STRUCTURAL EQUATION MODELING, *14*(4), 535–569

Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study

Karen L. Nylund
Graduate School of Education & Information Studies, University of California, Los Angeles

Tihomir Asparouhov
Muthén & Muthén

Bengt O. Muthén
University of California, Los Angeles

Mixture modeling is a widely applied data analysis technique used to identify unobserved heterogeneity in a population. Despite mixture models' usefulness in practice, one unresolved issue in the application of mixture models is that there is not one commonly accepted statistical indicator for deciding on the number of classes in a study population. This article presents the results of a simulation study that examines the performance of likelihood-based tests and the traditionally used Information Criterion (ICs) used for determining the number of classes in mixture modeling. We look at the performance of these tests and indexes for 3 types of mixture models: latent class analysis (LCA), a factor mixture model (FMA), and a growth mixture models (GMM). We evaluate the ability of the tests and indexes to correctly identify the number of classes at three different sample sizes ($n = 200$, 500, 1,000). Whereas the Bayesian Information Criterion performed the best of the ICs, the bootstrap likelihood ratio test proved to be a very consistent indicator of classes across all of the models considered.

Correspondence should be addressed to Karen L. Nylund, Graduate School of Education & Information Studies, University of California, Los Angeles, 2023 Moore Hall, Mailbox 951521, Los Angeles, CA 90095-1521. E-mail: knylund@ucla.edu

Mixture modeling techniques, such as latent class analysis (LCA; McCutcheon, 1987) and growth mixture modeling (GMM; Muthén & Asparouhov, 2007; Muthén & Shedden, 1999), are statistical modeling techniques that are becoming more commonly used in behavioral and social science research. In general, mixture models aim to uncover unobserved heterogeneity in a population and to find substantively meaningful groups of people that are similar in their responses to measured variables or growth trajectories (Muthén, 2004).

The use of mixture modeling has allowed for deeper investigation of a variety of substantive research areas. Examples of the use of LCA can be found throughout the behavioral and social sciences, such as the analysis of data on Antisocial Personality Disorder (Bucholz, Hesselbrock, Heath, Kramer, & Schuckit, 2000) exploring whether subtypes exist with respect to different symptoms, analysis of clinically diagnosed eating disorders identifying four symptom-related subgroups (Keel et al., 2004), and analysis of Attention Deficit/Hyperactivity Disorder (ADHD; Rasmussen et al., 2002) exploring typologies of activity disorders. For an overview of applications and recent developments in LCA, see the edited book by Hagenaars and McCutcheon (2002).

Factor mixture analysis (FMA) is a type of cross-sectional mixture analysis considered a hybrid model because it involves both categorical and continuous latent variables. The application of an FMA model to achievement data in Lubke and Muthén (2005) explored the interaction effect of gender and urban status on class membership in math and science performance. For more on FMA, see Muthén (2006), Muthén and Asparouhov (2006) and Muthén, Asparouhov, and Rebollo (2006).

GMM is a modeling technique that can be used to identify unobserved differences in growth trajectories. Different from LCA, which is a cross-sectional analysis, GMM is a longitudinal analysis that explores qualitative difference in growth trajectories. These developmentally relevant growth trajectories are based on differences in growth parameter means (i.e., intercept and slope). For example, the application of GMM to college drinking data identified five drinking trajectories that differed in their mean number of drinks per week and their change over the semester (Greenbaum, Del Boca, Darkes, Wang, & Goldman, 2005). Further, this article identified particular drinking trajectories that were more likely to later develop into problematic drinking patterns. In another study of alcohol abuse, the application of a two-part GMM tested the hypothesis of two different stages in alcohol development, where each group had its own transition points (Li, Duncan, Duncan, & Hops, 2001). For more on GMM, see Muthén et al. (2002), Muthén and Muthén (1998–2006), and Muthén and Shedden (1999).

Despite the usefulness to applied researchers, one unresolved issue in the application of mixture models is how to determine the number of classes (e.g., unobserved subgroups) in a study population, also known as class enumeration.

Currently, applied researchers use a combination of criteria to guide the decision on the number of classes in mixture modeling. Such criteria include the combination of statistical information criteria (IC), like Akaike's Information Criterion (AIC; Akaike, 1987) and Bayesian Information Criterion (BIC; Schwartz, 1978) as well as agreement with substantive theory. A variety of textbooks and articles suggest the use of the BIC as a good indicator for class enumeration over the rest (Collins, Fidler, Wugalter, & Long, 1993; Hagenaars & McCutcheon, 2002; Magidson & Vermunt, 2004). Simulation studies considering LCA models suggest that the adjusted BIC (Sclove, 1987) is superior to other IC statistics (Yang, 2006).

There are a variety of simulation studies that explore the issue of deciding on the number of classes in mixture modeling. These studies differ in the types of models considered (i.e., finite mixture models, cross-sectional vs. longitudinal mixture models, mixture structural equation models, etc.) and the indexes used (e.g., IC, likelihood-based indexes, etc.). In a recent article, Yang (2006) used a simulation study to explore the performance of IC in a set of LCA models with continuous outcomes, and as described earlier, indicated that the adjusted BIC was the best indicator of the information criteria considered. Tofighi and Enders (2006), in a simulation study that considered similar fit indexes to those used in this study (e.g., AIC, BIC, and Lo–Mendell–Rubin [LMR]), considered a limited number of GMMs and concluded that the adjusted BIC and the LMR are promising in determining the number of classes. The Tofighi and Enders (2006) study did not consider the bootstrap likelihood ratio test (BLRT). In other studies, the AIC has been shown to overestimate the correct number of components in finite mixture models (Celeux & Soromenho, 1996; Soromenho, 1993), whereas the BIC has been reported to perform well (Roeder & Wasserman, 1997). Jedidi, Jagpal, and DeSarbo (1997) found that among commonly used model selection criteria, the BIC picked the correct model most consistently in the finite mixture structure equation model. Other authors suggest using less common techniques that are not easily implemented in software, such as Bayesian-based graphical techniques to aid in deciding on the number of classes (Garrett & Zeger, 2000) and the Rudas–Clogg–Lindsay (RCL) index of lack of fit (Formann, 2003).

To date, there is not common acceptance of the best criteria for determining the number of classes in mixture modeling, despite various suggestions. This is a critical issue in the application of these models, because classes are used for interpreting results and making inferences. The goal of this simulation study is to investigate the performance of these tests and indexes, and to provide insight about which tool is the most useful in practice for determining the number of classes.

The commonly used log likelihood difference test, which assumes a chi-square difference distribution, cannot be used to test nested latent class models.

Although LCA models with differing numbers of classes are, in fact, considered nested models, the chi-square difference test in the form of the likelihood ratio test (LRT) is not applicable in this setting due to regularity conditions not being met. Described in more detail later, if one were to naively apply this method, the p value obtained would not be accurate (for this reason, we called this the naive chi-square [NCS]). Thus, when one computes the difference in likelihoods of a $k - 1$ class and a k class model, the difference is not chi-square distributed (McLachlan & Peel, 2000) and standard difference testing is not applicable.

As an alternative, Lo, Mendell, and Rubin (2001) proposed an approximation to the LRT distribution, which can be used for comparing nested latent class models. This test was based on previous work by Vuong (1989), which considered the application of the test for general outcome distributions. The LMR test compares the improvement in fit between neighboring class models (i.e., comparing $k - 1$ and the k class models) and provides a p value that can be used to determine if there is a statistically significant improvement in fit for the inclusion of one more class. Jeffries (2003) claimed that there is a flaw in the mathematical proof of the LMR test for normal outcomes. Early simulation studies in the original Lo et al. (2002) paper show that despite this supposed analytic inconsistency, the LMR LRT may still be a useful empirical tool for class enumeration. The application of the LMR in practice has been limited and the performance of this test for LCA and GMM models has not been formally studied.

Another likelihood-based technique to compare the nested LCA models is a parametric bootstrap method described in McLachlan and Peel (2000). This method, which we call the BLRT, uses bootstrap samples to estimate the distribution of the log likelihood difference test statistic. In other words, instead of assuming the difference distribution follows a known distribution (e.g., the chi-square distribution), the BLRT empirically estimates the difference distribution. Similar to the LMR, the BLRT provides a p value that can be used to compare the increase in model fit between the $k - 1$- and k class models. To date, the BLRT has not commonly been implemented in mixture modeling software, so it is not commonly used in LCA and GMM modeling applications.

This article helps to further the understanding of the performance of available tools used for determining the number of classes in mixture models in several ways. First, we explore how inaccurate one would be if the NCS difference test were to be applied when testing nested mixture models. We also explore the performance of the two alternative likelihood-based tests, the LMR LRT and the BLRT, and compare their performances to each other and to the traditional difference test if it were applied in the naive fashion. Further, for the sake of comparison, the more commonly used IC indexes are considered (AIC, CAIC, BIC, and adjusted BIC, all defined later). All are considered for a select set of

mixture models as a first attempt to understand the performance of the indexes and tests when considering a range of modeling settings. The limited set of models and modeling settings considered restrict the scope of the results of this study. We focus on only a few mixture models and limited model structures. Further, we do not consider situations where model assumptions are violated (e.g., for FMA and GMM models we do not allow for within-class nonnormality). As a result, we are restricted in our interpretation and generalization of the results to the limited settings we considered.

The structure of this article is as follows. The first section introduces the models and the class enumeration tools considered in this investigation. The second section describes the Monte Carlo simulation study, describing the population from which the data were drawn and the models used to fit the data. Results of the simulation are presented in the third section. The final section draws conclusions and suggests recommendations for use.

THE MIXTURE MODELS CONSIDERED

The Latent Class Model

Lazarsfeld and Henry (1968) introduced the LCA model as a way to identify a latent categorical attitude variable that was measured by dichotomous survey items. LCA models identify a categorical latent class variable measured by a number of observed response variables. The objective is to categorize people into classes using the observed items and identify items that best distinguish between classes. Extensions of the LCA model have allowed for a variety of interesting applications in a variety of substantive areas. Advances in the statistical algorithms used to estimate the models, and the statistical software for analyzing them, allow for many types of outcomes (binary, ordinal, nominal, count, and continuous) or any combination of them. Because of this flexibility of the possible combination of outcomes, in this article we do not distinguish between models with categorical or continuous outcomes and refer to them all as LCA models. Models with the combination of categorical and continuous outcomes are not considered in this article.

There are two types of LCA model parameters: item parameters and class probability parameters. For LCA models with categorical outcomes, the *item parameters* correspond to the conditional item probabilities. These item probabilities are specific to a given class and provide information on the probability of an individual in that class to endorse the item. The *class probability parameters* specify the relative prevalence (size) of each class.

The LCA model with r observed binary items, u, has a categorical latent variable c with K classes ($c = k$; $k = 1, 2, \ldots, K$). The marginal item probability

for item $u_j = 1$ is

$$P(u_j = 1) = \sum_{k=1}^{K} P(c = k)P(u_j = 1|c = k).$$

Assuming conditional independence, the joint probability of all the r observed items is

$$P(u_1, u_2, \ldots, u_r) = \sum_{k=1}^{K} P(c = k)P(u_1|c = k)P(u_2|c = k)\ldots P(u_r|c = k).$$

For LCA models with continuous outcomes, the item parameters are class-specific item means and variances. As with LCA models with categorical outcomes, the class probability parameters specify the relative prevalence of each class. The LCA model with continuous outcomes has the form

$$f(y_i) = \sum_{k=1}^{K} P(c = k)f(y_i|c = k).$$

Here, y_i is the vector of responses for individual i on the set of observed variables, and the categorical latent variable c has K classes ($c = k$; $k = 1, 2, \ldots, K$). For continuous outcomes, $\boldsymbol{y}_i$, the multivariate normal distribution is used for $f(y_i|c)$ (e.g., within class normality) with class-specific means and the possibility for class-specific variances. To preserve the local independence assumption, the within-class covariance matrix is assumed diagonal.[1] In summary, for LCA models with categorical outcomes, the class-specific item parameters are item probabilities and for LCA models with continuous outcomes, the class-specific item parameters are the item means and variances.

Figure 1 is a diagram of the general latent class model. Variables in boxes represent measured outcomes (u for categorical outcomes or y for continuous). The circled variable represents the latent variable, c, the unordered latent class variable with K categories. In this general diagram, the number of observed items and the number of classes are not specified. The conditional independence assumption for LCA models implies that the correlation among the us or ys is explained by the latent class variable, c. Thus, there is no residual covariance among the us or ys.

[1]In general, the within-class covariance structure can be freed to allow within-class item covariance.

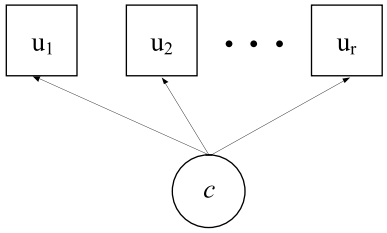


FIGURE 1 General latent class analysis model diagram.

Factor Mixture Analysis Model

Another type of mixture model we considered is a finite mixture model. This model is a hybrid latent variable model that includes both a continuous and categorical latent variable. Both latent variables are used to model heterogeneity in the observed items. In general, the categorical latent variable (i.e., the latent class variable) is used to identify distinct groups in the population and the continuous latent variable (i.e., the factor) can be used to describe a continuum that exists within the classes (e.g., a severity dimension). There are several different specifications of the FMA model, depicted in Figure 2. For example, FMA models can be specified with or without invariance of the measurement parameters for the latent classes and models can have different numbers of factor and classes. For more on these models and the comparisons to other techniques, see Muthén (2006) and Muthén and Asparouhov (2006).

The Growth Mixture Model

The last type of mixture model we considered is one example of the commonly used GMM (Muthén et al., 2002; Muthén, & Shedden, 1999). The GMM is

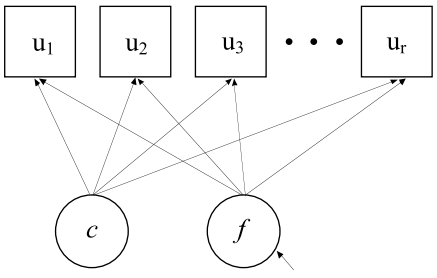


FIGURE 2 Factor mixture model diagram.

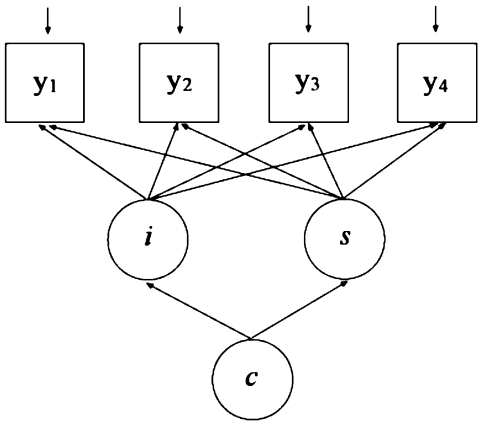


FIGURE 3 General growth mixture model with four continuous outcomes.

a type of longitudinal mixture model that is used to model heterogeneity in growth trajectories. Like LCA models, GMM identifies a categorical latent class variable. Instead of being identified by the outcomes themselves, in GMM, the latent class variable captures heterogeneity in the growth model parameters (i.e., intercept and slope). GMM and LCA models have the same class enumeration problem. It therefore makes sense to consider the performance of the fit indexes for GMMs as well.

Figure 3 is a diagram of a GMM with four repeated measures. The *y*s represent the continuous repeated measure outcomes. This particular model assumes linear growth, and thus has two continuous latent growth factors, the intercept (*i*) and slope (*s*), which are the growth parameters. These growth factors are considered random effects because we estimate both a mean and a variance of the latent variables. Lastly, we have the latent class variable, *c*, which is an unordered categorical latent variable indicated by the growth parameters. Because longitudinal mixture models are not the major focus of this article, we consider a GMM in a very limited setting as a first attempt to understand how these indexes perform for these models.

FIT INDEXES CONSIDERED

For the purpose of this article, three types of LRTs are considered: the traditional chi-square difference test (NCS), the LMR test, and the BLRT. The LRT is a commonly used technique that is used to perform significance testing on the difference between two nested models. The likelihood ratio (see, e.g.,

Bollen, 1989) is given by

$$LR = -2[\log L(\hat{\theta}_r) - \log L(\hat{\theta}_u)],$$

where $\hat{\theta}_r$ is the maximum likelihood (ML) estimator for the more restricted, nested model and $\hat{\theta}_u$ is the ML estimator for the model with fewer restrictions. The likelihood ratio has a limiting chi-square distribution under certain regularity assumptions, where the degree of freedom for the difference test equals the difference in the number of parameters of the two models. However, when the necessary regularity conditions are not met, the likelihood ratio difference does not have a chi-square distribution.

As mentioned before, however, this difference test, in its most commonly used form, is not applicable for nested LCA models that differ in the number of classes, as parameter values of the k class model are set to zero to specify the $k - 1$ class model. Specifically, we set the probability of being in the kth class to zero. By doing this, we are setting the value of a parameter at the border of its admissible parameter space (as probabilities range from 0–1), making this difference not chi-square distributed. Further, when making this restriction, the resulting k parameter space does not have a unique maximum. From here forward, this difference test is referred to NCS, as it is known that its application in this setting is inappropriate for the stated reasons.

The LMR differs from the NCS because the LMR uses an approximation of the distribution for the difference of these two log likelihoods (i.e., instead of using the chi-square distribution). The specific form of this test is provided in detail in Lo et al. (2001). The LMR test provides a p value, which indicates whether the $k-1$ class model is rejected in favor of the k class model. Similarly, the BLRT estimates the log likelihood difference distribution to obtain a p value, which, like the LMR, indicates if the $k - 1$ class model is rejected in favor of the k class model.

Because the BLRT has not traditionally been used widely for LCA and GMM models, we describe the process of obtaining the BLRT p value in detail. The method of obtaining the BLRT can generally be described in the following steps.

1. Initially estimate the $k - 1$ and k class models to provide the likelihoods for calculating the -2^* log likelihood difference.
2. Under the null $k-1$ class model, generate a bootstrap sample and calculate the -2^* log likelihood difference between the $k - 1$ and k class models.
3. Repeat this process independently many times and estimate the true distribution of the -2^* log likelihood difference.
4. Estimate the p value by comparing the distribution obtained in Step 3 with the -2^* log likelihood difference obtained in Step 1. This p value is then used to determine if the null $k-1$ class model should be rejected in favor of the k class model.

Consider a situation where we generate data using a model with three classes. We then use these data to compare the two- and three-class models, as well as the three- and four-class models. Using the bootstrapping method described earlier, we can empirically derive the distribution of the differences. Figure 4 displays two empirical distributions of the log likelihood differences between classes with differing numbers of classes. The panel on the left is a histogram of the distribution of differences between the two- and three-class models. Note that the difference distribution ranges from 2,300 to about 2,850. Because these data were generated to be a true three-class model, the BLRT p value should reject the null hypothesis of the two-class model in favor of the alternative three-class model. Using this distribution, we can use the observed difference value of 2839.14 to obtain a p value. The corresponding p value of 0.0 indicates that we would reject the null in favor of the alternative three-class hypothesis.

The panel on the right in Figure 4 is the estimated log likelihood difference distribution between the three- and four-class models. When comparing the three- versus four-class models, the null hypothesis in this model is three-class model and the alternative hypothesis is the four-class model. Note that the range for this difference distribution is between 25 and 60, which is a much smaller range than we observe in the panel on the right. For this comparison, we observe a log likelihood difference of 36.2947; placing that observed value on the bootstrapped distribution results in a p value of $p = .2947$. Thus, we fail to reject the null hypothesis, concluding that the three-class model is superior to the four-class model. Because the data were generated as a three-class model, in this setting, the BLRT correctly identifies the three-class model.

In addition to the three likelihood-based tests, several other fit criteria are considered. Specifically, we explore the performance of commonly used ICs. In their general form, IC indexes are based on the log likelihood of a fitted model, where each of the ICs apply a different penalty for the number of model parameters, sample size, or both. Because of the different penalties across the ICs,

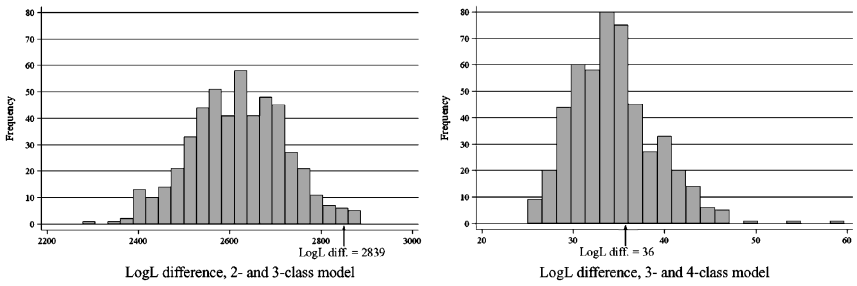


FIGURE 4 Log likelihood difference distribution for bootstrap estimates of the difference between two- and three-class model (left) and the three- and four-class model (right).

when using them it is possible that each of the ICs points toward a different class solution as the best model. The ICs considered in this study are the commonly used AIC, BIC, and adjusted BIC and the less commonly used consistent AIC (CAIC). The AIC is defined as

$$AIC = -2\log L + 2p,$$

where p is the number of free model parameters (Akaike, 1987). The CAIC (Bozdogan, 1987) a derivative of the AIC, includes a penalty for models having larger numbers of parameters using the sample size n and is defined as

$$CAIC = -2\log L + p(\log(n) + 1).$$

The BIC (Schwartz, 1978) is defined as

$$BIC = -2\log L + p\log(n).$$

The adjusted BIC, defined by Sclove (1987), replaces the sample size n in the BIC equation above with n^*:

$$n^* = (n + 2)/24.$$

The IC indexes mentioned earlier, the AIC, CAIC, BIC, and adjusted BIC, are used for comparison across several plausible models where the lowest value of a given IC indicates the best fitting model.

Latent class models can be tested against data in the form of frequency tables using a chi-square goodness-of-fit test, either Pearson or likelihood-ratio-based. For the purpose of this article, the chi-square goodness-of-fit test is not considered for two reasons. The chi-square distribution is not well approximated in situations with sparse cells, which commonly occurs when the model has many items. Our experience with models having sparse cells indicates that as the number of observed items increases, models are rarely rejected.

METHOD

The Monte Carlo Study

This study has four aims: (a) to understand the danger of incorrectly using the NCS, (b) to compare the performance of the three likelihood-based tests (the NCS, the LMR, and the BLRT), (c) to study common IC fit indexes, and (d) to compare the performance of the likelihood-based test to the IC indexes. The best way to achieve these aims is through a series of Monte Carlo simulations. By using Monte Carlo simulation techniques, we generate sample data with

known population parameters and evaluate the performance of these indexes under different modeling conditions. Specifically, we analyze these data with "alternative" models, ones that are not the truth (i.e., not the same model that generated the data), and evaluate the performance of the fit indexes and tests mentioned earlier.

There are two types of specifications that are allowed to vary in a simulation study: Monte Carlo variables and population specifications. The Monte Carlo variables include the choice of sample size and the number of replications. We chose three different sample sizes, specifically $n = 200$, 500, and 1,000. For the LCA models with categorical outcomes, 500 replications were generated for each sample size to ensure that there was sufficient reliability in the summary information calculated. Because of the increased time for estimating models, 100 replications were generated for the LCA models with continuous outcomes, and the FMA and GMM models for each of the sample sizes considered. The population variables, discussed in detail later, were the number of items in a model, item and class probabilities, and the true number of classes in the population.

Data Generation Models

The specification of the population parameters determines the generation of the sample data. As summarized in Table 1, LCA data with three different population

TABLE 1
Summary of Monte Carlo and Population Specifications for LCA, FMA, and GMM, Models Considered With Categorical and Continuous Outcomes

Population Specification						
Sample Size	*200, 500, 1000*					
Replications	*100 or 500*					
	Type of Outcome					
	Categorical and Continuous				*Categorical*	*Continuous*
Model Type	*Simple LCA*			*Complex LCA*	*FMA*	*GMM*
Number of items	8	8	15	10	8	4
Population number of classes	4	4	3	4	2	2
Class probabilities	Equal	Unequal	Equal	Unequal	Equal	Unequal

Note. LCA = latent class analysis; FMA = factor mixture analysis; GMM = growth mixture model.

model attributes were generated. The number of items, item probabilities, or class means and the number of classes in the population defined the three model populations for this simulation study. For models with binary outcomes, the conditional item probabilities were used to help distinguish the classes. For models with continuous outcomes, item means were used to identify classes, and item variances were held equal across items and classes. The true number of classes in a given population is *K,* and the four models considered are as follows:

- An 8-item model where $K = 4$ with equal class sizes.
- An 8-item model where $K = 4$ with unequal class sizes.
- A 15-item model where $K = 3$ with equal class sizes.
- A 10-item model where $K = 4$ with unequal class sizes.

These models were considered for both binary and continuous outcomes. The only exception is the 8-item model with unequal class sizes, where we considered only categorical outcomes were considered.

For the LCA models, we generated data under two model structures: simple and complex structures. These structures differ in the distribution of their conditional item probabilities or class means as well as their class probabilities (i.e., class size). *Simple structure* models are defined by having item probabilities or means that are particularly high or low for a given class so that these items discriminate among the classes. This structure is similar to a factor analysis model where there are unique items that identify each of the factors (i.e., no cross-loadings). In simple structure models, the class prevalence (i.e., class size) is the same across classes. The *complex structure,* on the other hand, does not have any single item that is particularly high or low for a specific class, thus there are not distinguishing items for any given class. Instead, one item can have high endorsement probability or class means for more than one of the latent classes.

Table 2 includes the model specifications for both the categorical and continuous outcome LCA models considered. This implies that for both the categorical and continuous LCA, the item parameters take on similar structures. Looking at the first item of the 8-item simple model (far left model in Table 2), we see .85 (2). This indicates that for the 8-item categorical LCA model, that item has an item probability of .85, which corresponds to a high probability of endorsement. For the 8-item model with continuous outcomes, that item has a mean of 2, specified by (2), where all item variances are set to 1. Note that the item has high probability (or high item mean) for Class 1 and low probabilities (e.g., zero means) for all remaining classes. As a result, each of the four classes in this model has two items that have high endorsement probabilities or item means. Similarly, the 15-item simple model is structured such that each class has five items

TABLE 2
Simple and Complex Structure Latent Class Analysis (LCA) Models With Equal and Unequal Class Sizes: Class and Item Probability/Item Mean Distributions for the LCA Models Considered[a]

8-item (Simple, Equal)					*8-item (Simple, Unequal)*					*15-item (Simple, Equal)*				*10-item (Complex, Unequal)*				
Item	*Class 1 25%*	*Class 2 25%*	*Class 3 25%*	*Class 4 25%*	*Item*	*Class 1 5%*	*Class 2 10%*	*Class 3 15%*	*Class 4 75%*	*Item*	*Class 1 33%*	*Class 2 33%*	*Class 3 33%*	*Item*	*Class 1 5%*	*Class 2 10%*	*Class 3 15%*	*Class 4 75%*
1	**0.85 (2)**	0.10 (0)	0.10 (0)	0.10 (0)	1	**0.85**	0.10	0.10	0.10	1	**0.85 (2)**	0.10 (0)	0.10 (0)	1	**0.85 (2)**	**0.85 (2)**	0.10 (0)	0.10 (0)
2	**0.85 (2)**	0.20 (0)	0.20 (0)	0.20 (0)	2	**0.85**	0.20	0.20	0.20	2	**0.85 (2)**	0.20 (0)	0.20 (0)	2	**0.85 (2)**	**0.85 (2)**	0.20 (0)	0.20 (0)
3	0.10 (0)	**0.85 (2)**	0.10 (0)	0.10 (0)	3	0.10	**0.85**	0.10	0.10	3	**0.85 (2)**	0.10 (0)	0.10 (0)	3	**0.85 (2)**	**0.85 (2)**	0.10 (0)	0.10 (0)
4	0.20 (0)	**0.85 (2)**	0.20 (0)	0.20 (0)	4	0.20	**0.85**	0.20	0.20	4	**0.85 (2)**	0.20 (0)	0.20 (0)	4	**0.85 (2)**	**0.85 (2)**	0.20 (0)	0.20 (0)
5	0.10 (0)	0.10 (0)	**0.85 (2)**	0.10 (0)	5	0.10	0.10	**0.85**	0.10	5	**0.85 (2)**	0.10 (0)	0.10 (0)	5	**0.85 (2)**	**0.85 (2)**	0.10 (0)	0.10 (0)
6	0.20 (0)	0.20 (0)	**0.85 (2)**	0.20 (0)	6	0.20	0.20	**0.85**	0.20	6	0.10 (0)	**0.85 (2)**	0.20 (0)	6	**0.85 (2)**	0.20 (0)	**0.85 (2)**	0.20 (0)
7	0.10 (0)	0.10 (0)	0.10 (0)	**0.85 (2)**	7	0.10	0.10	0.10	**0.85**	7	0.20 (0)	**0.85 (2)**	0.10 (0)	7	**0.85 (2)**	0.10 (0)	**0.85 (2)**	0.10 (0)
8	0.20 (0)	0.20 (0)	0.20 (0)	**0.85 (2)**	8	0.20	0.20	0.20	**0.85**	8	0.10 (0)	**0.85 (2)**	0.20 (0)	8	**0.85 (2)**	0.20 (0)	**0.85 (2)**	0.20 (0)
										9	0.20 (0)	**0.85 (2)**	0.10 (0)	9	**0.85 (2)**	0.10 (0)	**0.85 (2)**	0.10 (0)
										10	0.10 (0)	**0.85 (2)**	0.20 (0)	10	**0.85 (2)**	0.20 (0)	**0.85 (2)**	0.20 (0)
										11	0.20 (0)	0.20 (0)	**0.85 (2)**					
										12	0.10 (0)	0.10 (0)	**0.85 (2)**					
										13	0.20 (0)	0.20 (0)	**0.85 (2)**					
										14	0.10 (0)	0.10 (0)	**0.85 (2)**					
										15	0.20 (0)	0.20 (0)	**0.85 (2)**					

[a]Item probabilities for categorical LCA models are specified by the probability in each cell, and the class means for the continuous LCA are specified by the value in parentheses.

that distinguish it from the other classes. The classes of the LCA models with continuous outcomes were specified to have high separation, specifically the observed item means were set at the value 0 or 2, as specified in Table 2.

The complex model, as seen on the right side of Table 2, was structured using a real data example. The 10-item model is from an example in the ADHD literature. In particular, this model was characterized as having the most prevalent class considered a *normative class*—the largest class in which individuals do not exhibit problem behaviors (Class 4 on far right model in Table 2). Two other smaller classes are identified by having high probability on a subset of the items (Classes 2 and 3 in far right model of Table 2). Lastly, the smallest and most problematic class was characterized by having a high probability of endorsing all of the measured symptom items (Class 1 in far right model in Table 2). Note that for the 10-item complex structure model, any one item had a high probability or means for two classes (i.e., cross loading) and the prevalence (size) of the classes was not equal.

The FMA model considered in this study is just one example of this type of model. The FMA model is considered a generalized GMM model. For this article, the FMA model 8 observed binary items, with a true $k = 2$ class model and a single normal factor. The factor loadings and factor variances are assumed invariant across the classes and the factor means are constrained to be zero. The model specifications can be found in Table 3. This is an example of a simple FMA model, and thus a reasonable one to begin with.

TABLE 3
Factor Mixture Model Specification

Model Parameters	*Class 1*		*Class 2*	
Class size	50%		50%	
Factor variance	3.0		3.0	
Factor means	0		0	
Item	*Item Probability*	*Factor Loading*	*Item Probability*	*Factor Loading*
1	.1	1.0	**.9**	1.0
2	.1	1.0	**.9**	1.0
3	.1	1.0	**.9**	1.0
4	.1	1.0	**.9**	1.0
5	**.9**	1.0	.1	1.0
6	**.9**	1.0	.1	1.0
7	**.9**	1.0	.1	1.0
8	**.9**	1.0	.1	1.0

TABLE 4
Growth Mixture Model Specification

Model Parameters	*Class 1 Good Dev.*	*Class 2 Slow Dev.*
Class size	75%	25%
Growth parameters		
Intercept mean	2.00	1.00
Var(Intercept)	0.25	0.25
Slope mean	0.50	0.00
Var(Slope)	0.04	0.04
Var(Intercept, Slope)	0.00	0.00
Residuals		
Var(Y1)	0.15	0.15
Var(Y2)	0.20	0.20
Var(Y3)	0.20	0.20
Var(Y4)	0.35	0.35

The GMM considered was a simple structure model. The GMM has a true $k = 2$ class model with continuous outcomes and has linear growth parameters. Specifically, as seen in Table 4, one class of the GMM is characterized as having good development where the mean of the intercept growth factor is 2.0 and the mean slope factor is 0.5. This class is 75% of the sample. The second class, which is 25% of the sample, is characterized by having a lower intercept factor mean (i.e., 1.0) and slope factor mean of zero. The separation of the mean initial value of the two classes in the GMM model is two, a value that is thought to provide well-separated classes. The growth parameter and residual variances are specified to be equal across the two classes.

For the GMM and FMA models, we specified a one- through four-class model and considered 100 replications. For these models, the performance of the likelihood-based tests and indexes may be more sensitive to the specification of the alternative models. We specified alternative models to have similar model attributes as the generated population model. Specifically, for the GMM models, we constrained the residual variances to be equal across classes, the covariance between the intercept and slope parameters were invariant across classes, and the class prevalences were freely estimated. Further, for the alternative models, the slope and intercept means and variances were allowed to be class specific, where linear growth was assumed within each class.

Data Analysis Models

Data sets were generated according to the previously mentioned population model attributes and were analyzed using a series of models that differed in

the number of classes. Each of the replications and the three different sample sizes were analyzed using two- through six-class analysis models. The LMR and BLRT p values are provided for each replication that converged.

The simulation and analysis of the sample data were conducted using the Monte Carlo facilities in M*plus* Version 4.1 (L. Muthén & Muthén, 1998–2006). Within M*plus*, the population and analysis models are easily specified and summary information is provided across all completed replications. The LMR and BLRT p value can be obtained by specifying the output options of Tech 11 and Tech 14, respectively. Nonconvergence of any given replication occurs because of the singularity of the information matrix or an inadmissible solution that was approached as a result of negative variances. As a result, model estimates and summaries were not computed for these replications. This occurred in badly misspecified models (e.g., where $k = 3$ in the data generation and the analysis specified a six-class model) and occurred less than 1% of the time for the LCA models, and at most 5% for the FMA and GMM models. In the FMA and GMM model setting, nonconvergence was considered an indication of model misfit, and was used as evidence that the model with one fewer classes was superior.

It is widely known that mixture models are susceptible to converging on local, rather than global, solutions. To avoid this, it is often recommended that multiple start values for estimated model parameters be considered (McLachlan & Peel, 2000). Observing the same log likelihood obtained from multiple sets of start values increases confidence that the solution obtained is not a local maximum. M*plus* Version 4 has a random start value feature that generates a number of different random start sets, facilitating the exploration of possible local solutions. Before moving on to the larger simulation study, a variety of start value specifications was considered to ensure that a sufficient number of random starts were chosen to overcome local solutions.[2] Thus, there is little chance that the results of this study are based on local solutions. It should be noted that when using random start values in M*plus*, the program goes through the random start value procedure for both the $k-1$ and the k class models when calculating the LMR and BLRT p value.

One way to obtain the parametric BLRT p value is to generate and analyze, say, 500 replicated data sets and obtain a close approximation to the LRT distribution. This procedure, however, is computationally demanding. Typically, we need to know only whether the p value is above or below 5%, as this would determine whether the LRT test would reject or accept the null hypothesis. M*plus* uses the sequential stopping rule approach described in the Appendix. When the

[2]The number random starts for LCA models with categorical outcomes was specified to be "starts = 70 7;" in M*plus*. The models with continuous outcomes had differing numbers of random starts.

p value is close to 5%, this sequential procedure uses more replications than when the p value is far from 5%.

In simulation studies, the analysis models must be able to accurately recover the population parameters. If the models chosen to analyze the simulated data are not able to provide estimates that are close to the population parameters when correctly specified, results of the simulation have little meaning. For example, when four-class data are generated and analyzed with a four-class model, it is expected that the estimated parameter values are close to the population parameters that generated the data. The ability of the analysis models to recover the population parameters can be summarized by looking at the number of replications with confidence intervals that contain the true population parameter. These values are called *coverage estimates* and are considered for each one of the estimated parameters in the model. A coverage value of say, .93, for a given parameter would indicate that, across all replications, 93% of the model estimates fall within a 95% confidence interval of the population parameter value.[3] One rule of thumb is that coverage estimates for 95% confidence estimates should fall between .91 and .98 (L. Muthén & Muthén, 2002).

RESULTS

This section presents summaries from the simulation study. Results include information on coverage, a comparison of the three likelihood-based tests (NCS, LMR, and BLRT), the performance of the ICs (AIC, CAIC, BIC, and adjusted BIC) in identifying the correct model, and the ability of the LMR and BLRT to pick the correct model.

Coverage values for all the models (results are not presented here) were found to be very good (between .92 and .98).[4] The exceptions were in the modeling setting where there was a very small class (e.g., 5%). This occurs in the complex, 10-item categorical LCA model with sample size $n = 200$ and the 8-item categorical LCA with unequal classes with sample size $n = 200$. In this setting, the average coverage estimates for parameters in the smallest class (prevalence of 5%) were low (.54). These results were not surprising given that class size is only 10 (5% of 200). As sample size increases, coverage for this class increases in both settings to an average of .79 for $n = 500$ and .91 for $n = 1{,}000$.

[3]It is important to note that when coverage is studied, the random starts option of M*plus* should not be used. If it is used, label switching may occur, in that a class for one replication might be represented by another class for another replication, therefore distorting the estimate.

[4]The models that presented convergence problems were those that were badly misspecified. For example, for the GMM (true $k = 3$ class model) for $n = 500$, the convergence rates for the three-, four-, and five-class models were 100%, 87%, and 68%, respectively.

Likelihood-Based Tests

Results comparing the three LRTs are presented in terms of their ability to discriminate among neighboring class models. By *neighboring class models*, we are referring to models that differ by one class from the true population model. Thus, we focus on the performance of the LRTs by looking at the rate at which the indexes are able to discriminate between both the $k-1$ versus k class models and the k versus $k+1$ class models, where k is the correct number of classes. We narrow our focus in this way because we assume that a very misspecified model (e.g., one where we specify a two-class model and the population is a six-class model) would be easily identified by these tests. We are more concerned with the ability to reject models that are close to the true population model. Generally speaking, both the LMR and the BLRT methods are able to distinguish between the $k - g$ and k lass models (where $g < k$), which is how they are commonly used in practical applications.

Type I Error Comparison

Table 5 contains information on the Type I error rates for the LRTs considered (NCS, LMR, and BLRT). This information concerns the ability of the tests to "correctly" identify the k class model in comparison to the $k + 1$ class model. Table 5 gives the estimated Type I error (i.e., the probability of incorrectly rejecting a true model). If these tests worked perfectly, in that they correctly identified the k class model, we would expect the values in the cells to be approximately .05. The values in this table can be thought of as error rates. For example, looking at the 8-item categorical LCA model column for $n = 200$, we observe that the LMR's proportion rejected is .25. This means that for 25% of the replications for $n = 200$, the LMR incorrectly rejects the null four-class model in favor of the alternative five-class model. Knowing that the true population model has $k = 4$ classes for the 8-item model, we would not want to reject the null and hope that the overall rejection proportion would be around .05.

Looking at Table 5, we first notice that the NCS has an inflated Type I error rate across nearly all modeling settings. This is evident by noting that the error rates are consistently higher than the expected value of .05. The LMR has inflated Type I error rates for the LCA models with categorical outcomes and for the FMA and GMM models, but has adequate Type I error rates for the LCA with continuous outcomes. The BLRT works remarkably well, where we observe Type I error rates near or below .05 for all model structure and sample size combinations. When comparing the performance across LCA models for all sample sizes, the LMR consistently outperforms the NCS, and the BLRT outperforms both the NCS and LMR.

TABLE 5
Type I Error Rates for the Three Likelihood Ratio Tests: NCS, LMR, and BLRT

	8-Item Simple Structure						*15-Item Simple Structure*			*10-Item Complex Structure*		
	Equal Classes			*Unequal Classes*			*Equal Classes*			*Unequal Classes*		
	H0: 4-Class (True) H1: 5-Class			*H0: 4-Class (True) H1: 5-Class*			*H0: 3-Class (True) H1: 4-Class*			*H0: 4-Class (True) H1: 5-Class*		
n	*NCS*	*LMR*	*BLRT*	*NCS*	*LMR*	*BLRT*	*NCS*	*LMR*	*BLRT*	*NCS*	*LMR*	*BLRT*
Latent class analysis with categorical outcomes												
200	.33	.25	.05	.30	.11	.04	.97	.10	.08	.51	.17	.06
500	.41	.25	.04	.36	.19	.07	.98	.07	.05	.66	.19	.06
1,000	.48	.18	.05	.41	.12	.05	.99	.06	.05	.73	.21	.06
Latent class analysis with continuous outcomes												
200	.74	.11	.06				.60	.02	.04	.56	.03	.04
500	.76	.06	.03				.62	.06	.01	.47	.03	.03
1,000	.79	.06	.02				.75	.03	.02	.49	.02	.04

	Other Mixture Models					
	FMA 8-Item Binary Outcome			*GMM 4 Time-Point, Linear Growth*		
	H0: 2-Class (True) H1: 3-Class			*H0: 2-Class (True) H1: 3-Class*		
n	*NCS*	*LMR*	*BLRT*	*NCS*	*LMR*	*BLRT*
200	.48	.22	.10	.11	.07	.05
500	.49	.19	.08	.16	.12	.06
1,000	.51	.21	.07	.08	.16	.01

Note. NCS = naive chi-square; LMR = Lo–Mendell–Rubin; BLRT = bootstrap likelihood ratio test; FMA = factor mixture model; GMM = growth mixture model.

Focusing solely on the values in the NCS column, we notice that the error rate increases as sample size increases, which could be caused by known problems of the chi-square statistic when analyzing large samples. Comparing the similarly structured 8- and 15-item models for both continuous and categorical outcome models, the LMR performs better in LCA models with more items. Looking across the categorical LCA model results, both the NCS and BLRT perform best for the model with fewest items, the 8-item simple structure model. The GMM results show the LMR performs about the same as NCS, but the BLRT is the clear winner. Note that for GMMs, the LMR Type I error rate increases as sample size increases, a pattern not seen in other modeling settings. For the FMA model, the BLRT is the clear winner over the other tests. Overall, the

BLRT performs very well in terms of Type I error (e.g., all values are close to .05) across all model settings.

Power Comparisons

Table 6 is concerned with power (i.e., the probability that the test will reject the null hypothesis when it is false). Thus, values of at least .80 would indicate that the particular test worked well in finding the true k class model. Table 6 summarizes testing the $k - 1$ class model against the correct k class model. In contrast to the discussion of Type I error in Table 5, in this setting we want the null hypothesis to be rejected in favor of the alternative. It is meaningful to compare the power across approaches only when their Type I error rate is at

TABLE 6
Power Values for the Three Likelihood Ratio Tests: NCS, LMR, and BLRT

	8-Item Simple Structure						*15-Item Simple Structure*			*10-Item Complex Structure*		
	Equal Classes			*Unequal Classes*			*Equal Classes*			*Unequal Classes*		
	H0: 3-Class H1: 4-Class (True)			*H0: 3-Class H1: 4-Class (True)*			*H0: 2-Class H1: 3-Class (True)*			*H0: 3-Class H1: 4-Class (True)*		
n	*NCS*	*LMR*	*BLRT*	*NCS*	*LMR*	*BLRT*	*NCS*	*LMR*	*BLRT*	*NCS*	*LMR*	*BLRT*
	Latent class analysis with categorical outcomes											
200	1.00	.95	1.00	.80	.36	.53	.97	1.00	1.00	.98	.62	.84
500	1.00	1.00	1.00	.98	.75	.94	.98	1.00	1.00	1.00	.90	1.00
1,000	1.00	1.00	1.00	1.00	.96	1.00	.99	1.00	1.00	1.00	.98	1.00
	Latent class analysis with continuous outcomes											
200	1.00	1.00	1.00				.99	1.00	1.00	1.00	.67	.98
500	1.00	1.00	1.00				1.00	1.00	1.00	1.00	1.00	1.00
1,000	1.00	1.00	1.00				1.00	1.00	1.00	1.00	1.00	1.00

	Other Mixture Models					
	FMA 8-Item Binary Outcome			*GMM 4 Time-Point, Linear Growth*		
	H0: 1-Class H1: 2-Class (True)			*H0: 1-Class H1: 2-Class (True)*		
n	*NCS*	*LMR*	*BLRT*	*NCS*	*LMR*	*BLRT*
200	1.00	1.00	1.00	1.00	.85	.95
500	1.00	1.00	1.00	1.00	1.00	1.00
1,000	1.00	1.00	1.00	1.00	1.00	1.00

Note. NCS = naive chi-square; LMR = Lo–Mendell–Rubin; BLRT = bootstrap likelihood ratio test; FMA = factor mixture model; GMM = growth mixture model.

an acceptable rate. Because the Type I error rates for the NCS are consistently inflated (e.g., well above .05), it does not make sense to compare power rates for this test. Thus, we consider only the LMR and BLRT for the power comparisons. The LMR, however, does have inflated Type I error in certain settings.

Table 6 shows that for both the LMR and BLRT, there is sufficient power to detect the k class model in almost all of the models. The exceptions are for the LMR where the power is .36 for the categorical LCA with 8 items with unequal class sizes and is .62 for the 10-item LCA, $n = 200$ modeling with categorical outcomes, and .67 for the 10-item LCA, $n = 200$ modeling with continuous outcomes. The only setting where BLRT has lower than expected power is in the 8-item categorical LCA setting with $n = 200$ where there is a power of .53. The power quickly reaches an acceptable value as the sample size increases. Even in the 10-class modeling setting with small sample size, the BLRT has sufficient power. The LMR power is low for the GMM, but there are also high Type I error rates so power results are not meaningful.

Information Criteria

The likelihood-based results presented in Tables 5 and 6 are concerned with comparing neighboring class models. For the ICs, we compare values across a series of model specifications. The values in Table 7 were computed by comparing the AIC, CAIC, BIC, and adjusted BIC across all models (two- through six-class models), then identifying where the lowest values occurred across those models considered. For example, looking at the 8-item LCA model with categorical outcomes (which is a true four-class model) we note that for $n = 500$, the lowest values of AIC occurred at the four-class model 68% of the time and 100% of the time for CAIC, BIC, and adjusted BIC. A value of 100 in the bolded column indicates perfect identification of the k class model by one of the indexes.

Looking across all the models considered there are a few general trends worth noting. First, the AIC does not seem to be a good indicator for identifying the k class model for any of the modeling settings. The value in the bolded column indicates that, at best, the AIC correctly identifies the k class model 75% of the time when looking across all models and sample size considerations. Also, notice that accuracy decreases as sample size increases, a known problem with AIC because there is no adjustment for sample size. Further, we notice that when AIC is not able to identify the correct k class model, it is most likely to identify the $k + 1$ class model as the correct model. The CAIC is able to correctly identify the correct k class model close to 100% of the time for both the 8- and 15-item categorical LCA models and for most of the continuous item LCA, regardless of sample size, and performs well for both the FMA and GMM models, regardless of sample size. For the categorical 10-item, complex model, the CAIC performance was much worse, identifying the correct model only 1%

TABLE 7
Percentage of Times the Lowest Value Occurred in Each Class Model for the AIC, CAIC, BIC, and Adjusted BIC

		Latent Class Analysis with Categorical Outcomes																			
		AIC					*CAIC*					*BIC*					*Adjusted BIC*				
		Classes					*Classes*					*Classes*					*Classes*				
Model	*n*	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*
8-item	200	0	0	**75**	22	3	0	5	**95**	0	0	0	1	**99**	0	0	0	0	**83**	15	2
(Simple,	500	0	0	**68**	27	5	0	0	**100**	0	0	0	0	**100**	0	0	0	0	**100**	0	0
Equal)	1000	0	0	**60**	32	8	0	0	**100**	0	0	0	0	**100**	0	0	0	0	**100**	0	0
		Classes					*Classes*					*Classes*					*Classes*				
	n	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*
8-item	200	0	26	**48**	24	2	83	17	**0**	0	0	58	42	**0**	0	0	0	29	**58**	12	1
(Simple,	500	0	2	**67**	27	4	6	83	**11**	0	0	1	72	**27**	0	0	0	12	**88**	0	0
Unequal)	1000	0	0	**62**	33	5	0	26	**74**	0	0	0	12	**88**	0	0	0	0	**100**	0	0
		Classes					*Classes*					*Classes*					*Classes*				
	n	*2*	***3***	*4*	*5*	*6*	*2*	***3***	*4*	*5*	*6*	*2*	***3***	*4*	*5*	*6*	*2*	***3***	*4*	*5*	*6*
15-item	200	0	**43**	41	13	3	0	**100**	0	0	0	0	**100**	0	0	0	0	**62**	31	6	1
(Simple,	500	0	**32**	41	17	6	0	**100**	0	0	0	0	**100**	0	0	0	0	**99**	1	0	0
Equal)	1000	0	**31**	45	17	6	0	**100**	0	0	0	0	**100**	0	0	0	0	**100**	0	0	0
		Classes					*Classes*					*Classes*					*Classes*				
	n	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*
10-item	200	0	5	**67**	24	3	2	97	**1**	0	0	0	92	**8**	0	0	0	7	**76**	15	2
(Complex,	500	0	0	**55**	35	9	0	44	**56**	0	0	0	24	**76**	0	0	0	1	**98**	1	0
Unequal)	1000	0	0	**46**	39	14	0	1	**99**	0	0	0	0	**100**	0	0	0	0	**100**	0	0

(*continued*)

and 56% of the time for $n = 200$ and $n = 500$, respectively. For the larger sample size of $n = 1{,}000$, the CAIC performed well. The CAIC's adjustment for the number of parameters using sample size significantly improves its performance over the AIC.

The BIC and the adjusted BIC are comparatively better indicators of the number of classes than the AIC. Both the BIC and adjusted BIC are sample size adjusted and show improvement at identifying the true k class model as sample size increases. The BIC correctly identified the k class model close to 100% of the time for both the categorical and continuous 15-item and 8-item LCA models with equal class size. In the 10-item categorical LCA model with unequal class size, it only identified the correct model 8% of the time when $n = 200$ and for

TABLE 7
(*Continued*)

		Latent Class Analysis with Continuous Outcomes																			
		AIC					*CAIC*					*BIC*					*Adjusted BIC*				
		Classes					*Classes*					*Classes*					*Classes*				
Model	*n*	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*
8-item	200	0	0	**33**	35	32	0	0	**100**	0	0	0	0	**100**	0	0	0	0	**49**	31	20
(Simple	500	0	0	**42**	34	24	0	0	**100**	0	0	0	0	**100**	0	0	0	0	**95**	5	0
Structure)	1000	0	0	**29**	43	28	0	0	**100**	0	0	0	0	**100**	0	0	0	0	**100**	0	0
		Classes					*Classes*					*Classes*					*Classes*				
	n	*2*	***3***	*4*	*5*	*6*	*2*	***3***	*4*	*5*	*6*	*2*	***3***	*4*	*5*	*6*	*2*	***3***	*4*	*5*	*6*
15-item	200	1	**39**	44	16	0	1	**99**	0	0	0	1	**99**	0	0	0	1	**58**	32	9	0
(Simple	500	0	**42**	39	17	2	0	**99**	0	1	0	0	**99**	0	1	0	0	**99**	0	1	0
Structure)	1000	0	**29**	39	21	11	0	**100**	0	0	0	0	**100**	0	0	0	0	**100**	0	0	0
		Classes					*Classes*					*Classes*					*Classes*				
	n	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*	*2*	*3*	***4***	*5*	*6*
10-item	200	0	0	**65**	21	14	0	35	**65**	0	0	0	26	**74**	0	0	0	0	**75**	18	7
(Complex	500	0	0	**59**	30	11	0	0	**100**	0	0	0	0	**100**	0	0	0	0	**99**	1	0
Structure)	1000	0	0	**68**	23	9	0	0	**100**	0	0	0	0	**100**	0	0	0	0	**100**	0	0

		Factor Mixture Model with Categorical Outcomes															
		AIC				*CAIC*				*BIC*				*Adjusted BIC*			
		Classes				*Classes*				*Classes*				*Classes*			
Model	*n*	*1*	***2***	*3*	*4*	*1*	***2***	*3*	*4*	*1*	***2***	*3*	*4*	*1*	***2***	*3*	*4*
FMA	200	0	**59**	33	8	0	**94**	6	0	0	**100**	0	0	0	**69**	27	4
	500	0	**59**	27	14	0	**100**	0	0	0	**100**	0	0	0	**97**	3	0
	1000	0	**58**	28	14	0	**100**	0	0	0	**100**	0	0	0	**94**	3	3

		Growth Mixture Model with Continuous Outcomes															
		AIC				*CAIC*				*BIC*				*Adjusted BIC*			
		Classes				*Classes*				*Classes*				*Classes*			
Model	*n*	*1*	***2***	*3*	*4*	*1*	***2***	*3*	*4*	*1*	***2***	*3*	*4*	*1*	***2***	*3*	*4*
GMM	200	0	**58**	22	20	0	**98**	2	0	16	**84**	0	0	0	**66**	18	16
	500	0	**65**	24	11	0	**100**	0	0	0	**100**	0	0	0	**90**	9	1
	1000	0	**64**	24	12	0	**100**	0	0	0	**100**	0	0	0	**100**	0	0

Note. Columns in boldface type represent the true *k*-class for the given model. AIC = Akaike's Information Criterion; CAIC = Consistent Akaike's Information Criterion; BIC = Bayesian Information Criterion.

the categorical 8-item LCA with unequal class size, it was never able to identify the correct model for $n = 200$. For the 10-item categorical LCA, performance increased as n increased, reaching 100% correct identification when $n = 1{,}000$. For the categorical 8-item LCA with unequal class size, the performance of the BIC increases as sample size increases, eventually reaching 88% for $n = 1{,}000$. The BIC identified the correct model 74% of the time for the 10-item continuous LCA model, $n = 200$, and jumped to 100% for $n = 500$. In both the FMA and GMM modeling settings, BIC performed well across all sample sizes. The adjusted BIC performs relatively well across all models, but shows some weakness when the sample size is small. For $n = 200$ in the categorical LCA 15- and 10-item models, adjusted BIC correctly identified the k class model only 62% and 76% of the time, respectively. In general, when the adjusted BIC goes wrong, it tends to overestimate the number of classes. In summary, comparing across all the models and sample sizes, there seems to be strong evidence that the BIC is the best of the ICs considered.

Likelihood-Based Tests

Table 8 contains information on the performance of the three likelihood-based tests: the NCS, LMR, and BLRT. Similar to the results presented in Table 7, these results are based on comparing across different numbers of classes for a set of models. To calculate these values, we examined the p values resulting from specifying the two- through six-class models for each replication. Note that for these LMR and BLRT, when specifying a two-class model, we compare the one-class model to the two-class model. Note that we specified alternative models that range between two and six classes. Thus, for Table 8 the possible solution for the LMR and BLRT ranges from a one-class model to a five-class model for the LCA models, and range from one class to three classes for the FMA and GMM models. For the NCS, we obtain p values using the traditional chi-square distribution. The p value provided is used to assess if there is significant improvement between the specified model and a model with one less class. Looking at these p values, we identified the model selected based on the occurrence of the first nonsignificant p value ($p > .05$). A bolded value close to .95 would indicate that the test performed close to perfectly. The bolded number in the LMR column in the categorical LCA eight-class model for $n = 500$ indicates that 78% of the replications concluded that the correct four-class solution was the correct model.

Results in Table 8 indicate a clear advantage of the BLRT over the NCS and LMR. In almost all LCA model and sample size considerations, the BLRT is able to correctly identify the true k class model nearly 95% of the time. Looking at the NCS and considering the LCA models, we notice that the performance decreases as sample size increases and remains consistently poor. For the FMA

TABLE 8
Percentage of Time a Nonsignificant p Value Selected the Given Class Model for the NCS, LMR, and BLRT

		Latent Class Analysis with Categorical Outcomes															
		NCS						*LMR*					*BLRT*				
		Classes						*Classes*					*Classes*				
Model	*n*	*1*	*2*	*3*	***4***	*5*	*6*	*1*	*2*	*3*	***4***	*5*	*1*	*2*	*3*	***4***	*5*
8-item	200	0	0	0	**67**	29	4	4	8	6	**64**	18	0	0	0	**96**	4
(Simple,	500	0	0	0	**58**	33	9	0	0	0	**78**	22	0	0	0	**96**	5
Equal)	1000	0	0	0	**52**	36	13	0	0	0	**83**	17	0	0	0	**95**	5
		Classes						*Classes*					*Classes*				
	n	*1*	*2*	*3*	***4***	*5*	*6*	*1*	*2*	*3*	***4***	*5*	*1*	*2*	*3*	***4***	*5*
8-item	200	0	0	20	**50**	26	4	6	30	39	**21**	4	0	0	47	**49**	4
(Simple,	500	0	0	1	**62**	28	9	0	10	23	**53**	10	0	0	6	**87**	7
Unequal)	1000	0	0	0	**60**	31	9	0	7	4	**78**	11	0	0	0	**95**	5
		Classes						*Classes*					*Classes*				
	n	*1*	*2*	*3*	***4***	*5*	*6*	*1*	*2*	***3***	*4*	*5*	*1*	*2*	***3***	*4*	*5*
15-item	200	0	0	**3**	29	40	29	0	0	**90**	9	1	0	0	**92**	7	1
(Simple,	500	0	0	**2**	19	34	45	0	0	**93**	6	1	0	0	**95**	5	0
Equal)	1000	0	0	**1**	17	31	50	0	0	**94**	6	0	0	0	**95**	5	0
		Classes						*Classes*					*Classes*				
	n	*1*	*2*	*3*	***4***	*5*	*6*	*1*	*2*	*3*	***4***	*5*	*1*	*2*	*3*	***4***	*5*
10-item	200	0	0	2	**48**	41	10	5	9	34	**43**	9	0	0	16	**78**	6
(Complex,	500	0	0	0	**34**	45	21	1	4	9	**72**	14	0	0	0	**94**	6
Unequal)	1000	0	0	0	**26**	41	33	0	2	2	**80**	17	0	0	0	**94**	6

(*continued*)

model, the NCS performs consistently poorly, specifically identifying the correct model only about 60% of the time across all sample sizes. For the GMM model, however, the NCS performed consistently well, although not as well as the BLRT. In the categorical LCA 15-item models, the LMR performs almost as well as the BLRT, where between 90% and 94% of the time the LMR picked the correct number of classes. In this same setting, the NCS performs very poorly, identifying the correct model, at most, only 3% of the time. For LCA models with continuous outcomes, the LMR and BLRT perform very well for $n = 500$ and 1,000. For both the categorical and continuous LCA, in the 15-item models, the LMR performs rather well, where it identifies the correct model over 90% of the time for all sample sizes.

TABLE 8
(*Continued*)

Latent Class Analysis with Continuous Outcomes

		NCS						LMR					BLRT				
		Classes						Classes					Classes				
Model	*n*	*1*	*2*	*3*	***4***	*5*	*6*	*1*	*2*	*3*	***4***	*5*	*1*	*2*	*3*	***4***	*5*
8-item	200	0	0	0	**26**	37	37	0	5	0	**84**	9	0	0	0	**94**	6
(Simple	500	0	0	0	**24**	36	40	0	0	0	**94**	5	0	0	0	**97**	3
Structure)	1000	0	0	0	**21**	44	35	0	0	0	**94**	5	0	0	0	**98**	2
		Classes						Classes					Classes				
	n	*1*	*2*	***3***	*4*	*5*	*6*	*1*	*2*	***3***	*4*	*5*	*1*	*2*	***3***	*4*	*5*
15-item	200	0	1	**39**	41	17	2	0	0	**98**	2	0	0	0	**96**	3	1
(Simple	500	0	0	**38**	44	18	0	0	0	**94**	6	0	0	0	**99**	1	0
Structure)	1000	0	0	**25**	49	21	5	0	0	**97**	3	0	0	0	**98**	2	0
		Classes						Classes					Classes				
	n	*1*	*2*	*3*	***4***	*5*	*6*	*1*	*2*	*3*	***4***	*5*	*1*	*2*	*3*	***4***	*5*
10-item	200	0	0	0	**56**	34	10	0	21	30	**48**	1	0	0	2	**94**	3
(Complex	500	0	0	0	**46**	36	18	0	5	0	**93**	1	0	0	0	**97**	3
Structure)	1000	0	0	0	**49**	34	17	0	4	0	**94**	2	0	0	0	**96**	3

Factor Mixture Model with Categorical Outcomes

		NCS				LMR			BLRT		
		Classes				Classes			Classes		
Model	*n*	*1*	***2***	*3*	*4*	*1*	***2***	*3*	*1*	***2***	*3*
FMA	200	0	**60**	40	0	0	**80**	20	0	**87**	13
	500	0	**61**	39	0	0	**84**	16	0	**92**	8
	1000	0	**60**	40	0	0	**84**	16	0	**93**	7

Growth Mixture Model with Continuous Outcomes

		NCS				LMR			BLRT		
		Classes				Classes			Classes		
Model	*n*	*1*	***2***	*3*	*4*	*1*	***2***	*3*	*1*	***2***	*3*
GMM	200	0	**90**	10	0	14	**81**	5	5	**90**	5
	500	0	**84**	14	2	0	**90**	10	0	**94**	6
	1000	0	**92**	8	0	0	**85**	15	0	**99**	1

Note. NCS = naive chi-square; LMR = Lo–Mendell–Rubin; BLRT = bootstrap likelihood ratio test.

DISCUSSION AND CONCLUSIONS

This simulation considered the performance of the three likelihood-based tests (NCS, LMR, and BLRT) and commonly used fit indexes used for determining the number of classes in mixture models with both categorical and continuous outcomes. We set out to understand how misleading it would be if the NCS test was used for these models compared to the newly proposed LMR and the BLRT. We also considered the conventional IC indexes.

Comparing the Likelihood-Based Tests

The results of this study support the fact that a researcher using the NCS for testing the $k-1$ versus k class model would reject the true model too frequently. It is noted that the NCS is especially sensitive to sample size, and its performance actually worsens as sample size increases for LCA models. As mentioned before, we know that this test is not appropriate for testing nested mixture models. It could be that this assumption violation is amplified as sample size is increased. It is important to note that, as seen in Table 8, when the NCS goes wrong, it tends to overestimate the number of classes. As an alternative to the NCS, the other LRTs, the LMR and BLRT, were shown to have the ability to identify the correct model more accurately. The BLRT clearly performed better than the LMR in almost all modeling settings.

We consider only the power results for the LMR and BLRT because there are high Type I error rates for the NCS test. There is good power for both the LMR and BLRT for nearly all the models considered in this study. The BLRT has more consistent power across all sample sizes than the LMR. Slight differences in power across the BLRT and LMR do not indicate a significant distinction in performance between the two tests in terms of power (i.e., distinguishing between the k and $k-1$ class models).

Considering the LMR results in Table 8, we notice that, in general, for LCA models if the LMR incorrectly identifies a model, it tends to overestimate the number of classes. Thus, to the analyst using the LMR as a tool for class enumeration, when the p value indicates a nonsignificant difference for the LMR, one can feel somewhat confident that there is, at most, that number of classes, but that there might, in fact, be fewer. Overestimating the number of classes can be thought of as being better than underidentifying classes, as the true k class solution can still be extracted from a $k+1$ class solution. For example, if the true solution is $k=3$ and the model LMR identified a four-class solution, it could be the case that one of the classes in the four-class solution does not make substantive sense or there is a very small class that is hard to identify. This could result in the decision to not go with the four-class solution despite the LMR results, and instead settle on the true three-class solution. If LMR had

indicated that a two-class model was most fitting, it might be more difficult to somehow increase the number of classes and settle on the true $k = 3$ class model.

For LCA models, the BLRT has better accuracy for correctly identifying the true number of classes. The only setting in which the BLRT has a correct identification rate below 92% is in the categorical LCA 10-item setting when $n = 200$. As noted before, this was a difficult modeling setting, where there was one small class with very few observations. Further, for the LCA models with continuous outcomes, FMA and GMM models, the BLRT performs the best out of the likelihood tests presented.

As mentioned previously, both the LMR test and BLRT provide a p value that can be used to decide whether the $k - 1$-class model should be rejected in favor of the k class model. In practice, for a researcher fitting a series of LCA models, the LMR may result in p values that bounce around from being significant to nonsignificant and then back to significant again. This does not appear to be the case for the p value based on the BLRT. Findings show that once the BLRT p value is nonsignificant, it remains nonsignificant for the subsequent increased class models. Based on experience and the findings of this study, preliminary results suggest the first time the p value of the LMR is nonsignificant might be a good indication to stop increasing the number of classes.

Comparing the Information Criteria

Summary information for the AIC, CAIC, BIC, and adjusted BIC were included to aid in understanding the utility of these indexes in LCA and GMM modeling settings. The results are in agreement with previous research indicating the AIC is not a good indicator for class enumeration for LCA models with categorical outcomes (Yang, 2006). Based on the results, the AIC is not a good indicator of class for any of the models considered in this study. The CAIC showed a clear improvement over the AIC in both categorical and continuous LCA, but was sensitive to the combination of unequal class sizes and a small sample size for LCA models with categorical outcomes. The CAIC is not widely used for class enumeration, and more studies should look at its performance given its clear superiority to AIC in certain modeling situations.

Based on the results of this study, when comparing across all modeling settings, we conclude that the BIC is superior to all other ICs. For categorical LCA models, the adjusted BIC correctly identifies the number of classes more consistently across all models and all sample sizes. In this setting, the BIC performed well for the 8- and 15-item categorical LCA models with equal class sizes, but performance decreased for the two categorical LCA models with unequal class size (e.g., the 10-item, complex structure with unequal class size and the 8-item, simple structure with unequal class size). When considering continuous LCA,

however, the superiority of the BIC is more evident. The BIC more consistently identifies the correct model over the adjusted BIC, where the adjusted BIC drops as low as 49% for the 8-item model with $n = 200$. For both the FMA and GMM, the BIC performs well, where at worst it identifies the correct model 84% of the time. Based on these results, we conclude that the BIC is the most consistent IC among those considered for correctly identifying the number of classes. Table 7 shows that the BIC has sensitivity to small sample sizes, regardless of the type of model. Undoubtedly, there needs to be further study to facilitate a greater understanding of the impact of the structure on the performance of these ICs.

Comparing the BIC and BLRT

We can compare the results of Tables 7 and 8 to understand the performance of the BIC, the best performing of the ICs, to the BLRT, the best performing of the LRTs. For the LCA with categorical outcomes, the BLRT is more consistent at identifying the correct number of classes than the BIC, because at its worst, the BLRT identifies the correct number of classes 49% of the time. This is better than the BIC, which at its worst, is not able to identify the correct number of classes at all for the 8-item categorical outcome LCA model with unequal classes, $n = 200$. Considering the LCA with continuous outcomes, the BIC performs well, but it correctly identifies the correct model only 74% of the time for $n = 200$. In this setting, the BLRT is consistent; it identifies the correct model a remarkable 94% of the time. Comparing the results of Tables 7 and 8 for the FMA and GMM models, both the BIC and BLRT perform well. In this setting, BIC, at its worst, identifies the correct model 84% of the time for the GMM where $n = 200$; at its worst, the BLRT identifies the correct model 87% of the time for the FMA with $n = 200$. Thus, considering all the models in this study, the BLRT stands out as the most consistent indicator for the correct number of classes when comparing results from Tables 7 and 8.

Although the results presented here represent only a small subset of all mixture models, it is important to note that the FMA and GMM results are based on only one model of each. Thus, when comparing the results presented in Tables 7 and 8, the FMA and GMM results are not given as much weight as the LCA models. A recent simulation study by Tofighi and Enders (2006) more closely examines class enumeration issues for a wider range of GMM models.

In summary, the results of this study indicate a clear advantage of the BLRT test compared to NCS and LMR, and show that it can be used as a reliable tool for determining the number of classes for LCA models. The BIC was found to correctly identify the number of classes better than the other ICs considered, for LCA models and FMA and GMM models. The LMR performed better than the NCS, but not as well as the BLRT. If one had to choose one of the IC indexes,

the BIC would be the tool that seems to be the best indicator of the number of classes. The BLRT would be chosen over the LMR because of its consistency in choosing the correct class model. Overall, by comparing the results in Tables 7 and 8 across all models and sample sizes, the BLRT is the statistical tool that performs the best of all the indexes and tests considered for this article.

The BLRT, however, does have its disadvantages. It should be noted that when using the BLRT, the computation time increased 5 to 35 times in our examples. Another disadvantage of the BLRT approach is that it depends on distributional and model assumptions. The replicated data sets are generated from the estimated model and have the exact distributions as the ones used in the model. Thus, if there is a misspecification in the model or the distributions of the variables, the replicated data sets will not be similar in nature to the original data set, which leads to incorrect p value estimation. For example, if data within a class are skewed but modeled as normal, the BLRT p value might be incorrect. Outliers can also lead to incorrect p value estimates. In addition, the BLRT cannot currently accommodate complex survey data. Similarly, the various ICs depend on the model, distributional and sampling assumptions. On the other hand, the LMR is based on the variance of the parameter estimates, which are robust and valid under a variety of model and distributional assumptions and can accommodate complex survey data. Thus, the LMR may be preferable in such contexts. Our simulations, however, did not evaluate the robustness of the tests and more research on this topic is needed.

Recommendations for Practice

Due to the increased amount of computing time of the BLRT, it might be better to not request the BLRT in the initial steps of model exploration. Instead, one could use the BIC and the LMR p values as guides to get close to possible solutions and then once a few plausible models have been identified, reanalyze these models requesting the BLRT. Further, although it is known that the likelihood value cannot be used to do standard difference testing, the actual value of the likelihood can be used as an exploratory diagnostic tool to help decide on the number of classes in the following way.

Figure 5 displays four plots of log likelihood values for models with differing numbers of classes that can be used as a descriptive for deciding on the number of classes. Looking at the upper left panel, the 8-item LCA with categorical outcomes ($n = 1{,}000$), we see a pattern of the likelihood increasing by a substantial amount when moving from two classes to three classes, and also when we move from three classes to four classes. Then there is a flattening out when moving from four classes to five classes and similarly when moving from five classes to six classes. The flattening out of the lines between four and five classes suggests that there is a nonsubstantial increase in the likelihood when

you increase from four to five. In addition, we observed a flattening out that happens at the correct point for the 8-item model because we know that it is a true $k = 4$ class model. We see a similar pattern for almost all modeling settings. The 10-item, $n = 200$ model does not have as dramatic of a flattening out as the others, but as discussed before this is the most difficult modeling setting. This is also observed for the GMM model plot. Although there are only four plots included, the general findings suggest that the log likelihood plot fairly consistently identifies the correct model for the LCA models with $n = 500$ and $n = 1{,}000$. Although we know we cannot test these log likelihood values for difference testing using conventional methods, this plot is a way to use the likelihoods as a descriptive tool when exploring the number of classes.

It is important to note that the selection of models for this study did not allow conclusions to be made about the specific model and sample attributes' impact on the results. For example, we do not generalize about the performance of these tests and indexes for a particular type model structure, like the simple structure model. Rather, model and sample selection were motivated by wanting to explore a range of mixture models to understand the performance of the LRTs and the ICs. Future studies could aim to better understand the performance of the indexes for class enumeration and interrelation between the structure of the models, nature of the outcomes (categorical vs. continuous), and number of items.

CONCLUSIONS AND FUTURE DIRECTIONS

This study explored the performance of IC and likelihood-based tests to identify the correct number of classes in mixture modeling. Among the tools considered in this study, results indicated that the BLRT outperformed the others. Second best was the BIC, followed by the adjusted BIC. Although previous research has looked at a variety of fit indexes and tests for deciding on the number of classes, this article is one of the first to closely examine the performance of the BLRT method for these types of mixture models. By considering LCA models with both categorical and continuous outcomes as well as a limited number of FMA and GMM models, we expand the understanding of the performance of the ICs and likelihood-based tests beyond what has been examined before. These results, however, are merely a preview. More studies looking at the BLRT and its performance over a wider range of LCA models (e.g., including covariates, differently structured models, and the combination of continuous and categorical outcomes) should be considered before broad statements about its utility are made. Nevertheless, these results contribute to a further understanding of how to decide on the number of classes in mixture models by providing more insight into the performance of these indicators.

ACKNOWLEDGMENTS

Karen L. Nylund's research was supported by Grant R01 DA11796 from the National Institute on Drug Abuse (NIDA) and Bengt O. Muthén's research was supported by Grant K02 AA 00230 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA). We thank Mplus for software support, Jacob Cheadle for programming expertise, and Katherine Masyn for helpful comments.

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317–332.

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

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

Bucholz, K., Hesselbrock, V., Heath, A., Kramer, J., & Schuckit, M. (2000). A latent class analysis of antisocial personality disorder symptom data from a multi-centre family study of alcoholism. *Addiction, 95,* 553–567.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13,* 195–212.

Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research, 28,* 375–389.

Formann, A. K. (2003). Latent class model diagnostics-A review and some proposals. *Computational Statistics & Data Analysis, 41,* 549–559.

Garrett, E. S., & Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics, 56,* 1055–1067.

Greenbaum, P., Del Boca, F., Darkes, J., Wang, C.-P., & Goldman, M. (2005). Variation in the drinking trajectories of freshmen college students. *Journal of Consulting and Clinical Psychology, 73,* 229–238.

Hagenaars, J., & McCutcheon, A. (Eds.). (2002). *Applied latent class analysis models.* New York: Cambridge University Press.

Jedidi, K., Jagpal, H., & DeSarbo W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science, 16,* 39–59.

Jeffries, N. (2003). A note on "Testing the number of components in a normal mixture." *Biometrika, 90,* 991–994.

Keel, P., Fichter, M., Quadflieg, N., Bulik, C., Baxter, M., Thornton, L., et al. (2004). Application of a latent class analysis to empirically define eating disorder phenotypes. *Archives of General Psychiatry, 61,* 192–200.

Lazarsfeld, P., & Henry, N. (1968). *Latent structure analysis*. New York: Houghton-Mifflin.

Li, F., Duncan, T., Duncan, S., & Hops, H. (2001). Piecewise growth mixture modeling of adolescent alcohol use data. *Structural Equation Modeling, 8,* 175–204.

Lo, Y., Mendell, N., & Rubin, D. (2001). Testing the number of components in a normal mixture. *Biometrika, 88,* 767–778.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10,* 21–39.

Magidson, J., & Vermunt, J. (2004). Latent class models. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 175–198). Newbury Park, CA: Sage.

McCutcheon, A. C. (1987). *Latent class analysis*. Beverly Hills, CA: Sage.

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York: Wiley.

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345–368). Newbury Park, CA: Sage.

Muthén, B. (2006). Should substance use disorders be considered as categorical or dimensional? *Addiction, 101*(Suppl. 1), 6–16.

Muthén, B., & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors, 31,* 1050–1066.

Muthén, B., & Asparouhov, T. (2007). Growth mixture analysis: Models with non-Gaussian random effects. Forthcoming in Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (eds.), *Advances in Longitudinal Data Analysis.* Chapman & Hall/CRC Press.

Muthén, B., Asparouhov, T., & Rebollo, I. (2006). Advances in behavioral genetics modeling using M*plus*: Applications of factor mixture modeling to twin data. *Twin Research and Human Genetics, 9,* 313–324.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55,* 463–469.

Muthén, L., & Muthén, B. (1998–2006). *M*plus *user's guide* (4th ed.). Los Angeles: Muthén & Muthén.

Muthén, L., & Muthén, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4,* 599–620.

Rasmussen, E., Neuman, R., Heath, A., Levy, F., Hay, D., & Todd, R. (2002). Replication of the latent class structure of attention-deficit/hyperactivity disorder (ADHD) subtypes in a sample of Australian twins. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 43,* 1018–1028.

Roeder, K., & Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association, 92,* 894–902.

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

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

Soromenho, G. (1993). Comparing approaches for testing the number of components in a finite mixture model. *Computational Statistics, 9,* 65–78.

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

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57,* 307–333.

Yang, C. (2006). Evaluating latent class analyses in qualitative phenotype identification. *Computational Statistics & Data Analysis, 50,* 1090–1104.

APPENDIX

The BLRT is implemented in M*plus* Version 4.1 as follows. To minimize computing time for estimating the p value, a sequential stopping rule is used. The sequential stopping rule consists of a number of lower or upper stopping points (n_i, p_i). We continue to generate and analyze data until the stopping rule is in effect. The replication process is terminated at an upper stopping point (n_i, p_i) if after n_i replications the current estimate of the p value is greater or equal to p_i. The replication process is terminated at a lower stopping point (n_i, p_i) if after

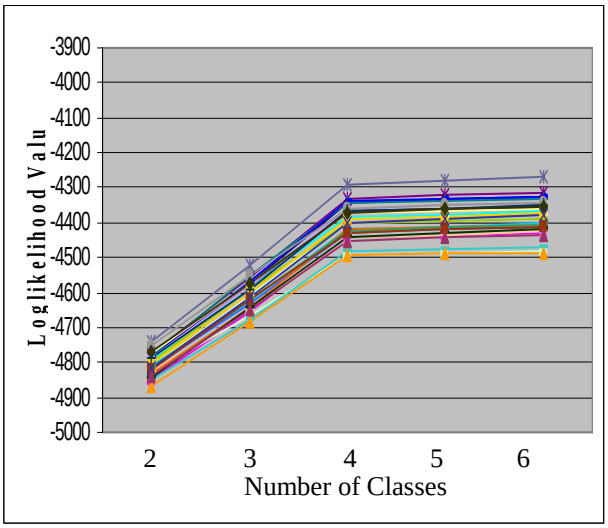
TABLE A.1
Probability for p Values Less Than 10%

True p Value	*Probability of Agreement Between Sequential Procedure and Infinite Replication Procedure*
1%	100%
2%	95%
3%	82%
4%	64%
6%	72%
7%	83%
8%	90%
9%	95%
10%	97%

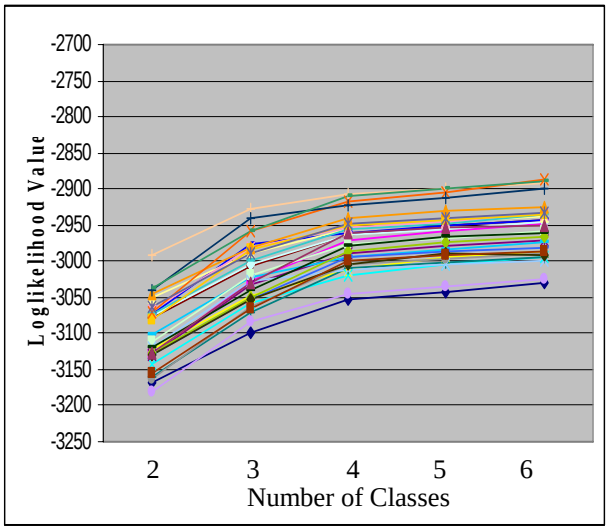
n_i replications the current estimate of the p value is less than or equal to p_i. The upper stopping points we utilize are as follows: $(n, 2/n)$, for $n = 2, 3$; $(n, 3/n), n = 4, \ldots, 9$; $(n, 4/n), n = 10, \ldots, 17$; $(n, 5/n), n = 18, \ldots, 26$; $(n, 6/n)$, for $n = 27, \ldots, 99$; and (100, 0). The lower stopping points we utilize are as follows: (49, 0), (78, 1/78). In addition, we use conditional lower stopping points (n_i, p, s_i). The replication process is terminated at a conditional lower stopping point if after n_i replications the current estimate of the p value is less than or equal to p_i and the LRT statistic is more than s standard deviation units away from the mean of the LRT distribution obtained from the first n_i replications. The conditional lower stopping points we utilize are as follows: (5, 0, 20), (10, 0, 10), and (20, 0, 5). By using these conditional stopping points, we implicitly assume that the LRT distribution does not severely deviate from a normal distribution; however, usually this is the case.

With certain probability, this stopping rule gives the same result in terms of rejecting or accepting the null hypothesis, as would the procedure based on infinitely many replications. The conditional lower stopping points make these probabilities dependent on the LRT distribution. If we exclude these stopping points from the stopping rule, however, we can obtain approximate probabilities of agreement. If the true p value is greater than 10%, this probability is close to 100%. If the true p value is less than 10%, the probability is given in Table A.1. When the p value is 5% or close to 5%, this procedure equally likely estimates a p value above or below 5% and thus the results should be interpreted as being inconclusive.

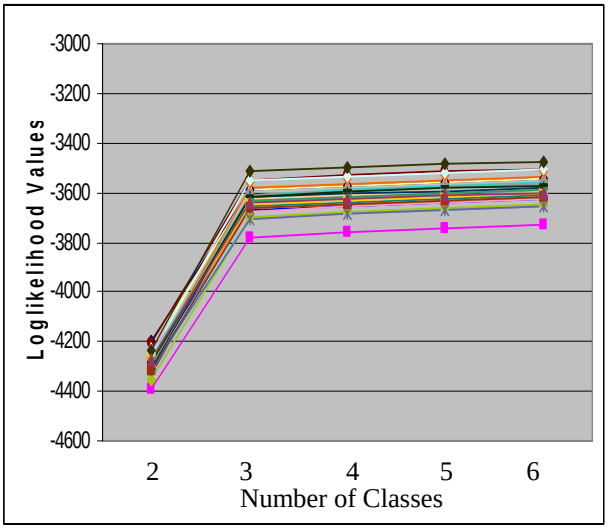
8-item, Equal Class LCA with
Categorical Outcomes (k = 4), n = 1,000



10-item, Unequal Class LCA with
Continuous Outcomes (k = 4), n = 200



15-item, Equal Class LCA with
Categorical Outcomes (k = 3), n = 500



4-item GMM with
Continuous Outcomes (k = 3), n = 500