see whether the method we described can be adapted to the case of multisample models. 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) proposes a modification of this definition wherein RMSEA is expressed as $\varepsilon = \sqrt{G}\sqrt{F/df}$, where G is the number of groups, but this new definition alters the original interpretation of RMSEA as the square root of discrepancy per degrees of freedom. We choose to retain the original definition of RMSEA, i.e., $\varepsilon = \sqrt{F/df}$, and thus the developments in Equations (10) and (11) are left unaltered even in the multisample case. Researchers wishing to adopt Steiger's (1998) definition may simply divide the right-hand side of Equation (10) by $\sqrt{G}$, and from there everything else remains the same.

Alternative Power Analysis Procedures for the No-Difference Test

The critical feature of the method for power analysis described earlier is the use of the RMSEA fit measure as the basis for establishing an effect size, and in turn a value of the noncentrality parameter of the distribution of T under the

alternative hypothesis. This RMSEA-based approach follows directly from previous work by MacCallum et al. (1996), but it is not the only possible method for establishing the noncentrality parameter. A more complete account of alternative procedures is given in MacCallum et al. (2006), so here we mention only one method that is quite different from our procedure.

This alternative procedure is essentially an extension of Satorra and Saris' (1985) power analysis procedure for testing the fit of a single model. Adapting their method to the case of two competing models, with A nested in B, the first step is to establish numerical values for all parameters in Model B. Given such a set of parameter values for Model B, the implied covariance matrix $\mathbf{\Sigma}_B$ can be easily computed, for instance, by simply fixing all parameter values in a structural equation modeling software application and taking the implied covariance matrix from the output as $\mathbf{\Sigma}_B$. Model A is then fit to $\mathbf{\Sigma}_B$, yielding a discrepancy function value designated F_A. The noncentrality parameter for the distribution of T under the alternative hypothesis is $\lambda = (N - 1)F_A$. (Note that in general $\lambda = (N - 1)(F_A - F_B)$, but that $F_B \equiv 0$ in the Satorra-Saris formulation.) From this point on, the computation of power proceeds exactly as defined earlier in Equation 11. Therefore, the alternative procedure due to Satorra and Saris simply uses a different approach for specifying λ. However, there are important conceptual differences between the two power analysis procedures. The method based on the work of Satorra and Saris treats Model B as correctly specified in the population, hence $F_B \equiv 0$, and Model A as misspecified in a way defined by the difference in specification of the two models. The parameters that differentiate the models are assigned numerical values that are treated as if they were true population values. By contrast, the RMSEA-based procedure that we proposed earlier does not treat either model as if it were correct in the population. Rather, both models can be viewed as being incorrect in the population (via specification of nonzero RMSEA values), a feature that is undoubtedly more consistent with the nature of models in the real world. In addition, in the procedure we propose, there is no need to assign numerical values to parameters that the models have in common, and the outcome of the power computation will not depend on which parameters differentiate the models.

A Null Hypothesis of Small-Difference

As we have argued earlier, the null hypothesis of no difference between two competing models is of limited practical value, and we also mentioned that the same issue is present in the context of testing the fit of a single model. The null hypothesis in a conventional LR test is that the model under scrutiny is exactly correct in the population, which is always false in practice for any parsimonious

model. To deal with this problem, Browne and Cudeck (1993) proposed a test of the 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's (1993) approach can be viewed as a direct application of The Good-Enough Principle (Serlin & Lapsley, 1985), which, when applied to the present context, basically holds that a range hypothesis of acceptable fit is preferable to a point hypothesis of perfect fit.

We suggest that in the context of model comparisons, the Good-Enough Principle can be applied constructively again, by considering a null hypothesis of small difference in population discrepancy function values. 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. Building on Browne and Cudeck's (1993) work, we make use of the RMSEA as a basis for establishing δ^*. By the Good-Enough Principle, 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 $\varepsilon_B^* = .05$). Once this pair of RMSEA values is determined, δ^* can be computed from the relationship between fit function values and the RMSEA laid out in Equation (10), i.e., $\delta^* = (d_A \varepsilon_A^{*2} - d_B \varepsilon_B^{*2})$. We still use T as the test statistic, but under this null hypothesis, T has a noncentral χ^2 distribution with $d = (d_A - d_B)$ degrees of freedom, and noncentrality parameter $\lambda^* = (N - 1)d^*$. Then the decision as to whether H_0 is rejected at level α becomes a matter of finding a critical value c^* from the aforementioned noncentral χ^2 distribution, such that

$$\alpha = 1 - G(c^*; d, \lambda^*), \tag{12}$$

where $G(c^*;d,\lambda^*)$ is the CDF of a noncentral χ^2 variable with d degrees of freedom, and noncentrality parameter λ^*. Rejection of the null hypothesis of small difference implies that the observed difference between the models is too large for us to believe that the true difference is small. Failure to reject will imply that the observed difference is small enough for us to believe that the true difference is small. Such an approach will also alleviate the sample size problem associated with the no-difference hypothesis test. SAS code for carrying out the necessary computations is provided in MacCallum et al. (2006).

To illustrate the potential utility of this approach in evaluating differences between models, we test a null hypothesis of a small difference in fit in an empirical example that utilizes multi-sample CSM analysis in a cross-cultural context. Kang, Shaver, Sue, Min, and Jing's (2003) study involved respondents from four countries, and total $N = 639$. On page 1603 the authors tested a series of nested models to determine whether a particular coefficient should be

constrained to be equal across groups. The χ^2 difference is $T = 9.05$ with $d = d_A - d_B = 761 - 758 = 3$. For the test of no difference, the critical value from a central χ^2 distribution is $c^* = 7.81$, so the decision is to reject the constraints imposed in Model A, meaning that the groups differ significantly with respect to this particular path coefficient. It would be interesting to look at the result of a test of small difference, say, $\varepsilon_A^* = .06$ and $\varepsilon_B^* = .05$, so $\delta^* = [(761)(.06)^2 - (758)(.05)^2] = .8446$ by Equation 10. We can then define the null hypothesis as $H_0 : (F_A^* - F_B^*) \leq .8446$. The reference distribution for T under this null hypothesis is noncentral χ^2 with $d = 3$ degrees of freedom and noncentrality parameter $\lambda^* = (639 - 1)(.8446) = 538.85$. At $\alpha = .05$, the critical value is $c^* = 619.99$, indicating clear failure to reject the null hypothesis of the small difference as represented by $\varepsilon_A^* = .06$ and $\varepsilon_B^* = .05$. Therefore the constraint in question may well be plausible (according to the Good-Enough Principle) and perhaps should not have been rejected based on the result of the test of no-difference alone.

Power Analysis for the Small-Difference Test

Given the preceding developments, it is straightforward to combine our power analysis procedure with the specification of the null hypothesis of small difference into 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.

We give a brief account of this procedure here; a more complete discussion can be found in MacCallum et al. (2006). The null hypothesis is that of a small difference in fit, that is, $H_0 : (F_A^* - F_B^*) \leq \delta_0^*$. The alternative hypothesis specifies that the difference $(F_A^* - F_B^*)$ be greater than δ_0^*, i.e., $H_1 : (F_A^* - F_B^*) = \delta_1^*$, where $\delta_1^* > \delta_0^*$. As before, we suggest establishing useful values of δ_0^* and δ_1^* by selecting two pairs of RMSEA values and obtaining δ_0^* and δ_1^* using Equation 10. 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. 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. From then on, the procedure follows the same general template as the method described earlier, with the simple modification that now both the null and the alternative distributions (as shown in Figure 3.1) are noncentral χ^2. Specifically, the distribution of the test statistic T under H_0 will be noncentral χ^2, with $d = (d_A - d_B)$ degrees of freedom and noncentrality parameter $\lambda_0 = (N - 1)\delta_0^*$. The distribution under H_1 will be noncentral χ^2 with the same degrees of freedom and noncentrality parameter $\lambda_1 = (N - 1)\delta_1^*$. Given a specified level of α, a critical value can be

determined from 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 3.1 earlier. Again, SAS code for carrying out the necessary computations is provided in MacCallum et al. (2006).

Concluding Remarks

There are two broad issues that we wish to emphasize to close this section on power analysis and specification of the null hypothesis when performing comparisons of nested models. The first issue is the choice of pairs of RMSEA values. Essentially the results of any application of any of the methods we described are contingent on the particular RMSEA values that the user selects. Here we can offer only some general principles. For a more thorough discussion of this issue we refer the reader to MacCallum et al. (2006). For specifying RMSEA values for testing a null hypothesis of a small difference in fit, the user should regard the Good-Enough Principle (Serlin & Lapsley, 1985) as the objective, and pick RMSEA values for Models A and B that represent a difference so small that the user is willing to ignore it. In the context of power analysis, the relevant general principle would be to choose values that represent a difference that the investigator would wish to have a high probability of detecting. In practice, users will need to rely on guidelines for the use of RMSEA as mentioned earlier (Browne & Cudeck, 1993; Steiger, 1994), as well as the characteristics of the models under comparison.

The second issue has to do with the assumptions involved in our developments. All of the methodological developments presented thus far rely on well known distribution theory and its assumptions. Specifically, we make extensive use of the assumptions that ensure the chi-squaredness of the LR test statistic T, for both the central and noncentral cases. These include multivariate normality, the standard set of regularity conditions on the likelihood to carry out asymptotic expansions, and the population drift assumption (Steiger et al., 1985). As always, however, such assumptions never hold exactly in the real world, so the user should always be cautious in the application of these methods in data analysis and should watch for potential pitfalls due to assumption violations. MacCallum et al. (2006) discuss the consequences of such violations.

MODEL SELECTION AND MODEL COMPLEXITY

Model Selection and Generalizability

In the preceding section we provide and illustrate methods for comparing rival models in terms of a noncentrality-based fit index, RMSEA. We suggest that

this strategy is appropriate for statistically comparing the fit of rival, parametrically nested models, but the procedure depends in part on the researcher's judgment of appropriate choices for ε_A^* and ε_B^*, or what, in the researcher's judgment, constitutes the smallest difference in fit that it would be interesting to detect. In practice, a model can demonstrate good fit for any number of reasons, including a theory's proximity to the objective truth (or *verisimilitude*; Meehl, 1990), random chance, simply having many free parameters, or by possessing a structure allowing parameters to assume values which lead to good model fit for many different data patterns—even those generated by other processes not considered by the researcher. In other words, models can demonstrate close fit to data for reasons other than being "correct," even if one grants that true models are possible to specify (we do not), so good fit should represent only one criterion by which we judge a model's usefulness or quality.

Another criterion of model success that has found much support in mathematical psychology and the cognitive modeling literature is *generalizability* (or replicability). The idea here is that it is not sufficient for a model to show good fit to the data in hand. If a model is to be useful, it should *predict* other data generated by the same latent process, or capture the regularities underlying data consisting of signal and noise. If a model is highly complex, refitting the model to new data from scratch will not advance our knowledge by much; if a model's structure is complex enough to show good fit to one data set, it may be complex enough to show good fit to many other data sets simply by adjusting its parameters. In other words, pure goodness of fit represents fit to signal *plus* fit to noise. However, if model parameters are fixed to values estimated in one setting, and the model still demonstrates good fit in a second sample (i.e., if the model *cross-validates* well), the model has gained considerable support. A model's potential to cross-validate well is its generalizability, and it is possible to quantify generalizability based only on knowledge of the model's form and of its fit to a given data set. By quantifying a model's potential to cross-validate, generalizability avoids problems associated with good fit arising from fitting error or from a model's flexibility. It also does not rely on unsupportable assumptions regarding a model's absolute truth or falsity. Therefore, generalizability is arguably a better criterion for model retention than is goodness of fit per se (Pitt & Myung, 2002).

Earlier we stated that adopting a model selection perspective requires a fundamental shift in how researchers approach model evaluation. Traditional hypothesis testing based on LR tests results in a dichotomous accept–reject decision without quantifying how much confidence one should place in a model, or how much relative confidence one should place in each member of a set of rival models. In model comparison, on the other hand, no null hypothesis is tested (Burnham & Anderson, 2004). The appropriate sample size is not selected

based on power to reject hypotheses of exact or close fit (obviously, since no such hypotheses are tested), but rather to attain acceptable levels of precision of parameter estimates. Rather than retaining or discarding models on a strict accept–reject basis, models are ranked in terms of their generalizability, a notion that combines fit with parsimony, both of which are hallmark characteristics of a good model.

The model selection approach does not require that any of the rival models be correct, or even (counterintuitively) that any of the models fit well in an absolute sense. The process is designed in such a way that researchers will gravitate toward successively better models after repeated model comparisons. The more such comparisons a particular model survives, the better its track record becomes, and the more support it accrues. Therefore, it is incumbent upon scientists to devise models that are not only superior to competing models, but also perform well in an absolute sense. Such models will, in the long run, possess higher probabilities of surviving risky tests, facilitate substantive explanation, predict future data, and lead to the formulation of novel hypotheses. But, again, the model selection strategy we advocate does not require that any of the competing models be correct or even close to correct in the absolute sense.

Adjusting Fit for Complexity

With rare exceptions, traditional fit indices in CSM are based on the LR test statistic described earlier in Equation (7). In recognition of the limitations of the raw χ^2, most indices employ some correction for model complexity. For example, the Δ_2 index proposed by Bollen (1989) subtracts df from the denominator as an adjustment for complexity:

$$\Delta_2 = \frac{\chi_b^2 - \chi_m^2}{\chi_b^2 - df_m} \tag{13}$$

where χ_b^2 is the fit associated with a baseline model and χ_m^2 and df_m are associated with the hypothesized model. The adjustment has the effect of penalizing fit for the number of free parameters. Many fit indices contain similar adjustments. For example, RMSEA divides by df as a way of distributing lack of fit across all parameter constraints. In this way, RMSEA penalizes fit due to unnecessary free parameters.

However, complexity is not governed completely by the number of free parameters (MacCallum, 2003; Pitt, Myung, & Zhang, 2002; Preacher, 2003, in press). The corrections employed in most indices carry the implicit assumption that all free parameters contribute equally to a model's ability to fit data (or, that all model constraints contribute equally to *lack* of fit). Yet it is easy to see how, in a loose sense, some parameters may be more important than others

in a given model. For example, constraining a covariance parameter linking otherwise disparate sections of a model to zero would probably limit a model's potential to fit data more than would constraining, say, a factor loading to zero. More generally, parameters that appear in many equations for implied covariances likely influence complexity more so than do parameters that appear in fewer equations. Fortunately, information theory offers some alternatives to traditional fit indices that avoid quantifying complexity as if it were a strict linear function of the number of parameters.

Information-Theoretic Criteria

In contrast to model selection methods rooted in Bayesian or frequentist traditions, much research points to information theory as a likely source for the optimal model selection criterion. Selection criteria based on information theory seek to locate the one model, out of a pool of rival models, which shows the optimal fidelity, or signal-to-noise ratio; this is the model that demonstrates the best balance between fit and parsimony. This balance was termed *generalizability* earlier. Several popular model selection criteria were either derived from, or are closely related to, information theory. The most popular such criteria are the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwartz, 1978). Excellent treatments of AIC and BIC can be found elsewhere (e.g., Burnham & Anderson, 2002, 2004; Kuha, 2004).

Many information-based criteria may be construed as attempts to estimate the Kullback–Leibler (K–L) distance. The K-L distance is the (unknown) information lost by representing the true latent process with an approximating model (Burnham & Anderson, 2004). Even though we cannot compute the K–L distance directly because there is one term in the K–L distance definition that is not possible to estimate, we can approximate *relative* K–L distance in various ways by combining knowledge of the data with knowledge of the models under scrutiny. Of great importance for model comparison, the ability to approximate relative K–L distance permits the ranking of models in terms of their estimated verisimilitude, tempered by our uncertainty about the degree of approximation. In other words, using information-based criteria, models can be ranked in terms of estimated generalizability.

Minimum Description Length and the Normalized Maximum Likelihood

Information-based criteria such as AIC and BIC are used with great frequency in model comparisons and with increasing frequency in applications of CSM. However, they suffer from at least two major drawbacks. First, they employ

complexity adjustments that are functions only of the number of free model parameters. Second, they implicitly require the strong assumption that a correct model exists. We focus instead on a newer criterion that remains relatively unknown in the social sciences, yet we feel has great promise for application in model selection. This is the principle of *minimum description length* (MDL: Grünwald, 2000; Myung, Navarro, & Pitt, 2005; Rissanen, 1996, 2001; Stine, 2004). The MDL principle involves construing data as compressible strings, and conceiving of models as compression codes. If models are viewed as data compression codes, the optimal code would be one that compresses (or simply represents) the data with the greatest fidelity. With relevance to the limitations of criteria such as AIC and BIC, the MDL principle involves no assumption that a true model exists. If one accepts that a model's proximity to the truth is either undefined (i.e., that the notion of a true model is merely a convenience and bears no direct relation to reality) or is at any rate impossible to determine, then the MDL principle offers a viable alternative to traditional methods of model selection. Excellent discussions of the MDL principle can be found in Grünwald (2000), Grünwald, Myung, and Pitt (2005), Hansen and Yu (2001), and Markon and Krueger (2004). Three quantifications of the MDL principle are *normalized maximum likelihood* (NML), *Fisher information approximation* (FIA), and *stochastic information complexity* (SIC). NML is quantified as:

$$\text{NML} = \frac{L(y|\hat{\theta})}{\int\limits_{S} L(z|\hat{\theta}(z))dz}, \tag{14}$$

or the likelihood of the data given the model divided by the sum of all such likelihoods. FIA is quantified as:

$$\text{FIA} = -\ln L\left(y|\hat{\theta}\right) + \frac{q}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int\limits_{\Theta}\sqrt{|I\left(\theta\right)|}d\theta, \tag{15}$$

an approximation to the negative logarithm of NML that makes use of the number of free parameters (q) and the determinant of the Fisher information matrix, $I(\theta)$. SIC, an approximation to FIA that is typically more tractable in practice, is quantified as:

$$\text{SIC} = -\ln L\left(y|\hat{\theta}\right) + \frac{1}{2}\ln\left|nI\left(\theta\right)\right|. \tag{16}$$

The Appendix (see Quant.KU.edu) contains more detailed discussion of these criteria. NML, FIA, and SIC all represent model fit penalized by the model's average ability to fit any given data.

NML is similar in spirit to selection criteria such as AIC and BIC in several respects, save that preferable models are associated with higher values of NML but with lower values of AIC or BIC.[1] All of these criteria can be framed as functions of the likelihood value adjusted for model complexity, although the complexity correction assumes different forms for different criteria. NML differs from criteria like AIC and BIC mainly in that not every parameter is penalized to the same extent. NML imposes an adjustment commensurate with the degree to which each free parameter increases complexity, as reflected in the model's general data-fitting capacity. Consequently, NML does not assume (as do AIC and BIC) that each parameter contributes equally to goodness of fit. Therefore, both parametric and structural components of complexity are considered. A major additional advantage of NML (which it shares with AIC and BIC) is that it does not require rival models to be nested. Thus, if two competing theories posit different patterns of constraints, such models can be directly compared using criteria derived from information theory.

Applying MDL in Practice

To illustrate how the MDL principle may be employed in practice, we present two brief examples from the applied literature. In both examples we compute NML; in the second, we supplement NML with computation of SIC because original data were available with which to compute the $|nI(\theta)|$ term. Neither the denominator term in NML (see Equation [A1]) nor the structural complexity term in FIA (see Equation [A2]) can be computed directly in the context of CSM. Numerical integration techniques are typically applied instead. To facilitate computation of NML, we simulated the data space by generating large numbers of random uniform correlation matrices (R) using Markov chain Monte Carlo (MCMC) methods.[2] These matrices were uniform in the sense that all possible R matrices had equal *apriori* probabilities of being generated. All models were fit to all simulated matrices, and the likelihoods were averaged to form the denominator of the NML formula.[3] The numerators were supplied by simply noting the likelihood value associated with the converged solution for each model applied to real data.

Example 1. Our first demonstration makes use of models provided by In-

[1] In fact, BIC may be obtained as a special case of MDL if structural complexity is neglected (Myung et al., 2005).

[2] Fortran 90 code is available from the first author upon request. See Preacher (2003) for details.

[3] An average was used rather than a sum because (a) computation of NML is more manageable and intuitive using the mean likelihood rather than a sum of likelihoods and (b) the rank ordering of models is not affected. Solutions with estimation errors were omitted from this averaging.

gram, Betz, Mindes, Schmitt, and Smith (2001) in a study of the effects and correlates of unsupportive social interactions.

Part of their study involved comparison of five rival confirmatory factor models, depicted in Figure 3.2, which we denote Models $I1$ – $I5$. The four primary factors in each model represent dimensions of the Unsupportive Social Interactions Inventory (USII). Each factor was measured by three 2-item parcels, for a total of $p = 12$ variables. After removing one outlier, the five models were fit to data from $N = 221$ introductory psychology students. Based on LR tests, the authors selected Model $I5$ as superior to its rivals.

NML was computed for each model. The empirical likelihoods were obtained by employing the following formula using the χ^2 statistics reported in Ingram et al. (2001, Table 4):

$$L(y|\hat{\theta}) = e^{\frac{\chi^2}{-2(N-1)}} \tag{17}$$

The complexity estimates were obtained by fitting each model to 10,000 random $\mathbf{R}$ matrices and computing the mean obtained likelihood. Computation of complexity was based only on proper solutions with no convergence errors. The resulting NML, along with the number of solutions on which computation was based, can be found in Table 3.1.

Whereas the authors chose Model $I5$ as the preferred model based on LR tests ($I5$ showed significantly better fit in terms of χ^2 than did the next-worst fitting model), application of NML indicates a preference for Model $I2$. The higher order factor model was characterized by the highest NML in the set of models compared, implying that $I2$ has greater potential for replicating in future samples than its rivals. Although Model $I5$ demonstrated the best absolute fit, it did so at the price of having a more complex structure and more free parameters.

TABLE 3.1
NML Estimates for the Five Factor Models Compared by Ingram et al. (2001)

Model	Empirical χ^2	*df*	NML	Solutions without Estimation Errors
I1 One Factor	325.34	54	2.867	9,999
I2 Higher Order	155.38	50	3.528	9,870
I3 Four Factor	152.45	48	3.223	9,848
I4 Higher Order with One Cross-Loading	145.83	49	3.003	9,699
I5 Four Factor with One Cross-Loading	131.42	47	3.188	9,716

TABLE 3.2
NML and SIC Estimates for the Three Models Compared by Larose et al. (2002)

Model	Empirical χ^2	*df*	NML	SIC	Solutions without estimation Errors
L1 Cognitive Bias	17.07	8	1.759	66.294	9,319
L2 Social Networks	24.57	7	1.520	71.088	9,221
L3 Cognitive-Network	4.94	6	1.404	76.448	9,225

Note. Because the x^2 values obtained through reanalysis differed slightly from those reported by Larose et al. (2002), we report the values we obtained. These differences are likely to due rounding error in Larose et al.'s reported correlations.

This finding has implications for the conclusions drawn by Ingram et al. (2001). Because the authors used a model selection approach that does not consider the relaive complexities of rival models, the model that showed the best absolute fit was also the one with the highest complexity, or the best *apriori* expected fit. In essence, the chosen model capitalized on an unfair advantage. In contrast, a selection criterion that appropriately adjusts fit for complexity selected a model with a better balance of fit and parsimony.

Example 2. Our second example draws on three covariance structure models compared by Larose, Guay, and Boivin (2002). The authors were primarily interested in comparing the Cognitive Bias Model and Social Network Model, two models proposed to explain variability in a Loneliness latent variable using Attachment Security, Emotional Support, and Social Support. These two models (which we denote $L1$ and $L2$) are presented in the first two panels of Figure 3.3. Based on results indicating that both models fit the data well and were thus viable explanations for the observed pattern of effects, the authors devised a third model combining features of the first two, dubbed the Cognitive-Network Model ($L3$ in Figure 3.3).

All three models were found to fit the data well using self-report measures ($N = 125$), and to fit even better using friend-report measures. In both cases, the Cognitive-Network Model was found to fit the data significantly better than either the Cognitive Bias Model or the Social Network Model. Following procedures already described, we reevaluated Larose et al.'s models (fit to self-report data) using NML. Results are reported in Table 3.2. Because raw data were available in their article, we are also able to provide estimates of SIC.

Contrary to the authors' findings, both NML and SIC indicate that the Cognitive Bias Model performs better than either the Social Networks Model or the proposed Cognitive-Network Model in terms of generalizability. Combining features of two already well-fitting models does not necessarily grant a scientific advantage when the resulting model is more complex than either of its

FIGURE 3.2
Rival models investigated by Ingram et al. (2001).

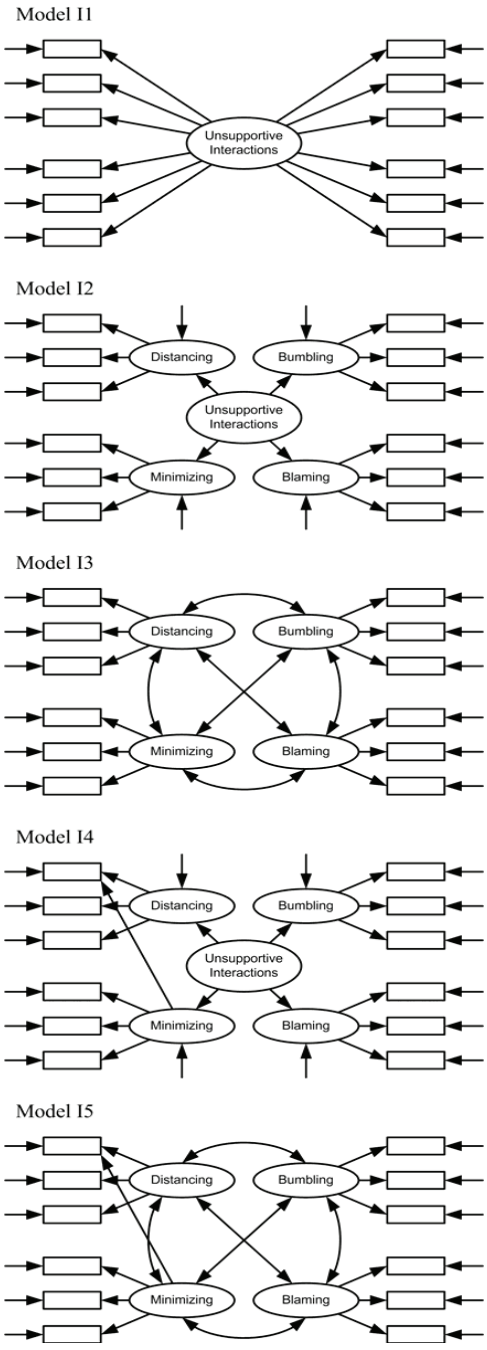
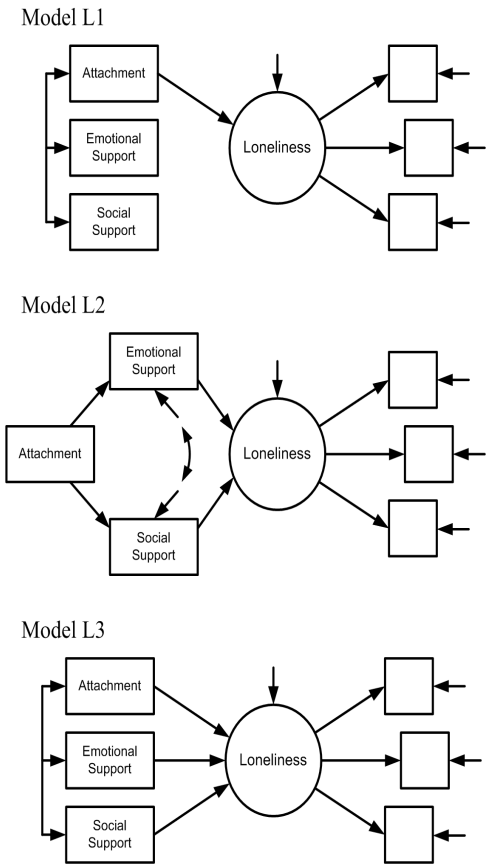


FIGURE 3.3
Rival models investigated by Larose et al. (2002).



competitors. In this instance, as in the previous example, the chosen model was selected primarily because it showed better absolute fit; this better fit was due in part to the fact that the Cognitive-Network Model was more complex than its competitors. An implication of this finding is that, whereas the Cognitive-Network Model may fit the given data set better than the Cognitive Bias Model and the Social Networks Model in absolute terms, it has a lower likelihood of generalizing well to future data.

Summary

Like other information-theoretic selection criteria, MDL does not require rival models to be parametrically nested. Nor does its use require the assumption that a true model exists. Furthermore, MDL considers more sources of complexity than simply a model's number of parameters. In sum, we feel that the MDL principle has great potential for use in model comparisons in CSM.

Limitations

Of course, NML is not a panacea. Three limitations of NML are that it is difficult to compute, it relies on the assumptions of maximum likelihood, and it involves often arbitrary bounds on the data space. The first limitation will be overcome as processor speeds increase and as NML becomes included in standard model estimation packages. In the meantime, the more tractable MDL approximation, SIC (Rissanen, 1989), can be used if the numerical integration necessary for NML proves too time-intensive. As for the second limitation, it is unknown how robust MDL methods are to violations of ML assumptions. This would be a fruitful avenue for future research.

The third limitation is more challenging because it requires the researcher to make a subjective decision regarding boundaries on the data space. We restricted attention to correlation matrices for simplicity. We recognize that many modeling applications require covariance matrices rather than correlation matrices (and sometimes also mean vectors). For example, virtually any application in which models are fit to multiple groups simultaneously, such as in factorial invariance studies, requires the use of covariance matrices. Growth curve modeling requires covariance matrices and mean vectors. Lower and upper boundaries must be imposed on generated means and variances if such data are required, and these choices constitute even more subjective input. It is generally agreed that data generated for the purpose of quantifying model complexity should be uniformly representative of the data space (Dunn, 2000), yet choices regarding the range of data generation may exert great influence on the ranking of competing models. It is thus important that reasonable bounds be

investigated to ensure reasonable and stable model rankings. A discussion of the implications for arbitrary integration ranges can be found in Lanterman (2005).

DISCUSSION

We have proposed two alternatives to traditional methods of comparing covariance structure models. Both alternatives were suggested in response to limitations of the popular LR test; the most severe limitation is that the hypothesis tested by the LR test (that two models have identical fit) is never true in practice, so investigating its truth or falsity would seem to be a questionable undertaking (MacCallum et al., 2006). The first alternative procedure posits a modified null hypothesis such that the difference in fit between two nested models is within tolerable limits. The second alternative we discuss is to compare rival (not necessarily nested) models in terms of relative generalizability using selection indices based on the MDL principle. Both methods encourage a model comparison approach to science that is likely to move the field in the direction of successively better models.

There are interesting parallels between the strategies proposed here and a framework for model assessment proposed by Linhart and Zucchini (1986) and elaborated upon by Cudeck and Henly (1991) in the context of CSM. Because it relies on RMSEA to specify null and alternative hypotheses, the first approach (using RMSEA to specify hypotheses of close fit) can be seen as way to compare nested models in terms of their *approximation discrepancy*, or lack of fit in the population. In other words, this method is a way to gauge models' relative nearness to the objectively true data-generating process, or their relative verisimilitudes. The second method of model comparison makes use of the MDL principle to facilitate comparison of models in terms of their relative generalizabilities, or abilities to predict future data arising from the same generating process. This strategy can be seen as a way to compare models (nested or non-nested) in terms of their *overall discrepancy*, tempering information about lack of fit with lack of confidence due to sampling error. When N is large, enough information is available to support highly complex models if such models are appropriate. When N is small, uncertainty obliges us to conservatively select less complex models until more information becomes available (Cudeck & Henly, 1991). Thus, NML and similar criteria are direct applications of the parsimony principle, or Occam's razor.

The parallels between the measures of verisimilitude and generalizability on one hand, and the Linhart–Zucchini and Cudeck–Henly frameworks on the other, perhaps deserve more attention in future research. High verisimilitude and high generalizability are both desirable characteristics for models to possess, but selecting the most generalizable model does not necessarily imply that

the selected model is also closest to the objective truth. Therefore we do not advocate choosing one approach or the other, or even limiting attention to these two strategies. Rather, we suggest combining these strategies with existing model evaluation and selection techniques so that judgments may be based on as much information as possible. Regardless of what strategy the researcher chooses, the strongest recommendation we can make is that researchers should, whenever circumstances permit it, adopt a model selection strategy rather than to evaluate single models in isolation. The methods illustrated here are viable alternatives to the standard approach, and can be applied easily in many modeling settings involving longitudinal and/or ecological data.

ACKNOWLEDGMENTS

This work was funded in part by National Institute on Drug Abuse Grant DA16883 awarded to the first author while at the University of North Carolina at Chapel Hill. We thank Daniel J. Navarro for providing helpful comments.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (p. 267). Budapest, Hungary: Akademiai Kiado.

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

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

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

Browne, M. W., MacCallum, R. C., Kim, C., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, *7*, 403-421.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261-304.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, *109*, 512-519.

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.

Dunn, J. C. (2000). Model complexity: The fit to random data reconsidered. *Psychological Research*, *63*, 174-182.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review*, *93*, 216-229.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*, 133-152.

Grünwald, P., Myung, I. J., & Pitt, M. A. (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: The MIT Press.

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.

Ingram, K. M., Betz, N. E., Mindes, E. J., Schmitt, M. M., & Smith, N. G. (2001). Unsupportive responses from others concerning a stressful life event: Development of the Unsupportive Social Interactions Inventory. *Journal of Social and Clinical Psychology*, *20*, 173-207.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 user's reference guide*. Uppsala: Scientific Software International.

Kang, S., Shaver, P. R., Sue, S., Min, K., & Jing, H. (2003). Culture-specific patterns in the prediction of life satisfaction: Roles of emotion, relationship quality, and self-esteem. *Personality and Social Psychology Bulletin*, *29*, 1596-1608.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, *33*, 188-229.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91-196). Cambridge, England: Cambridge University Press.

Lanterman, A. D. (2001). Schwarz, Wallace and Rissanen: Intertwining themes in theories of model selection. *International Statistical Review*, *69*, 185-212.

Lanterman, A. D. (2005). Hypothesis testing for Poisson vs. geometric distributions using stochastic complexity. In P. D. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: Theory and applications* (pp. 99-123). Cambridge, MA: The MIT Press.

Larose, S., Guay, F., & Boivin, M. (2002). Attachment, social support, and loneliness in young adulthood: A test of two models. *Personality and Social Psychology Bulletin*, *28*, 684-693.

Linhart, H., & Zucchini, W. (1986). *Model selection*. New York: Wiley.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, *38*, 113-139.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*, 19-35.

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.

Markon, K. E., & Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics*, *34*, 593-610.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*, 108-141.

Myung, J. I., Navarro, D. J., & Pitt, M. A. (2005). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167-179.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *TRENDS in Cognitive Sciences*, *6*, 421-425.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472-491.

Platt, J. R. (1964, October 16). Strong inference. *Science*, *146(3642)*, 347-353.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Preacher, K. J. (2003). *The role of model complexity in the evaluation of structural equation models.* Unpublished doctoral dissertation. Ohio State University, Columbus, OH.

Preacher, K. J. (in press). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*, 40-47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, *47*, 1712-1717.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464.

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

Shorey, H. S., Snyder, C. R., Yang, X., & Lewin, M. R. (2003). The role of hope as a mediator in recollected parenting, adult attachment, and mental health. *Journal of Social and Clinical Psychology*, *22*, 685-715.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78-90.

Steiger, J. H. (1994). *Structural equation modeling (computer program).* In Statistica/w, version 4.5. 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. C. (1980). Statistically based tests for the number of factors.

Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

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

Stine, R. A. (2004). Model selection using information theory and the MDL principle. *Sociological Methods & Research*, *33*, 230-260.

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

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.

Weakliem, D. L. (2004). Introduction to the special issue on model selection. *Sociological Methods & Research*, *33*, 167-187.

Zhang, S. (1999). *Applications of geometric complexity and the minimum description length principle in mathematical modeling of cognition.* Unpublished doctoral dissertation. Ohio State University, Columbus, OH.