Journal of Modern Applied Statistical Methods

Volume 8 | Issue 1 Article 15

5-1-2009

The Comparison of Model Selection Criteria When Selecting Among Competing Hierarchical Linear Models

Tiffany A. Whittaker
The University of Texas at Austin, tiffany.whittaker@mail.utexas.edu

Carolyn F. Furlow
Georgia State University, carolyn.f.furlow@saic.com

Recommended Citation
Whittaker, Tiffany A. and Furlow, Carolyn F. (2009) "The Comparison of Model Selection Criteria When Selecting Among Competing Hierarchical Linear Models," *Journal of Modern Applied Statistical Methods*: Vol. 8 : Iss. 1 , Article 15.
DOI: 10.22237/jmasm/1241136840

Journal of Modern Applied Statistical Methods
May 2009, Vol. 8, No. 1, 173-193

1538 – 9472/09/$95.00

The Comparison of Model Selection Criteria When Selecting Among Competing Hierarchical Linear Models

Tiffany A. Whittaker
The University of Texas at Austin

Carolyn F. Furlow
Georgia State University

Little is known about the use and accuracy of model selection criteria when selecting among a set of competing multilevel models. The practices of applied researchers and the performance of five model selection criteria are examined when selecting the correct multilevel model using simulation techniques.

Key words: Hierarchical linear modeling, multilevel modeling, model selection criteria.

Introduction

Researchers are typically interested in comparing the fit of various theoretically plausible models to data. Hierarchical Linear Modeling (HLM), or multilevel modeling, has become a widely used tool to aid in the explanation of predictive theoretical models within the social and behavioral sciences. As is common with other statistical techniques (e.g., multiple linear regression, structural equation modeling), there exist various criteria for model comparison and selection within the HLM arena. Little is known, however, about the accuracy of various selection criteria within the HLM arena.

The purpose of this article is twofold: (1) to examine the current practices of researchers in the field when comparing and selecting hierarchical linear models; and (2) to examine the performance of various model selection techniques with respect to selecting the correct hierarchical linear model from a group of competing models.

Tiffany Whittaker is an Assistant Professor in the Department of Educational Psychology in the College of Education. Email: tiffany.whittaker@mail.utexas.edu. Carolyn Furlow is a statistician with the Science Applications International Corporation. Email: carolyn.f.furlow@saic.com.

Model Comparison and Selection Criteria in the HLM Arena

One method for the comparison of nested hierarchical linear models is the Chi-square difference test. The Chi-square difference test incorporates the deviance statistic in its calculation. The deviance statistic is given as:

$$-2[LL_{current\ model} - LL_{saturated\ model}] \quad (1)$$

where $LL_{current\ model}$ is the Log Likelihood (LL) value obtained from fitting the proposed model to the data. $LL_{saturated\ model}$ is the Log Likelihood value of fitting the best possible fitting model, the saturated model, to the data which results in a LL value of zero. Consequently, the deviance statistic reduces to $-2LL$.

The difference between two nested models' deviance statistics, which is asymptotically distributed as a Chi-square (χ^2) statistic, may be used to determine if a significant difference between the two models exists when adding or eliminating model parameters:

$$\chi^2_{-2LLdifference} = -2LL_{restricted} - 2LL_{unrestricted}, \quad (2)$$

where $-2LL_{restricted}$ is the deviance statistic for the nested, less parameterized (restricted) model and $-2LL_{untrestricted}$ is the deviance statistic for

the more parameterized, less restricted (unrestricted) model, with corresponding degrees of freedom equal to the difference in the number of parameters estimated (q) in each model:

$$df_{\text{-2LLdifference}} = q_{\text{restricted}} - q_{\text{unrestricted}}. \quad (3)$$

When the $\chi^2_{\text{-2LLdifference}}$ indicates a significant difference between two hierarchically related models, the nested model with less parameters has been oversimplified. That is, the less parameterized (nested) model has significantly decreased the overall fit of the model when compared to the model with more parameters. In this situation, then, the more parameterized model would be selected over the less parameterized model. On the other hand, when the $\chi^2_{\text{-2LLdifference}}$ test is not significant, the two models are comparable in terms of overall model fit. In this situation, the less parameterized would most likely be selected over the more parameterized model in support of parsimony.

When hierarchical linear models are non-nested, the $\chi^2_{\text{-2LLdifference}}$ test is an inappropriate method to assess significant model fit differences because neither of the two models can serve as a baseline comparison model. Still, there are instances in which different theoretical models posited to support the data are non-nested. In this situation, information criteria may be used for model comparison and selection. The benefit of using information criteria in the model selection process is that they may be used to compare and select among a set of nested and/or non-nested models.

The most popular information criterion is Akaike's (1973) information criterion (AIC) which compensates for the number of parameters in the model to encourage parsimony:

$$\text{AIC} = -2LL + 2q, \quad (4)$$

where -2LL is the deviance statistic for a given model and q is the number of parameters estimated in the given model. When comparing two competing models, the model with the lowest AIC value would be selected as the model demonstrating better fit than its comparison model.

The AIC is asymptotically efficient, meaning that it will select the best finite dimensional model (closest to the correct/true model) if the correct/true model is infinite dimensional. The AIC, however, has often been criticized for lack of consistency (Bozdogan, 1987; Hannon & Quinn, 1979; Hurvich & Tsai, 1989; Schwarz, 1978). Consistent model selection criteria select the correct/true model reliably (probabilities close to or at 1) when the correct/true model exists among the set of competing models. In addition, the AIC has been shown to incorrectly select more highly parameterized models, particularly when the ratio of estimated parameters to sample size is large (Hurvich & Tsai, 1989). Consequently, additional information criteria, which have extended the AIC to account for both model complexity and sample size, have been proposed.

Although various information criteria exist, this paper will focus on the information criteria readily available in current versions of SAS's PROC MIXED (version 9.2; SAS Institute Inc., 2007) and/or SPSS when using the Mixed Models command (version 16.0; SPSS Inc., 2007). SAS's PROC MIXED is a commonly used multilevel software program and with the recent addition of the Mixed Models command in SPSS, it too should become increasingly used when conducting multilevel analyses. Both software programs are able to provide more than one information criterion in the output. These include the Bayesian information criterion (BIC; Schwarz, 1978):

$$\text{BIC} = -2LL + \ln(N)q; \quad (5)$$

Hannon and Quinn's (1979) information criterion, *which is only available in SAS* (HQIC):

$$\text{HQIC} = \text{-}2LL + 2q\ln(\ln(N)); \quad (6)$$

and Bozdogan's (1987) consistent AIC (CAIC):

$$\text{CAIC}(k) = -2LL + [\ln(N) + 1]q; \quad (7)$$

where ln is the natural log and N is the sample size. While the BIC, HQIC, and CAIC were proposed to be more consistent model selection criteria, Hurvich and Tsai (1989) proposed a criterion that extends the AIC to correct for its tendency to overfit models (select highly parameterized models) which is still asymptotically efficient, called the finite sample corrected AIC (AICC):

$$\text{AICC} = -2LL + 2qN/(N-q-1). \quad (8)$$

There remains debate concerning which model selection feature is best (efficiency versus consistency). Some may argue that models are simply approximations of the truth and that researchers will never know if the true model exists in their set of competing models, supporting the use of efficient model selection criteria. Others, however, may argue that they are able to measure all relevant variables and thus have the correct model in their set of competing models, supporting the use of consistent model selection criteria.

The point of this article is not to argue in favor of either efficiency or consistency as it depends upon the context and the discipline (Burnham & Anderson, 2002; McQuarrie & Tsai, 1998; Shi & Tsai, 2002). Nonetheless, the current paper will assess the performance of these five model selection criteria (both efficient and consistent ones) in terms of selecting the correct multilevel model from among a set of competing incorrect models. This performance standard does support the definition of consistency; unfortunately, it is difficult to assess the performance of these model selection criteria otherwise.

To our knowledge, there is no study that has compared the performance of all five of these information criteria with respect to selecting the correct model among a set of competing models in the HLM arena. The most recent and relevant study was conducted by Gurka (2006) who examined the performance of the AIC, AICC, BIC, and CAIC in terms of selecting the correct multilevel growth curve model under various conditions, including different sample sizes, total variances, ICC values, model misspecification, criteria calculation, and estimation methods.

The model selection criteria were assessed in three different scenarios: 1) the ability to select the correct fixed effects given a compound symmetric covariance structure; 2) the ability to select the correct random effects given the fixed effects in the model; and 3) the ability to select the correct fixed and random effects in the model. Overall, the results indicated that the BIC and CAIC tended to outperform both the AIC and AICC. In addition, the AICC tended to outperform the AIC when selecting the correct model. None of the criteria performed well under the small sample size condition (with 25 cases at level-2 and 3 observations within each case). All four criteria performed well when selecting the correct random effects model (in more than 90% of the replications), regardless of total variance and ICC conditions. When selecting the correct fixed effects only and the correct fixed and random effects models, the criteria performed worse as the ICC values increased with the larger total variance conditions.

The impetus behind Gurka's (2006) study was the interest in comparing these criteria under different estimation methods available in multilevel software packages. The five model selection criteria presented in Equations 4 through 8 above are calculated under full information maximum likelihood (FIML) estimation as opposed to restricted maximum likelihood (REML) estimation in which the calculations change a bit with respect to N and q. When using FIML, the likelihood function contains both the fixed effects and the random effects (Raudenbush & Bryk, 2002). REML, however, rests on the assumption that fixed effects are uncertain and should be estimated separately from the random effects. It has been argued that deviance statistics, as well as the information criteria, of different models can be compared when the models differ only in their random effects under REML estimation while the deviance statistics, as well as the information criteria, of different models can be compared when the models differ in their fixed effects or their random effects under FIML estimation (Verbeke & Molenberghs, 2000).

Gurka (2006) questioned why the information criteria calculated under REML estimation could not be used in the model

selection process when comparing models containing fixed effects. As a result, he compared the performance of the four model selection criteria under FIML and REML estimation conditions. The findings indicated that the selection criteria performed better or equally well under REML estimation compared to FIML estimation when selecting the fixed effects model. As Gurka (2006) noted, the question as to whether the information criteria of fixed effects models may be compared under REML estimation should be examined further. In spite of this, the current paper does not examine this question. Instead, FIML estimation will solely be used as the models compared in the current paper differ with respect to their fixed and random effects.

SAS and SPSS differ with respect to the calculations of the BIC and the CAIC. More specifically, sample size in SAS is equal to the number of observations at level-2 (*m*) whereas sample size in SPSS is equal to the total number of observations at level-1 (*N*) when calculating the BIC and the CAIC under FIML estimation. The AICC, however, is calculated identically in both SAS and SPSS, using the total number of observations at level-1 (*N*) in the calculation. In cross-sectional designs, it seems reasonable to use the number of observations at level-1 as *N* in the calculation of these criteria.

In contrast, it seems more reasonable to use the number of observations at level-2 (*m*) in the calculation of these criteria in growth curve modeling designs. Additionally, Raudenbush and Liu (2000) reported that in their research on power with HLM designs, the sample size at level-2 was typically more important for power than the sample size at level-1. This research would also seem to indicate the utility of using *m* in the calculation of these criteria. Gurka (2006) also examined the performance of the model selection criteria (AICC, BIC, and CAIC) when using *N* versus *m* in their calculation. The results indicated that the criteria tended to perform better in terms of selecting the correct model when they were calculated using the number of observations at level-2 (*m*) as opposed to the number of observations at level-1 (*N*) under FIML estimation.

To summarize, there is no study, to our knowledge, that has examined all five model selection criteria (AIC, AICC, BIC, CAIC, and HQIC) simultaneously within the HLM arena. Gurka (2006) recently examined all of these criteria, with the exception of the HQIC. Hence, it is unknown how the HQIC will compare with the remaining model selection criteria examined in his study under different conditions. In addition, Gurka used a fairly simple correct model, both in terms of fixed and random effects, with only two predictors included in the model, and the single slope coefficient from level-1 was not allowed to randomly vary at level-2. This is unfortunate as researchers commonly allow slopes to vary randomly and the capability to model random slopes is a major advantage of multilevel modeling.

Researchers are also typically interested in examining more complex models that include more than just two predictors. Consequently, it is unclear how the model selection criteria will perform when comparing a set of simple models versus more complicated models. In addition, because Gurka was interested in how the criteria perform under a growth curve modeling context, the sample sizes used in his study were not reflective of those found in typical HLM designs where individuals are nested within groups. Thus, the purpose of this article is to examine the performance of all five model selection criteria in terms of selecting the correct multilevel model (with slopes allowed to randomly vary) under various conditions, including criteria calculation, model complexity, model misspecification, number of groups at level-2, number of participants per group, parameter magnitude, and ICCs.

Content Analysis

In order to evaluate the use of model selection criteria within the HLM arena, a content analysis was conducted. When conducting the content analysis, several different characteristics were assessed. More specifically, interest was placed on 1) the frequency with which model selection criteria are used by applied researchers when selecting among competing hierarchical models; 2) the types of model selection criteria used by applied researchers in the model comparison/selection process; and 3) if model selection criteria were

used, what multilevel software package was used when conducting the analyses.

Content Analysis Procedure

To assess these characteristics, a search in PsycInfo was conducted using the following search terms: "HLM," "Hierarchical Linear Modeling," "Multilevel Modeling," and "Random Effects Modeling." All applied articles using HLM techniques published between January 2002 and March 2007 were collected. Two hundred twenty articles were collected as a result of this search. These 220 articles were examined in order to collect information concerning the three characteristics mentioned above (see Table 1 for complete information on the content analysis characteristics).

Content Analysis Results

Of the 220 articles reviewed, the authors of 45 articles reported using some form of model selection criteria whereas the authors of 175 articles did not report using model selection criteria. The most commonly used model selection criteria was the Chi-square difference test followed by both the AIC and the BIC used together. Neither the AICC, the CAIC, nor the HQIC was used in any of these reviewed articles. The articles in which model selection criteria were used were also reviewed to determine what type of multilevel software package was used to conduct the analyses. The most popular software used was HLM (Raudenbush, Bryk, & Congdon, 2007) followed by MLwiN (Rasbash, Charlton, Browne, Healy, & Cameron, 2005) and SAS's PROC MIXED (SAS Institute Inc., 2003). LISREL (Jöreskog & Sörbom, 2006), MIXREG (Hedeker & Gibbons, 1999), and Mplus (Muthén & Muthén, 2007) were used less frequently. The authors of the remaining 20 articles did not report which software package was used. For a list of all the articles collected in this study, please contact the first author.

Methodology

Simulation Study

A Monte Carlo simulation study was conducted in order to examine the performance of five different model selection criteria when selecting the correct multilevel model from a group of competing multilevel models. The performance of these criteria was examined under varied conditions, including criteria calculation, model complexity, model misspecification, number of groups at level-2, number of participants per group, parameter magnitude, and the intraclass correlation (ICC) value.

Model Selection Criteria Calculation

The five model selection criteria (AIC, AICC, BIC, CAIC, HQIC) were examined under all conditions. To compare whether m or N is best in the calculation of the criteria (except the AIC as sample size is not used in its calculation), the AICC, BIC, CAIC, and HQIC were calculated in all conditions using the number of observations at level-2 (m; as calculated in SAS) and the number of observations at level-1 (N; as calculated in SPSS), resulting in the following nine model selection criteria: the AIC, the $AICC_m$, the $AICC_N$, the BIC_m, the BIC_N, the $CAIC_m$, the $CAIC_N$, the $HQIC_m$, and the $HQIC_N$. Although the Chi-square difference test was more commonly used by applied researchers, as demonstrated by the content analysis, a number of the misspecified models (described below) examined in this study were non-nested, rendering the Chi-square difference test ineffectual across all possible model comparisons. Therefore, the Chi-square difference test was not used as one of the model selection criteria.

Model Complexity

To examine whether the model selection criteria would perform differently when selecting among a simple set of multilevel models versus a more complex set of multilevel models, a simple generating model and a complex generating model were used. The simple generating model (Simple Model 1) consisted of a two-level model in which one predictor is included at both the participant-level (level-1) and the group-level (level-2) and is as follows:

$$
\begin{aligned}
&Level-1: \\
&Y_{ij} = \beta_{0j} + \beta_{1j} X1 + r_{ij} \\
&Level-2: \\
&\beta_{0j} = \gamma_{00} + \gamma_{01} W1 + u_{0j} \\
&\beta_{1j} = \gamma_{10} + u_{1j}
\end{aligned} \tag{9}
$$

In the simple generating model, the parameters from level-1 were all allowed to randomly vary at level-2. However, there was no cross-level interaction between X1 and W1. The variance/covariance matrix at level-2 associated with this model is:

$$
Var\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} = \begin{pmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{pmatrix}. \tag{10}
$$

The complex generating model (Complex Model 1) consisted of the same variance/covariance structure as the simple model but included a more complex fixed effects structure in which two predictors were included at both the participant-level (level-1) and the group-level (level-2):

$$
\begin{aligned}
&Level-1: \\
&Y_{ij} = \beta_{0j} + \beta_{1j} X1 + \beta_{2j} X2 + r_{ij} \\
&Level-2: \\
&\beta_{0j} = \gamma_{00} + \gamma_{01} W1 + \gamma_{02} W2 + u_{0j} \\
&\beta_{1j} = \gamma_{10} + \gamma_{11} W1 + \gamma_{12} W2 + u_{1j} \\
&\beta_{2j} = \gamma_{20} + \gamma_{21} W1 + \gamma_{22} W2
\end{aligned} \tag{11}
$$

While the simple model did not include a cross-level interaction, there were four cross-level interaction terms estimated in the complex model.

Model Misspecification

The simple and complex hierarchical linear model sets consisted of eight different nested and non-nested models, including the correct simple and complex generating model, respectively. The models examined were misspecified by incorrectly adding a parameter, incorrectly removing a parameter, or incorrectly adding and removing a parameter from the correct model. For the simple model set, the seven models were misspecified as follows:

a) by including a cross-level interaction between X1 and W1, γ_{02} (Simple Model 2);
b) by dropping X1 from the model which results in the loss of the level-2 equation for the prediction of β_{1j} (Simple Model 3);
c) by dropping W1, γ_{01}, from the model (Simple Model 4);
d) by dropping u_{1j} from the model, thus the corresponding variance, τ_{11}, and covariance, τ_{10} were not estimated (Simple Model 5);
e) by dropping u_{0j} from the model, thus the corresponding variance, τ_{00}, and covariance, τ_{10} were not estimated (Simple Model 6);
f) by dropping u_{1j} from the model and including the cross-level interaction between X1 and W1, γ_{02} (Simple Model 7); and
g) by dropping u_{0j} from the model and including the cross-level interaction between X1 and W1, γ_{02} (Simple Model 8).

Of all of these misspecified models, Model 2 is the more parameterized, incorrect nested model, Models 3 – 6 are less parameterized, incorrect nested models, and Models 7 – 8 are non-nested, incorrect models.

For the complex model set, the seven models were misspecified as follows:

a) by including u_{2j}, thus estimating the corresponding variance, τ_{22}, and covariance, τ_{20} (Complex Model 2);
b) by including an interaction between X1 and X2 that was fixed at level-2, γ_{30} (Complex Model 3);

Table 1: Characteristics of the Applied HLM Articles Reviewed
(January 2002 – March 2007)

Characteristic	Frequency
Reported Use of Model Selection Criteria	45
Model Selection Criteria Used	
Chi-Square Difference Test	35
AIC	2
BIC	1
AIC with BIC	3
Chi-Square Difference Test with AIC	2
Chi-Square Difference Test with BIC	1
Chi-Square Difference Test with AIC & BIC	1
HLM Software Used	
HLM	10
MLwiN	7
SAS PROC MIXED	4
LISREL	2
MIXREG	1
Mplus	1
Did Not Specify	20
Did Not Report Use of Model Selection Criteria	175

c) by including an interaction between W1 and W2 in the intercept equation, γ_{03} (Complex Model 4);
d) by dropping u_{1j}, from the model, thus the corresponding variance, τ_{11}, and covariance, τ_{10}, were not estimated (Complex Model 5);
e) by dropping the cross-level interaction between X1 and W2, γ_{12} (Complex Model 6);
f) by dropping u_{1j} from the model and including u_{2j} (Complex Model 7); and
g) by dropping W2 from the intercept equation, γ_{02}, and including u_{2j} (Complex Model 8).

Of all of these misspecified models, Models 2 – 4 are more parameterized, incorrect nested models, Models 5 – 6 are less parameterized, incorrect nested models, and Models 7 – 8 are non-nested, incorrect models.

Number of Groups at Level-2 and Participants per Group

The number of groups modeled at level-2 was varied to be either 20 or 40 to represent small to moderate sizes. Within each group, the sample size was varied to be either 15 or 30 participants to represent fairly small to moderate to large total sample sizes (300, 600, and 1,200, respectively).

Parameter Magnitude

The magnitude of all of the slope coefficients was varied to equal .5 or .7 to represent moderate to large magnitudes. The overall intercept (γ_{00}) remained constant at a value of 1 and the intercept values for the slope

equations (γ_{10} and γ_{20} in the complex model and only γ_{10} in the simple model) remained constant at a value of .5.

Intraclass Correlation (ICC) Value

The conditional intraclass correlation (ICC), which represents the proportion of the residual variance between groups remaining after including explanatory variables, was varied to equal either .1 or .3. The level-1 residual variance was set to equal .5. The level-2 variance components, τ_{00} and τ_{11}, were set to be equal to one another and their values were dictated by the ICC and the level-1 variance. This resulted in level-2 variances equal to 0.055555556 with an ICC of .1 and 0.214285714 with an ICC of .3. The level-2 covariance term, τ_{01}, was assumed to be equal to 0.

Simulation Study Procedure

SAS (version 9.1) was used to generate raw data according to the correct simple and complex generating models (see Equations 9 and 11) under the 16 combinations of different number of groups, participants per group, parameter magnitude, and ICC value conditions, resulting in 32 conditions. For each of the 32 conditions, 1,000 sets of raw data were generated. Each variable was generated to be standard normal. Once each data set was generated, all eight models (one correct and seven misspecified) were fit to the data using full information maximum likelihood (FIML) estimation in SAS's PROC MIXED procedure. The nine model selection criteria under examination were calculated for each of the models. The number of times each criteria selected each of the models was then documented.

Results

The selection rates of the nine criteria are presented in Tables 2 – 9. The simple model selection rates are presented in Tables 2 – 5 and the complex model selection rates are presented in Tables 6 – 9. None of the criteria performed well in the smallest total sample size (20 groups x 15 participants per group = 300) and low ICC value (.1) conditions, regardless of parameter magnitude (see Tables 2 and 6). Overall, however, the accuracy of the selection criteria with respect to selecting the correct hierarchical linear model tended to increase as total sample size and ICC values increased. Further, the criteria generally performed better when selecting the correct model from the simple multilevel model set than when selecting the correct model from the complex multilevel model set.

Parameter magnitude did not have an effect on all of the selection criteria in all of the conditions. In general, the criteria tended to perform similarly in both low and high parameter magnitude conditions. Still, it did have an effect on the performance of the $AICC_m$ in two conditions. More specifically, the $AICC_m$ selected the correct model more frequently in the high parameter condition when group size was equal to 20 and the ICC value was high in the complex model set (see Tables 6 – 7).

The AIC and the $AICC_N$ were the least accurate selection criteria. These criteria never correctly selected the Simple or Complex Model 1 in more than 84% or 62% of the replications in any one condition, respectively. When the AIC or the $AICC_N$ did not select the correct multilevel model, they tended to select the more parameterized, misspecified models.

The next least accurate criterion was the $HQIC_m$, which never selected the Simple or Complex Model 1 in more than 89% or 73% of the replications in a condition, respectively. The $HQIC_N$ outperformed its *m*-calculated counterpart in all but four conditions (see Tables 2 and 4). Still, while the $HQIC_N$ selected the Simple Model 1 in more than 90% of the replications in more than half of the conditions, it never selected the Complex Model 1 in more than 90% of the replications in any condition. When the $HQIC_m$ or $HQIC_N$ did not select the correct multilevel model, they tended to incorrectly select the more parameterized, misspecified models.

The next least accurate criterion was the BIC_m. It correctly selected Simple Model 1 in more than 90% of the replications in half of the conditions but never correctly selected Complex Model 1 in more than 90% of the replications in

Table 2: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 300 Total Participants (20 Groups X 15 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Simple Model Set

	AIC	$AICC_m$	$AICC_N$	BIC_m	BIC_N	$CAIC_m$	$CAIC_N$	$HQIC_m$	$HQIC_N$
Parameter Magnitude = .5, ICC = .1									
M1	57.8	31.1	57.8	50.0	20.1	37.4	14.4	57.3	43.3
M2	11.8	0.0	10.8	3.7	0.0	0.9	0.0	9.9	1.8
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	10.5	27.3	11.0	17.4	33.7	24.1	36.9	11.4	21.5
M6	10.4	32.6	10.9	18.6	39.6	28.5	43.7	11.7	23.7
M7	7.0	7.6	7.0	8.4	5.9	7.7	4.7	7.2	8.2
M8	2.5	1.4	2.5	1.9	0.7	1.4	0.3	2.5	1.5
Parameter Magnitude = .5, ICC = .3									
M1	81.5	92.0	82.8	88.4	89.9	91.1	85.8	83.7	90.3
M2	17.6	1.9	16.2	9.5	1.9	5.3	0.7	15.2	6.8
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.9	4.9	1.0	1.7	6.2	2.8	9.8	1.1	2.4
M5	0.0	0.3	0.0	0.0	0.7	0.1	1.1	0.0	0.0
M6	0.0	0.4	0.0	0.2	0.9	0.4	1.9	0.0	0.2
M7	0.0	0.3	0.0	0.1	0.3	0.2	0.6	0.0	0.2
M8	0.0	0.2	0.0	0.1	0.1	0.1	0.1	0.0	0.1
Parameter Magnitude = .7, ICC = .1									
M1	59.3	30.9	59.2	49.5	18.4	35.7	12.3	58.0	42.9
M2	10.8	0.2	9.8	3.9	0.2	1.3	0.1	9.2	2.3
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	10.9	28.1	11.2	17.5	35.6	25.7	39.9	12.0	21.4
M6	10.5	31.8	10.9	19.2	39.7	28.7	43.2	11.9	24.2
M7	6.3	7.8	6.6	8.1	5.4	7.3	4.0	6.8	7.8
M8	2.2	1.2	2.3	1.8	0.7	1.3	0.5	2.1	1.4
Parameter Magnitude = .7, ICC = .3									
M1	82.8	97.3	83.6	90.1	96.2	94.1	94.9	83.9	92.3
M2	17.0	1.4	16.2	9.4	1.4	5.1	0.8	15.7	7.0
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.5	0.0	0.1	0.8	0.3	1.3	0.0	0.3
M5	0.1	0.3	0.1	0.2	0.9	0.3	1.6	0.1	0.2
M6	0.0	0.1	0.0	0.0	0.3	0.0	1.1	0.0	0.0
M7	0.1	0.4	0.1	0.2	0.4	0.2	0.3	0.3	0.2
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 is the more parameterized, incorrect nested model. M3 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

Table 3: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 600 Total Participants (20 Groups X 30 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Simple Model Set

	AIC	$AICC_m$	$AICC_N$	BIC_m	BIC_N	$CAIC_m$	$CAIC_N$	$HQIC_m$	$HQIC_N$
Parameter Magnitude = .5, ICC = .1									
M1	79.8	82.1	80.2	83.2	70.9	82.6	64.6	80.9	83.1
M2	16.5	1.3	16.1	8.2	0.6	4.0	0.3	14.8	4.8
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.1	0.0	0.0	0.1	0.1	0.2	0.0	0.0
M5	0.8	4.0	0.8	2.2	8.8	3.3	11.4	1.1	3.1
M6	1.5	9.0	1.5	4.3	16.4	7.5	20.4	1.9	6.5
M7	1.2	2.9	1.2	1.8	2.8	2.3	2.8	1.2	2.3
M8	0.2	0.6	0.2	0.3	0.4	0.2	0.3	0.1	0.2
Parameter Magnitude = .5, ICC = .3									
M1	81.7	93.5	82.2	88.5	91.3	91.4	87.7	83.4	90.7
M2	17.7	2.5	17.2	10.1	1.5	5.8	1.0	16.0	6.7
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.6	4.0	0.6	1.4	7.0	2.8	11.1	0.6	2.6
M5	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .1									
M1	78.0	82.3	78.6	81.5	70.7	82.9	65.2	79.4	82.8
M2	18.0	1.6	17.3	10.7	1.0	4.8	0.5	16.0	6.1
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	1.1	4.9	1.1	2.5	10.8	3.9	14.1	1.4	3.4
M6	1.1	7.4	1.1	3.0	14.3	5.1	17.6	1.5	4.6
M7	1.3	3.2	1.4	1.9	2.8	2.7	2.4	1.3	2.5
M8	0.5	0.6	0.5	0.4	0.4	0.6	0.2	0.4	0.6
Parameter Magnitude = .7, ICC = .3									
M1	82.6	97.9	83.2	91.4	97.9	94.9	97.7	84.3	94.0
M2	17.4	2.0	16.8	8.6	1.5	5.1	1.1	15.7	6.0
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.1	0.0	0.0	0.6	0.0	1.2	0.0	0.0
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 is the more parameterized, incorrect nested model. M3 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

Table 4: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 600 Total Participants (40 Groups X 15 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Simple Model Set

	AIC	$AICC_m$	$AICC_N$	BIC_m	BIC_N	$CAIC_m$	$CAIC_N$	$HQIC_m$	$HQIC_N$
Parameter Magnitude = .5, ICC = .1									
M1	79.8	83.0	80.4	78.9	59.0	73.1	51.2	81.0	78.9
M2	16.0	6.6	15.3	4.2	1.1	2.2	0.4	10.3	4.0
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	1.7	3.8	1.7	6.1	17.4	10.3	21.4	3.4	6.2
M6	1.2	4.1	1.2	7.9	20.7	11.4	25.2	3.0	7.9
M7	0.8	1.8	0.9	2.2	1.5	2.4	1.5	1.4	2.3
M8	0.5	0.7	0.5	0.7	0.3	0.6	0.3	0.9	0.7
Parameter Magnitude = .5, ICC = .3									
M1	81.9	91.6	82.5	94.0	98.8	96.7	99.1	88.1	94.0
M2	18.1	8.4	17.5	6.0	1.2	3.3	0.7	11.9	6.0
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .1									
M1	80.1	84.0	80.5	82.0	61.0	76.9	51.4	82.2	82.0
M2	16.5	8.8	15.7	5.1	0.4	2.0	0.1	11.7	4.9
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	1.2	2.9	1.2	5.6	15.5	8.1	20.4	2.7	5.6
M6	1.3	2.9	1.5	6.1	21.2	11.1	26.3	2.4	6.3
M7	0.6	1.1	0.8	1.2	1.6	1.5	1.7	0.8	1.2
M8	0.3	0.3	0.3	0.0	0.3	0.4	0.1	0.2	0.0
Parameter Magnitude = .7, ICC = .3									
M1	82.0	90.5	82.4	93.2	98.1	95.7	98.6	87.0	93.2
M2	18.0	9.5	17.6	6.8	1.9	4.3	1.4	13.0	6.8
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 is the more parameterized, incorrect nested model. M3 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

Table 5: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 1200 Total Participants (40 Groups X 30 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Simple Model Set

	AIC	$AICC_m$	$AICC_N$	BIC_m	BIC_N	$CAIC_m$	$CAIC_N$	$HQIC_m$	$HQIC_N$
Parameter Magnitude = .5, ICC = .1									
M1	83.5	91.4	83.8	93.8	97.9	96.1	97.7	88.4	94.7
M2	16.4	8.4	16.1	5.8	0.8	3.3	0.4	11.4	4.9
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	0.0	0.0	0.0	0.1	0.4	0.1	0.5	0.0	0.1
M6	0.1	0.1	0.1	0.2	0.6	0.3	1.1	0.1	0.2
M7	0.0	0.1	0.0	0.1	0.3	0.2	0.3	0.1	0.1
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .5, ICC = .3									
M1	83.2	91.5	83.6	93.6	98.9	96.2	99.3	88.1	94.2
M2	16.8	8.5	16.4	6.4	1.0	3.8	0.5	11.9	5.8
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .1									
M1	82.3	89.9	82.5	92.8	98.2	95.1	97.8	87.0	93.5
M2	17.5	9.9	17.3	6.9	1.0	4.5	0.5	12.8	6.2
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	0.1	0.1	0.1	0.1	0.3	0.1	0.4	0.1	0.1
M6	0.1	0.1	0.1	0.2	0.4	0.2	1.2	0.1	0.2
M7	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .3									
M1	82.7	90.2	83.0	92.1	98.8	95.3	99.2	87.3	92.9
M2	17.3	9.8	17.0	7.9	1.2	4.7	0.8	12.7	7.1
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 is the more parameterized, incorrect nested model. M3 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

Table 6: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 300 Total Participants (20 Groups X 15 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Complex Model Set

	AIC	$AICC_m$	$AICC_N$	BIC_m	BIC_N	$CAIC_m$	$CAIC_N$	$HQIC_m$	$HQIC_N$
Parameter Magnitude = .5, ICC = .1									
M1	41.8	1.0	45.0	45.9	30.1	43.3	24.3	44.6	45.7
M2	8.9	0.0	6.2	2.2	0.2	0.5	0.0	6.4	0.7
M3	11.0	0.0	9.3	5.9	1.0	2.6	0.6	9.4	3.8
M4	19.8	0.0	18.2	12.4	3.2	7.5	2.0	18.2	9.5
M5	14.9	96.9	18.0	31.6	65.0	44.5	72.8	18.1	38.6
M6	0.0	1.9	0.0	0.0	0.1	0.0	0.1	0.0	0.0
M7	3.6	0.2	3.3	2.0	0.4	1.6	0.2	3.3	1.7
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .5, ICC = .3									
M1	56.1	16.8	60.0	72.9	84.1	80.3	82.1	59.7	78.1
M2	9.5	0.0	7.8	4.1	0.2	1.4	0.0	8.0	2.5
M3	14.3	0.0	12.8	7.6	1.7	5.2	1.1	12.8	6.1
M4	18.4	0.0	17.4	12.5	3.6	7.8	2.4	17.5	9.5
M5	0.1	21.1	0.2	0.6	2.3	0.8	3.6	0.2	0.6
M6	1.4	62.1	1.6	2.2	8.1	4.4	10.8	1.6	3.1
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.2	0.0	0.2	0.1	0.0	0.1	0.0	0.2	0.1
Parameter Magnitude = .7, ICC = .1									
M1	40.8	2.0	43.5	45.8	35.9	43.5	30.2	43.3	45.4
M2	8.2	0.0	6.6	2.0	0.1	0.4	0.0	6.7	0.8
M3	10.9	0.0	9.4	5.4	0.5	2.7	0.5	9.4	4.0
M4	19.7	0.2	18.0	11.5	3.0	7.5	2.0	18.1	9.1
M5	17.6	97.4	19.9	33.1	60.1	44.3	67.3	20.0	38.9
M6	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	2.8	0.1	2.6	2.2	0.4	1.6	0.0	2.5	1.8
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .3									
M1	55.6	48.0	60.4	72.5	90.7	83.0	91.9	60.0	77.7
M2	9.9	0.0	7.5	2.6	0.1	0.7	0.0	7.7	1.2
M3	14.3	0.0	13.3	10.0	2.6	5.6	1.6	13.3	8.0
M4	19.9	0.0	18.5	14.4	4.4	9.8	3.1	18.6	12.3
M5	0.2	22.4	0.2	0.4	1.6	0.6	2.2	0.3	0.6
M6	0.1	29.6	0.1	0.1	0.5	0.3	1.1	0.1	0.2
M7	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 – M4 are more parameterized, incorrect nested models. M5 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

any one condition. When the BIC_m did not select the correct model, it tended to select both the less parameterized and more parameterized, misspecified models, depending upon the condition. More specifically, it tended to select the less parameterized, misspecified models when ICC values were low with smaller total sample sizes and the more parameterized, misspecified models when ICC values were high with larger total sample sizes.

Interestingly, the $AICC_m$ tended to outperform the $AICC_N$ in all but two conditions in the simple model set, but only in approximately half of the conditions in the complex model set. More specifically, the $AICC_m$ did not outperform the $AICC_N$ until total sample size reached a moderate size (20 groups x 30 participants per group = 600), high parameter magnitude (.7), and high ICC value (.3) (see Tables 7 – 9). The $AICC_m$ correctly selected Simple Model 1 in more than 90% of the replications in more than half of the conditions and correctly selected Complex Model 1 in more than 90% of the replications in a little less than half of the conditions. When the $AICC_m$ did not select the correct model, it tended to incorrectly select the less parameterized, misspecified models.

The $CAIC_m$ performed fairly comparably to the AICC*m*, selecting the Simple Model 1 in more than 90% of the replications in a little more than half of the conditions but correctly selected Complex Model 1 in more than 90% of the replications in a little less than half of the conditions. When the $CAIC_m$ did not select the correct model, it tended to select the more parameterized and less parameterized, misspecified models depending upon the condition. For example, it tended to select the less parameterized, misspecified models when ICC values were low with smaller sample sizes and the more parameterized, misspecified models when ICC values were high with larger sample sizes.

The BIC_N and $CAIC_N$ performed the most accurately and fairly similarly. While the BIC_N did perform slightly better than the $CAIC_N$, these differences were generally small. The BIC_N correctly selected Simple Model 1 in more than 90% of the replications in a little more than half of the conditions and correctly selected Complex Model 1 in more than 90% of the replications in half of the conditions. The BIC_N outperformed its *m*-calculated counterpart in more than half of the conditions. Nonetheless, when the BIC_m outperformed the BIC_N in the remaining conditions, the ICC value was low (see Tables 2 – 4, 6, and 8). When the BIC_N did not select the correct model, it generally tended to incorrectly select the less parameterized models.

The $CAIC_N$ correctly selected Simple Model 1 in more than 90% of the replications in half of the conditions and correctly selected Complex Model 1 in more than 90% of the replications in a little more than half of the conditions. The $CAIC_N$ outperformed the $CAIC_m$ in a little more than half of the conditions. Similar to the BIC, when the $CAIC_m$ outperformed the $CAIC_N$ in the remaining conditions, the ICC value tended to be low (see Tables 2 – 4, and 6 - 8), with the exception of two conditions (see Tables 2 and 3). When the $CAIC_N$ did not select the correct model, it tended to incorrectly select the less parameterized, misspecified models. It should be mentioned that when the *m*-calculated BIC and CAIC outperformed their *N*-calculated counterparts, the differences were quite large, particularly within the Simple Model set. In contrast, when the *N*-calculated BIC and CAIC outperformed their *m*-calculated counterparts, the differences were not as large.

It must be noted that the results presented are based on 1,000 replications in which all of the eight simple and complex models did not encounter any estimation problems. Hence, replications in which any model encountered a problem involving a non-positive definite variance component matrix or a convergence problem were discarded.

Additional replications were conducted until 1,000 replications in which problems did not exist were reached (see Table 10 for a summary of replications needed and percentage of usable replications in each generating condition). Less estimation problems were encountered when running the simple models than when running the complex models. Overall, fewer problems were encountered as total sample size and ICC values increased. Non-positive definite covariance matrix and

Table 7: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 600 Total Participants (20 Groups X 30 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Complex Model Set

	AIC	$AICC_m$	$AICC_N$	BIC_m	BIC_N	$CAIC_m$	$CAIC_N$	$HQIC_m$	$HQIC_N$
Parameter Magnitude = .5, ICC = .1									
M1	54.8	27.3	57.1	70.8	75.4	77.8	72.3	58.8	76.3
M2	12.2	0.0	10.9	4.8	0.3	2.0	0.0	9.8	2.7
M3	12.1	0.0	11.7	6.9	0.9	3.7	0.2	11.0	4.5
M4	18.7	0.0	17.9	12.3	2.4	7.7	1.5	17.7	8.9
M5	2.0	68.5	2.2	4.9	20.9	8.5	25.9	2.5	7.3
M6	0.0	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.2	0.0	0.2	0.3	0.1	0.3	0.1	0.2	0.3
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .5, ICC = .3									
M1	53.4	33.1	55.5	72.2	88.8	82.3	86.9	57.5	80.1
M2	14.4	0.0	13.3	4.6	0.4	2.0	0.0	12.0	2.6
M3	13.1	0.0	12.6	8.4	1.2	5.2	0.9	12.2	5.6
M4	18.4	0.0	17.8	12.6	2.0	7.8	1.4	17.3	9.1
M5	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.7	66.3	0.8	2.0	7.6	2.5	10.8	1.0	2.4
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.2	0.0	0.2	0.0	0.0	0.2
Parameter Magnitude = .7, ICC = .1									
M1	51.4	33.3	53.5	68.0	75.9	76.4	72.6	55.4	75.6
M2	12.8	0.0	11.8	5.7	0.1	1.5	0.0	11.2	2.1
M3	12.7	0.0	11.9	7.8	0.8	4.4	0.4	11.2	5.1
M4	20.1	0.0	19.7	13.7	2.2	8.4	1.2	19.0	9.6
M5	2.2	66.4	2.3	4.0	20.6	8.5	25.7	2.4	6.6
M6	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.8	0.0	0.8	0.8	0.4	0.8	0.1	0.8	1.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .3									
M1	53.4	73.6	56.7	72.6	95.3	84.6	96.6	58.7	81.4
M2	13.6	0.0	12.1	5.2	0.4	2.0	0.1	11.0	2.6
M3	12.5	0.0	11.7	7.7	0.8	3.4	0.2	11.2	4.7
M4	20.5	0.0	19.5	14.5	2.9	9.9	1.9	19.1	11.2
M5	0.0	1.5	0.0	0.0	0.0	0.0	0.1	0.0	0.0
M6	0.0	24.9	0.0	0.0	0.6	0.1	1.1	0.0	0.1
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 – M4 are more parameterized, incorrect nested models. M5 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

Table 8: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 600 Total Participants (40 Groups X 15 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Complex Model Set

	AIC	$AICC_m$	$AICC_N$	BIC_m	BIC_N	$CAIC_m$	$CAIC_N$	$HQIC_m$	$HQIC_N$
Parameter Magnitude = .5, ICC = .1									
M1	56.3	79.1	57.8	76.5	70.7	75.7	63.1	67.6	76.7
M2	12.1	0.6	11.5	2.4	0.1	0.9	0.0	6.5	2.4
M3	12.9	1.7	12.5	3.4	0.4	1.7	0.3	8.6	3.3
M4	15.4	3.2	14.6	5.7	1.1	3.5	0.7	11.0	5.6
M5	2.1	14.4	2.2	10.6	26.9	17.3	35.3	5.0	10.6
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	1.2	1.0	1.4	1.4	0.8	0.9	0.6	1.3	1.4
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .5, ICC = .3									
M1	55.9	92.4	57.9	83.5	97.2	91.5	98.4	69.2	83.8
M2	13.2	0.8	12.1	3.8	0.2	1.4	0.0	7.6	3.6
M3	14.8	3.1	14.2	5.1	1.0	3.2	0.5	10.9	5.1
M4	16.1	3.7	15.8	7.6	1.5	3.9	0.8	12.3	7.5
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.1	0.0	0.3	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .1									
M1	57.0	81.1	59.5	77.7	71.0	77.6	64.2	67.5	77.7
M2	12.0	0.4	10.8	1.6	0.0	0.7	0.0	6.8	1.6
M3	13.3	3.1	12.6	5.0	1.4	3.1	0.8	9.6	4.9
M4	14.7	2.4	14.0	5.5	1.0	2.5	0.7	11.0	5.5
M5	2.6	12.2	2.6	9.7	26.4	15.8	34.1	4.4	9.8
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.4	0.8	0.5	0.5	0.2	0.3	0.2	0.7	0.5
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .3									
M1	56.1	91.0	58.8	83.9	96.7	90.2	98.3	70.0	83.9
M2	15.5	0.9	13.8	3.0	0.1	1.1	0.0	8.4	3.0
M3	11.8	3.1	11.3	5.2	1.6	3.1	0.8	8.6	5.2
M4	16.6	5.0	16.1	7.9	1.6	5.6	0.9	13.0	7.9
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 – M4 are more parameterized, incorrect nested models. M5 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

Table 9: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 1,200 Total Participants (40 Groups X 30 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Complex Model Set

	AIC	$AICC_m$	$AICC_N$	BIC_m	BIC_N	$CAIC_m$	$CAIC_N$	$HQIC_m$	$HQIC_N$
Parameter Magnitude = .5, ICC = .1									
M1	59.4	92.1	60.3	85.4	97.6	91.7	97.8	71.4	86.8
M2	13.2	0.8	12.8	3.2	0.0	1.0	0.0	7.9	2.9
M3	12.3	3.0	12.1	4.7	0.6	3.0	0.3	9.6	4.1
M4	15.1	3.9	14.8	6.6	1.1	4.1	0.4	11.1	6.0
M5	0.0	0.2	0.0	0.1	0.7	0.2	1.5	0.0	0.2
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .5, ICC = .3									
M1	58.6	92.1	59.6	85.3	97.2	91.6	97.7	71.8	87.2
M2	13.4	0.9	13.1	2.6	0.2	1.1	0.1	7.5	2.1
M3	13.0	2.7	12.4	5.3	0.7	2.9	0.6	9.3	4.9
M4	15.0	4.3	14.9	6.8	1.4	4.3	0.9	11.4	5.8
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.5	0.1	0.7	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .1									
M1	55.3	91.2	56.6	82.7	96.6	90.1	96.7	68.5	84.7
M2	14.7	1.3	14.2	3.7	0.1	1.7	0.0	9.0	3.3
M3	13.0	2.8	12.8	5.9	0.9	3.2	0.3	9.7	5.1
M4	17.0	4.4	16.4	7.5	1.2	4.7	1.1	12.6	6.7
M5	0.0	0.3	0.0	0.2	1.2	0.3	1.9	0.1	0.2
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .3									
M1	59.3	93.7	61.1	87.6	97.6	92.9	98.7	72.7	89.2
M2	12.5	0.5	11.7	1.8	0.2	1.0	0.0	6.6	1.4
M3	14.5	2.3	13.9	5.0	0.6	2.5	0.3	10.0	4.2
M4	13.7	3.5	13.3	5.6	1.6	3.6	1.0	10.7	5.2
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 – M4 are more parameterized, incorrect nested models. M5 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

convergence problems could both be encountered for different models within the same replication. Convergence problems were encountered more frequently when trying to fit Complex Model 2 (which incorrectly included a random effect), Complex Model 7 (which was missing a random effect and incorrectly included a random effect), and Complex Model 8 (which was missing a fixed effect and incorrectly included a random effect).

Conclusion

The current study examined the performance of the AIC, AICC, BIC, CAIC, and the HQIC when selecting the correct multilevel model under different criteria calculation, model complexity, model misspecification, number of groups at level-2, number of participants per group, parameter magnitude, and intraclass correlation (ICC) value conditions. Several of the study factors, either in isolation or in conjunction with another factor or factors, did affect the performance of the model selection criteria. For example, none of the model selection criteria performed well with respect to selecting the correct model when total sample size and ICC values were small and the performance of the model selection criteria improved as total sample size and ICC value increased.

The criteria generally performed more accurately when selecting the correct model from the simple model set than the complex model set. This seems reasonable given that adding or dropping parameters from a simple model would result in a more highly misspecified model than when adding or dropping parameters from a complex model. For example, dropping a random slope from a model in which there is only 1 random slope would result in a more highly misspecified model than dropping a random slope from a model in which there are 2 or more random slopes. Thus, the criteria would be more likely to select the correct model from among a set of severely misspecified models in the simple model set than a set of moderately misspecified models in the complex model set.

Although parameter magnitude did not appear to have a great impact on the performance of the model selection criteria, it did impact the $AICC_m$ in two conditions. That is, the $AICC_m$ performed more accurately in the high parameter magnitude condition when group size was 20 with a high ICC value in the complex model set. Again, this appears to be an isolated occurrence as parameter magnitude did not generally affect the remaining criteria.

The efficient model selection criteria, the AIC and the $AICC_N$, did not perform as well as the remaining, consistent criteria. This is to be expected when the definition of the performance standard, such as the one used in this study, is consistency (i.e., the selection of the correct model from among a set of competing models). These results corroborate the findings in Gurka's (2006) study.

To date, the HQIC, to the best of our knowledge, has not been examined in the relevant literature. Thus, the performance of the HQIC under various conditions and in comparison to the remaining criteria was of interest in the current study. The results indicated that while the HQIC performed more accurately than the AIC and the $AICC_N$, it did not perform more accurately than the BIC, CAIC, or the $AICC_m$ when selecting the correct model.

The $AICC_m$ proved to be a contender, not only outperforming its *N*-calculated counterpart in almost all conditions, but also performing comparably to the $CAIC_m$, next to the most accurately performing criteria (BIC_N and $CAIC_N$). Gurka (2006) also found that the $AICC_m$ performed adequately. Gurka (2006) recommended the use of the BIC_m and the $CAIC_m$ based on his findings, however, the BIC_N and $CAIC_N$ outperformed their *m*-calculated counterparts in several conditions in the current study. When the BIC_m and the $CAIC_m$ did outperform their *N*-calculated counterparts, the ICC value was low. Also, the differences in the rates of choosing the correct model were appreciably higher for the *m*-calculated criteria in these conditions, particularly within the Simple Model set.

The results of the current study did not determine which one model selection criterion will perform optimally in every situation encountered. It is clear, however, that the BIC, the CAIC, as well as the AICC*m*, generally out-

Table 10: Non-Positive Definite Variance Component Matrix and Convergence Problems Encountered as a Function of Generating Condition

	Simple		Complex	
Condition	Replications Needed	% Usable Replications	Replications Needed	% Usable Replications
20 x 15; Parameter = .5; ICC = .1	1110	90.1	1733	57.7
20 x 15; Parameter = .5; ICC = .3	1000	100.0	1318	75.9
20 x 15; Parameter = .7; ICC = .1	1118	89.4	1794	55.7
20 x 15; Parameter = .7; ICC = .3	1002	99.8	1397	71.6
20 x 30; Parameter = .5; ICC = .1	1018	98.2	1184	84.5
20 x 30; Parameter = .5; ICC = .3	1000	100.0	1096	91.2
20 x 30; Parameter = .7; ICC = .1	1009	99.1	1185	84.4
20 x 30; Parameter = .7; ICC = .3	1000	100.0	1117	89.5
40 x 15; Parameter = .5; ICC = .1	1006	99.4	1072	93.3
40 x 15; Parameter = .5; ICC = .3	1000	100.0	1019	98.1
40 x 15; Parameter = .7; ICC = .1	1010	99.0	1096	91.2
40 x 15; Parameter = .7; ICC = .3	1000	100.0	1035	96.6
40 x 30; Parameter = .5; ICC = .1	1000	100.0	1012	98.8
40 x 30; Parameter = .5; ICC = .3	1000	100.0	1001	99.9
40 x 30; Parameter = .7; ICC = .1	1000	100.0	1013	98.7
40 x 30; Parameter = .7; ICC = .3	1000	100.0	1004	99.6

performed the remaining criteria examined. Still, the performance of these criteria was dependent upon the conditions examined in the current study. None of the criteria performed very well in the smallest total sample size with low ICC value conditions. Thus, in this situation, researchers may want to employ the BIC_m and the $CAIC_m$ along with the AIC, regardless of model complexity. When total sample sizes are larger with higher ICC values, the BIC_N, $CAIC_N$, and the $AICC_m$ together may be used to select among a set of multilevel models. Researchers should be cautioned, however, that the $AICC_m$ performs less accurately when the competing

models are complex, unless the number of groups is large.

The models and conditions examined in the current study do not reflect all possible models and conditions found when analyzing real-world data. Hence, it is difficult to generalize the findings to every situation that may be encountered by applied researchers. Future research still needs to be conducted in order to more fully understand the characteristics of model selection criteria in the HLM arena. For example, the models examined in this study were limited to two levels; it would be interesting to examine how well these selection criteria perform with three-level models, particularly when calculating the criteria using *N*, *m* at level-2, and *m* at level-3.

Future research should also examine the sensitivity of the model selection criteria to non-normally distributed data as well as data that are missing at level-1, level-2, or both. Based on Gurka's (2006) finding that the model selection criteria worked well when models were misspecified by fixed effects using REML estimation, future research could also examine how well these criteria work using REML estimation under additional conditions.

In recent years, HLM has grown widely popular in its use. Indeed, our search in PsycInfo between January 2002 and March 2007 for articles in which HLM was used uncovered 220 articles. Our content analysis also indicated that model selection criteria were used in the model selection/comparison process in 45 of the 220 articles, with only 10 of those consisting of information criteria. Thus, most HLM research does not incorporate any type of model selection criteria. This could be a result of a lack of literature informing researchers as to the performance of these criteria and a lack of literature pointing to the necessity of these criteria when deciding between several competing models. In addition, while major software packages like SAS and SPSS include a number of information criteria in their output, other packages that estimate multilevel models, such as HLM 6 (Raudenbush, Bryk & Congdon, 2007) and MLwiN (Rasbash, et al., 2000), do not provide any information criteria in their output. While the deviance statistic is provided in these software packages, applied researchers may be less likely, or aware of, the different information criteria available. This may also possibly be contributing to the lack of utilization of these criteria in the applied literature. Therefore, the current study provides valuable information concerning the existing practices of applied researchers when comparing and selecting among hierarchical models as well as the performance of existing and alternative criteria when selecting among hierarchical models.

References

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

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(*3*), 345-370.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodal inference: A practical information-theoretic approach.* NY: Springer.

Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *The American Statistician*, *60*(*1*), 19-26.

Hannon, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, *Series B*, *41*(*2*), 190-195.

Hedeker, D., & Gibbons, R. D. (1999). MIXREG (Version 1.2). [Computer Software]. Chicago, IL: Hedeker & Gibbons.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, *72*(*2*), 297-307.

Jöreskog, K. G., & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc.

McQuarrie, A. D., & Tsai, C. L. (1998). *Regression and time series model selection.* River Edge, NJ: World Scientific Publishing Co.

Muthén, L. K., & Muthén, B. O. (2007). Mplus (Version 4.2). [Computer Software]. Los Angeles, CA: Muthén & Muthén.

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2005). MLwiN (Version 2.02). [Computer Software]. Bristol, UK: University of Bristol, Centre for Multilevel Modeling.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raudenbush, S., Bryk, A., & Congdon, R. (2007). HLM (Version 6.04) [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(*3*), 199-213.

SAS Institute Inc. (2007). SAS (Version 9.2) [Computer Software]. Cary, NC: SAS Institute Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464.

Shi, P., & Tsai, C. L. (2002). Regression model selection - a residual likelihood approach. *Journal of the Royal Statistical Society*, *Series B*, *64*(*2*), 237-252.

SPSS Inc. (2007). SPSS (Version 16.0) [Computer Software]. Chicago, IL: SPSS Inc.

Verbeke, G., & Molenerghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.