Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

Selecting Path Models in SEM: A Comparison of Model Selection Criteria

Li-Chung Lin, Po-Hsien Huang & Li-Jen Weng

To cite this article: Li-Chung Lin, Po-Hsien Huang & Li-Jen Weng (2017) Selecting Path Models in SEM: A Comparison of Model Selection Criteria, Structural Equation Modeling: A Multidisciplinary Journal, 24:6, 855-869, DOI: 10.1080/10705511.2017.1363652

To link to this article: https://doi.org/10.1080/10705511.2017.1363652

Structural Equation Modeling: A Multidisciplinary Journal, 24: 855–869, 2017

ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2017.1363652

Selecting Path Models in SEM: A Comparison of Model Selection Criteria

Li-Chung Lin, Po-Hsien Huang, and Li-Jen Weng
National Taiwan University

Model comparison is one useful approach in applications of structural equation modeling. Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are commonly used for selecting an optimal model from the alternatives. We conducted a comprehensive evaluation of various model selection criteria, including AIC, BIC, and their extensions, in selecting an optimal path model under a wide range of conditions over different compositions of candidate set, distinct values of misspecified parameters, and diverse sample sizes. The chance of selecting an optimal model rose as the values of misspecified parameters and sample sizes increased. The relative performance of AIC and BIC type criteria depended on the magnitudes of the parameter misspecified. The BIC family in general outperformed AIC counterparts unless under small values of omitted parameters and sample sizes, where AIC performed better. Scaled unit information prior BIC (SPBIC) and Haughton's BIC (HBIC) demonstrated the highest accuracy ratios across most of the conditions investigated in this simulation.

Keywords: AIC, BIC, model comparison, model selection criterion, path model, structural equation modeling

Structural equation modeling (SEM) has been a popular statistical method in psychology and social sciences research (Guo, Perron & Gillespie, 2009; Hershberger, 2003). Among the approaches adopted for this methodology, model comparison is a highly useful strategy (Jöreskog, 1993; MacCallum, 1995, 2003). The review by MacCallum and Austin (2000) indicated that 53% of SEM studies used this strategy. With this strategy, alternative models based on competing theories or conflicting findings are evaluated to select an optimal model that balances the trade-off between model goodness of fit and model complexity. However, selecting an optimal model can be difficult in statistical analysis (Bozdogan, 1987). As a result, model selection criteria are frequently used to single out the best model from the alternatives. Among the criteria being developed, Akaike's information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978) are the two most representative ones (see Shao, 1997).

Both AIC and BIC are formulated as the sum of negative log-likelihood and a penalty term that increases with the number of parameters in a given model. The negative log-likelihood represents the goodness of fit of a proposed model with a smaller value signifying a better fit. The penalty term shows the complexity of a model and the smaller it is, the more parsimonious a model is. Thus a model with the minimal value of AIC or BIC among all the competing models indicates an optimal balance between model fit and model complexity and would be selected. Although the two criteria are commonly used in choosing an optimal structural equation model, studies examining their performance seem limited and a comprehensive investigation on the behaviors of various model selection criteria under a wide range of conditions is called for (Bollen, Harden, Ray, & Zavisca, 2014; Haughton, Oud, & Jansen, 1997; Homburg, 1991; Vrieze, 2012). This study therefore considers the factors that have been shown to affect the performance of the model selection criteria and includes several criteria beyond those examined in prior SEM simulations. This article is organized

Po-Hsien Huang is now at Department of Psychology, National Cheng Kung University.

Correspondence should be addressed to Li-Jen Weng, Department of Psychology, National Taiwan University, Taipei 106, Taiwan. E-mail: ljweng@ntu.edu.tw

Supplemental data for this article can be accessed at www.tandfonline.com/HSEM.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/HSEM.

as follows. AIC, BIC, and their extensions are briefly introduced first. Existing simulation studies on the performance of model selection criteria in SEM are then reviewed. It then depicts the simulation design and summarizes the results. The discussion and recommendations for applying the model selection criteria are presented in the final section.

BRIEF DESCRIPTIONS OF AIC, BIC, AND THEIR EXTENSIONS

This section introduces the theoretical bases of AIC, BIC, and their extensions, including AIC3 (Sclove, 1987), consistent AIC (CAIC; Bozdogan, 1987), consistent AIC with Fisher information (CAICF; Bozdogan, 1987), adjusted BIC (ABIC; Sclove, 1987), Haughton's BIC (HBIC; Haughton, 1988), Kashyap's BIC (KBIC; Kashyap, 1982), information matrix-based Bayesian information criterion (IBIC; Bollen, Ray, Zavisca, & Harden, 2012), and scaled unit information prior BIC (SPBIC; Bollen et al., 2012). Even though these criteria can be directly applied to SEM to select an optimal model, to our knowledge, no studies have simultaneously compared the behaviors of all these criteria in SEM and a systematic investigation of their relative performances is still needed (Dziak, Coffman, Lanza, & Li, 2012).

Both AIC and BIC were proposed under the framework of maximum likelihood (ML) estimation. ML is also the most commonly used estimation method in SEM applications (Jackson, Gillaspy, & Purc-Stephenson, 2009; Kline, 2011). In SEM, the $p \times p$ population covariance matrix Σ among p observed variables is assumed to be a function of a k-dimensional parameter vector θ embedded in the true population model; that is, $\Sigma = \Sigma(\theta)$. For a sample of N observations randomly drawn from a multivariate normal distribution with population covariance matrix Σ, the ML estimator $\hat{\theta}$ could be obtained by maximizing the log-likelihood function,

$$\log L(\theta) = -\frac{N}{2}\left\{p\log(2\pi) + \log|\Sigma(\theta)| + tr\left[S\Sigma^{-1}(\theta)\right]\right\}, \quad (1)$$

where $|\Sigma(\theta)|$is the determinant of $\Sigma(\theta)$, S is a consistent estimate of Σ, and $\Sigma^{-1}(\theta)$ is the inverse of $\Sigma(\theta)$. Maximizing $\log L(\theta)$ is equivalent to minimizing the popular ML fitting function $F[S, \Sigma(\theta)]$ in SEM asymptotically,

$$F[S, \Sigma(\theta)] = \log|\Sigma(\theta)| + tr\left[S\Sigma^{-1}(\theta)\right] - \log|S| - p. \quad (2)$$

Under the null hypothesis $\Sigma = \Sigma(\theta)$ and multinormality, $T = (N-1)F\left[S, \Sigma\left(\hat{\theta}\right)\right]$ is asymptotically distributed as a chi-square distribution with $p(p+1)/2 - k$ degrees of freedom to evaluate the plausibility of the proposed k-dimensional model. When the model is not severely misspecified, T asymptotically follows a noncentral chi-square distribution with $p(p+1)/2 - k$ degrees of freedom and noncentrality parameter $(N-1)F\left[\Sigma, \Sigma\left(\tilde{\theta}\right)\right]$ with $\tilde{\theta}$ being the value of θ minimizing $F[\Sigma, \Sigma(\theta)]$. $F\left[\Sigma, \Sigma\left(\tilde{\theta}\right)\right]$, the minimum function value of fitting the proposed model to Σ, quantifies the error of approximation of the proposed model to the population covariance matrix (Steiger, Shapiro, & Browne, 1985).

In model selection, J models each with k_j distinct parameters, denoted as $\Sigma_1(\theta_1)$, $\Sigma_2(\theta_2)$, …, $\Sigma_J(\theta_J)$, are compared. An ML estimate of θ_j, $\hat{\theta}_j$, can be obtained via minimizing the ML fitting function $F\left[S, \Sigma_j(\theta_j)\right]$. In this study, a model is considered optimal among a set of alternatives if it attains the smallest $F\left[\Sigma, \Sigma_j\left(\tilde{\theta}_j\right)\right]$ with the fewest number of parameters. The quantity $F\left[\Sigma, \Sigma_j\left(\tilde{\theta}_j\right)\right]$ signals the extent of approximation error of model j in the population and the number of parameters represents the degree of simplicity of a given model. As a result, for the set of models with the least degree of model errors in the population, the most parsimonious model is considered the optimal model in this study.

AIC was originally developed as an estimator for the Kullback–Leibler (KL) information of a model under consideration (Akaike, 1974), quantifying the discrepancy between the model considered and the true model. Akaike showed that selecting a model with minimum KL information is equivalent to choosing a model that minimizes the criterion

$$\mathrm{AIC}_j = -2\log L_j\left(\hat{\theta}_j\right) + 2k_j. \quad (3)$$

The negative log-likelihood $-2\log L_j\left(\hat{\theta}_j\right)$ measures the goodness of fit of model j to the data and $2k_j$ can be taken as a penalty term on the complexity of the considered model. Adding parameters could improve model fit and reduce the magnitude of $-2\log L_j\left(\hat{\theta}_j\right)$. Yet at the same time the value of AIC can be elevated due to the increase in k_j, especially when the added parameters do not provide sufficient improvement in model fit. When the dimensionality of the true model is infinite and when the true model is not in the candidate set, AIC is efficient in selecting a model that yields a prediction error as small as the best model asymptotically (Kuha, 2004; Shibata, 1976, 1983; Vrieze, 2012).

Shortly after Akaike (1974), Schwarz (1978) reported an alternative approach to model selection based on the Bayesian framework, showing that finding a model with the highest posterior probability is equivalent to maximizing the marginal probability of data. The log of this marginal probability for model j, $\log p_j(D)$, can be approximated by

the second-order Taylor series expansion. This approximation, when multiplied by −2, can be represented as

$$-2\log p_j(D) = -2\log L_j\left(\hat{\theta}_j\right) - 2\log p_j\left(\hat{\theta}_j\right) - k_j \log(2\pi) + k_j \log(N) + \log\left|I\left(\hat{\theta}_j\right)\right| - O\left(N^{-\frac{1}{2}}\right), \quad (4)$$

where $p_j\left(\hat{\theta}_j\right)$ is the prior probability of θ_j assumed to be a unit information prior, $I\left(\hat{\theta}_j\right)$ is the Fisher information matrix under model j, and $O\left(N^{-\frac{1}{2}}\right)$ is the approximation error. For details please refer to Haughton (1988) and Raftery (1995). As the sample size approaches infinity, the approximation is dominated by $-2\log L_j\left(\hat{\theta}_j\right)$ and $k_j \log(N)$. Schwarz therefore proposed a new criterion by retaining these two terms as shown in the common form of BIC,

$$\mathrm{BIC}_j = -2\log L_j\left(\hat{\theta}_j\right) + k_j \log(N). \quad (5)$$

Previous studies have shown that if the true model is of finite dimension and is included in the candidate set, BIC is consistent in the sense that it chooses the true model with a probability of one as the sample size approaches infinity (Haughton et al., 1997; Kuha, 2004; Nishii, 1984). Kass (1993) showed that BIC can be seen as an approximation to the log of a Bayes factor multiplied by 2. A Bayes factor, signifying the degree of superiority of one model over the other, is an effective index for comparing two candidate models under a Bayesian framework (Kass & Raftery, 1995; Raftery, 1995). However, the computation of Bayes factors is difficult because it requires the specification of the prior distribution of θ_j. BIC, with no need to specify a prior distribution, eases this difficulty and is thus regarded as a convenient alternative for comparing models (Bollen et al., 2012).

The negative log-likelihood $-2\log L_j\left(\hat{\theta}_j\right)$ can be rewritten as $N\{p\log(2\pi) + \log|\mathrm{S}| + p\} + \frac{N}{N-1}T_j$. Given a data set, the first term is a constant and can be omitted. Because T_j is commonly used in SEM, we replace $-2\log L_j\left(\hat{\theta}_j\right)$ by T_j in calculating AIC and BIC as

$$\mathrm{AIC}_j = T_j + 2k_j, \quad (6)$$

and

$$\mathrm{BIC}_j = T_j + k_j \log(N). \quad (7)$$

After the development of AIC and BIC, their extensions have been proposed based on different grounds to enhance the chance of selecting an optimal model. These extensions can also be written as a sum of T_j and a penalty term as introduced in the following ections.

AIC-Related Adjustments

Three extensions of AIC are presented in Equations 8 through 10. Sclove (1987) suggested that a range of penalty terms could be considered to expedite the search for optimal models. For example, AIC3, replacing $2k_j$ by $3k_j$ in AIC, considers a more stringent penalty than AIC. This adjusted index was found to outperform AIC in selecting the correct numbers of factors and latent classes (Dziak et al., 2012; Yang & Yang, 2007). In the meantime, Bozdogan (1987) provided a thorough theoretical account of AIC and analytically introduced two consistent extensions based on the fundamental principles of Akaike. The two extensions, CAIC and CAICF shown in Equations 9 and 10, also penalize overparameterization more stringently than AIC. As can be seen, CAIC, although belonging to the AIC family, resembles BIC. Bozdogan derived the asymptotic properties of these two criteria and found both outperforming AIC in selecting the true regression model.

$$\mathrm{AIC3}_j = T_j + 3k_j \quad (8)$$

$$\mathrm{CAIC}_j = T_j + [\log(N) + 1]k_j \quad (9)$$

$$\mathrm{CAICF}_j = T_j + [\log(N) + 2]k_j + \log\left|I\left(\hat{\theta}_j\right)\right| \quad (10)$$

BIC-Related Adjustments

Equations 11 through 15 present five extensions of BIC with θ_j^* being the prior mean of the model parameters under model j and $I_o\left(\hat{\theta}_j\right)$ being the observed information matrix of $\hat{\theta}_j$. Crediting to the work of Rissanen (1978) and Boekee and Buss (1981), Sclove (1987) presented ABIC, an adjusted BIC shown in Equation 11, based on the principle of the shortest description length in selecting a model balancing model fit and complexity. The ABIC was also known as the sample-size-adjusted BIC (e.g., Yang, 2006). Dziak et al. (2012) and Yang (2006) found that ABIC performed better than BIC in choosing the correct numbers of factors and latent classes. Haughton (1988) extended the work of Schwarz (1978) on linear models to curve models and suggested an alternative consistent criterion for model selection by retaining the third term $-k_j\log(2\pi)$ in Equation 4, leading to the HBIC shown in Equation 12. HBIC reduces the magnitude of penalty in BIC and improves the performance of BIC in selecting an optimal confirmatory factor analysis (CFA) model (Haughton et al., 1997).

$$\mathrm{ABIC}_j = T_j + \log\left(\frac{N+2}{24}\right)k_j \quad (11)$$

$$\mathrm{HBIC}_j = T_j + \log\left(\frac{N}{2\pi}\right)k_j \quad (12)$$

$$\mathrm{KBIC}_j = T_j + k_j \log(N) + \log\left|I\left(\hat{\theta}_j\right)\right| \quad (13)$$

$$\mathrm{IBIC}_j = T_j + k_j \log\left(\frac{N}{2\pi}\right) + \log\left|I\left(\hat{\theta}_j\right)\right| \quad (14)$$

$$\mathrm{SPBIC}_j = T_j + k_j \left(1 - \log\left[\frac{k_j}{\left(\hat{\theta}_j - \theta_j^*\right)^T I_o\left(\hat{\theta}_j\right)\left(\hat{\theta}_j - \theta_j^*\right)}\right]\right) \quad (15)$$

Kashyap (1982) and Bollen et al. (2012) proposed other extensions that retained the additional terms in the asymptotic approximation of $-2\log p_j(D)$. Kashyap developed a criterion KBIC as displayed in Equation 13 to select a time series model with the minimum average probability of error and argued that the fifth term $\log\left|I\left(\hat{\theta}_j\right)\right|$ in Equation 4 needed to be retained except at a very large N. Kashyap noted that KBIC behaved the same as BIC asymptotically in that both can select the true model if it is included in the candidate set. Nevertheless, in the simulation of Bollen et al. (2014), KBIC failed to improve the performance of BIC in selecting the true structural equation model at sample sizes smaller than 1,000.

Bollen et al. (2012) introduced two variants of BIC to better approximate the Bayes factors for improving the selection of optimal regression models. IBIC, as shown in Equation 14, incorporates two additional terms, $-k_j\log(2\pi)$ and $\log\left|I\left(\hat{\theta}_j\right)\right|$, in Equation 4. SPBIC, as illustrated in Equation 15, uses scaled unit information prior instead of unit information prior as a response to the arguments of Weakliem (1999). Weakliem contended that the assumption of unit information prior of θ_j in deriving BIC was unrealistic and the penalty of BIC was so heavy that it tended to select an oversimplified model. As a result, Bollen et al. (2012) developed these two extensions to improve the performance of BIC and showed that IBIC and SPBIC in general had higher accuracy ratios in selecting the true regression model than BIC in small samples.

The criteria briefly summarized consider different penalty terms and might choose distinct models from a given set of alternative models. A criterion with a larger penalty tends to select a more parsimonious model than a criterion with a smaller penalty. For example, a model with fewer parameters is likely to be selected by CAIC relative to AIC, AIC3, BIC, HBIC, or ABIC. Yet the inclusion of the information matrix in penalty makes it difficult to predict the performances of CAICF, KBIC, IBIC, and SPBIC. Simulation studies are thus needed to empirically examine their behaviors. The next section reviews related prior simulations to highlight the factors that could affect the performance of model selection criteria in SEM. Improvements in the research designs over past simulations were also noted to bring forward the aspects considered in this study beyond previous investigations.

SIMULATIONS ON MODEL SELECTION CRITERIA IN SEM

Prior simulation studies assess the effects of sample size, the composition of candidate sets, and the value of misspecified parameters on the performance of model selection criteria in SEM (Bollen et al., 2014; Haughton et al., 1997; Homburg, 1991; Vrieze, 2012). The composition of candidate sets can be further classified along two dimensions: the inclusion or exclusion of the true model, and the inclusion or exclusion of overfitting models in which extra unneeded parameters are added to the true model. Misspecified parameters are a result of model misspecification. Two types of model misspecification have been examined in previous simulations. Omitting necessary parameters from or adding unnecessary parameters to the true population model constitutes the common type of model misspecification simulated (Bollen et al., 2014; Burnham & Anderson, 2004; Homburg, 1991; Vrieze, 2012). These models involve a nested relationship with the true model. Adding or releasing constraints on model parameters, as in Haughton et al. (1997), can also be classified into this category for the misspecified model and the true population model displays a nested relationship. The other type of misspecification involves specifying a model with a different structure from the true model in that the two models do not exhibit any nested relationship (Bollen et al., 2014). Misspecified parameters resulting from both types of model misspecification are considered in this study.

In his pioneer work, Homburg (1991) examined the chance of discovering the correct population models by AIC, BIC, and cross-validation indexes (Cudeck & Brown, 1983) in sample sizes of 50 to 1,000. Data were generated from two population structural models of 11 and 12 indicators with standardized population structural coefficients ranging from .28 to .55. The true model, overfitting models with extra unneeded parameters, and models omitting necessary parameters from the true model were included in the candidate set. The probability of selecting the true model over 10 replications by AIC and BIC improved as the sample size increased up to 750 and leveled off, and BIC performed the best in general. Homburg's work provides valuable insights into the behaviors of AIC and BIC in SEM. However, two aspects of the study design could be refined. First, more than 10 successful replications could be conducted to enhance the reliability of the results. Second, the condition in which the true model is not included in the candidate set, a situation commonly encountered in reality (Cudeck & Henly, 1991), should be considered to test the generality of its findings. Haughton et al. (1997) and Bollen et al. (2014) modified Homburg's research design and

considered the condition in which the true model was not in the candidate sets, and the effect of values of misspecified parameters was also touched on by Haughton et al.

Haughton et al. (1997) compared the performance of AIC, CAIC, BIC, HBIC, BICR[1], and several fit indexes under different sample sizes, compositions of candidate sets, and values of the misspecified residual variances. Three orthogonal three-factor CFA models with two indicators of equal loading and equal residual variance for each factor were used to generate data at sample sizes of 100, 400, 1,000, and 6,000. The three true models differed in the variation of the variance of measurement errors, showing large (.25, .50, .75), intermediate, and small (.45, .50, .55) differences in residual variances. Accordingly, constraining all the residual variances to be equal signaled a greater degree of model misspecification for the true model with large difference in residual variances. Two simulations with different compositions of candidate sets were conducted. The first simulation included the true model, overfitting models freeing the equality constraints on residual variances, and models restricting all six residual variances to be equal in the candidate set. The second simulation contained only approximate models. The models fitted to the data generated were the same as in the first experiment but further constraining all the factor loadings to be equal. The model with the smallest noncentrality parameter value and the fewest number of parameters among the alternatives was considered the best. The ratio of selecting the true or the best model over 500 replications tended to improve with increasing sample sizes and larger differences in residual variances. Note that the comparison of the three true models differing in variation of residual variances could be deemed as an attempt to assess the effect of size of misspecified parameters. In most conditions, HBIC performed the best and AIC yielded a higher accuracy ratio than the remaining model selection criteria under small variation of residual variances except at N = 6,000. In general, fit indexes did not perform as well as model selection criteria in selecting an optimal model.

Haughton et al. (1997) furthered Homburg's (1991) work in considering candidate sets that excluded both the true model and overfitting models, in comparing accuracy ratios from different values of misspecified parameters, and in examining more model selection criteria. The simulation design in Haughton et al., in our opinion, could be modified to strengthen the generalizability of the findings. First, models beyond orthogonal factors could be used to improve the ecological validity of the simulated models for orthogonal factors are uncommon in psychological research (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Moreover, because each orthogonal factor had only two indicators, equality constraints had to be imposed to ensure the identification of the models being evaluated. The generalizability of the findings hence might be limited. Second, nonnested models could be included in a candidate set. All the candidate models in Haughton et al. (1997) displayed a nested relationship to the true model with the misspecified models either freeing parameters that were constrained to be equal in the true model, or confining parameters of different values in the true model to be equal. Third, a full range of magnitudes of the misspecified parameters should better be considered and manipulated explicitly. The values of misspecified parameters have been shown to affect the relative performance of AIC and BIC. For example, Burnham and Anderson (2004) showed that AIC outperformed BIC in selecting the true regression model when the values of omitted population parameters were small or moderate. Vrieze (2012) also explored the effects of values of misspecified parameters on the performance of AIC and BIC in selecting the true orthogonal CFA model. Four sample sizes from 500 to 5,000 and omitted population unstandardized factor loadings ranging from 0 to .60 were considered. Larger misspecified parameter values resulted in higher accuracy ratios in selecting the true model, and the relative performance of AIC or BIC was found to be contingent on the value of the omitted parameters. AIC was more likely to select the true model than BIC for small misspecified parameter values (lower than .30), whereas BIC performed better than AIC as the value of omitted parameters increased (larger than .40). The effect of the value of misspecified parameters on model selection criteria is worthy of further pursuit.

Bollen et al. (2014) sought to test the generality of the findings from Homburg (1991) and Haughton et al. (1997) by looking at larger structural equation models consisting of 9 and 15 observed variables with misspecified approximation models marked with minor to major deviation from the true model. The performance of AIC, BIC, HBIC, SPBIC, and various fit indexes under different sample sizes and compositions of candidate sets were examined. The values of population parameters were fixed and not manipulated. Sample sizes in this study ranged from 100 to 5,000. The population standardized loadings with all the observed and latent variables set to unit variance were specified as .70 for primary factor loadings and .21 for complex factor loadings, and the standardized structural coefficients were set at .60. Similar to Haughton et al. (1997), two simulations were conducted with the true model included in (Part I) or excluded from (Part II) the candidate set. Overfitting models with extra unnecessary parameters were always included in the candidate sets. The candidate models for Part I included the true model, overfitting models, models with necessary

[1] Dudley and Haughton (1997) developed an adjustment of BIC, BICR, by retaining the second term $2\log p_j\left(\hat{\theta}_j\right)$, the third term $-\ k_j\ \log(2\pi)$, and the fifth term $\log\left|I\left(\hat{\theta}_j\right)\right|$ in Equation 4. Haughton et al. (1997) indicated that selecting a suitable prior distribution for θ_j was difficult and found BICR to perform worse than BIC across all the simulation conditions. This study therefore did not include BICR.

parameters dropped, models with both added and dropped parameters, and nonnested models of a different number of factors and structures. The candidate models in Part II were the same as those in Part I, except the true model was excluded. For Part I, the accuracy ratios of model selection criteria over 1,000 replication samples improved as sample sizes increased up to 1,000 and leveled off. SPBIC and HBIC performed the best among the criteria evaluated, and model selection criteria outperformed fit indexes. In Part II, BIC and its extensions tended to select overfitting models as the sample size increased. For small samples, the probability of selecting a model with omitted parameters was higher for BIC and IBIC than for SPBIC and HBIC.

Bollen et al. (2014) extended prior SEM simulations on model selection criteria (Haughton et al., 1997; Homburg, 1991) by considering larger structural equation models and a variety of alternative models. The generality of their findings could have been enhanced if some candidate sets had excluded overfitting models and if the values of the misspecified parameters could have covered a wider range. Overfitting models with unnecessary parameters were always included in the candidate sets of their simulations. However, Chakrabarti and Ghosh (2011) showed in a regression example that AIC outperformed BIC when the candidate set included the true model and excluded overfitting models. Such a condition was not considered in Bollen et al. (2014) and should be investigated to thoroughly understand the behaviors of AIC, BIC, and their extensions. In addition, with the values of the misspecified parameters fixed, the effects of parameter value on the relative performances of the model selection criteria as demonstrated in Vrieze (2012) were not examined either, but merit further investigation to extend the generalizability of their findings to models where the misspecified parameters are of diverse values.

This study hence attempts to improve the study design of previous simulations in four aspects. First, four types of candidate sets were constructed based on the simultaneous consideration of two factors: (a) the inclusion or exclusion of the true model, and (b) the inclusion or exclusion of overfitting models. In light of the results of Chakrabarti and Ghosh (2011), AIC and its extensions are expected to have higher accuracy ratios than BIC and its family when the candidate sets include the true model and exclude models adding unnecessary parameters. BIC and its extensions are expected to perform better under the other three conditions. Second, the effect of the values of misspecified parameters was examined by extending the work of Vrieze (2012) to structural models. Based on the results reported by Burnham and Anderson (2004) and Vrieze (2012), AIC type criteria are anticipated to outperform BIC and its extensions under small values of misspecified parameters, whereas BIC type criteria are expected to perform better as the values of misspecified parameters increase.

Third, this study considered AIC3, CAICF, and ABIC, three measures not evaluated in prior SEM simulations. These three criteria have been shown to improve the performance of AIC and BIC in selecting the correct numbers of factors (Dziak et al., 2012) and latent classes (Yang, 2006; Yang & Yang, 2007), or the true regression model (Bozdogan, 1987). In factor analysis and latent class models, AIC3 outperformed AIC in sample sizes larger than 100 ($N \geq 100$ in Dziak et al., 2012, and $N \geq 200$ in Yang & Yang, 2007), and ABIC improved the performance of BIC, particularly in small or moderate sample sizes ($N \leq 300$ in Dziak et al., 2012, and $N \leq 700$ in Yang, 2006). CAICF was also found to outperform AIC in selecting the true regression model for sample sizes larger than 100 and reduced the probability of selecting an overfitting model at large sample sizes (Bozdogan, 1987). If these three criteria could enhance the chance of selecting an optimal exploratory factor model, latent class model, and regression model over AIC and BIC, a comprehensive evaluation of their performance in SEM is warranted. Based on previous findings, AIC3 and CAICF are expected to outperform AIC, except for small sample sizes, and ABIC is expected to be superior to BIC in small to moderate sample sizes.

Finally, path models were used in this study in responding to the urging of McDonald (2010) that measurement model and path model should be sequentially and separately evaluated in SEM applications. After reevaluating the empirical uses of SEM in psychology journals, McDonald and Ho (2002) raised the concern that model evaluation in SEM could be distorted by fitting the full structural models to the data because a good fit of a measurement model might conceal the misfit of a path model. The need to separately assess the measurement component and the path component in a structural equation model was then advocated. Nearly a decade later, O'Boyle and Williams (2011) examined organizational research using latent variables and found that in approximately half of the studies (47%) they reviewed, a misfit of the relations among latent variables as indicated by a path model was masked by a good fit of the measurement model. In the meantime, McDonald (2010) noted that despite their recommendation, most of the recent SEM studies continued to ignore the problem they raised in model evaluation. This negligence prompted him to restate the importance of assessing the fit of a path model after evaluating the fit of the measurement model. Note that most prior simulations we reviewed (Bollen et al., 2014; Haughton et al., 1997; Homburg, 1991; Vrieze, 2012) used either measurement models or full structural models. Whether model selection criteria can select an optimal path model is still an open question. An investigation of the performance of model selection criteria on path models is hence called for so that McDonald and colleagues' recommendation can be followed when competing models are compared. This study therefore employed path models in the simulations to fill in the gap in the literature.

In sum, this study expanded the design of previous simulations to assess the performance of AIC, BIC, and their extensions in selecting an optimal path model from the candidate set. Model selection criteria beyond those assessed in previous SEM simulations were examined, including AIC3, CAICF, and ABIC. Factors shown to affect the behaviors of the model selection criteria were simultaneously considered, including the sample size, the inclusion or exclusion of true model in the candidate set, overfitting models being among the alternatives or not, and the value of misspecified population parameters. As a result, this study can be expected to provide useful guidelines to facilitate optimal model selection in SEM.

METHOD

Monte Carlo simulation was conducted to examine the impact of the composition of the candidate set, the values of misspecified population parameters, and the sample size on the probability of selecting an optimal path model from a candidate set by various model selection criteria. In light of the importance of mediating variables in psychological studies (Baron & Kenny, 1986) and scientific research (Kenny, 2008), this study employed a mediation path model shown in Figure 1a as the population model (M_0). Considering the statistical and practical significance of the parameters (Paxton, Curran, Bollen, Kirby, & Chen, 2001), the values of ϕ_{21}, γ_{11}, γ_{21}, γ_{42}, and β_{31} in population model M_0 were fixed at .40. The magnitudes of γ_{12}, γ_{31}, β_{21}, and β_{41} were varied to evaluate the influence of the sizes of misspecified parameters on behaviors of the criteria. The sizes of residual variances were deliberately chosen when necessary so that all the observed variables had unit variances of 1.

Variables Manipulated in Simulations

Compositions of the candidate sets

Following the principle of constructing alternative models in Bollen et al. (2014), nine alternative path models (M_1–M_9) in addition to the true model (M_0) were specified, as summarized in Table 1. Model M_1 omitted one parameter from M_0. Models M_2 and M_3 added extra unnecessary parameters to M_0 that both had minimum function values of zero as the true model and were thus regarded as overfitting models. Models M_4 to M_6 added one unneeded parameter and omitted certain nonzero parameters in M_0. Models M_7 to M_9 represented models of different structural relationships among the variables, as shown in Figure 1.

To fully represent the model comparison scenarios encountered in practice, four types of candidate sets were constructed by simultaneously considering two factors: (a) the inclusion or exclusion of the true model (*T*), and (b) the inclusion or exclusion of overfitting models with extra unneeded parameters added to the true model (*O*). Models from M_0 to M_9 were deliberately chosen and assembled to form each type of candidate set. Candidate Set 1, consisting of M_0, M_1, and M_4 to M_9, represented the condition in which the set included the true model but excluded overfitting models. This condition has not been studied in prior studies. Overfitting models M_2 and M_3 were added to Candidate Set 1 to form Candidate Set 2 to examine whether the true model could be selected over overfitting models when present.

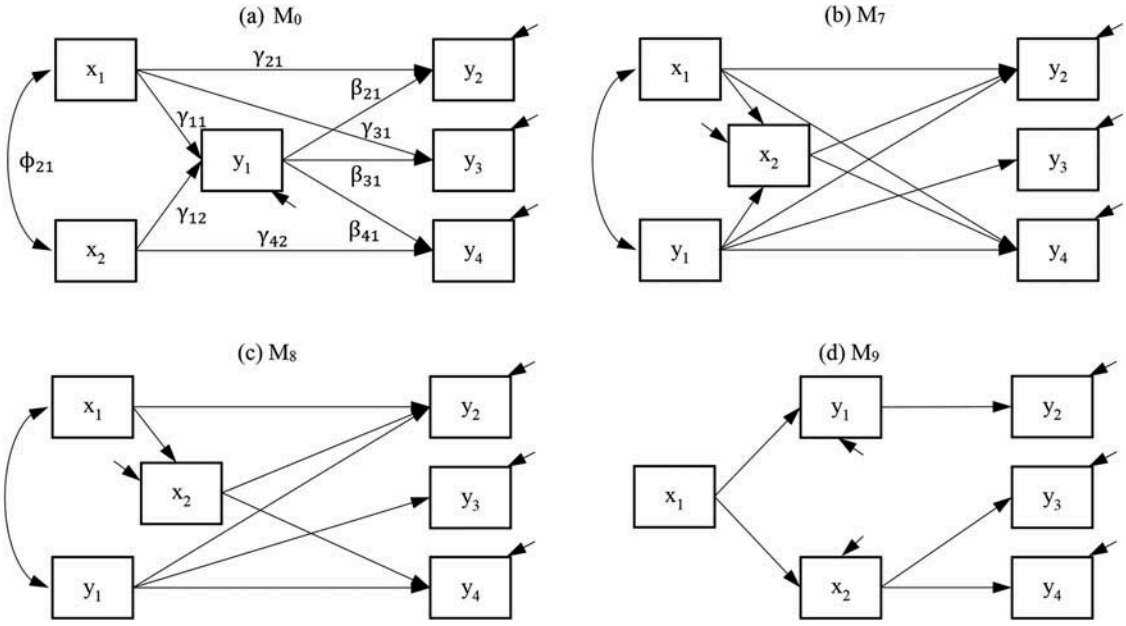

FIGURE 1 Path diagrams of true model M_0 and nonnested candidate models M_7–M_9.

TABLE 1
Characteristics of Candidate Models With Associated Population Minimum Function Value $F((\Sigma, \Sigma(\tilde{\theta}_j))$ and RMSEA[a] at Five Values of Misspecified Parameters

Model Description	*k*	*Omitted Parameters*	*Extra Parameters*	*F ((Σ, Σ(θ̃ⱼ)) Under Values of Misspecified Parameters at* .1	.2	.3	.4	.5
Nested misspecified path models								
M_1 Omit 1 coefficient	14	γ_{31}		0.01 (0.04)	0.04 (0.08)	0.10 (0.12)	0.20 (0.17)	0.38 (0.23)
M_2 Add two coefficients	17		β_{32}, β_{42}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
M_3 Add three coefficients	18		γ_{41}, γ_{22}, γ_{32}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
M_4 Omit one coefficient and add one coefficient	15	γ_{31}	γ_{22}	0.01 (0.04)	0.04 (0.08)	0.10 (0.13)	0.20 (0.18)	0.38 (0.25)
M_5 Omit three coefficients and add one coefficient	13	γ_{31}, γ_{12}, β_{41}	γ_{41}	0.03 (0.06)	0.13 (0.13)	0.30 (0.19)	0.58 (0.27)	1.06 (0.36)
M_6 Omit four coefficients and add one coefficient	12	γ_{31}, γ_{12}, β_{21}, β_{41}	γ_{41}	0.04 (0.07)	0.17 (0.14)	0.40 (0.21)	0.78 (0.29)	1.44 (0.40)
Nonnested misspecified path models								
M_7 Switch variables (x_2 and y_1)	16	β_{31}	γ_{22}, γ_{32}	0.01 (0.05)	0.04 (0.09)	0.10 (0.14)	0.20 (0.20)	0.38 (0.28)
M_8 Switch variables (x_2 and y_1)	14	γ_{31}, γ_{12}, β_{31}	γ_{22}, γ_{32}	0.02 (0.05)	0.09 (0.11)	0.21 (0.17)	0.42 (0.25)	0.77 (0.33)
M_9 Switch the variable x_2	11	ϕ_{21}, γ_{31}, γ_{12}, γ_{42}, β_{21}, β_{41}	β_{42}, β_{52}, ζ_{55}	0.38 (0.20)	0.53 (0.23)	0.77 (0.28)	1.15 (0.34)	1.79 (0.42)

Note. RMSEA = root mean square error of approximation; *k* = number of parameters.
[a]RMSEA is shown in parentheses.

Both Candidate Sets 1 and 2 included the true model M_0. Yet in reality investigators seldom know what the population true model is (Cudeck & Henly, 1991; McDonald, 2010). True model M_0 was therefore removed from both sets to form Candidate Sets 3 and 4. As a result, Candidate Set 3, with both true and overfitting models excluded, could be used to examine whether an approximate model with the smallest $F(\Sigma, \Sigma_j(\tilde{\theta}_j))$, the minimum function value when fitting a model to the population covariance, and the fewest number of parameters could be selected. For this particular composition of alternative models, model M_1 was considered optimal. Candidate Set 4, composed of M_1 to M_9, simulated the situation where all the models were approximations and overfitting models were among the alternatives. Model M_2 was considered optimal for this set because it fully explained the population covariance matrix with the fewest number of parameters.

Values of misspecified parameters (V)

The parameter values of γ_{12}, γ_{31}, β_{21}, and β_{41} in the true model were systematically varied to explore the effect of the size of misspecified parameters on the behaviors of model selection criteria. The magnitudes of these four parameters were assumed to be equal and set to vary from .1 to .5 at an increment of .1. Table 1 displays the population minimum function value $F[\Sigma, \Sigma_j(\tilde{\theta}_j)]$ and the corresponding root mean square error of approximation (RMSEA; Steiger & Lind, 1980) at every level of misspecified parameter values under each model. The RMSEAs indicated that the five levels could represent minor to large degrees of model misspecification except for M_2, M_3, and M_9 (Browne & Cudeck, 1992). This study extended the manipulation of the size of omitted parameters in CFA models (Vrieze, 2012) to path models and considered a wider range of misspecified parameter values than Homburg (1991) and Bollen et al. (2014).

Sample sizes (N)

The likelihood of selecting an optimal model by model selection criteria has been shown to improve as the sample size increases (Bollen et al., 2014; Haughton et al., 1997; Homburg, 1991; Vrieze, 2012). Previous simulations have examined sample sizes between 100 and 6,000 and found the accuracy ratios of model selection criteria for a sample size of 750 to be similar to those for a sample size of 1,000 (Homburg, 1991) and likely to level off subsequently (Bollen et al., 2014; Haughton et al., 1997; Vrieze, 2012). This study therefore focused on sample sizes of less than 1,000 and considered six levels of sample size at 100, 200, 300, 400, 800, 1,000, and 2,000. This range of *N* appears to

reflect the common sample sizes observed in empirical SEM applications (Baumgartner & Homburg, 1996; Guo et al., 2009; Shah & Goldstein, 2006).

Data Generation and Analysis

All the data were generated and analyzed using the *sem* package version 3.1–1 (Fox, Nie, & Byrnes, 2013) in R. Multivariate normal data were simulated according to the true model with prespecified parameter values and sample sizes. All 10 models (M_0–M_9) were then fitted to the generated data by the ML method. A data set that resulted in nonconvergence or improper solutions in any of the 10 analyses was removed and replaced with additionally generated data until 1,000 data sets yielded proper solutions on all 10 models. Model selection criteria AIC, BIC, and their extensions, including AIC3, CAIC, CAICF, KBIC, IBIC, HBIC, SPBIC, and ABIC, were calculated for each fitted model based on the formula described previously. The computation of SPBIC in this study followed the procedure in Bollen et al. (2014). There were a total of 140 conditions (4 compositions of candidate set under two features × 5 values of omitting parameters × 7 sample sizes) in this study. Under each one of the 140 conditions, whether a model selection criterion selected the optimal model (M_0 for Candidate Sets 1 and 2, M_1 for Candidate Set 3, and M_2 for Candidate Set 4) was recorded. When a criterion failed to select the optimal model, the model it selected was also documented. Four-way analysis of variance (ANOVA) was conducted for each model selection criterion to examine the influence of manipulated factors on its success (coded as 1) or failure (coded as 0) in selecting an optimal model with the effect size indicated by eta squared ($\eta^2 = SS_{effect}/SS_{total}$). Note that the composition of candidate sets was represented by two factors in the ANOVAs: the inclusion or exclusion of the true model in the set and the inclusion or exclusion of overfitting models in the set. For each composition of candidate set, the percentage that every model selection criterion selected the optimal model over 1,000 replications under any combination of value of misspecified parameters and sample size was calculated. These accuracy ratios were further compared to reveal the relative performances of the 10 model selection criteria investigated.

RESULTS

All the analyses in the original 350,000 analyses (5 values of misspecified parameters × 7 sample sizes × 10 models × 1,000 data sets) reached converged solutions within 30 iterations. Only one improper solution was found for the sample size of 100 paired with a misspecified parameter value of .5. An additional data set was generated. The eta squared presented in Table 2 indicated that the values of misspecified parameter exerted the largest effect on whether the optimal model could be selected by various model selection criteria, followed by sample sizes and their interaction. As to be discussed, the large influences of the magnitudes of misspecified parameters mainly resulted from the poor performance of the criteria at small values of misspecified parameters. Whether the candidate set contained true or overfitting models did not exhibit substantial effects on the performance of the criteria. However, to systematically compare these findings to those from previous studies

TABLE 2
η^2 for Each of the Model Selection Criterion at Manipulated Factors

	AIC	AIC3	CAIC	CAICF	BIC	KBIC	IBIC	HBIC	SPBIC	ABIC
T	0.025	0.022	0.005	0.005	0.007	0.004	0.004	0.013	0.014	0.019
O	0.041	0.025	0.019	0.024	0.019	0.023	0.021	0.020	0.013	0.028
V	0.050	**0.130**	***0.392***	***0.469***	***0.356***	***0.469***	***0.467***	***0.229***	***0.224***	**0.090**
N	0.020	0.045	**0.092**	**0.115**	**0.085**	**0.103**	**0.091**	**0.075**	**0.078**	**0.088**
T × O	0.002	0.005	0.019	0.024	0.018	0.023	0.021	0.012	0.008	0.001
T × V	0.001	0.000	0.001	0.003	0.001	0.003	0.003	0.000	0.001	0.000
T × N	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.001	0.001
O × V	0.000	0.002	0.011	0.008	0.010	0.009	0.010	0.007	0.006	0.002
O × N	0.000	0.000	0.001	0.002	0.001	0.001	0.001	0.002	0.001	0.004
V × N	0.034	**0.071**	**0.087**	**0.071**	**0.083**	**0.070**	**0.071**	**0.061**	**0.072**	0.026
T × O × V	0.002	0.004	0.011	0.008	0.011	0.009	0.010	0.010	0.005	0.006
T × O × N	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.000	0.000	0.005
T × V × N	0.001	0.001	0.008	0.015	0.007	0.012	0.010	0.004	0.003	0.002
O × V × N	0.000	0.002	0.017	0.026	0.013	0.024	0.022	0.005	0.003	0.003
T × O × V × N	0.001	0.002	0.017	0.026	0.013	0.024	0.022	0.004	0.002	0.001

Note. η^2 greater than .059 are shown in bold and η^2 over .138 are further displayed in italics. AIC = Akaike's information criterion; CAIC = consistent Akaike's information criterion; CAICF = consistent AIC with Fisher information; BIC = Bayesian information criterion; KBIC = Kashyap's information criterion; IBIC = information matrix-based criterion; HBIC = Haughton Bayesian information criterion; SPBIC = scaled unit information prior Bayesian information criterion; ABIC = adjusted Bayesian information criterion; T = including or excluding the true model; O = including or excluding overfitting models; V = values of misspecified population parameters.

employing different setups of alternative models (Bollen et al., 2014; Haughton et al., 1997; Homburg, 1991; Vrieze, 2012), the results from every composition of candidate set are presented accordingly.

Figures 2 through 4 present the percentages at which each model selection criterion selected the optimal model in every type of candidate set, except Set 3. The figure for Set 3 is presented online due to its similarity to Set 2 and space limitations. The likelihood of selecting an optimal model was shown to increase with higher values of misspecified parameters and larger sample sizes, regardless of the alternative models contained in the set across all the model selection criteria evaluated. In agreement with previous studies (Bollen et al., 2014; Haughton et al., 1997; Homburg, 1991; Vrieze, 2012), the increase in the accuracy ratio appeared trivial for sample sizes over 1,000.

Accuracy Ratios in Candidate Set 1 (M_0, M_1, M_4–M_9)

Candidate Set 1 included the true model but excluded overfitting models. This composition of alternative models has not been investigated in prior SEM simulations. As shown in Figure 2, the true model was consistently selected by all the criteria, and the accuracy ratio approached 100% when misspecified parameter values were .3 or larger. The overall accuracy ratio tended to be higher than for the other three types of candidate sets. If the true model was missed, overly simple models, M_1 and M_6, were likely to be selected. As expected, AIC type criteria outperformed their BIC counterparts, especially when small values of misspecified parameters were paired with small sample sizes. For smaller values of misspecified parameters, AIC performed better than its three variants, and ABIC performed the best among BIC type criteria, followed closely by SPBIC and HBIC. AIC3 and CAICF failed to bring forth enhancement over AIC in this composition of candidate models, whereas ABIC improved the performance of BIC across all sample sizes and magnitudes of misspecified parameters. It is interesting to note that AIC in general yielded the highest accuracy ratio among all the model selection criteria for this composition of candidate models, followed by ABIC.

Accuracy Ratios in Candidate Set 2 (M_0, M_1–M_9)

Both the true model and overfitting models were included in Candidate Set 2. This composition has been studied in most previous simulations, including Homburg (1991), Haughton et al. (1997), Vrieze (2012), and Bollen et al. (2014). The accuracy ratios of the examined model selection criteria under this composition rose significantly as the value of misspecified parameters increased and started to level off as the value of misspecified parameters went beyond .3. BIC type criteria in general showed a higher tendency toward selecting the true model than AIC counterparts unless the small misspecified parameter value of .1 was paired with sample sizes less than 400, as displayed in Figure 3. Overall, SPBIC, HBIC, BIC, and CAIC performed better than the remaining criteria with relatively high accuracy ratios unless at a small misspecified parameter value of .1. Over parsimonious models M_1 and M_6 were likely to be selected when the true model failed to be singled out. Consistent with the previous findings (Dziak et al., 2012;

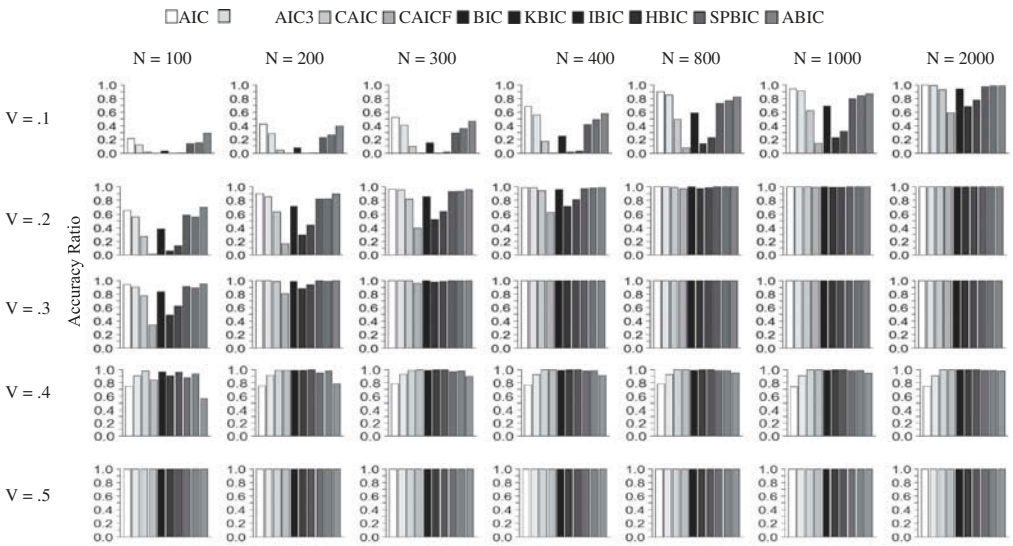

FIGURE 2 Accuracy ratios of selecting the optimal model by each model selection criteria at various sample sizes (*N*) and values of misspecified parameters (V): Candidate Set 1. AIC = Akaike's information criterion; CAIC = consistent Akaike's information criterion; CAICF = consistent AIC with Fisher information; BIC = Bayesian information criterion; KBIC = Kashyap's information criterion; IBIC = information matrix-based criterion; HBIC = Haughton Bayesian information criterion; SPBIC = scaled unit information prior Bayesian information criterion; ABIC = adjusted Bayesian information criterion.

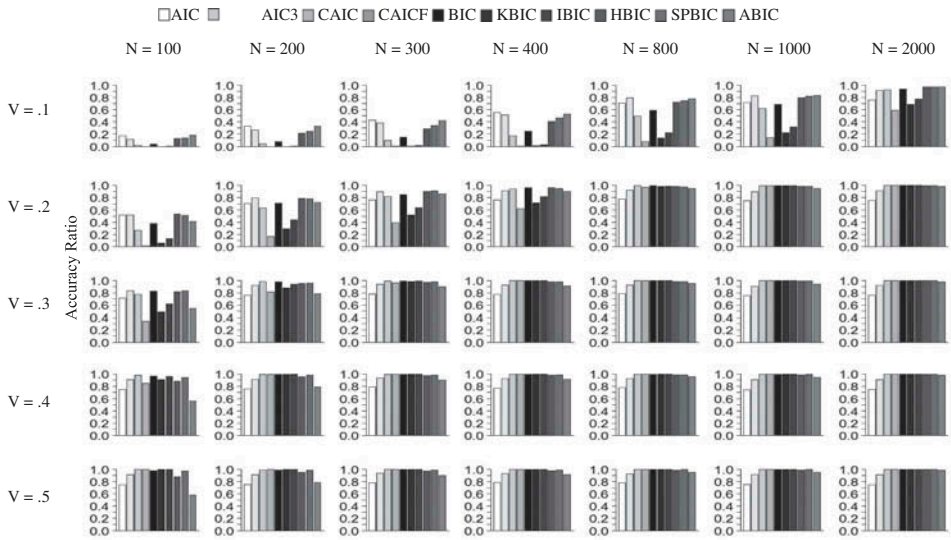

FIGURE 3 Accuracy ratios of selecting the optimal model by each model selection criteria at various sample sizes (*N*) and values of misspecified parameters (V): Candidate Set 2. AIC = Akaike's information criterion; CAIC = consistent Akaike's information criterion; CAICF = consistent AIC with Fisher information; BIC = Bayesian information criterion; KBIC = Kashyap's information criterion; IBIC = information matrix-based criterion; HBIC = Haughton Bayesian information criterion; SPBIC = scaled unit information prior Bayesian information criterion; ABIC = adjusted Bayesian information criterion.

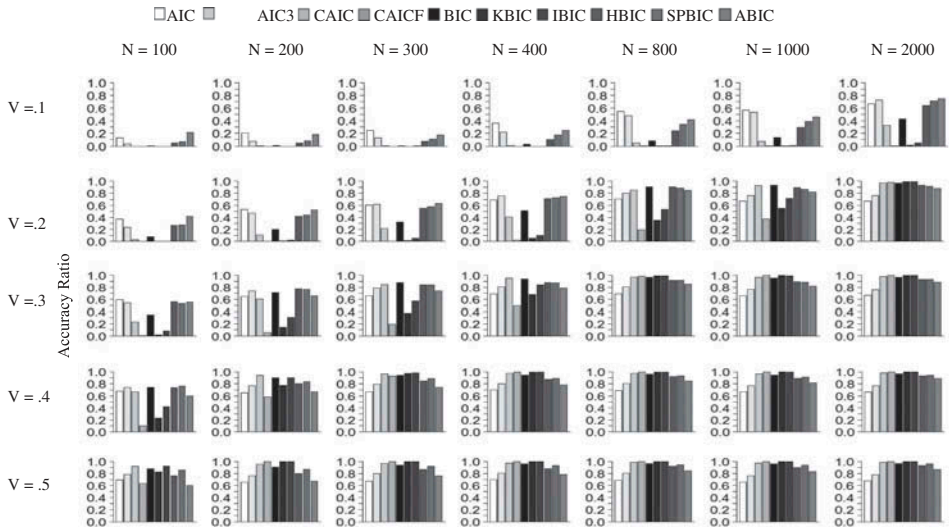

FIGURE 4 Accuracy ratios of selecting the optimal model by each model selection criteria at various sample sizes (*N*) and values of misspecified parameters (V): Candidate Set 4. AIC = Akaike's information criterion; CAIC = consistent Akaike's information criterion; CAICF = consistent AIC with Fisher information; BIC = Bayesian information criterion; KBIC = Kashyap's information criterion; IBIC = information matrix-based criterion; HBIC = Haughton Bayesian information criterion; SPBIC = scaled unit information prior Bayesian information criterion; ABIC = adjusted Bayesian information criterion.

Yang & Yang, 2007), AIC3 brought improvement over AIC. Yet it failed to consistently select the true model, even for a sample size of 2,000. As expected by Bozdogan (1987), CAICF reduced the tendency to select an overfitting model in large sample sizes. ABIC improved the performance of BIC for $N \leq$ 1,000 only when the misspecified parameter value was as small as .1. Among the model selection criteria examined, SPBIC and HBIC performed the best for Candidate Set 2, followed by BIC and CAIC.

The simulation conditions in Homburg (1991) and in Part I of Bollen et al. (2014) were similar to Candidate Set 2, but with only moderate to large values of misspecified parameters. The superior performance of BIC type criteria over the AIC family found in their studies was replicated. This study also considered small values of misspecified parameters as in Vrieze (2012) and repeated the finding that AIC type criteria outperformed BIC counterparts under such conditions. Consistent with the report of Vrieze, the

relative performance of AIC and BIC type criteria in this candidate set was found contingent on the magnitude of the misspecified parameters. This phenomenon was not revealed in previous SEM simulations on model selection criteria in which small values of misspecified parameters were not considered.

Accuracy Ratios in Candidate Set 3 (M_1, M_3–M_9)

Candidate Set 3, excluding both the true model and the overfitting models, consisted of approximate models only. The second simulation in Haughton et al. (1997) was similar to this set with true and overfitting models excluded. However, nonnested models were not considered in their study and were incorporated in Set 3. The results from Set 3 (see supplemental material) resembled those in Set 2 with slightly higher accuracy ratios. The relative performance of AIC and BIC type criteria was again shown to be conditioned on the magnitude of misspecified parameters. BIC type criteria, except for KBIC and IBIC, behaved similarly and outperformed AIC counterparts for misspecified parameter values larger than .1. In contrast, AIC and AIC3 performed better than all the BIC type criteria except ABIC at small values of the misspecified parameter. This study examined more model selection criteria than Haughton et al. and replicated their findings on the superiority of BIC type criteria over the AIC family when the values of misspecified parameters were moderate to large.

Most BIC type criteria and the consistent versions of AIC selected the optimal model at large samples, whereas AIC and AIC3 might select overly complex models (M_4 or M_7) even with a sample size of 2,000. As expected, AIC3 improved the performance of AIC at moderate to large sample sizes. The behavior of CAICF, although better than AIC at large samples as anticipated, appeared unstable and varied widely across simulation conditions. ABIC, as in Candidate Set 2, outperformed BIC only when sample sizes smaller than 800 were paired with a small misspecified parameter value of .1. For Candidate Set 3, across the 10 criteria evaluated, HBIC, SPBIC, BIC, and CAIC emerged with better performance in most conditions, and AIC, AIC3, and ABIC performed the worst.

Accuracy Ratios in Candidate Set 4 (M_1–M_9)

Candidate Set 4 included overfitting models but excluded the true model. The composition of alternative models in Part II of Bollen et al. (2014) was similar to this set, but it had values of misspecified parameters that were moderate to large. As shown in Figure 4, the overall likelihood of selecting the optimal model in Set 4 (M_2) was lower than in the other three candidate sets, especially under small values of misspecified parameters, for which nearly all the criteria failed to select M_2 unless under large samples. When the misspecified parameter value was as small as .1, a situation not considered in Bollen et al., AIC and AIC3 outperformed the BIC family and ABIC performed better than the rest of the BIC type criteria. As the values of misspecified parameters increased, BIC type criteria, except KBIC and IBIC, tended to perform better than their AIC counterparts, a result consistent with the finding in Part II of Bollen et al. (2014). Similar to Candidate Sets 2 and 3, the moderating effect of the magnitude of misspecified parameters on the relative performance of AIC and BIC type criteria was again observed. Most BIC type criteria and consistent versions of AIC selected the optimal model for large sample sizes, whereas AIC and AIC3 were likely to favor overly complex models (M_3). AIC3 improved the performance of AIC at moderate to large samples, and CAICF outperformed AIC under large sample sizes. ABIC improved the performance of BIC at a small misspecified parameter value of .1 across all the sample sizes. Overall, SPBIC, HBIC, CAIC, and BIC had higher accuracy ratios than the other criteria examined for Set 4 when neither the true model nor overfitting models were among the alternatives.

Summary

The examined criteria performed the best in Candidate Set 1, followed by Sets 2 and 3, and had the lowest accuracy ratios in Set 4. As expected, AIC type criteria outperformed BIC type criteria when the alternative models included the true model and excluded overfitting models as in Set 1, and BIC type criteria performed better under the other three compositions of candidate models, except for small values of misspecified parameters. Other than Candidate Set 1, the relative performance of the criteria in both families was contingent on the values of the misspecified parameters, as Vrieze (2012) demonstrated. Consistent with the expectation, BIC type criteria showed higher accuracy ratios than their AIC counterparts, unless under small values of misspecified parameters where the AIC family performed better. Three model selection criteria —AIC3, CAICF, and ABIC—not considered in previous simulations were examined in this study. As expected, AIC3 and CAICF improved the performance of AIC for moderate to large sample sizes in general. The expectation that ABIC outperformed BIC for small to moderate sample sizes was partly supported. ABIC did improve the performance of BIC for small to moderate sample sizes in Candidate Set 1. However, for the other three sets, the better performance of ABIC was observed only for small misspecified parameter values. Overall, among the model selection criteria we evaluated, SPBIC and HBIC performed the best across most conditions of this simulation followed closely by BIC and CAIC.

DISCUSSION

This study extended previous simulations on the behaviors of model selection criteria in SEM (Bollen et al., 2014;

Haughton et al., 1997; Homburg, 1991; Vrieze, 2012) to investigate the relative performances of AIC, BIC, and their extensions in selecting an optimal path model in various conditions. Different compositions of candidate sets, distinct values of misspecified parameters, and diverse sample sizes along with multiple criteria were considered. The relative performances of the examined model selection criteria were found to depend on the values of misspecified parameters in most conditions, a finding revealed in Vrieze (2012), but not in other studies. The findings also resembled those from Burnham and Anderson (2004) on the regression model that the misspecified parameter values influenced the relative performance of AIC and BIC. Except in Candidate Set 1 where the alternative models included the true population model and excluded any overfitting models, for misspecified parameters to be of moderate to large sizes as usually considered in past simulations (Bollen et al., 2014; Homburg, 1991), BIC type criteria, other than ABIC, outperformed the AIC family. In contrast, AIC and AIC3 yielded higher accuracy ratios than the BIC counterparts when the misspecified parameters were of small magnitudes. The superior performance of BIC type criteria demonstrated in previous studies was perhaps due to the neglect of small misspecified parameters in the simulation designs. When alternative models do not substantially deviate from the population model, BIC type criteria tend to overpenalize model complexity and result in selecting oversimplified models.

The composition of Candidate Set 1 has not been investigated in previous simulations. It is worth noting that the AIC family consistently performed better than the BIC family under this set that included the true model but excluded overfitting models. The other three candidate sets have been examined in prior research (Bollen et al., 2014; Haughton et al., 1997; Homburg, 1991; Vrieze, 2012) and we replicated previous findings in these conditions in which the BIC type criteria outperformed the AIC counterparts. Although the effects of the candidate set composition were marginal in this study, this study bridges the gap in the literature to comprehensively consider possible compositions of candidate models and sheds light on the conditions in which AIC type criteria might be preferred over the BIC family. AIC has also been shown to yield higher accuracy ratios than BIC under small values of misspecified parameters, as BIC tended to select an overly parsimonious model due to its high magnitude of penalty. On the other hand, except for Candidate Set 1, as the misspecified parameter values increased, BIC performed better than AIC, and AIC had a tendency to select an overly complex model due to its small size of penalty.

Most extensions of BIC were developed by considering additional terms in $-2\log p_j(D)$ beyond $-2\log L_j\left(\hat{\theta}_j\right)$ and $k_j \log(N)$. In contrast to Kashyap (1982) and Bollen et al. (2012), this study did not find it beneficial to include the term $\log\left|I\left(\hat{\theta}_j\right)\right|$. KBIC and IBIC performed poorly, as did the AIC type criterion CAICF, which contains $\log\left|I\left(\hat{\theta}_j\right)\right|$ as well. Adding the Fisher information matrix in the criteria tended to severely penalize model complexity and resulted in selecting overparsimonious models. Based on the findings, model selection criteria with $\log\left|I\left(\hat{\theta}_j\right)\right|$ were not recommended.

This study covers a variety of conditions. It is worth noting that over all the scenarios, SPBIC and HBIC performed the best, followed by BIC and CAIC. The superior performances of SPBIC and HBIC were also recognized in the studies of Haughton et al. (1997) and Bollen et al. (2014). Between the best two, HBIC might be preferable to SPBIC for its simplicity in computation. When comparing several competing models, researchers rarely know in advance the magnitudes of the parameters that are possibly misspecified, and whether the candidate set contains the true population model or any overfitting models. In practice, using model selection criteria that have been shown to yield high accuracy ratios in selecting optimal models over diverse settings will reduce the chance of making inappropriate inferences from the analyses.

The current simulations used path models instead of measurement models or full structural equation models. Previous studies on the performance of model selection criteria in SEM employed either measurement models (Haughton et al., 1997; Vrieze, 2012) or full structural equation models (Bollen et al., 2014; Homburg, 1991). Path models were deliberately chosen in this study to fill in the gap in the literature and to respond to the recommendation of McDonald and colleagues. McDonald and Ho (2002) and McDonald (2010) advocated the separate assessment of measurement models and path models to prevent the misfit of a path model to be concealed by a good fit of a measurement model. With HBIC standing out in the selection of measurement models in Haughton et al. (1997), SPBIC and HBIC surpassing other criteria in selecting full structural models in Bollen et al. (2014), and these two criteria again performing the best in selecting path models in our study, we recommend SPBIC and HBIC for model comparison in SEM.

The generalizability of the results from this study is, however, limited to the conditions simulated, as in most Monte Carlo simulations. Future studies covering a wide variety of conditions should be conducted to evaluate whether these findings would generalize to other research scenarios such as different sets of models or different distributions of observed variables. Even with limited generalizability, we expand our study design beyond previous research and provide useful suggestions on the choice of model selection criteria for the future practice of model comparison in SEM.

FUNDING

This research was supported in part by grants NSC 99-2410-H-002-083-MY3 and NSC 102-2410-H-002-053 from the National Science Council in Taiwan.

REFERENCES

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, *AC-19*, 716–723. doi:10.1109/TAC.1974.1100705

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, *13*, 139–161. doi:10.1016/0167-8116(95)00038-0

Boekee, D. E., & Buss, H. H. (1981). Order estimation of autoregressive models. In *Proceedings of the 4th Aachen colloquium: Theory and application of signal processing* (pp. 126–130).

Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). BIC and alternative Bayesian information criteria in the selection of structural equation models. *Structural Equation Modeling*, *21*, 1–19. doi:10.1080/10705511.2014.856691

Bollen, K. A., Ray, S., Zavisca, J., & Harden, J. J. (2012). A comparison of Bayes factor approximation methods including two new methods. *Sociological Methods & Research*, *41*, 294–324. doi:10.1177/0049124112452393

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370. doi:10.1007/BF02294361

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

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

Chakrabarti, A., & Ghosh, J. K. (2011). AIC, BIC, and recent advances in model selection. *Handbook of the Philosophy of Science*, *7*, 583–605. doi:10.1016/B978-0-444-51862-0.50018-6

Cudeck, R., & Brown, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavior 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. doi:10.1037/0033-2909.109.3.512

Dudley, R. M., & Haughton, D. (1997). Information criteria for multiple data sets and restricted parameters. *Statistica Sinica*, *7*, 265–284.

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). *Sensitivity and specificity of information criteria* (Tech. Rep. No. 12–119). University Park, PA: The Methodology Center, The Pennsylvania State University.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *3*, 272–299. doi:10.1037/1082-989X.4.3.272

Fox, J., Nie, Z., & Byrnes, J. (2013). Sem: Structural equation models. R package version 3.1-1. Retrieved from http://CRAN.R-project.org/packages/sem

Guo, B., Perron, B. E., & Gillespie, D. F. (2009). A systematic review of structural equation modeling in social work research. *Bristish Journal of Social Work*, *39*, 1556–1574. doi:10.1093/bjsw/bcn101

Haughton, D. M. A. (1988). On the choice of model to fit data from an exponential family. *The Annals of Statistics*, *16*, 342–355. doi:10.1214/aos/1176350709

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

Hershberger, S. L. (2003). The growth of structural equation modeling: 1994–2001. *Structural Equation Modeling*, *10*, 35–46. doi:10.1207/S15328007SEM1001_2

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

Jackson, D. L., Gillaspy, J. A., Jr., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*, 6–23. doi:10.1037/a0014694

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

Kashyap, R. L. (1982). Optimal choice of ar and ma parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *4*, 99–104.

Kass, R. E. (1993). Bayes factors in practice. *The Statistician*, *42*, 551–560. doi:10.2307/2348679

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

Kenny, D. A. (2008). Reflections on mediation. *Organizational Research Methods*, *11*, 353–358. doi:10.1177/1094428107308978

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

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

MacCallum, R. C. (1995). Model specification: Procedures, strategies, and related issues. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 16–36). Newbury Park, CA: Sage.

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

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201–224. doi:10.1146/annurev.psych.51.1.201

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science*, *5*, 675–686. doi:10.1177/1745691610388766

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82. doi:10.1037/1082-989X.7.1.64

Nishii, B. R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, *12*, 758–765. doi:10.1214/aos/1176346522

O'Boyle, E. H., Jr., & Williams, L. J. (2011). Decomposing model fit: Measurement vs. theory in organizational research using latent variables. *Journal of Applied Psychology*, *96*, 1–12. doi:10.1037/a0020539

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*, 287–312. doi:10.1207/S15328007SEM0802_7

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

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.

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

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333–343. doi:10.1007/BF02294360

Shah, R., & Goldstein, S. M. (2006). Use of structural equation modeling in operations management research: Looking back and forward. *Journal of Operations Management*, *24*, 148–169. doi:10.1016/j.jom.2005.05.001

Shao, J. (1997). An asymptotic theory for model selection. *Statistics Sincia*, *7*, 221–264.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, *63*, 117–126. doi:10.1093/biomet/63.1.117

Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics*, *35*, 415–423. doi:10.1007/BF02480998

Steiger, J. H., & Lind, J. M. (1980). *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*, 253–264. doi:10.1007/BF02294104

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*, 228–243. doi:10.1037/a0027127

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

Yang, C. C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics & Data Analysis*, *50*, 1090–1104. doi:10.1016/j.csda.2004.11.004

Yang, C. C., & Yang, C. C. (2007). Separating latent classes by information criteria. *Journal of Classification*, *24*, 183–203. doi:10.1007/s00357-007-0010-1