Psychological Methods
2012, Vol. 17, No. 1, 1–14

1082-989X/12/$12.00 DOI: 10.1037/a0026804

The Problem of Model Selection Uncertainty in Structural Equation Modeling

Kristopher J. Preacher
Vanderbilt University

Edgar C. Merkle
University of Missouri

Model selection in structural equation modeling (SEM) involves using selection criteria to declare one model superior and treating it as a best working hypothesis until a better model is proposed. A limitation of this approach is that sampling variability in selection criteria usually is not considered, leading to assertions of model superiority that may not withstand replication. We illustrate that selection decisions using information criteria can be highly unstable over repeated sampling and that this uncertainty does not necessarily decrease with increases in sample size. Methods for addressing model selection uncertainty in SEM are evaluated, and implications for practice are discussed.

Keywords: model selection, structural equation modeling, information criteria, BIC

The goal of model selection in structural equation modeling (SEM) is to find a useful approximating model that (a) fits well, (b) has easily interpretable parameters, (c) approximates reality in as parsimonious a fashion as possible, and (d) can be used as a basis for inference and prediction. A large literature on model selection exists (Bentler & Mooijaart, 1989; Bollen & Long, 1993; Browne & Cudeck, 1992; Cudeck & Browne, 1983; MacCallum, 2003; Mulaik et al., 1989; Preacher, 2006; Preacher, Cai, & MacCallum, 2007; Raykov & Marcoulides, 1999); the reader is referred to book-length treatments by Burnham and Anderson (2002) and Linhart and Zucchini (1986).

An important—yet unspoken—presumption in model selection is that the selection process is largely devoid of error. That is, researchers typically base this all-important decision, at least partly, on a fit statistic or selection criterion without communicating the degree of precision in that statistic or criterion, which in turn implies a high degree of confidence in the decision. However, model selection methods are prone to what Burnham and Anderson (2002) called *model selection uncertainty*, or sampling variability in the decisions based on selection criteria. That is, even if researchers conscientiously follow expert recommendations for model selection, they nevertheless usually fail to consider sampling variability in the selection criteria themselves. Selection criteria are, after all, statistics based on sample data and necessarily must vary over repeated sampling. Ignoring this important source of uncertainty can lead to the rejection of potentially good approximating models, the adoption of poor models, and invalid inference. The degree to which this uncertainty is problematic in SEM remains an open question.

In this article we discuss problems stemming from sampling variability in selection indices. We first review model selection criteria and show, via simulation, that model selection decisions using information criteria (specifically the Bayesian information criterion, or BIC) can be highly unstable over repeated sampling, even in large samples. We then explore and evaluate three available approaches that could be used to address or quantify model selection uncertainty. Finally, the practical implications of model selection uncertainty in SEM are discussed.

The Bayesian Information Criterion

Because we focus so heavily on BIC as a stand-in for the class of information-based model selection criteria, we devote this section to a brief discussion of the definition and characteristics of BIC. Perhaps the most popular selection criterion is the G statistic reported by all SEM software, where $\hat{G} = n\hat{F}_{\mathrm{ML}} = (N - 1)\hat{F}_{\mathrm{ML}}$, $\hat{F}_{\mathrm{ML}} = \ln|\hat{\boldsymbol{\Sigma}}| - \ln|\mathbf{S}| + \mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S}) - p$ is the maximum likelihood discrepancy function value at convergence, $\hat{\boldsymbol{\Sigma}}$ and $\mathbf{S}$ are the model-implied and observed covariance matrices, and p is the number of variables. If normality and certain regularity conditions are met (Steiger, Shapiro, & Browne, 1985), $\hat{G}$ approximately follows a noncentral χ^2 distribution with $df = p(p + 1)/2 - q$ (where q is the number of estimated parameters) and noncentrality parameter λ (and thus $\hat{G}$ is often termed the χ^2 statistic). Tests based on $\hat{G}$ have several well-known limitations, including sensitivity to sample size, rigid adherence to a tradition of testing nil hypotheses, and failure of the noncentral χ^2 distribution to describe the behavior of $\hat{G}$ in applied situations (Yuan, 2008; Yuan, Hayashi, & Bentler, 2007). An important additional limitation is that $\hat{G}$ does not incorporate information about model parsimony. For sequences of nested models,[1] more complex (less parsimonious) models fit better than simpler ones by the $\hat{G}$ criterion, essentially giving all of the emphasis to fit at the total expense of parsimony.

A number of other indices are based on principles from information theory and were designed explicitly for comparing multiple models, principally Akaike's information criterion (AIC; Akaike, 1973) and the BIC (Raftery, 1995; Schwarz, 1978). These indices

This article was published Online First January 23, 2012.

Kristopher J. Preacher, Department of Psychology and Human Development, Vanderbilt University; Edgar C. Merkle, Department of Psychological Sciences, University of Missouri.

We thank Sonya Sterba and Taehun Lee for helpful comments.

Correspondence concerning this article should be addressed to Kristopher J. Preacher, Department of Psychology and Human Development, Vanderbilt University, 230 Appleton Place, Peabody #552, Nashville, TN 37203. E-mail: kris.preacher@vanderbilt.edu

[1] Model A is said to be nested within Model B if Model A (or a model equivalent to it) can be obtained by placing constraints on Model B.

have shown great promise for application in SEM. For example, Steiger and Lind (1980) found that BIC performed well at selecting the correct number of factors in exploratory factor analysis when the number of factors was small. Cudeck and Browne (1983) found AIC and BIC to perform comparably to a cross-validation procedure, with AIC slightly more liberal and BIC slightly more conservative than cross-validation. Homburg (1991) applied AIC, BIC, and cross-validation measures to SEM and found both AIC and BIC to perform well at identifying the data-generating model. Haughton, Oud, and Jansen (1997) found information-based indices (AIC, BIC, and variants) to greatly outperform several traditional fit indices—*G*, nonnormed fit index, incremental fit index, comparative fit index, adjusted goodness-of-fit index, critical-/*n*/, root-mean-square error of approximation (RMSEA), parsimony fit index, parsimony goodness-of-fit index—in simulations that both included and excluded the data-generating model. Among the information-based indices, variants of BIC excelled at choosing the model closest to the data-generating model without overfitting. Other authors have criticized AIC and BIC on the basis of poor performance, dependence on *N*, and the tendency for researchers to misunderstand or misapply these indices (e.g., Mulaik, 2009). In what follows, we choose BIC as a representative model selection criterion, with the understandings that (a) as with any model selection criterion, BIC has limitations; (b) the phenomenon we demonstrate applies to other criteria as well; and (c) optimal model assessment does not stop at BIC and includes diagnostic checks, residual analyses, measures of absolute fit, and so on.

Assume a set of equally plausible models $\{M_1, M_2, \ldots, M_k\}$, one of which is true. BIC will then select the model with the largest posterior probability of being true given the observed data. For general models, the posterior probability of M_i can be written as

$$P(M_i|\mathbf{Y}) \propto P(M_i) \int P(\mathbf{Y}|M_i, \theta_i) P(\theta_i) d(\theta_i), \quad (1)$$

where $\mathbf{Y}$ is the data matrix, θ_i is the parameter vector for model i, $P(M_i)$ is the prior probability of M_i, $P(\mathbf{Y}|M_i, \theta_i)$ is the likelihood associated with M_i, and $P(\theta_i)$ is the prior distribution of M_i's parameter vector. A Taylor series expansion of this equation, along with some simplification, yields the following approximation:

$$\ln P(M_i|\mathbf{Y}) = \ln P(\mathbf{Y}|M_i, \theta_i) - \frac{q_i}{2} \ln N + O(1), \quad (2)$$

where q_i is the number of parameters for M_i, N is the sample size, and $O(1)$ indicates that this approximation converges to the true posterior probability plus a constant. Multiplying by -2 yields the common form of BIC (Raftery, 1995; Schwarz, 1978) as

$$\text{BIC}_i = -2\ln P(\mathbf{Y}|M_i, \theta_i) + q_i \ln N. \quad (3)$$

The $q_i \ln N$ term can be viewed as a penalty term for model complexity due to the number of estimated parameters, weighted by the natural logarithm of sample size. In applications, the log-likelihood in Equation 3 is evaluated at the maximum likelihood estimate $\hat{\theta}_i$.

The formulation in Equation 3 is slightly different from that commonly used in SEM, given by

$$\text{BIC}_i = \hat{G}_i + q_i \ln N. \quad (4)$$

In using $\hat{G}_i$ instead of $-2\ln P(\mathbf{Y}|M_i, \hat{\theta}_i)$, the saturated model implicitly plays a role in the calculation. This is because $\hat{G}_i$ can be written as

$$-2\ln P(\mathbf{Y}|M_i, \hat{\theta}_i) + 2\ln P(\mathbf{Y}|M_s, \hat{\theta}_s), \quad (5)$$

where M_s is the saturated model with associated parameter estimates $\hat{\theta}_s$. This formulation is typically used when one is modeling only the observed covariance matrix and not the raw data, although similar issues exist regardless of whether one uses the covariance matrix or the raw data.

BIC has a number of nice properties that many researchers believe make it a good selection criterion for SEM. BIC is easy to compute from standard SEM output and can be used to compare nonnested models (Kass & Raftery, 1995; Raftery, 1995). It trades off fit and complexity, thus incorporating Occam's razor. Furthermore, the difference in BIC for two models (ΔBIC) approximates the Bayes factor (defined in the next paragraph), thus providing a rational alternative to null hypothesis testing. Finally, when the true probability model is in the set to be compared, BIC asymptotically selects it, whereas other criteria do not. Conversely, when the true model is not in the set to be compared, BIC asymptotically selects the closest approximating model (Wasserman, 1997). The latter two points are typically discussed in the context of probability models (typically of raw data) as opposed to covariance structure models that do not include mean parameters. Further, the last point implies that, when the true model is not in the set, BIC will tend to select saturated models as N goes to infinity. Other merits and limitations of BIC as a selection criterion are discussed by Burnham and Anderson (2002, 2004), Kass and Raftery (1995), Kuha (2004), Mulaik (2009), Raftery (1993, 1995, 1999), and Wasserman (1997).

In addition to the advantages just described, guidelines exist for BIC that are similar to the interpretation of effect size (BIC can be understood as a parsimony-corrected measure of relative effect size). It is difficult to interpret raw values of BIC for individual models. However, the derivations described earlier imply that the difference in BIC between two models M_i and M_j ($\Delta\text{BIC}_{ij} = \text{BIC}_i - \text{BIC}_j = \hat{G}_i - \hat{G}_j + [q_i - q_j]\ln N = \Delta\hat{G}_{ij} + \Delta q_{ij} \ln N$) approximates $-2\ln(B_{ij})$, where B_{ij} is the Bayes factor:

$$B_{ij} = \frac{P(M_i|\mathbf{Y})/P(M_j|\mathbf{Y})}{P(M_i)/P(M_j)}. \quad (6)$$

B_{ij} is the ratio of the posterior odds of model i to model j and the prior odds of model i to model j (Wasserman, 1997). Values of $B_{ij} < 1$ favor model j, as do values of $\Delta\text{BIC} > 0$.

There are two principal sets of guidelines for interpreting effect magnitude for ΔBIC. The first set appeared as Jeffreys' (1961, Appendix B) *grades of evidence* for nested models in terms of Bayes factors. Various versions of these guidelines have appeared (e.g., Raftery, 1993). Raftery (1995, p. 129) and Kass and Raftery (1995, p. 777) noted that Jeffreys' guidelines provide "a rough descriptive statement about standards of evidence in scientific investigation." Based on their own experience, Raftery (1995) and Kass and Raftery (1995) preferred a second, more conservative scale for interpreting Bayes factors (and corresponding ΔBICs) in social research (see Table 1). It is this more conservative scale that we use to define meaningful differences between models. Lee and Song (2001, 2008) also suggested this scale for use in model selection involving linear, nonlinear, and two-level SEMs. We use this scale to help address the following questions: In finite samples, how much uncertainty exists in model selection based on BIC, and what can be done to address it?

Table 1
Raftery's (1995) and Kass and Raftery's (1995) Scale for Interpreting Bayes Factors

B_{ij}	ΔBIC	Strength of evidence for model i
$B_{ij} < 1$	ΔBIC > 0	None
$1 < B_{ij} < 3$	$-2 <$ ΔBIC < 0	Weak
$3 < B_{ij} < 20$	$-6 <$ ΔBIC < -2	Positive
$20 < B_{ij} < 150$	$-10 <$ ΔBIC < -6	Strong
$150 < B_{ij}$	ΔBIC < -10	Very strong

Note. B_{ij} = Bayes factor; ΔBIC = difference in Bayesian information criterion between two models.

Model Selection Uncertainty in SEM

The study of sampling variability in BIC (and consequently in ΔBIC) is clouded by the fact that there is not always a clearly defined population value. That is, if BIC is computed as BIC $= \hat{G} + q\ln N$ for a theoretically infinite population, it would equal infinity even for perfectly fitting models. We instead used the expected value of the sample BIC given N (i.e., over infinite repeated samples of size N). We called this quantity the *expected BIC* (EBIC), the expected value of BIC based on asymptotic theory:

$$\text{EBIC} = E(\text{BIC}) = E(\hat{G} + q\ln N) = E(\hat{G}) + E(q\ln N)$$
$$= df + nF_0 + q\ln N, \quad (7)$$

where $E(\cdot)$ is the expectation operator and $E(\hat{G}) = df + \lambda$ for noncentral χ^2 distributions with noncentrality parameter $\lambda = nF_0$ (where F_0 is the discrepancy function obtained by fitting a model to the population covariance matrix; Browne, 1984; Steiger, 2000). EBIC is "asymptotic" with respect to number of samples, not the sample size, and represents the average value of BIC we would obtain after infinite repeated sampling from the same population. The resulting EBIC values are used in the simulations seen later, where they are shown to agree closely with the means of the BIC sampling distributions.

Illustration of Model Selection Uncertainty

To illustrate selection uncertainty in the SEM context, consider the nine-variable path models in Figure 1. Model D was used to generate the data (using 0.2 for unstandardized path coefficients, .8 for residual variances, and 1.0 for the single exogenous variance) under the assumption of multivariate normality. EBIC values were obtained for each generating model by fitting Models A, B, and C to the model-implied covariance matrix of the fully specified Model D, yielding values of F_0 in Equation (7) for each model and sample size. Model A moved from being strongly preferred at $N =$ 80 to 500 to being ranked in the middle at $N =$ 1,500 and last at $N =$ 5,000 (see Table 2). Whereas the differences in EBIC for these models are often substantial according to Raftery's (1995) criteria (depending on N), much larger BIC differences are often observed in practice.[2]

Say a researcher wishes to compare Models A, B, and C in Figure 1 using empirically obtained data. Unknown to the researcher, Model D is the true data-generating process, but typically the data-generating model does not number among the models being compared. The researcher uses a sample size of $N = 500$, which well exceeds power requirements for models ranging in df from 20 to 27 (MacCallum, Browne, & Sugawara, 1996) and likely would be considered a large sample. Suppose the empirical BICs for these models, when fit to the sample, are $\text{BIC}_\text{A} = 276.73$, $\text{BIC}_\text{B} = 270.15$, and $\text{BIC}_\text{C} = 305.19$, implying that Model B is preferred to Model A, which in turn is preferred to Model C (compactly, B≫A≫C, where "≫" means "is ranked in preference to"). The difference between Models A and B exceeds Raftery's (1995) conservative criterion for "strong evidence" of a difference between models, and Model C differs "very strongly" from both.

What happens if another sample of the same size is drawn from the population, and the models are again fitted to the new data? Suppose the researcher now finds that $\text{BIC}_\text{A} = 281.88$, $\text{BIC}_\text{B} = 312.88$, and $\text{BIC}_\text{C} = 302.13$, this time implying A≫C≫B. Now Model B, instead of being the winner, is relegated to last place (by a large margin), and Model A is selected. This time, Models A and C differ very strongly, and both differ very strongly from Model B by Raftery's (1995) criteria. Clearly, these contradictory results pose a problem for the researcher. To complicate matters, in practice there is typically only one opportunity to fit competing models to data because only one sample is collected, and therefore there is no way of knowing whether a follow-up study would demonstrate the same or different rankings among models. In addition, many situations call for the comparison of more than three models, opening up the possibility of many erroneous, misleading, or inconsistent rankings.

The problem, of course, is that selection criteria and fit statistics are just that—statistics. As with any statistic, there is uncertainty in the estimation of selection criteria. It is to be expected that a model fitted to small repeated samples drawn from the same population will show unstable relative model fit (in comparison to other candidate models), sometimes assuming values representing superior fit and other times values reflecting inferior fit. The same model fit to larger samples may show very stable fit from sample to sample. As we show, even though the rank order of models with respect to BIC may eventually stabilize with increasing N, the variability of the BIC index increases with N because N is a component of the BIC formula. Just as one would never base conclusions on an observed mean difference without considering sampling variability in that difference, it makes little sense to report selection indices without considering sampling variability (Browne & Cudeck, 1992; Steiger & Lind, 1980). Not only is it the case that selection indices like BIC can vary widely across repeated sampling—so can pairwise differences in BIC (ΔBIC) and, as a consequence, the model rankings used to make decisions in model selection. This phenomenon is demonstrated in the next section.

It may appear inconsistent to examine the behavior of BIC, whose derivation is fundamentally Bayesian, in frequentist models. However, BIC is commonly reported as a model selection statistic in SEM, and it is undeniable that BIC will fluctuate from sample to sample. It is therefore natural to expect variability in

[2] We note that model selection with BIC is known to change with sample size (e.g., Atkinson, 1978). Here we emphasize a different phenomenon—variability in the selected model at a given N.

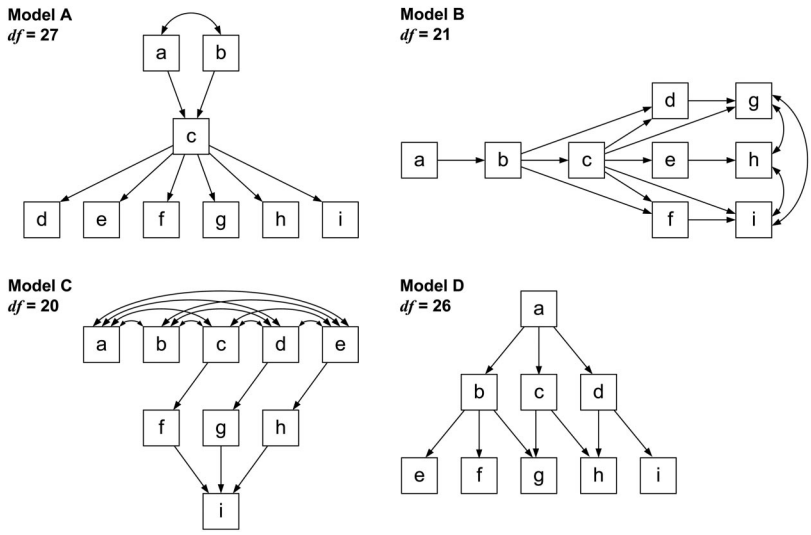

Figure 1. Models A, B, and C are compared in the illustration and simulation. Model D is the data-generating model. Variance and residual variance parameters are excluded for simplicity.

model selection decisions based upon BIC, regardless of the specific model estimation procedures.

To better appreciate the extent to which model selection results can vary over repeated sampling, we conducted a simulation that permitted us to examine variability in BIC for individual models, in ΔBIC for pairs of models, and in the model rankings themselves. The simulation procedure and results are reported in the next section. Following the simulation, we explore and evaluate three methods of quantifying model selection uncertainty in SEM, applying each in turn to BIC. We then discuss some practical implications of our findings.

Simulation

We used Model D to generate sample data (using Mplus 5.2), manipulating sample size (N = 80; 120; 200; 500; 1,500; 5,000). For simplicity, we restricted attention to normally distributed data. One thousand samples were simulated from Model D at each N. We used LISREL Version 8.8 to fit Models A, B, and C to each simulated data set and recorded the minimum fit function $\hat{G}$ to serve as a basis for computing BIC.

Table 2
EBIC Values for the Models Fitted to the Population Covariance Matrix Generated by Model D

	Model		
N	A	B	C
80	**125.77**	142.96	148.51
120	**143.14**	161.20	168.24
200	**172.47**	190.46	200.21
500	**264.50**	276.23	295.09
1,500	536.05	**515.19**	562.49
5,000	1,438.92	**1,288.15**	1,432.37

Note. The lowest EBIC at each N is in boldface. EBIC = expected Bayesian information criterion.

Figure 2 shows the sampling distributions of BIC_A for the N = 80 and N = 5,000 conditions. These distributions convey the sampling variability in BIC at two extremes of sample size. The sampling variance of BIC is the same as that of the noncentral χ^2 distribution, given as $\text{var(BIC)} = 2(\text{df}+4\lambda)$, where $\lambda = nF_0$ (Mulaik, 2009). Thus, models with worse fit (more noncentrality) demonstrate more variability in BIC. In addition, unlike most statistics, BIC becomes more, not less, variable with sample size because the $\hat{G}$ component generally increases with N (unless the model is perfectly specified or saturated). Even at N = 80, the interquartile range of BIC (135.08 – 119.02 = 16.06) is considered very large by Raftery's (1995) criteria. Increasing N makes the sampling distribution more variable. Figure 3 shows sampling distributions for ΔBIC for Models B and C at N = 80 and N = 5,000, displaying a similar high degree of spread in the statistic used to make comparisons between competing models.

We also tallied the proportion out of 1,000 trials in which each possible permutation of model preference (A≫B≫C and so on) was selected. The results are reported in Figure 4. From Figure 4 it can be seen that, over repeated sampling, A≫B≫C emerges as the top ranking approximately two thirds of the time for N = 80 to 200, with A≫C≫B in second place, implying that Model A is selected around 95% of the time for those sample sizes. At N = 500, there is a 72% chance that Model A will be selected, a 24% chance that Model B will be selected, and even a small probability that Model C will be selected. As N increases to 1,500 and then 5,000 (not depicted), B≫A≫C and then B≫C≫A emerge as the top rankings. Within most examined sample sizes, there is considerable uncertainty in rankings; it is the degree of this uncertainty that we find troubling. Model selection uncertainty does not decrease with increasing sample size but rather fluctuates with N as the balance of fit and parsimony changes. This phenomenon has

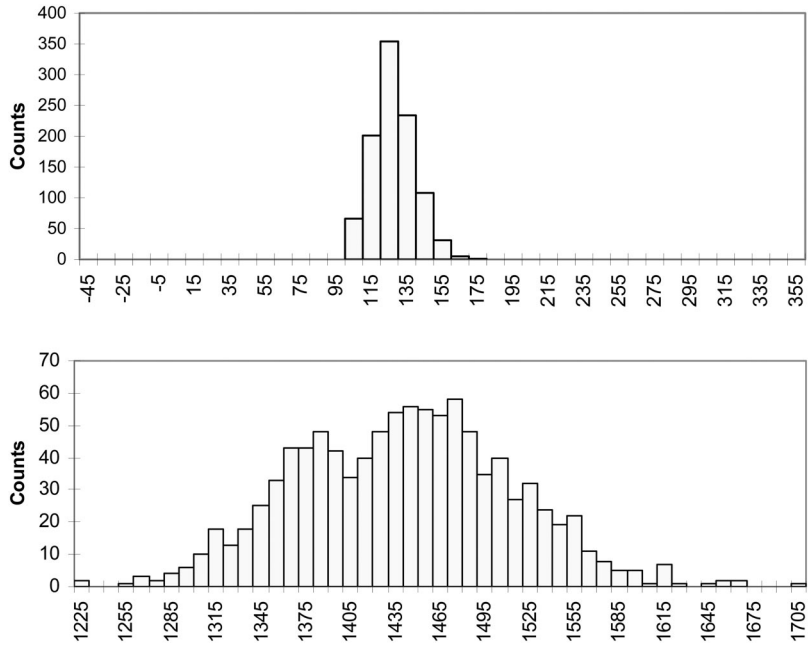

Figure 2. Sampling distributions of Bayesian information criterion for Model A at $N = 80$ (top) and $N = 5{,}000$ (bottom).

been noted before with respect to the use of AIC for model selection in SEM (McDonald, 1989). We are not concerned with the fact that BIC selects different models at different sample sizes but rather with the fact that BIC does not consistently select a single model at a given sample size.

Not only is sampling variability in model rankings present, but its magnitude is nontrivial. Table 3 contains proportions of samples in which the two most highly ranked models in a given ranking are separated by more than 6 ΔBIC units, Raftery's (1995) conservative criterion for strong evidence in favor of the top-ranked model. At least for the models examined in the simulation, model selection uncertainty would seem to present a serious problem, the extent of which would go unrecognized in a single-sample empirical study.

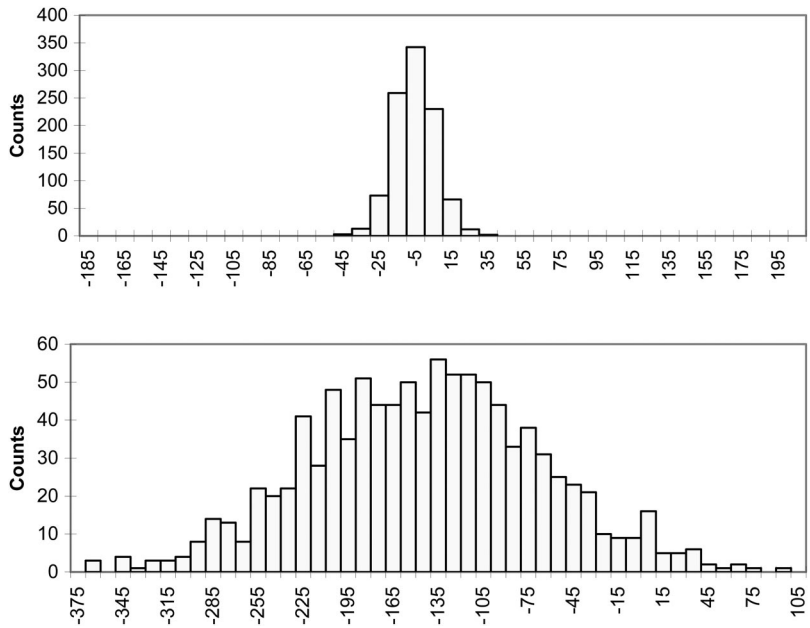

Figure 3. Sampling distributions of the difference in Bayesian information criterion for Models B and C at $N = 80$ (top) and $N = 5{,}000$ (bottom).

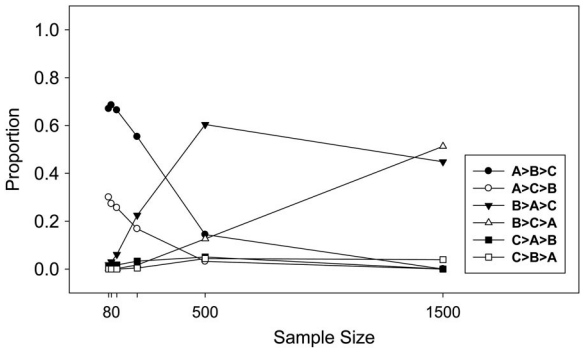

Figure 4. Proportions of simulated samples yielding each model ranking.

Methods for Addressing Model Selection Uncertainty in SEM

In the previous section we demonstrated that, under quite common circumstances, there can be a large degree of uncertainty in the model selection process. Despite this uncertainty, in applications of SEM and in other domains, it is common practice to simply report the selection criteria corresponding to the competing models and use them to select a single "winner" model. Little thought is given to reporting facets of the sampling distribution of BIC other than the estimated mean, perhaps because BIC and other criteria are not typically thought of as parameters to be estimated. Obviously the researcher would like a strategy for communicating the degree of selection uncertainty to readers, which often may imply that no candidate model is clearly better than the other(s). In this section we explore and evaluate three methods of quantifying this uncertainty, applying each in turn to BIC.

Quantifying Variability in Selection Criteria

The first strategy we explored involved quantifying the variability in the BIC index itself. The ideas in this section build on two traditions. First, Steiger and Lind (1980) proposed and discussed methods for constructing confidence intervals for noncentrality-based fit indices and selection criteria. An iterative procedure is used to locate the values of the noncentrality parameter that give rise to noncentral χ^2 distributions whose $(50 \pm \frac{1}{2}\omega)$th percentile values equal $\hat{G}$. These values form the ω% confidence interval (CI) for the population G.[3] Once obtained, the limits may be transformed into limits for any noncentrality-based criterion, such as RMSEA or BIC. For example, if the lower and upper limits G_L and G_U are added to $q\ln N$, the result is a ω% CI for the expected BIC, as seen in the following:

$$\mathrm{CI}_{1-\alpha} = \{G_L + q\ln N \leq \mathrm{EBIC} \leq G_U + q\ln N\}. \quad (8)$$

Basically, the distribution of BIC is a noncentral χ^2 distribution shifted to the right (or penalized) by $q\ln N$. The more parameters a model has, the further to the right this distribution is shifted. This is a simple idea, yet an analytic confidence interval for EBIC seems not to have been reported before now. Even in SEM software that outputs BIC, no CIs are reported, despite their ease of computation. One potential reason for this is that, as mentioned previously, the "population" BIC is not defined. The lack of a population BIC poses a problem only if the researcher wishes to test a null hypothesis about it, a practice we see little point in pursuing: If the goal is to quantify the sampling variability of BIC given N, a CI can still be quite useful. In this case, the EBIC would seem to be a more reasonable quantity to want to estimate. Nevertheless, we could find no examples of a CI being reported for EBIC in the applied literature, nor could we find an explicit expression for the CI of EBIC.

Second, nonparametric bootstrapping has been proposed as a method for constructing CIs for selection criteria like AIC and BIC, as well as other fit indices (Bollen & Stine, 1992; Burnham & Anderson, 2002; Raykov, 2001). Nonparametric bootstrapping requires access to raw data and mimics the sampling process that is ordinarily assumed in parametric statistical procedures. In bootstrapping, B "resamples" of size N are drawn with replacement from the original sample of size N, and the set of competing models are fitted to all resampled data sets. The statistic of interest (generically θ^*) is recorded each time. These statistics form the empirical sampling distribution of θ. The $(50 \pm \frac{1}{2}\omega)$th percentile values of the bootstrap distribution of θ^* define the limits of an approximate ω% CI, with the approximation improving as B increases. The only assumptions necessary for bootstrapping are that the rows of the data matrix be independent and that the data be "representative" of the population.

Whereas the traditional, nonparametric bootstrap can yield CIs that provide good coverage of EBIC, it does not appropriately characterize the variability in BIC. This is because sample data tend to be more discrepant from the model, compared with population data (e.g., Bollen & Stine, 1992; Steiger, 2000; Yuan & Hayashi, 2003). Borrowing from Cudeck and Henly's (1991) terminology, the bootstrap should estimate the *discrepancy of estimation* (i.e., that between the model and the population). Use of sample data, however, provides an estimate of *overall discrepancy* that is larger than the discrepancy of estimation. As a result, both the expectation and variance of bootstrap $\hat{G}$ distributions are incorrect (the expectation can be corrected by simply adding a constant to all values of $\hat{G}$, but the variance cannot be corrected in this way). To correct for this, Yuan, Hayashi, and Yanagihara (2007) provided a method (hereafter termed the *YHY bootstrap*) for transforming the raw data so that the bootstrap discrepancies are closer to the discrepancy of estimation. This transformation involves first finding a covariance matrix that is closer to the population covariance matrix than the sample covariance matrix and then transforming the raw data so that they correspond to the new covariance matrix. The resulting bootstrapped $\hat{G}$ distribution then provides a better approximation to the discrepancy of estimation.

We hypothesized that analytic CIs for BIC (see Equation 8) would perform well at recovering the sampling properties of BIC, as reflected in CI coverage. Based on the success of bootstrapping in numerous other contexts, we also hypothesized that both types of bootstrapping (traditional nonparametric and the YHY method)

[3] This interval assumes multivariate normality and is only first-order accurate. Yuan, Hayashi, and Bentler (2007) further showed that, in applied situations, a specific normal distribution may better describe the behavior of $\hat{G}$ than does the noncentral χ^2 distribution. This distribution may also be used to construct CIs for EBIC.

Table 3
Proportions of Simulated Samples in Which the First and Second Selected Models Differ Strongly According to Raftery's (1995) Criteria

N	A≫B≫C	A≫C≫B	B≫A≫C	B≫C≫A	C≫A≫B	C≫B≫A
80	**.619**	**.272**	.006	.001	.001	.000
120	**.632**	**.247**	.005	.000	.004	.000
200	**.594**	**.227**	.022	.001	.003	.000
500	**.444**	**.135**	**.129**	.008	.019	.003
1,500	**.100**	.020	**.547**	**.115**	.038	.032
5,000	.000	.000	**.448**	**.506**	.000	.030

Note. Proportions exceeding .05 are in boldface.

would perform well at recovering the sampling variability in BIC. Figure 5 depicts example bootstrap distributions of BIC for Models A, B, and C ($N = 200$) along with 95% CIs. Coverage for the three types of CIs is presented in Table 4. The intervals generally show excellent coverage, though the nonparametric bootstrap has conservative coverage probabilities at lower sample sizes. This likely is due to the fact that, as mentioned earlier, the nonparametric bootstrap yields a $\hat{G}$ distribution exhibiting more variability than it should.

As straightforward and intuitive as these three general methods are, they fail to fully address the issue at hand. That is, characterizing the BIC distributions for rival models says little about how the models should be evaluated relative to one another. Beyond the fact that these methods do not directly address differences among models, it is difficult even to glean such information from examining the distributions. BIC statistics (indeed, any index or criterion based on $\hat{G}$) can be highly correlated for adjacent nested models (Steiger & Lind, 1980; Steiger et al., 1985). They can also be highly correlated for nonnested models applied to the same data. For example, considering our $N = 1{,}500$ condition, the correlations among BICs for Models A, B, and C were $r_{A,B} = .729$, $r_{A,C} = .621$, and $r_{B,C} = .337$. These correlations complicate direct comparison of BIC distributions for rival models because the observations are not independent (in fact they often are highly dependent)—the relative ordering (or overlap) of the models' BIC distributions may give a misleading impression of how the models will be ranked relative to one another over repeated sampling, depending on the degree of correlation. That is, overlapping BIC distributions could imply (a) that the models each yield the lower BIC for some of the samples and thus are indistinguishable via BIC or (b) that one model consistently outperforms the other model by a small amount. Whereas the first conclusion may be more intuitive than the second, one cannot distinguish between the two when the distributions are correlated to an unknown extent. Hence, as useful and informative as CIs for BIC may be, one must look elsewhere to more fully quantify model selection uncertainty.

Quantifying Variability in Pairwise Differences in Selection Criteria

If the goal of model selection is ultimately to select one model from a pool of competing rival models, one must operate at the level of comparisons among models. In this case, it seems sensible to apply the same analytic and nonparametric methods developed in the previous section to ΔEBIC_{ij} for all pairs of rival models i and j rather than to BIC for each individual model. As with $\hat{G}$, $\Delta\hat{G}$ for adjacent nested models (i nested in j) follows a noncentral χ^2 distribution with $\Delta df = (df_i - df_j)$ and noncentrality parameter $\Delta\lambda = (\lambda_i - \lambda_j)$. In the special case of nested models, sequential $\Delta\hat{G}$ statistics are asymptotically independent (Steiger et al., 1985). When the noncentral χ^2 distribution with Δdf and $\Delta\lambda$ is invoked, one obtains

$$\text{CI}_{1-\alpha} = \{\Delta G_L + \Delta q \ln N \leq \Delta\text{EBIC} \leq \Delta G_U + \Delta q \ln N\}. \quad (9)$$

In the more general and common case in which the rival models are not nested, the situation is not so convenient and the problem requires bootstrapping to become tractable. Figure 6 displays bootstrap distributions (both traditional and YHY) of ΔEBIC for Models A, B, and C ($N = 200$), along with 95% CIs. An issue arises when applying the YHY method to bootstrap differences between BICs. The method yields a unique data transformation for each candidate model, so that one no longer uses the same data to estimate each model. From the distributions in Figure 6, it can be seen that the YHY distributions exhibit more variability than do those from the traditional bootstrap.

We obtained ordinary percentile 95% CIs for ΔEBIC. Using 1,000 intervals constructed from 1,000 bootstrap resamples each (for both the traditional bootstrap and the YHY bootstrap), we found the CI coverages reported in Table 5. Coverage for the traditional bootstrap CIs was quite good, erring on the side of conservatism due to the extra variability in the bootstrap distributions that was discussed previously. CI coverage for the YHY bootstrap CIs was very conservative, reflecting the fact that the unique data transformations for each model introduced extra variability (more than that introduced by use of the nonparametric bootstrap) into the CIs.

CIs for ΔEBIC yield insight into the degree of uncertainty with which model i is ranked with respect to model j, for all i and j. However, as with CIs for individual model EBICs, CIs for ΔEBIC still do not directly address the main problem: quantifying uncertainty in the choices made on the basis of sample BICs and ΔBICs (unless there are only two models to compare). Moreover, even though sequential $\Delta\hat{G}$s for adjacent nested models are independent (Steiger et al., 1985), when the models are nonnested the ΔBICs can be quite highly correlated. Considering our $N = 1{,}500$ condition, the correlations are $r_{AB,AC} = .056$, $r_{AB,BC} = -.618$, and $r_{AC,BC} = .751$. This degree of correlation would preclude making direct comparisons of ΔBIC for two sets of

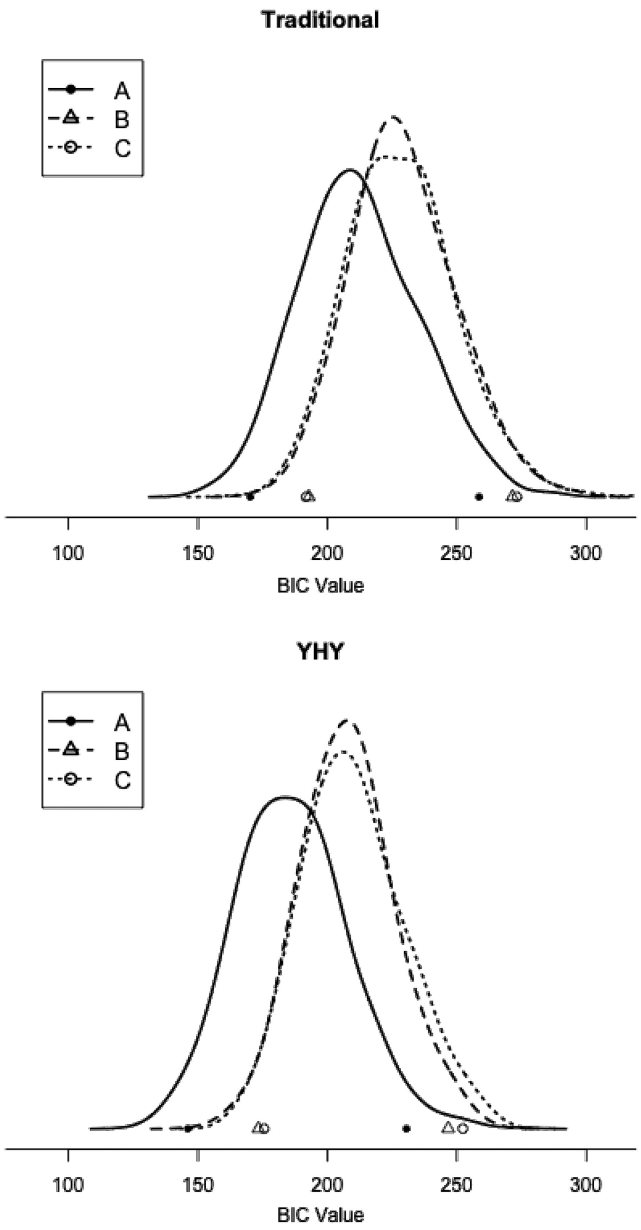

Figure 5. Bootstrap sampling distributions of Bayesian information criterion (BIC; kernel smoothed) for Models A, B, and C. The symbols •, △, and ○ denote 95% confidence limits for $EBIC_A$, $EBIC_B$, and $EBIC_C$, respectively. Traditional = nonparametric bootstrap; YHY = Yuan, Hayashi, and Yanagihara (2007) bootstrap; EBIC = expected BIC.

models (e.g., comparing ΔBIC_{AB} with ΔBIC_{BC} would be questionable if the models are not nested). Again, as useful and informative as these methods are, one must search for other means to quantify model selection uncertainty.

Bootstrap Distributions of Model Rankings

Whereas bootstrapping is useful for quantifying variability in BIC and ΔBIC, it does not quite address the problem we emphasize: variability in model rankings made on the basis of BIC. Model selection is often carried out by choosing the model with the lowest criterion (and ignoring magnitudes of the criteria), making criterion rankings most relevant. In this section we evaluate the nonparametric bootstrap as a method of approximating—from a single sample—the expected sampling variability in model rankings themselves. We did not consider the YHY procedure due to the increased variability in BIC differences noted in the previous section.

To study sampling variability in model rankings, we created a large number of bootstrap samples, fitted each rival model to each resample, and recorded the proportions of trials favoring each possible model ranking. Of interest is the degree to which these proportions resemble the "gold standard" proportions in Figure 4, the pattern of proportions expected over many draws from the same population. Because only a single sample is available in most research applications, the ability to reproduce this variability from a single sample would be highly convenient.

Given the demonstrated success of bootstrapping in other settings, we had three hypotheses regarding the performance of bootstrapping in recovering the variability in model rankings. First, we expected bootstrapping to provide a good match to the pattern of proportions observed for the six possible rankings of Models A, B, and C across repeated sampling from the same population (see Table 3 and Figure 4). A fundamental tenet of bootstrap methods is that a random sample may be treated as if it were the population, in the sense that repeated sampling with replacement mimics the sampling process, giving rise to sampling distributions of almost any desired statistic. Second, we expected the variability in rankings to be recovered more faithfully for larger sample sizes, as larger random samples naturally resemble the population more closely. Third, we expected bootstrapping to perform better in recovering variability in rankings for models with larger EBIC differences (by Raftery's, 1995, criteria). To test the third hypothesis, we specified a different set of rival models with larger EBIC differences.

To test our first and second hypotheses, we repeated our earlier simulation, except that data sets were generated not from a population model but rather by bootstrapping from single data sets (one for each *N*) generated from Model D. This we regard as a "moderate" ΔBIC condition. Results are reported in Table 6 and Figure 7. Our expectation was that the proportions depicted in Figure 7 and reported in Table 6 would strongly resemble those depicted in Figure 4 and reported in Table 3. We also expected that the resemblance between the two sets of proportions would increase with increasing *N* because larger samples generally resemble the population more so than do smaller samples. Figure 7 and Table 6 resemble Figure 4 and Table 3 only superficially—not as closely as was expected for 1,000 resamples. Not only is the variability in model rankings in Figure 7 and Table 6 different from that in Figure 4 and Table 3, but the bootstrap analysis converges to a different model ranking ($B \gg A \gg C$ vs. $B \gg C \gg A$) at N = 5,000. To rule out the possibility that we selected particularly unrepresentative samples from which to draw bootstrap resamples at each *N*, we repeated the simulation using additional samples generated by Model D, with similarly variable results. Thus, our first two hypotheses were not supported.

Our third hypothesis was that the resemblance between bootstrap ranking proportions and genuine sampling-based ranking proportions would improve as the compared models become more

Table 4
Coverage for 95% Analytic, Traditional Bootstrap, and YHY Bootstrap Confidence Intervals (CIs) for BIC

	Analytic			Traditional bootstrap			YHY bootstrap		
N	BIC_A	BIC_B	BIC_C	BIC_A	BIC_B	BIC_C	BIC_A	BIC_B	BIC_C
80	.951	.955	.950	.983	.982	.979	.975	.980	.966
120	.948	.949	.949	.980	.982	.978	.952	.954	.954
200	.953	.944	.942	.977	.971	.973	.940	.939	.940
500	.953	.960	.947	.957	.960	.974	.954	.953	.955
1,500	.946	.961	.943	.960	.966	.945	.955	.950	.950
5,000	.952	.953	.953	.943	.943	.946	.943	.949	.943

Note. Analytic CI coverage is based on 1,000 samples for each *N*. Bootstrap CI coverage is based on 1,000 bootstrap resamples from each of 1,000 random samples for each *N*. Subscripts indicate models. YHY = Yuan, Hayashi, and Yanagihara (2007); BIC = Bayesian information criterion.

distinct. We chose a second set of smaller models that differed more substantially in terms of BIC. This we regard as an "extreme" ΔBIC condition. Specifically, we used a hypothesized model and three plausible alternative models from a study of protective factors against substance use among Asian youths (Kim, Zane, & Hong, 2002). The models, all of which have $df = 3$, are depicted in Figure 8. All four models represent hypotheses about the causal relationships among youth–parent relations (Family), vulnerability to negative peer pressure (Peer), school-based social comfort (School), and substance use (Substance Use). The hypothesized model (Model H) was found to fit very well, whereas the other three (Models E, F, and G) fitted very poorly. We used Kim et al.'s (2002) well-fitting hypothesized model as the data-generating model, with parameters set equal to estimates provided by the authors. Using Model H to generate the population covariance matrix led to the EBIC values in Table 7. Notice that the BIC differences for these models are substantially larger (again, depending on *N*) than those for Models A, B, and C.

Results are given in Figures 9 and 10, where the results in Figure 9 were obtained by generating model rankings from 1,000 samples and the results in Figure 10 were obtained by generating model rankings from 1,000 bootstrap resamples taken from one sample. Over repeated sampling from the population, Model F differed strongly from Models E and G for $N \geq 80$. The model selection process resolved quickly to the F≫E≫G ranking in most trials by $N = 80$, with some degree of support for the E≫F≫G ordering for $N \leq 200$. Using bootstrap data to recover this variability in ranking shows a close match to the rankings found from repeatedly sampling from the population, with F≫E≫G quickly emerging as the preferred ranking. We repeated this process with two additional samples and obtained a similarly good match, save for some variability in rankings for $N < 200$. We conclude that the performance of bootstrapping at recovering variability in rankings is more impressive in this example than in the nine-variable example but that (a) the performance of bootstrapping at recovering variability in model rankings in small samples is still questionable and (b) the good performance in larger samples is illusory, as there is little variability in rankings to recover.

Summary

To summarize, we found that analytic and bootstrap CIs perform well in quantifying variability in BIC and that traditional, nonparametric bootstrap CIs perform well for ΔBIC. Second, we hypothesized that bootstrapping from a single sample would yield ranking proportions similar to those found from repeated samples from the same population. This hypothesis was not supported well in the nine-variable example. Third, we expected to better recover variability in rankings at higher *N*s. This hypothesis was supported somewhat but still not very well even at high *N*s. Fourth, we expected better recovery for model sets that differed more strongly in terms of BIC. This hypothesis was supported. Bootstrapping can quantify BIC uncertainty in SEM when the models are clearly distinguishable in terms of BIC and when the sample size is sufficiently large.

Why does the traditional, nonparametric bootstrap perform poorly for model ranks? As Bollen and Stine (1992; p. 207) stated, "the success of the bootstrap depends on the sampling behavior of a statistic being the same when the samples are drawn from the empirical distribution and when they are taken from the original population." This condition is met when bootstrapping BIC and ΔBIC. In the case of BIC rankings, however, bootstrap resampling does not yield the appropriate proportions. We attribute this phenomenon to two issues: *initial samples* and *small differences*.

Regarding initial samples, consider the sampling distributions generated by bootstrapping from repeated samples from the same population. Whereas the 95% CIs obtained from these bootstrap distributions may capture the ΔEBIC approximately 95% of the time, the initial sample problem is that ranking proportions derived from bootstrap resamples are highly dependent on the particular sample initially drawn from the population. For example, the first sample generated from Model D at $N = 120$ yielded a $\Delta EBIC_{BC}$ 95% CI of {−41.94, 3.62}, implying that Model C is more often chosen as the better model. If many bootstrap samples are generated from that sample, Model C is bound to be strongly favored in the majority of them (C≫B). If we had instead used the second simulated sample to serve as a basis for bootstrapping—for which the 95% CI was {–17.39, 43.30}—the outcome likely would have been very different (B≫C).

When two models have very similar BICs (i.e., there is a small difference) for a given sample, the initial sample problem is magnified: Ranking proportions based on bootstrap resamples will be more sensitive to the initial sample than when the two models have very different BICs. Consider the case of Figure 4. The fact that there is a great deal of uncertainty in how Models A and C ought to be ranked at $N = 1{,}500$ indicates that these two models

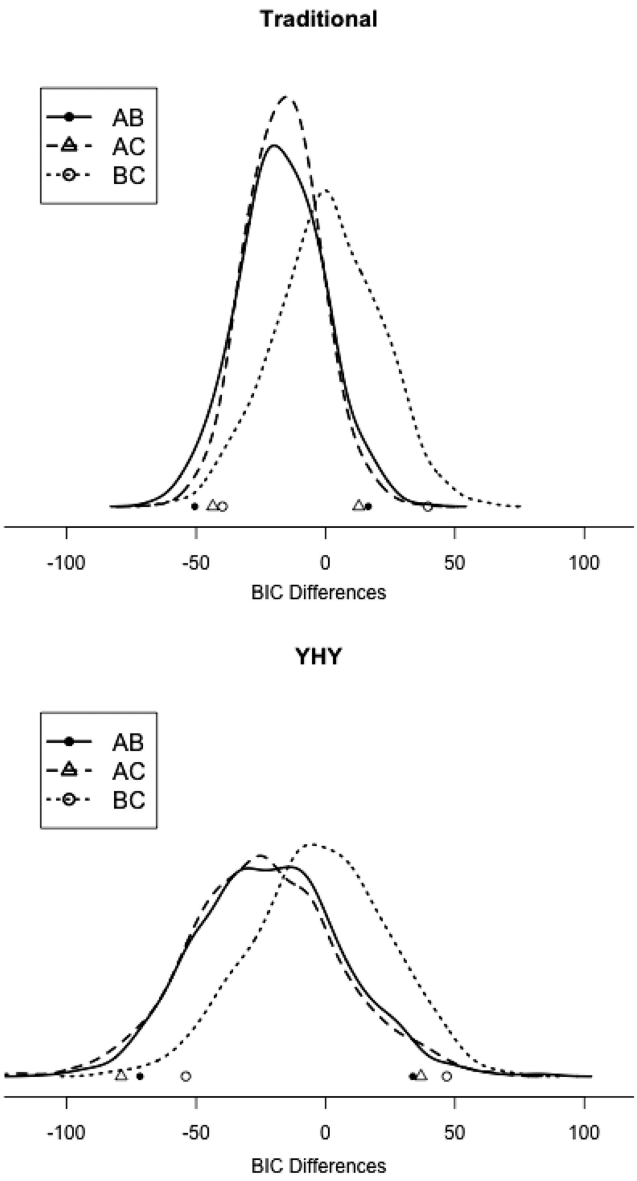

Figure 6. Bootstrap sampling distributions of the difference in Bayesian information criterion (BIC; kernel smoothed) for Models A, B, and C. The symbols •, △, and ○ denote 95% confidence limits for $\Delta EBIC_{AB}$, $\Delta EBIC_{AC}$, and $\Delta EBIC_{BC}$, respectively. Traditional = nonparametric bootstrap; YHY = Yuan, Hayashi, and Yanagihara (2007) bootstrap; ΔEBIC = difference in expected BIC.

have very similar BICs at that sample size and hence that the sampling distribution of ΔBIC_{AC} is centered near zero. We would like to be able to recover these proportions (.448, .513), but this is precisely the sort of situation that leads to the least confidence in bootstrap proportions. A little less than half of the samples would give rise to bootstrap distributions favoring B≫A≫C, and about half would favor B≫C≫A. Few would yield proportions similar to those in Figure 4. The only situation in which one could expect consistently good recovery is the degenerate case in which nearly 100% of the rankings are the same. In other words, bootstrap ranking proportions perform best where they are least needed. CIs for EBIC and ΔEBIC may be trustworthy, but it is highly problematic to read too much into the proportion of bootstrap samples for which one model fits better than another.

A third reason bootstrapping may fail to recover ranking proportions, especially at smaller *N*s, may derive from the distributional characteristics of $\hat{G}$ in small samples. Specifically, it has been shown that G tends to be overestimated in small samples (Curran, Bollen, Paxton, Kirby, & Chen, 2002). Thus, it may be profitable to examine small-sample corrections for $N < 200$ or so (see, e.g., Herzog & Boomsma, 2009).

Discussion

We have illustrated that model selection uncertainty (a) is rarely quantified or explored in any depth yet (b) can be substantial under common circumstances, leading to overconfidence in the selected model. We demonstrated by example that BIC, ΔBIC, and rankings based on BIC can be highly variable, even at large sample sizes. Dichotomous decisions based on continuous data inevitably introduce an additional degree of uncertainty. But if that uncertainty is not quantified and communicated, a potentially major source of error goes unnoticed. We also demonstrated that some seemingly reasonable approaches for quantifying this inherent uncertainty in model selection either fail to fully address the problem or actually misrepresent that variability. We close the article with a general discussion of the bootstrap's failure, other relevant statistical methods, and study limitations.

The failure of the bootstrap in model selection is not without precedent. For example, Freedman, Navidi, and Peters (1988) found that bootstrapping performs poorly at variable selection in exploratory regression analysis. They concluded that it is desirable to collect additional data and to test the model's replicability in a new sample. We come to the same conclusion; there really is no substitute for replication. Splitting the sample and attempting cross-validation may be inadvisable in this case because, aside from the uncertainty introduced by shrinking the sample, (a) this procedure requires a large initial sample, which may be impractical to obtain, and (b) changing *N* may easily result in a different rank ordering of the models simply because changing *N* reweights

Table 5
Coverage for 95% Percentile Bootstrap Confidence Intervals (CIs) for ΔBIC

	Traditional bootstrap			YHY bootstrap		
N	ΔBIC_{AB}	ΔBIC_{AC}	ΔBIC_{BC}	ΔBIC_{AB}	ΔBIC_{AC}	ΔBIC_{BC}
80	.985	.980	.970	1.000	1.000	.998
120	.983	.981	.977	.999	.999	.995
200	.974	.962	.955	.999	.998	.980
500	.955	.962	.956	1.000	.999	.993
1,500	.948	.943	.937	.999	.999	.987
5,000	.950	.954	.950	.999	1.000	.986

Note. Bootstrap CI coverage is based on 1,000 bootstrap resamples from each of 1,000 random samples for each *N*. ΔBIC = difference in Bayesian information criterion between two models; YHY = Yuan, Hayashi, and Yanagihara (2007).

Table 6
Proportions of Bootstrap Resamples in Which the First and Second Selected Models Differ Strongly According to Raftery's (1995) Criteria

N	A≫B≫C	A≫C≫B	B≫A≫C	B≫C≫A	C≫A≫B	C≫B≫A
80	**.563**	**.256**	.012	.000	.007	.002
120	**.454**	**.424**	.005	.000	.017	.002
200	**.398**	**.281**	.023	.003	**.062**	.004
500	**.224**	.021	**.469**	.009	.006	.001
1,500	**.299**	.015	**.483**	.002	.003	.002
5,000	.022	.002	**.898**	**.057**	.001	.005

Note. Proportions exceeding .05 are in boldface.

BIC's penalty for complexity. Future research could also investigate Bayesian approaches to model selection in SEM. Bayesian methods allow one to easily compute the sampling variability of functions of model parameters (such as model selection criteria). Further, Lee (2007) discussed methods for comparing models via Bayes factors, the computation of which traditionally has been intractable in the context of SEM. These advances may well lead to better model selection methods, although one must also worry about the influence of prior distributions on the model selection criteria.

Our conclusions are naturally circumscribed by our particular choices of models, selection criteria, sample sizes, and so on. Our models were chosen deliberately to bear very different substantive implications. In practice, however, the models to be compared will often differ only in minor respects. We maintain that model sets exhibiting greater similarity will also demonstrate greater selection uncertainty; contrasting our first illustration (Models A–D) and second illustration (Models E–H) provides one small example of this phenomenon, although it must be borne in mind that both model sets contain model differences that qualify as substantial by Raftery's (1995) criteria. In other words, we chose what amounts to a best-case scenario, and the situation is likely worse in practice.

We limited attention to BIC as the selection criterion investigated. In emphasizing BIC, we did not mean to imply that BIC is the only or best criterion for model selection in SEM. BIC has well-recognized limitations, including reliance on specific priors on model parameters and its use of sample size as the sole indicator of the informativeness of the data (Weakliem, 1999). Future work should examine selection uncertainty using other criteria.

We believe that the general principles outlined in this article apply to varying degrees to other selection criteria and fit indices as well. A number of selection criteria and fit indices are conceptually quite similar to BIC and would be expected to behave in a similar fashion. For example, AIC and its variants, as well as a number of variants of BIC, have functional forms that are similar to BIC's; that is, they contain a component reflecting absolute model fit and a component penalizing complexity. Yanagihara and Ohmoto (2005) provided distributional theory and approximate CIs for AIC for regression; their work could be extended for use in SEM.

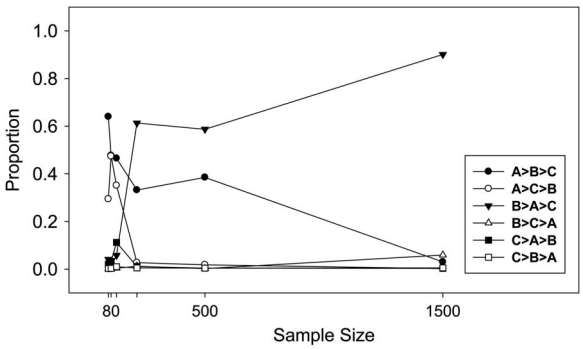

Figure 7. Proportions of nonparametric bootstrap resamples in the "moderate ΔBIC" example falling into the six possible BIC rankings. ΔBIC = difference in Bayesian information criterion.

Conclusion

Model selection criteria are being used with increasing frequency in psychology and other social sciences, particularly in new modeling areas where conventional fit indices are not yet available in software. Examples include latent class analysis (Ny-

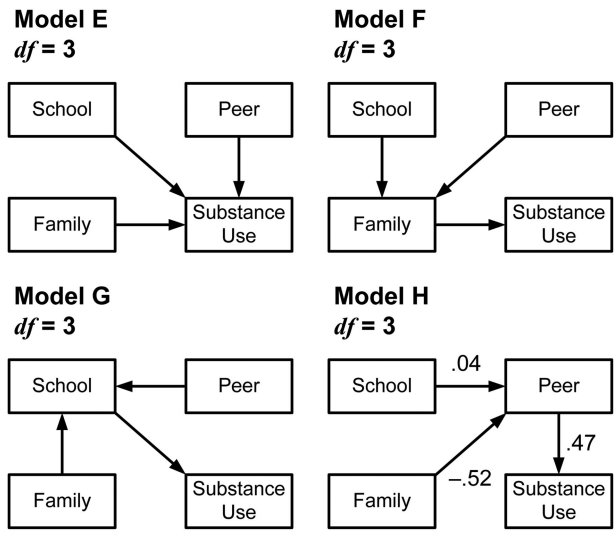

Figure 8. Models E, F, and G are compared in the illustration. Model H is the data-generating model. Variance and residual variance parameters are excluded for simplicity. Exogenous covariances were constrained to zero in Kim et al. (2002).

Table 7
BIC Values for the Models Fitted to the Population Covariance Matrix Generated by Model H

	Model		
N	E	F	G
80	58.75	**48.65**	78.27
120	74.29	**59.08**	103.69
200	103.26	**77.82**	152.43
500	204.91	**141.13**	328.20
1,500	530.06	**338.44**	900.41
5,000	1,649.57	**1,010.57**	2,884.66

Note. The lowest BIC at each *N* is in boldface. BIC = Bayesian information criterion.

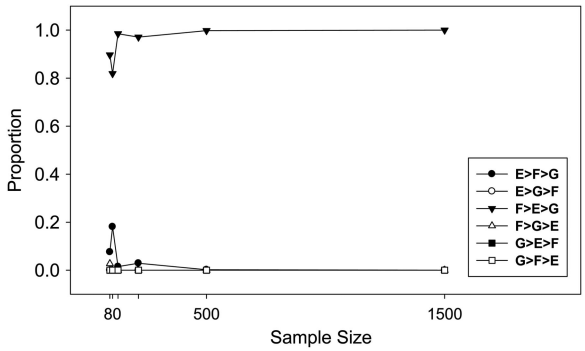

Figure 10. Proportions of bootstrap resamples in the "extreme ΔBIC" example falling into the six possible BIC rankings. ΔBIC = difference in Bayesian information criterion.

lund, Asparouhov, & Muthén, 2007), growth mixture modeling (Nagin, 1999; Tofighi & Enders, 2007), and multilevel structural equation modeling (Lee & Song, 2001), although there have been encouraging recent developments (Yuan & Bentler, 2007). Our purpose here is not to offer a general solution to the problem of quantifying the degree of selection uncertainty in SEM; rather, our points have been twofold. First, we wanted to demonstrate that selection uncertainty is a pervasive and serious problem in applications of model selection in SEM. One practical consequence of this observation is that many researchers may closely follow the recommendations of SEM experts yet still select suboptimal models due to (a) the large degree of variability in selection indices and (b) the strong dependency of model rankings on sample size. Collecting large samples and specifying models with solid theoretical support may represent good practice for a number of reasons but may offer little protection against the kind of variability in model selection decisions discussed and illustrated here. This largely unnoticed source of variability would seem to pose serious questions for model comparison and provides yet another argument in favor of replicating prior research whenever possible.

Second, given the large degree of selection uncertainty in SEM, it would seem to be a high priority to find ways to routinely quantify and consider this uncertainty, given the lack of available remedies. We explored some reasonable ways in which this uncertainty might be quantified. Each method addresses some aspect of the variability inherent in using BIC to make model selection decisions, and we recommend that they be used in combination with sample BICs in specific situations. Analytic confidence limits or, alternatively, confidence limits obtained from the YHY bootstrap procedure, should be reported with sample BICs. The latter would be preferred when the assumption of normality is tenuous. Confidence limits from the nonparametric bootstrap, on the other hand, can be used to quantify variability in ΔBIC. Caution should be used when making model selection decisions on the basis of BIC and associated CIs, however. As Schenker and Gentleman (2001) generally demonstrated, making decisions based on CI overlap is overly conservative. As applied to CIs for BIC, this implies that (a) if bootstrap or analytic CIs for BICs indicate overlap between models, little can be concluded about the superiority of some models over others, and (b) if the CIs do not overlap, one can have more certainty in selecting one model as best. Just as for general statistical decision making, however, researchers should recognize that overlap/nonoverlap in CIs for BIC or ΔBIC does not guarantee that the model selection decision would be stable over repeated samples.

Whereas the bootstrap CIs provide useful information, all the illustrated methods ultimately fall short of accurately and adequately summarizing the variability in model selection one would expect to see over repeated samples from the same population. Bootstrapping failed to adequately recover variability in model rankings at all sample sizes when the differences between models were moderate to large. Consequently, we cannot recommend bootstrapping as a general approach for quantifying selection uncertainty.

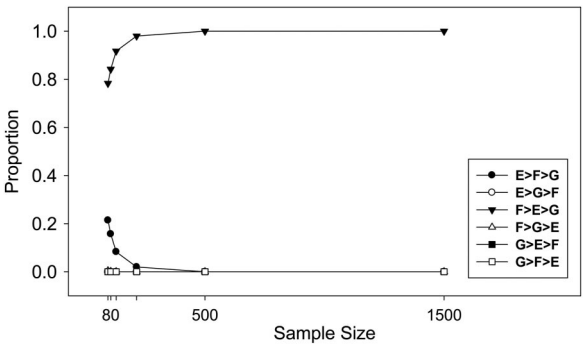

Figure 9. Proportions of simulated samples in the "extreme ΔBIC" example falling into the six possible BIC rankings. ΔBIC = difference in Bayesian information criterion.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akadémiai Kiadó.

Atkinson, A. C. (1978). Posterior probabilities for choosing a regression model. *Biometrika, 65,* 39–48. doi:10.1093/biomet/65.1.39

Bentler, P. M., & Mooijaart, A. (1989). Choice of structural model via parsimony: A rationale based on precision. *Psychological Bulletin, 106,* 315–317. doi:10.1037/0033-2909.106.2.315

Bollen, K. A., & Long, J. S. (Eds.). (1993). *Testing structural equation models.* Newbury Park, CA: Sage.

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research, 21,* 205–229. doi:10.1177/0049124192021002004

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83. doi:10.1111/j.2044-8317.1984.tb00789.x

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21,* 230–258. doi:10.1177/0049124192021002005

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

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33,* 261–304. doi:10.1177/0049124104268644

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147–167. doi:10.1207/s15327906mbr1802_2

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. doi:10.1207/S15327906MBR3701_01

Freedman, D. A., Navidi, W. C., & Peters, S. C. (1988). On the impact of variable selection in fitting regression equations. In T. K. Dijkstra (Ed.), *On model uncertainty and its statistical implications* (pp. 1–16). Berlin, Germany: Springer-Verlag. doi:10.1007/978-3-642-61564-1_1

Haughton, D. M. A., Oud, J. H. L., & Jansen, R. A. R. G. (1997). Information and other criteria in structural equation model selection. *Communications in Statistics: Simulation and Computation, 26,* 1477–1516. doi:10.1080/03610919708813451

Herzog, W., & Boomsma, A. (2009). Small-sample robust estimators of noncentrality-based and incremental model fit. *Structural Equation Modeling, 16,* 1–27. doi:10.1080/10705510802561279

Homburg, C. (1991). Cross-validation and information criteria in causal modeling. *Journal of Marketing Research, 28,* 137–144. doi:10.2307/3172803

Jeffreys, H. (1961). *Theory of probability.* Oxford, England: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. doi:10.2307/2291091

Kim, I. J., Zane, N. W. S., & Hong, S. (2002). Protective factors against substance use among Asian American youth: A test of the peer cluster theory. *Journal of Community Psychology, 30,* 565–584. doi:10.1002/jcop.10022

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research, 33,* 188–229. doi:10.1177/0049124103262065

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach.* Chichester, England: Wiley. doi:10.1002/9780470024737

Lee, S.-Y., & Song, X.-Y. (2001). Hypothesis testing and model comparison in two-level structural equation models. *Multivariate Behavioral Research, 36,* 639–655. doi:10.1207/S15327906MBR3604_07

Lee, S.-Y., & Song, X.-Y. (2008). Bayesian model comparison of structural equation models. In D. B. Dunson (Ed.), *Random effect and latent variable model selection* (pp. 121–150). New York, NY: Springer. doi:10.1007/978-0-387-76721-5_6

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

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38,* 113–139. doi:10.1207/S15327906MBR3801_5

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. doi:10.1037/1082-989X.1.2.130

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97–103. doi:10.1007/BF01908590

Mulaik, S. A. (2009). *Linear causal modeling with structural equations.* Boca Raton, FL: Chapman & Hall/CRC.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105,* 430–445. doi:10.1037/0033-2909.105.3.430

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods, 4,* 139–157. doi:10.1037/1082-989X.4.2.139

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14,* 535–569. doi:10.1080/10705510701575396

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41,* 227–259. doi:10.1207/s15327906mbr4103_1

Preacher, K. J., Cai, L., & MacCallum, R. C. (2007). Alternatives to traditional model comparison strategies for covariance structure models. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 33–62). Mahwah, NJ: Erlbaum.

Raftery, A. E. (1993). Bayesian model selection in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 163–180). Newbury Park, CA: Sage.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111–163. doi:10.2307/271063

Raftery, A. E. (1999). Bayes factors and BIC: Comment on "A critique of the Bayesian information criterion for model selection." *Sociological Methods & Research, 27,* 411–427. doi:10.1177/0049124199027003005

Raykov, T. (2001). Approximate confidence interval for difference in fit of structural equation models. *Structural Equation Modeling, 8,* 458–469. doi:10.1207/S15328007SEM0803_6

Raykov, T., & Marcoulides, G. A. (1999). On desirability of parsimony in structural equation model selection. *Structural Equation Modeling, 6,* 292–300. doi:10.1080/10705519909540135

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *American Statistician, 55,* 182–186. doi:10.1198/000313001317097960

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. doi:10.1214/aos/1176344136

Steiger, J. H. (2000). Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduk and Glaser. *Structural Equation Modeling, 7,* 149–162. doi:10.1207/S15328007SEM0702_1

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

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50,* 253–263. doi:10.1007/BF02294104

Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in a growth mixture model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Mixture models in latent variable research* (pp. 317–341). Greenwich, CT: Information Age.

Wasserman, L. (1997). *Bayesian model selection and model averaging* (Working Paper No. 666). Carnegie Mellon University, Department of Statistics.

Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research, 27,* 359–397. doi:10.1177/0049124199027003002

Yanagihara, H., & Ohmoto, C. (2005). On distribution of AIC in linear regression models. *Journal of Statistical Planning and Inference, 133,* 417–433. doi:10.1016/j.jspi.2004.03.016

Yuan, K.-H. (2008). Noncentral chi-square versus normal distributions in describing the likelihood ratio statistic: The univariate case and its multivariate implication. *Multivariate Behavioral Research, 43,* 109–136. doi:10.1080/00273170701836729

Yuan, K.-H., & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology, 37,* 53–82. doi:10.1111/j.1467-9531.2007.00182.x

Yuan, K.-H., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology, 56,* 93–110. doi:10.1348/000711003321645368

Yuan, K.-H., Hayashi, K., & Bentler, P. M. (2007). Normal theory likelihood ratio statistic for mean and covariance structure analysis under alternative hypotheses. *Journal of Multivariate Analysis, 98,* 1262–1282. doi:10.1016/j.jmva.2006.08.005

Yuan, K.-H., Hayashi, K., & Yanagihara, H. (2007). A class of population covariance matrices in the bootstrap approach to covariance structure analysis. *Multivariate Behavioral Research, 42,* 261–281. doi:10.1080/00273170701360662

Received February 6, 2010
Revision received September 1, 2011
Accepted September 9, 2011 ■