The Journal of Experimental Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/vjxe20

Multilevel Latent Class Analysis: Parametric and Nonparametric Models

W. Holmes Finch[a] & Brian F. French[b]
[a] Ball State University
[b] Washington State University
Published online: 20 Sep 2013.

To cite this article: W. Holmes Finch & Brian F. French (2014) Multilevel Latent Class Analysis: Parametric and Nonparametric Models, The Journal of Experimental Education, 82:3, 307-333, DOI: 10.1080/00220973.2013.813361

To link to this article: http://dx.doi.org/10.1080/00220973.2013.813361

THE JOURNAL OF EXPERIMENTAL EDUCATION, *82*(3), 307–333, 2014

ISSN: 0022-0973 print/1940-0683 online
DOI: 10.1080/00220973.2013.813361

Multilevel Latent Class Analysis: Parametric and Nonparametric Models

W. Holmes Finch
Ball State University

Brian F. French
Washington State University

Latent class analysis is an analytic technique often used in educational and psychological research to identify meaningful groups of individuals within a larger heterogeneous population based on a set of variables. This technique is flexible, encompassing not only a static set of variables but also longitudinal data in the form of growth mixture modeling, as well as the application to complex multilevel sampling designs. The goal of this study was to investigate—through a Monte Carlo simulation study—the performance of several methods for parameterizing multilevel latent class analysis. Of particular interest was the comparison of several such models to adequately fit Level 1 (individual) data, given a correct specification of the number of latent classes at both levels (Level 1 and Level 2). Results include the parameter estimation accuracy as well as the quality of classification at Level 1.

Keywords *latent class, multilevel, nonparametric*

LATENT CLASS ANALYSIS (LCA), introduced in 1968 by Lazaesfeld and Henry, is a popular analytic technique for identifying latent groups in a population, typically using a set of observed dichotomous or polytomous response variables. An examination of the ERIC database in September of 2011, for example, using the keywords *latent class analysis* revealed 380 articles in areas as diverse as educational psychology, educational leadership, special education, child abuse and neglect, and substance abuse where this technique has been used. Use of LCA in this sample of articles included the examination of latent patterns of risk behavior in middle school students (Hedden, Whitaker, & von Thomsen, 2011), the identification of adolescent smoking typologies with hierarchical structured data (Henry & Muthén, 2010), the identification of groups of health professionals in terms of health status (Hernandez, Blavo, Hadigan, Perez, & Hage, 2010) and methodological studies regarding the most appropriate approach for applying LCA (e.g., Halpin & Mauaun, 2010). This cursory review provides a glimpse of how LCA is used to examine important applied questions as well as investigate methodological questions to inform application.

Address correspondence to W. Holmes Finch, Department of Educational Psychology, Ball State University, Muncie, IN 47306, USA. E-mail: whfinch@bsu.edu; or, Brian F. French, Department of Educational Leadership and Counseling Psychology, 353 Cleveland Hall, Washington State University, Pullman, WA 99164, USA. E-mail: frenchb@wsu.edu

The application of LCA also is being extended to examine multilevel data typically found in educational settings (student in classrooms) to more unique settings such as in the medical arena to account for patient case mix or nesting of types of patients treated by a unit (Gilthorpe, Harrison, Downing, Forman, & West, 2011), and to the evaluation of long-term care facilities (Montanari, Ranalli, & Eusebi, 2010). Given interest in identifying latent classes in a variety of research setting where data are naturally structured in a multilevel manner, there is little doubt that the applications of multilevel latent class analysis (MLCA) and methodological studies of MLCA will flourish in the next several years.

LCA is typically conducted in an exploratory manner in which a priori hypotheses regarding the number or nature of the latent classes underlying the data may not be given (Hoijtink, 2001). In such cases, several proposed models are applied, each differentiated by the number of latent classes, and the relative fit of these are compared in order to identify the optimal number of classes. Comparative fit is assessed using indexes designed to determine which model best approximated the observed data. Of particular utility in this regard is the sample size adjusted Bayesian information criterion (aBIC; Tofighi & Enders, 2007), the Lo-Mendell-Rubin (Lo, Mendell, & Rubin, 2001; Nylund et al., 2007) test, and the Bootstrap Likelihood Ratio Test (McLachlan & Peel, 2000). Beyond these tests and criterion, the selection of an optimal model also needs to incorporate the substantive coherence of its latent classes (Bauer & Curran, 2004), just as interpretability is a key criterion in determining the optimal number of factors in a factor analysis solution, for example. That is, the final solution should be defensible on the basis of the types of individuals that have been grouped together, in terms of the indicators used in the LCA, as well as other potentially pertinent variables, such as demographic characteristics.

The process of identifying the correct latent class solution becomes more complex in the presence of multilevel data where the assumption of independence is violated. Much work in this area (e.g., Lukociene, Varriale, & Vermunt, 2010; Lukociene & Vermunt, 2002) has focused on methods for correctly determining the number of latent clusters at different levels in a multilevel context. In contrast, the goal of the present simulation study was to examine the performance of different approaches for parameterizing the MLCA models with regard to the fit of the data given a correct specification of the number of latent classes at Levels 1 and 2. In the following paragraphs, we describe the basic LCA model and alternative parameterizations that account for the presence of a multilevel structure. We outline the method we followed in order to answer our research questions, and then describe our results. Last, we offer conclusions regarding the implication of these results for practice as well as areas that would benefit from additional methodological investigations.

The Latent Class Model

The one-level LCA model is described by McCutcheon (2002) and is reviewed briefly. As an example, assume that data have been collected for four observed dichotomous variables, *X1*, *X2*, *X3*, and *X4* and a latent categorical variable *Y* exists which accounts for relationships among these four observed variables. The LCA model linking the latent and observed variables is as follows:

$$\pi_{ijklt}^{X1X2X3X4Y} = \pi_t^{Y}\pi_{it}^{X1|Y}\pi_{jt}^{X2|Y}\pi_{kt}^{X3|Y}\pi_{lt}^{X4|Y}$$

where

π_t^{Y} = Probability that a randomly selected individual will be in latent class t of latent variable Y

$\pi_{it}^{X1|Y}$ = Probability that a member of latent class t will provide a response of i for observed variable $X1$

$\pi_{jt}^{X2|Y}$ = Probability that a member of latent class t will provide a response of j for observed variable $X2$

$\pi_{kt}^{X3|Y}$ = Probability that a member of latent class t will provide a response of k for observed variable $X3$

$\pi_{lt}^{X3|Y}$ = Probability that a member of latent class t will provide a response of l for observed variable $X4$

The LCA model in Equation 1 asserts that the observed variables are conditionally independent of one another given a particular level of Y (Goodman, 2002). The model allows for the estimation of two types of parameters: (a) the probability of a particular response for an observed variable conditional on latent class membership, and (b) the probability of being in a specific latent class, t.

Multilevel LCA (MLCA)

A primary assumption underlying standard LCA models is that individual observations are independent of one another, given their membership in a specific latent class (Vermunt & Magidson, 2002). However, in many applications, particularly in educational research, participants (Level 1) are sampled in clusters such as classrooms or communities (Level 2), inducing correlated data among observations from the same clusters (Asparouhov & Muthén, 2008). When such data structures are present, models that account for this dependence should be used to avoid parameter estimation inaccuracy, biased standard errors, and inflated Type I error rates for hypothesis tests regarding the parameters (Asparouhov & Muthén, 2008; Vermunt, 2003, 2008). A number of such models have been described in the literature, along with examples of their application in practice (e.g., Gilthorpe et al., 2011; Montanari et al., 2010). However, less methodological attention has been given to their performance in regards to parameter estimation, and the quality of their classification of individuals at Level 1. The primary goal of this study was to ascertain how well each of these models could recover parameter values at Level 1, as well as the quality of group classification. Henry and Muthén (2010) provided a thorough description of these models. Therefore, what follows is a condensed review of the MLCA models that we used in this study (for more details on the models, see Henry and Muthén, 2010).

MLCA Models

Each of the four multilevel latent class models that are considered in this study appears in Figure 1. In the following paragraphs, we provide a conceptual discussion of these models. The reader interested in a more technical description is encouraged to refer to past studies[1]. For this discussion, we assume a MLCA model with six within-cluster indicator variables and three latent classes. This model can generalize to any number of indicators and any number of latent classes. However, for simplicity and clarity of explanation we limit discussion to this example case. Last, this graphical presentation of the models is based on a similar description provided in Henry and Muthén (2010).

All MLCA models have a Level 1 (within-cluster) component (e.g., students) nested within a Level 2 (between-cluster) component (e.g., schools). For the parametric model, which appears graphically in panel 1 of Figure 1, the six indicator variables are used to determine the within-cluster latent class solution. In turn, each of the within-cluster latent classes has a random mean that is itself a latent variable and is allowed to vary between clusters. The mean for a particular latent class can be thought of as the proportion of individuals who belong to that specific latent class. These means also have associated with them random error, denoted in Figure 1 as U. This random error represents the inherent uncertainty in estimating latent class membership using sample data. In other words, for each individual in the sample, a latent class membership is estimated on the basis of their values for the indicator variables. However, these latent class assignments are made with a degree of uncertainty so that some (hopefully a small number) individuals will be placed in a class incorrectly. In turn, the estimated proportion of individuals in each class based on the sample will therefore not be equal to the actual population latent class membership because of the aforementioned misclassification of some individuals. This inaccuracy, or error, is represented by U, and may differ between clusters. In summary, the parametric model represents the data by a standard latent class model within clusters, each class of which has a different proportion of individuals belonging to it, and this proportion is allowed to differ between clusters.

The parametric one-factor model (Figure 1, panel 2) also represents the within-cluster latent class solution as a function of the observed indicator variables. However, in this case, rather than simply estimating the latent class proportions and then utilizing them independently to characterize the structure of the data at Level 2, the proportions of the within-cluster latent classes are associated with a single common factor through a factor analysis. In other words, the proportions of membership for each latent class within each of the clusters are taken to be indicators of a single overall unobserved variable that represents the combined latent class structure within each cluster. It is important to note that this factor is not interpretable as a latent class itself, but rather represents mathematically the combined latent class proportions. This parametric one-factor model assumes that the proportions of the within-cluster latent classes are highly correlated with one another (Henry & Muthén, 2010), and can thus be well represented by a single common factor. Another way to think about this notion of correlated latent class proportions is that within a given cluster, we might expect to see certain combinations of latent classes appear more frequently with one another, and this specific mixing of classes will be represented in the factor loadings linking class proportions to the single factor. Furthermore, the model assumes that there is no error variation associated with each latent class proportion separately as we saw in the parametric model, but rather only one error term, U, for the common

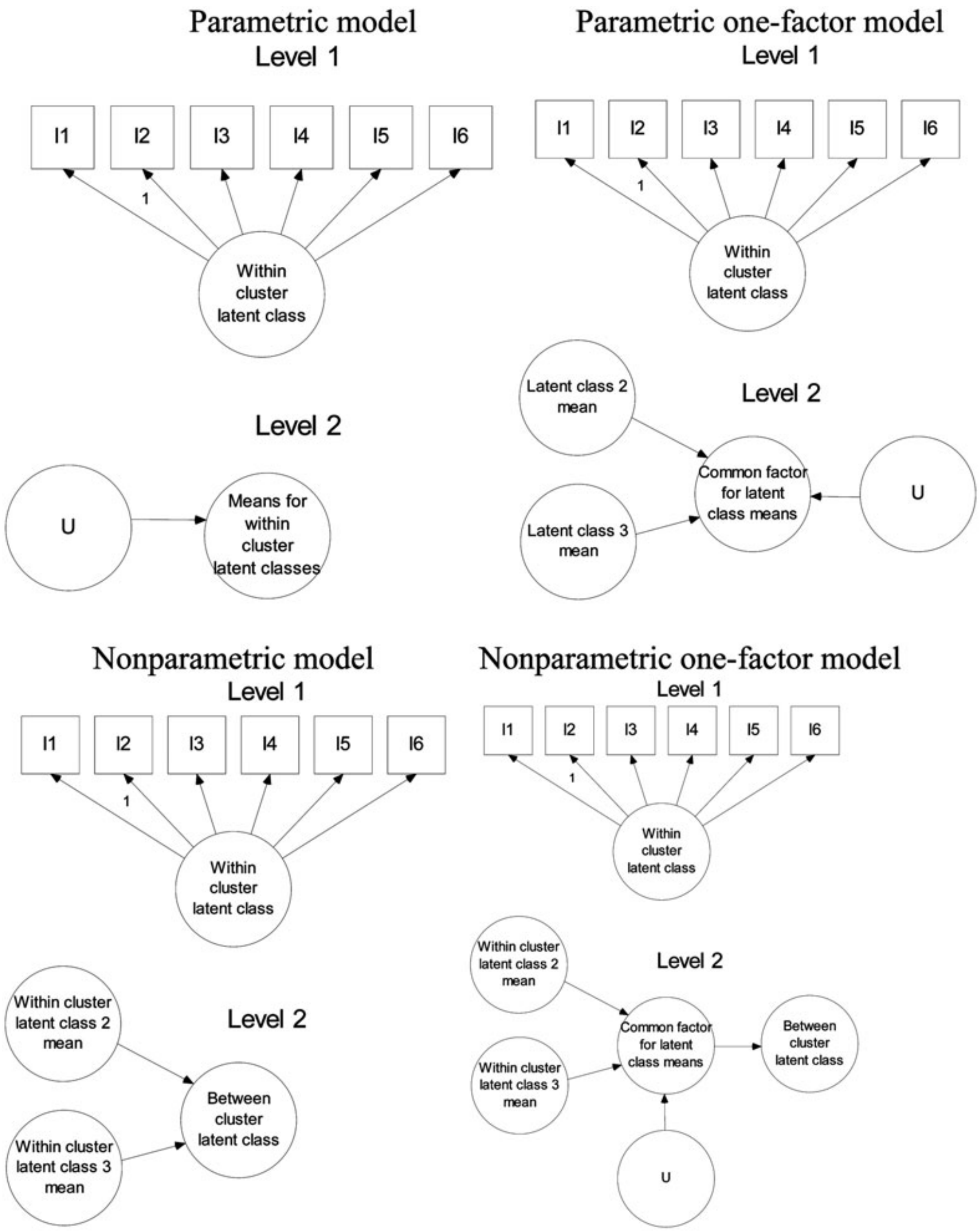


FIGURE 1 Multilevel latent class models. There will be T–1 latent class means, where T = number of latent classes. Latent class 1 will always have a mean of 0.

factor representing the combined latent structure. The parametric one-factor model has as an advantage the reduction of dimensionality at Level 2 from separate latent class proportions to a single factor, making it computationally less demanding than the parametric model. However, as Henry and Muthén (2010) noted, there has not been an empirical investigation regarding the relative performance of this simplified model with other representations of the MLCA. Such a comparison is one of the main goals of this study.

In contrast with the two parametric models, which express the between-cluster structure in the form of within-cluster latent class proportions that vary from cluster to cluster, the nonparametric approach model (panel 3 in Figure 1) expresses between-cluster structure in terms of latent classes at Level 2. Just as the observed indicator variables measured for each individual in the sample are used to model the within-cluster latent class structure, the proportions of the within-cluster latent classes themselves serve as the indicators of the between-cluster latent class variable. In other words, each individual at Level 1 (e.g., student) is placed into one of the within-cluster latent classes on the basis of their values for the indicator variables. Next, each cluster (e.g., school) is placed into one of the between-cluster latent classes on the basis of the proportion of individuals in that cluster who are members of each Level 1 latent class. As a result, individuals within each cluster are assigned to a Level 1 latent class, and each cluster is assigned to a Level 2 latent class.

The nonparametric one-factor model (Figure 1, panel 4) includes a between-cluster common factor, which, in turn, is used to characterize the between-cluster latent class variable. The between-cluster common factor is identical to that in the parametric one-factor model, and the between-cluster latent class variable is identical to that in the standard nonparametric model. In this case, the clusters are assigned to Level 2 latent classes on the basis of their values on the single factor, rather than on the individual latent class proportions, as was the case for the nonparametric model. The reduced dimensionality of the one-factor nonparametric model makes it somewhat less computationally demanding compared with the nonparametric model, which is a potential advantage for problems involving large samples and many indicator variables (Henry & Muthén, 2010).

Goal of the Present Study

Although there is a broad array of literature focusing on standard LCA models (e.g., Hagenaars & McCutcheon, 2002; Huang, Wang, & Hsu, 2011; Kaplan & Keller, 2011; Schrepp, 2003), relatively little work has assessed the performance of the MLCA models described earlier. Moreover, no simulation studies could be identified that have evaluated the performance of the various parametric and nonparametric models. Henry and Muthén (2010) provided an excellent example applying the MLCA models to existing data and report that the parametric model with a common factor for the Level 1 indicators provided the best fit to the observed data. However, information from a simulation investigation would assist the practitioner in deciding which model to apply given the data at hand. Therefore, the goal of the present simulation study was to build on earlier work in the area of MLCA (e.g., Asparouhov & Muthén, 2008; Henry & Muthén, 2010; Lukociene, Varriale & Vermunt, 2010) by focusing on the ability of the parametric and nonparametric models (and their Level 2 single-factor variants) to adequately fit the data given a correct specification of the number of latent classes at both levels in a simulation study where conditions are controlled. That is, we investigated the situation where a researcher has used the methods described in Lukociene and colleagues (2010) to correctly identify the number of Level 1 and 2 classes and attempt to shed light on how well the different MLCA parameterizations estimate Level 1 parameters and group membership under various conditions.

METHOD

In this simulation study (1,000 replications per condition) performance of the MLCA models was compared in terms of model fit, quality of group classification and recovery of model parameters under a variety of conditions, assuming that the correct number of classes has been identified. Data were simulated and modeled using Mplus 5.2 (Muthén & Muthén, 2008), with a population intraclass correlation (ICC) of 0.15 for the indicator variables. The ICC in educational research tends to range approximately from 0.05 to 0.25 based on large national databases (Hedges & Hedberg, 2007). Moreover, it is the middle range of ICC values that seem to be where results reveal the effects of multilevel data structure (e.g., Finch & French, 2011; French & Finch, 2010) whereas the extreme low (0.05) and high (0.45) ICC values tend to act in very predictable ways. Thus, to keep the study manageable in terms of number of conditions to run and interpret and have findings that reside where interest may most likely be, the ICC value was fixed at 0.15. All of the simulated indicators were dichotomous in nature with a population correlation ranging between 0.3 and 0.5, the value of which for a given pair was determined randomly. This range of values was selected so as to reflect a nontrivial relationship among the indicators, without making them collinear with one another.

Manipulated Factors

Level of Latent Class and Covariate

We included within (Level 1) and between (Level 2) latent class variables in the study. Latent class types were not mixed so that a given replication had only a set of Level 1 or a set of Level 2 cluster indicators, but not both. In addition, a normally distributed covariate associated with the latent class variable was simulated as Level 1 or Level 2, corresponding to the level of latent class.

Number of Level 2 Units (Clusters)

The number of clusters was set at 25, 75, 150 and 200. These values were selected because they encompass the range of clusters reported in published studies using MLCA or similar multilevel latent variable models (e.g., Henry & Muthén, 2010; Komro, Tobler, Maldonado-Molina, & Perry, 2010; Palardy, 2008).

Number of Latent Classes and Indicator Variables

Data were simulated with two or three latent classes at Level 1 or Level 2, with classes being of the same size. All indicator variables in the simulations were dichotomous, with the number being 5, 10, or 15, reflecting values used in practice (Henry & Muthén, 2010; Rindskopf, 2006; Vermunt, 2003, 2008). The threshold values for the indicators were based on those used in previous simulation work (i.e., Lukociene & Vermunt, 2010), such that in the two latent class condition, one group had a 0.8 probability of endorsing each indicator, whereas the other group had a 0.2 probability of indicator endorsement. In the three latent class cases, these patterns were repeated for two of the groups, whereas the third group had a 0.8 probability of endorsing the first half of the indicators and a 0.2 probability of endorsing the second half. These conditions were intended to reflect moderately separated classes (Lukociene & Vermunt, 2010). In addition, using similar values should assist with generalizability across studies and conditions.

Sample Size Per Cluster

The sample size per cluster was set at 10, 20, and 35 individuals, with all clusters being equal in size. In combination with the number of clusters, these values produced total sample sizes that were in range of the published studies referenced earlier, ranging from the smallest size of 250 (25 clusters with 10 participants each) to the largest of 7,000 (Kuntsche, Kuendig, & Gmel, 2008; Johnson, Burlingame, Olsen, Davies, & Gleave, 2005; Liu, Wang, Zhang, & Lee, 2011). In addition, these values do represent what an applied researcher may encounter in settings ranging from settings such as Head Start centers or preschool classrooms to large classroom in a high schools to small sections in a university setting or even an a organizational unit. The sample size per cluster was held constant in this study to determine how the various models performed in what might be viewed an ideal case. However, once this question is answered, future research should expand upon the present work by varying the sample size per cluster.

MLCA Models

Each replication was modeled using a naive LCA model ignoring multilevel data, parametric, parametric one-factor, nonparametric, and nonparametric one-factor model, as described earlier. Thus, we examined a total of five models for comparison of best fit when accounting for and not accounting for the simulated multilevel structure.

Simulation Outcome Variables

The outcomes of interest were (a) identification of the best fitting model based on the sample size aBIC, as per recommendations (Nylund, Asparouhov, & Muthén, 2007); (b) overall classification quality in the form of entropy; (c) the posterior probability of group membership at Level 1; (d) coverage rates for Level 1 and Level 2 indicator threshold parameters; and (e) coverage rates for the coefficient for a Level 1 or Level 2 covariate. The following is a brief description of how each of these statistics is interpreted in practice and how they were used in this study.

The aBIC is a measure used to determine which of a set of models provides the best fit to a set of data. The aBIC is based on a measure of unexplained variance from a model, with a penalty for model complexity. In practice, a researcher will select the model with the lowest aBIC value. For the purposes of this study, the aBIC was used in each replication to identify which of the five models being compared provided the best fit. Thus, the outcome variable was the proportion of replications for which each of the models provided the best fit to the data.

The second outcome of interest was entropy, which indicates the degree of uncertainty in the posterior classification of individuals into latent classes. Entropy in LCA is essentially a weighted average of the posterior probability of each individual's membership in the latent class to which they have been assigned (Ramaswamy, DeSarbo, Reibstein, & Robinson, 1993), where higher posterior probability values correspond to a more confident prediction of class membership. Values can range between 0 and 1, with values close to 1 indicative of higher certainty, or confidence in the classification of individuals to the latent classes. For this study, entropy was recorded for each replication for each model, and then compared across models to determine which provided the greatest certainty in classification across latent classes.

Entropy serves as a holistic measure of classification quality, and is a useful index in practice to examine the posterior probability of membership in each of the latent classes for a given solution. Higher values of these probabilities for a given latent class indicate that individuals assigned to that class were assigned with a relatively low probability for error (i.e. higher confidence in latent class membership). In contrast, lower posterior probabilities suggest greater uncertainty for assigning latent class membership. The mean posterior probabilities for each latent class for each replication were compared across models to ascertain which of them provided the greatest certainty in classification for each class separately.

In addition to overall model fit, and classification quality, two additional outcomes are of interest to researchers conducting LCA. The first is the coverage rate for threshold parameters of the categorical indicator variables. Thresholds are directly related to the probability of a given value for a categorical variable. In the current study, the observed categorical variables were all dichotomous (i.e., each had one threshold value). The threshold is directly associated with the probability that an individual will produce a value of 1 for the variable (e.g., give a correct response on a test item, or endorse a yes/no item). In a Monte Carlo study, the probability of endorsing a given item is simulated with a known value for the threshold (see the description of how these probabilities were simulated using threshold parameters). Therefore, it is possible to use the sample estimate of the threshold produced by a given model for a given replication of the simulation as a direct measure of how accurately the model can reproduce the probability of a specific item response within a latent class. Thus, an additional outcome of interest was the coverage rate for a 95% confidence interval on the threshold for each observed variable within each latent class. This value is obtained by calculating the 95% confidence interval for a threshold for each model for each simulation replication. When a model is working well, this confidence interval should include the actual data generating (population) threshold in 95% of the simulation replications. The observed proportion of replications for which the population value is actually contained in these intervals is known as the coverage rate, and serves as one of the outcomes in the study. The coverage rate should be within sampling error of 0.95.

The final outcome of interest was the coverage rate for the parameter linking a continuous covariate with latent class membership. As with the individual variable thresholds, the relationship between the covariate was set in the simulation, in this case at a value of 1. For each replication this relationship was then modeled using the simulated data, and the 95% confidence interval was calculated. The outcome of interest was the proportion of replications for which these confidence intervals included the actual population parameter value of 1 (i.e., the coverage rate for the parameter linking the covariate with the latent class variable).

Analysis of Simulation Results

To ascertain which of the manipulated factors were related to the simulation outcome variables, a generalized logits model for the best fitting model (dependent variable = best fitting model), and repeated measures factorial analysis of variance were used (dependent variable = entropy, overall posterior probability accuracy, and coverage rates for thresholds and covariate). The choice to use a repeated measures factorial analysis of variance for identifying significant main effects and interactions of the manipulated variables was based on the structure of the data, which included both within and between-subjects factors in the analysis. In this case, the subjects were the individual simulations, which consisted of unique combinations of the between-subjects conditions. Each of these simulations had outcome variables of interest for the five methods of estimating the models, the parametric, nonparametric, parametric one-factor, and nonparametric one-factor models. In addition, there were also several between-subjects factors, including sample size, number of clusters, number of indicators, number of latent classes, and level of data. This combination of within and between-subjects (simulations) factors led us to select a repeated measures factorial analysis of variance.

Assumptions of the various analyses were assessed and found to have been met. Given the large number of replications per condition (1,000), the effect size (η^2), rather than the hypothesis test results, was used to identify main effects and interactions that were worthy of investigation. Specifically, an η^2 value of 0.10 or greater was required for an effect to be deemed important, in order to ensure that only terms contributing a minimum of 10% to the variation in the outcome variables were discussed. With regard to the logistic regression analysis of aBIC, odds ratios for the main effects and interactions were used to identify important terms. Our use of the combination of descriptive information, graphical representation, and effect sizes follows recommendations for Monte Carlo studies in the structural equation modeling environment (Paxton et al., 2001).

Verification of Simulations

An important aspect of conducting simulation research is the verification that the simulated data match the intended data structure. Thus, several preliminary analyses were conducted in the development phase of the simulation programs in Mplus. First, to ensure that the population ICC of 0.15 was simulated properly, multiple replications were sampled for each number of clusters, sample size per cluster, and number of latent class conditions. The estimated ICCs for these samples matched the population value ($M = 0.15$, $SD = 0.03$). Second, to ensure that the indicators were being generated with the appropriate interindicator correlation values, we condcuted sampling of replications. Across replications, indicator variable pairs consistently ranged between the population simulating values of 0.3 and 0.5. Third, we verified item response probabilities, following the same procedure as mentioned earlier, and matched the generating values. This process verified that both the intended multilevel structure of the data in form of the ICC, as well as the indicator threshold/response probabilities and correlations were generated as intended. This provided confidence that the simulations were executed as intended.

RESULTS

aBIC

A generalized logits model was used to identify the manipulated factors, and their interactions, that had an effect on the aBIC statistic in terms of its identification of the optimal model from among those studied here. To identify those terms that are most salient to understanding the accuracy of model selection using aBIC, we examined the odds ratios for each main effect and interaction of the manipulated factors. In addition, we included for discussion in his article those with a confidence interval not including 1. For each replication, the model with the lowest aBIC was categorized as the best fitting. Thus, the outcome of the generalized logits model was the LCA model that fit best, where the individual observations in the analysis consisted of the replications across all manipulated conditions. On the basis of the odds ratios from the multinomial logits model, the interactions and main effects worthy of discussion included level of latent class (level) by number of indicators, (indicators) OR = 24.33, model type (model) by sample size, OR = 2.085, and model by number of clusters, (clusters) OR = 6.85. Table 1 contains the proportion of replications for which each model provided the best fit for level by indicators, model by sample size, and model by clusters. Especially notable results are in bold.

With regard to level (appearing in each cell and separated by a forward slash) by indicators, for the within-cluster case the naive method provided the best fit most frequently, particularly in the presence of fewer indicator variables. With 15 indicators, the nonparametric (nonparm) approach provided the best fit almost as frequently as did the naive model. In contrast, for the between-cluster level, the parametric one-factor model (parmF1) consistently provided the best fit regardless of the number of indicators. In fact, this method provided the best fit to the data

TABLE 1
Proportion of Each Method Providing Best Fit by Model, Number of Indicators, Number of Clusters, Sample Size, and Level of Latent Class (Within/Between)

Model	*Naive*	*Parametric*	*Nonparametric*	*Parametric with one factor*	*Nonparametric with one factor*
Number of indicators					
5	**0.75** / 0	0.13 / 0	0 / 0	0.13 / **0.92**	0 / 0.08
10	**0.67** / 0	0 / 0	0 / 0	0.12 / **0.88**	0.21 / 0.13
15	**0.42** / 0	0.08 / 0.13	0.37 / 0	0.13 / **0.83**	0 / 0.04
Number of clusters					
25	**0.78** / 0	0 / 0.06	0.17 / 0	0 / **0.83**	0.06 / 0.11
75	**0.56** / 0	0 / 0	0.11 / 0	0 / **1.00**	0.06 / 0
150	**0.44** / 0	0.17 / 0.11	0.06 / 0	0.22 / **0.72**	0.11 / 0.17
200	**0.37** / 0	0.11 / 0	0.17 / 0	0.28 / **0.94**	0.07 / 0.06
N					
10	**0.58** / 0	0.08 / 0.04	0.13 / 0	0.13 / **0.88**	0.08 / 0.08
20	**0.54** / 0	0.13 / 0	0.13 / 0	0.17 / **0.88**	0.03 / 0.12
35	**0.71** / 0	0 / 0.08	0.13 / 0	0.09 / **0.88**	0.07 / 0.04

Note. Especially notable results appear in boldface.

in 83% or more replications across number of replications. In contrast, the naive model never provided the best fit in the between-cluster case.

In terms of the model by sample size interaction, the naive model provided optimal model fit in the within level condition, with the proportion of times it provided such optimal fit increasing concomitantly with increases in sample size per cluster. In the between-cluster variable level, parmF1 always yielded the optimal model based on aBIC. With respect to the model by clusters interaction, for within-cluster variables as the number of clusters increased the proportion of times that the naive model provided the lowest aBIC declined, whereas the proportion of lowest aBIC for the parmF1 model increased. The parmF1 provided the optimal model across the number of clusters when the variables were between clusters.

In summary, researchers whose data are comprised of within-cluster indicator variables should expect the aBIC criterion to identify the naive model as optimal in most cases, although its primacy in this regard may decline somewhat for problems involving relatively more clusters and more indicator variables. When the indicators are between cluster, the parmF1 model will most likely provide optimal fit, based on the aBIC.

Entropy

We used a repeated measures factorial analysis of variance with type of MLCA model as the within replications factor and number of latent classes, number of clusters, sample size per cluster, number of indicators, and level of latent class as between replications factors to ascertain which main effects and interactions of the manipulated study conditions significantly influenced entropy, a measure of group classification quality as described earlier. We examined the effect sizes for the main effects and interactions of the factors manipulated in the study. The highest order interaction with an effect size greater than 0.10 was model by number of clusters by number of indicator variables by level of latent class, $\eta^2 = 0.161$. All other main effects and interactions with an effect size more than 0.10 were subsumed in this interaction. Figure 2 includes mean entropy values, and the mean posterior probability of Group 1 membership by model, number of clusters, number of indicators and level of latent class. Note that the posterior probability means for the other groups were very comparable to that of Group 1, and are thus not included in these results for the sake of brevity. Also, as a reminder, entropy ranges between 0 and 1 with larger values indicating greater certainty regarding classification of individual cases, as do higher mean posterior probabilities for individual groups.

The results in Figure 2 reveal that for the within-cluster case, more indicators were associated with higher entropy and higher mean posterior probabilities. In other words, having more indicator variables yielded greater certainty in terms of group classification for within-cluster models. In addition, except for the five indicators condition, mean entropy and posterior probabilities were very similar across models. In contrast, for the between-cluster case, there were marked differences in both mean entropy and posterior probability across the models. For example, the naive model consistently had the lowest values of both variables, particularly when there were a larger number of clusters. In other words, the naive approach ignoring the multilevel structure of the data resulted in less certainty regarding classification of individuals, particularly when there were more clusters present. The two nonparametric models, in contrast, displayed the highest mean entropy and posterior probabilities for the between-cluster condition, across number of indicators and number of clusters. For all models, the entropy and posterior probabilities were

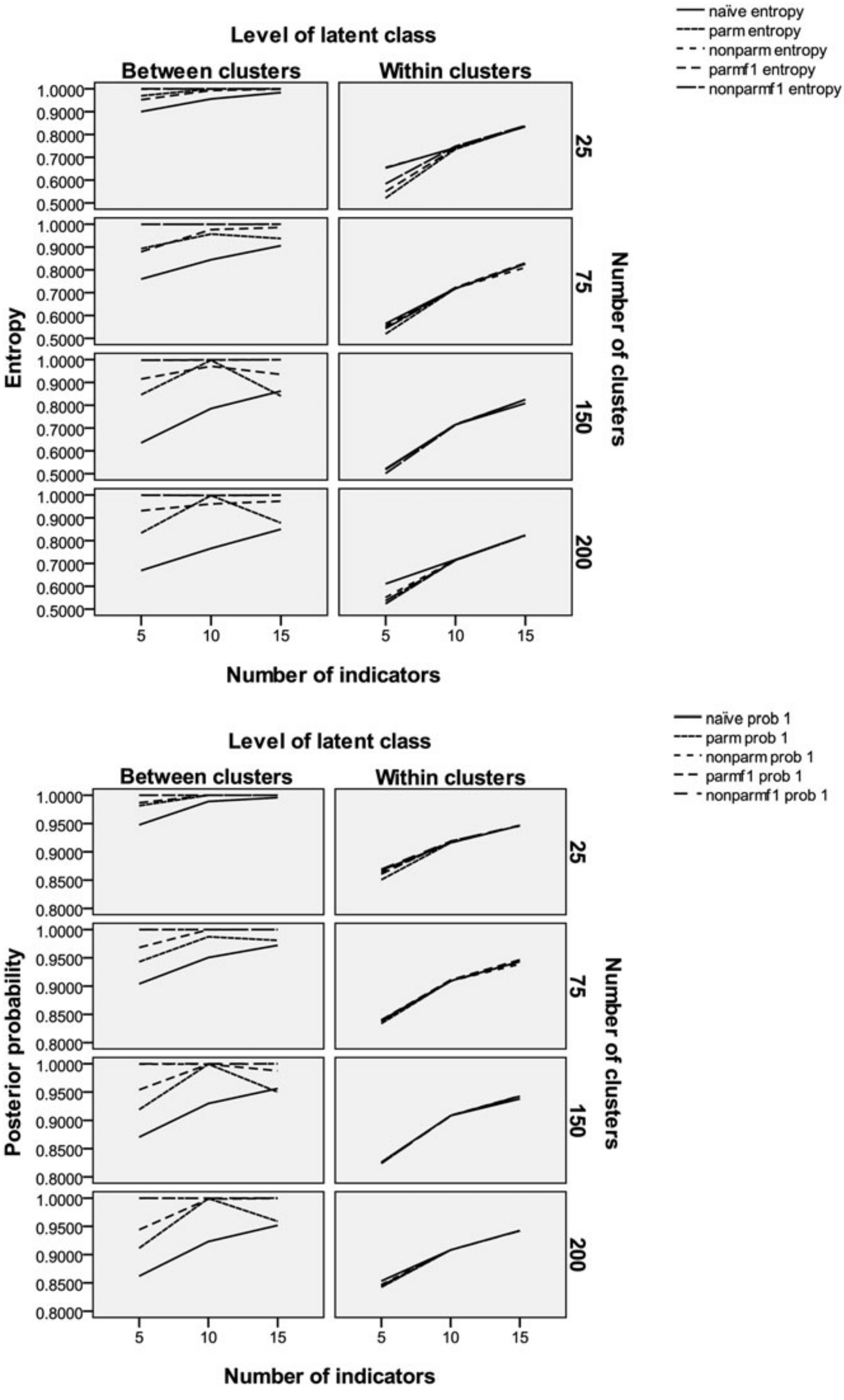


FIGURE 2 Mean entropy and posterior probability for Group 1 membership by model, number of clusters, number of indicators and level of latent class.

lower in the within-condition case than in the between-condition case. In short, both of the nonparametric approaches provided the least uncertain classification results in the between-cluster case, and all models were more uncertain in the within-condition case than in the between-condition case.

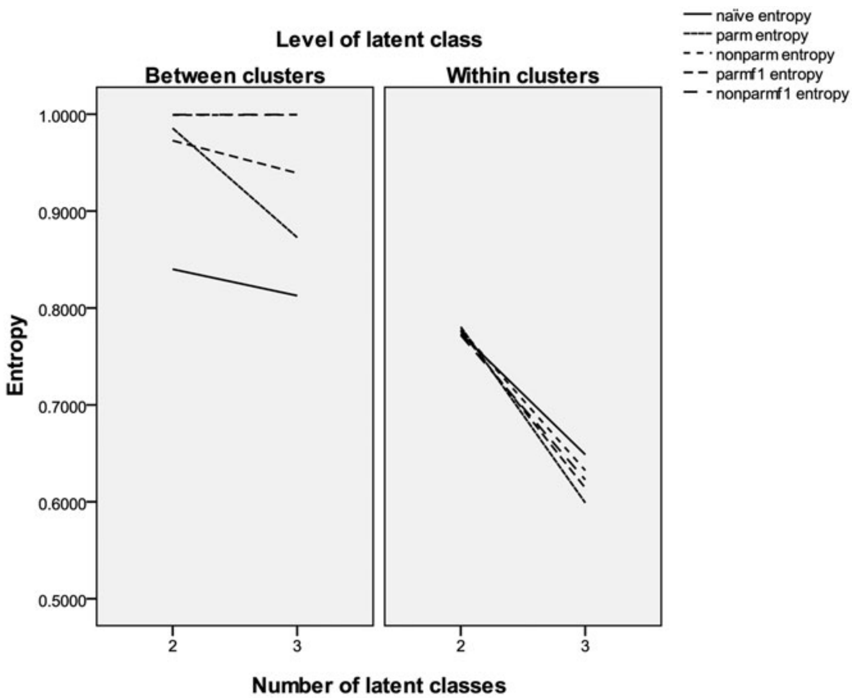
The entropy and mean posterior probabilities for the models by number of latent classes and level of latent class appear in Figure 3. For all models, these values were lower in the three class within-cluster condition. In contrast, for the between-cluster situation, the number of latent classes did not affect mean entropy or posterior probabilities for the nonparametric models.

These results suggest that researchers working with more within-cluster indicator variables will generally achieve greater classification certainty, as measured by entropy and the posterior probabilities. In addition, the MLCA model that is used will not greatly affect such classification certainty. In contrast, when the indicators are between cluster, the naive model will generally yield the least accurate predictions (particularly for a large number of clusters), and the nonparametric and nonparametric one-factor the most accurate, as measured by entropy and posterior probabilities. Last, when more latent classes are present in the population (something the researcher will not know with certainty), researchers should expect classification certainty to be somewhat lower, particularly when they are using within-cluster variables to obtain their predictions. This is an important issue for the practitioner to keep in mind if he or she believes that a large number of latent classes are present in the population.

Threshold Parameter Coverage

The repeated measures factorial analysis of variance with threshold coverage rates as the response variable found the following interactions to have effect sizes of 0.10 or greater: model by number of clusters by number of indicators by level of latent class, $\eta^2 = 0.147$; model by sample size per cluster by level of latent class, $\eta^2 = 0.457$; model by number of clusters by level of latent class, $\eta^2 = 0.152$. All other important main effects and interactions were subsumed in these terms, and therefore is not discussed. When a model is working well, the coverage rates for the thresholds should be very close to 0.95, which is the nominal rate calculated for each model. Observed coverage rates well below 0.95 are likely indicative of bias in the parameter estimates. Figure 4 includes the threshold coverage rates by model, level of latent class, number of latent classes, and number of indicators.

In the within-cluster condition all of the models performed similarly well in terms of threshold parameter coverage, with rates very close to the nominal 0.95 value regardless of the numbers of classes and indicators. The 0.95 nominal coverage rate is denoted by a horizontal line in Figure 4. Therefore, when dealing only in the within-cluster variable condition, all of the models were able to accurately estimate the threshold parameter values for the individual items across study conditions. On the other hand, for the between-cluster condition, the threshold coverage rates for most of the models were lower than the nominal rate of 0.95, with generally better coverage associated with more indicators. This effect of number of indicators in the between-clusters condition was most pronounced for the parametric and nonparametric models that included factors (parametric one-factor and nonparametric one-factor). In contrast, more indicator variables did not have a dramatic effect on the coverage rates for the naive model, which had the lowest rates in general. Last, the nonparametric and parametric models without factors yielded the highest coverage rates in the between-clusters case. In short, having more indicator variables made all



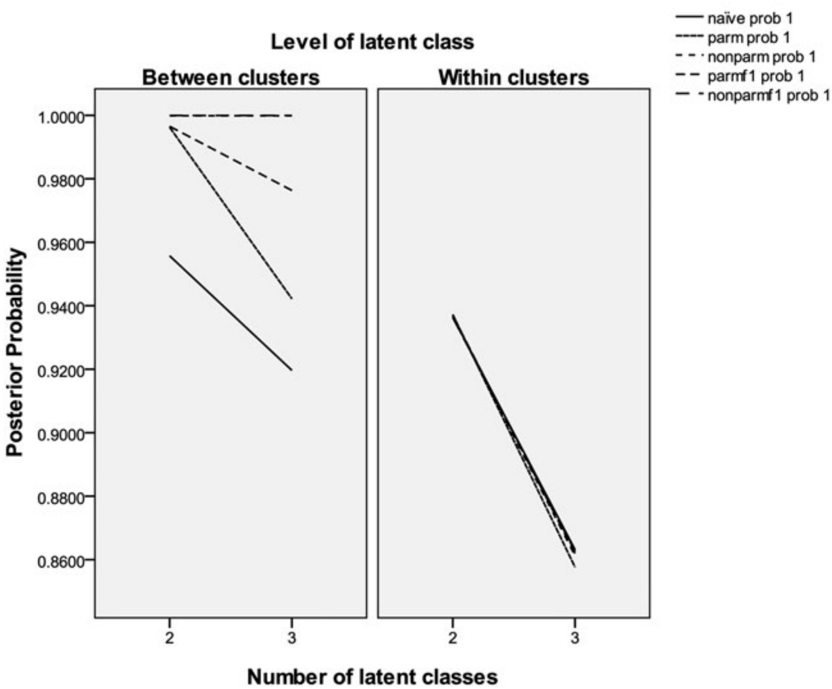


FIGURE 3 Mean entropy and posterior probability for Group 1 membership by model, number of latent classes, and level of latent class.

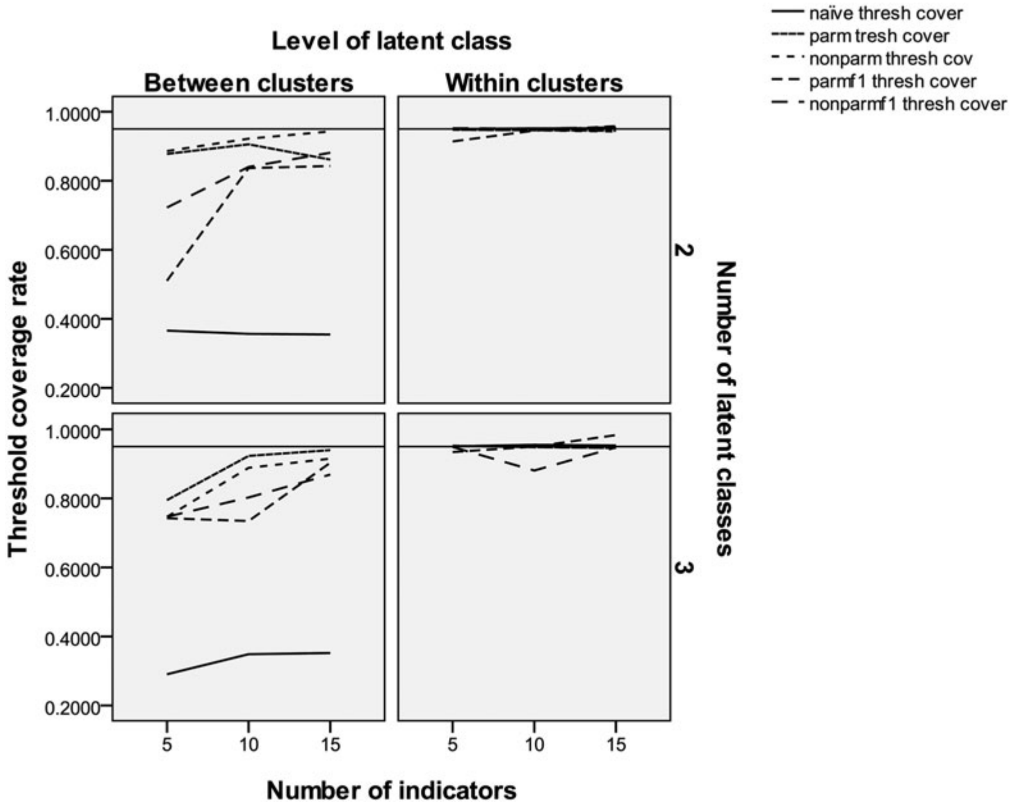


FIGURE 4 Mean threshold parameter coverage rates by model, number of indicators, number of latent classes, and level of latent class.

of the models more accurate in parameter estimation, with the exception of the naive model, and the nonparametric and parametric models with no factors were the most accurate of all.

The threshold coverage rates by level of latent class, sample size and model appear in Figure 5, with the nominal 0.95 rate marked by a reference line. Sample size had little effect on these rates in the within-clusters case, whereas for the between-cluster condition there is no consistent pattern of association between sample size and coverage for any of the models except the naive and parametric approaches. For each of these models, the larger samples were associated with lower coverage rates, with this effect being most pronounced for the naive model. The threshold coverage rates by number of clusters, level of latent class, and model appear in Figure 6. Again, there was little effect of number of clusters on coverage rates for any of the models in the within-cluster condition. On the other hand, for between-cluster variables an increasing number of clusters was associated with somewhat higher coverage in the naive and parametric one-factor methods, whereas for the nonparametric one-factor model coverage rates declined with an increased number of clusters.

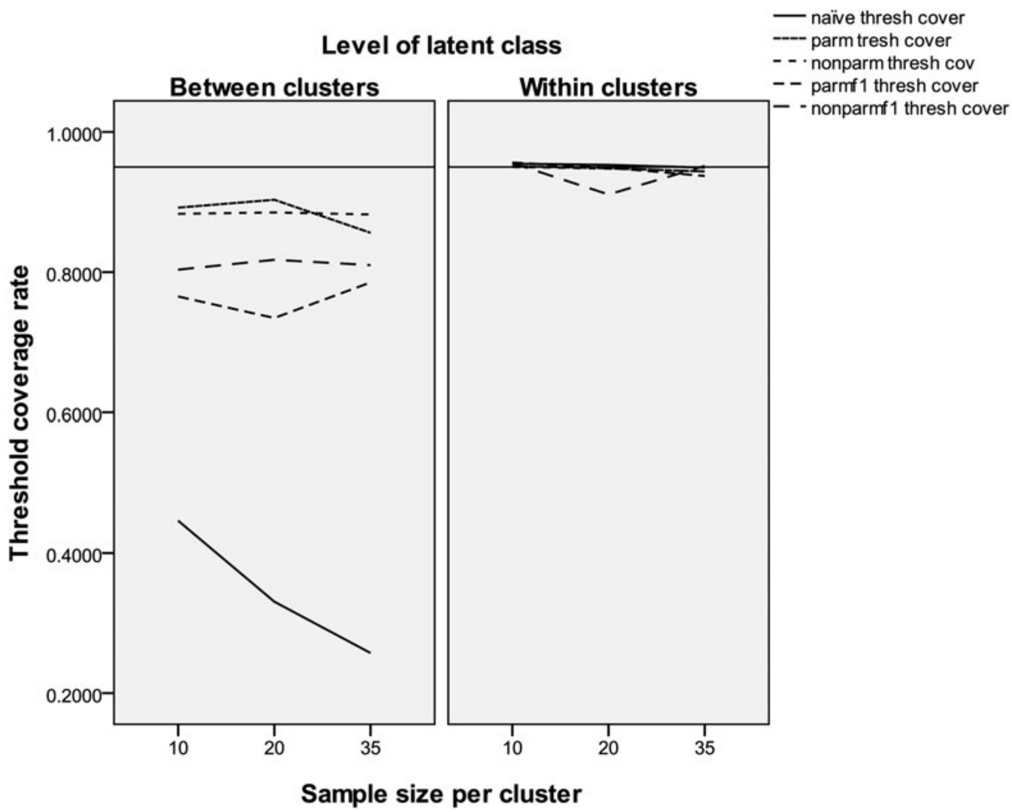


FIGURE 5 Mean threshold parameter coverage rates by model, sample size per cluster, and level of latent class.

Researchers using only within-cluster variables to obtain their predictions can expect to obtain good coverage of threshold parameter values. In other words, the 95% confidence interval for the thresholds will indeed have coverage at approximately 0.95, regardless of the model used. On the other hand, when the variables are between clusters, researchers should note that threshold coverage rates will be below the nominal 0.95, with the highest rates belonging to the nonparametric and parametric approaches, and the lowest to the naive model. In addition, when researchers have access to more clusters, the coverage rates in the between-cluster case will generally be higher.

Covariate Parameter Coverage

The repeated measures factorial analysis of variance identified three interactions with effect sizes greater than 0.10 among the manipulated variables when the outcome variable was the coverage rate of the relationship between the covariate and the latent class variable: model by number of latent classes by number of indicators by level of latent class, $\eta^2 = 0.411$; model by number of clusters by number of indicators by level of latent class, $\eta^2 = 0.187$; model by sample size per cluster by level of latent class, $\eta^2 = 0.251$. All other important main effects and interactions were

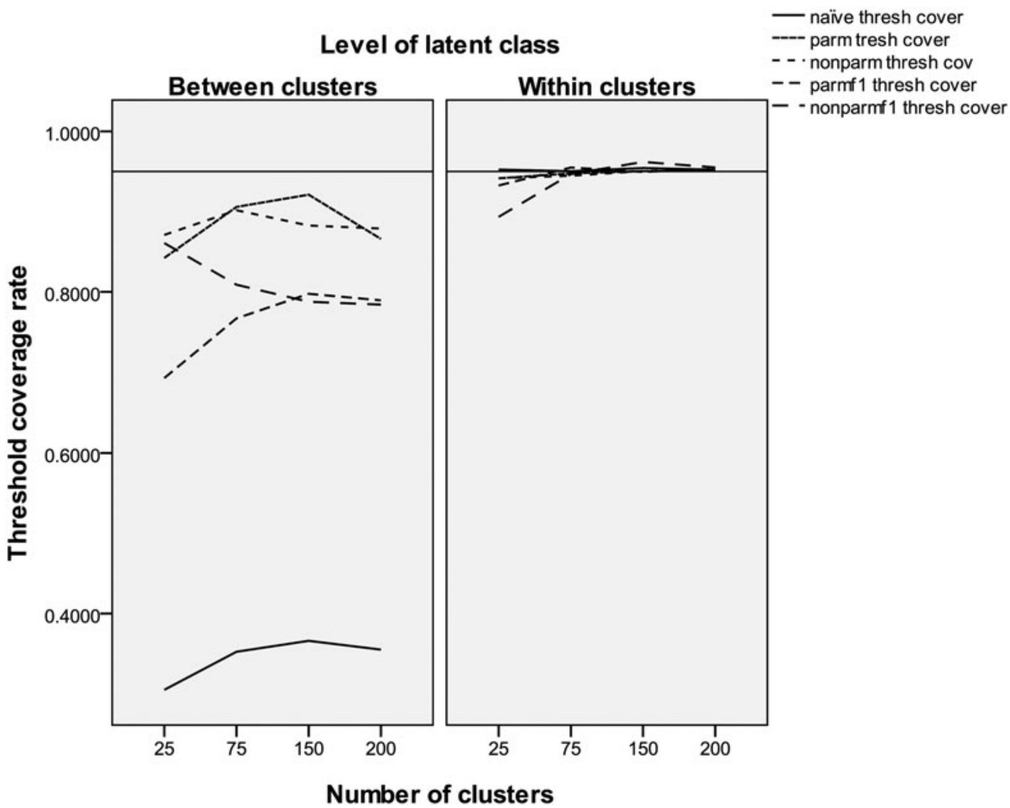


FIGURE 6 Mean threshold parameter coverage rates by model, number of clusters, and level of latent class.

subsumed in these terms. The covariate coverage rates by model, number of indicators, number of latent classes and level of latent class can be found in Figure 7, with a reference line at the 0.95 nominal coverage rate.

As with the threshold coverage rates, in the within-cluster condition the covariate coefficient coverage rates were similar for all of the models and near the nominal 0.95 level. That is, all of the models provided accurate estimates of the relationship between the covariate and the latent class variable for within-cluster variables. In contrast, coverage rates in the between-clusters condition diverged greatly by type of model, with the lowest values belonging to the parametric one-factor and naive models. The parametric method had among the highest covariate coverage rates across the number of indicators and the number of latent classes, with somewhat higher values for more indicators. The two nonparametric approaches also had higher coverage rates for a larger number of indicators, particularly in the three latent class condition. In other words, the parametric model provided the most accurate estimate of the relationship between the covariate and latent class variable, with the two nonparametric approaches yielding similarly accurate results with more indicators and latent classes. The naive and parametric one-factor model provided the least accurate estimate of the covariate relationship across conditions.

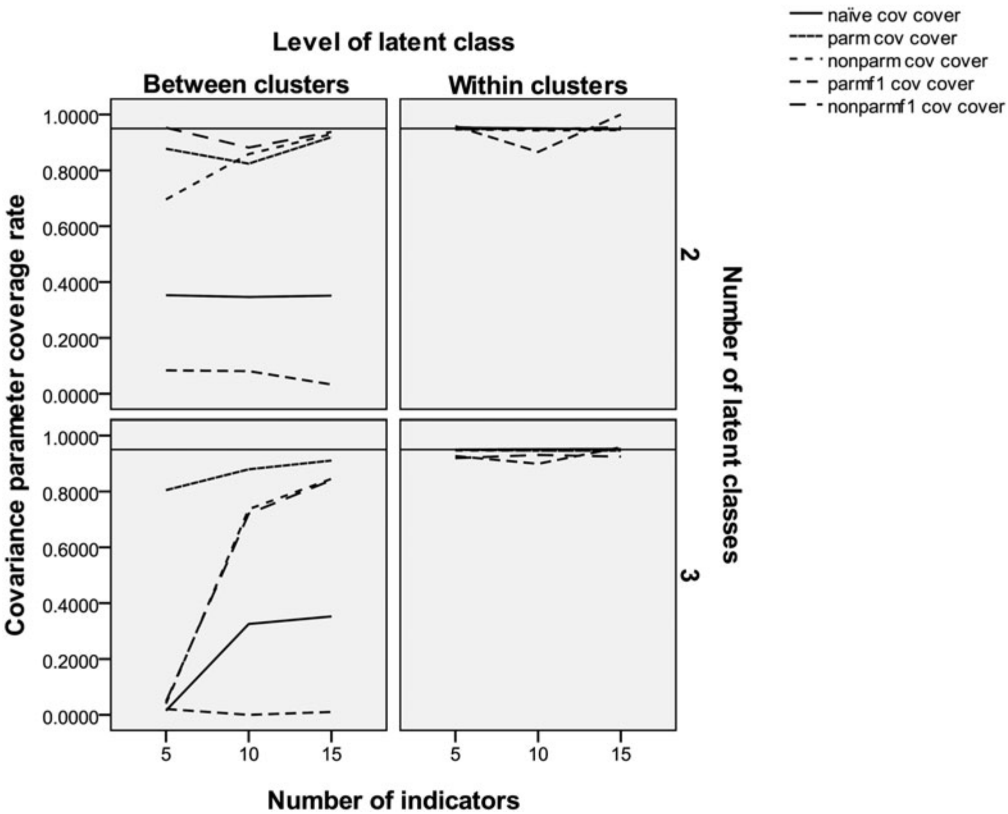


FIGURE 7 Mean covariate parameter coverage rates by model, number of indicators, number of latent classes, and level of latent class.

Figure 8, with a reference line at the nominal 0.95 level, contains covariate parameter coverage rates for each model by level of latent class, number of clusters, and number of indicators. In the within-cluster case the models had similar coverage rates, near the nominal level, except for parametric one-factor with 10 indicators and 25 clusters. In the between-cluster condition, the parametric and nonparametric techniques, as well as nonparametric one-factor, had similar coverage rates for 15 and 75 clusters with 10 or 15 indicators. However, for a larger number of clusters the parametric model had somewhat higher coverage rates, even for a larger number of indicators. Last, Table 2 contains covariate coverage rates for each model by level of latent class and sample size. The particularly important values are in bold. For the naive model in the between-cluster case larger sample sizes per cluster were associated with lower coverage rates, which is similar to results for the threshold parameter. As seen in other results, all of the models had higher coverage rates in the within-cluster condition, with the parametric one-factor model having the lowest coverage rates in the between-cluster condition, and the parametric model having the highest such rates.

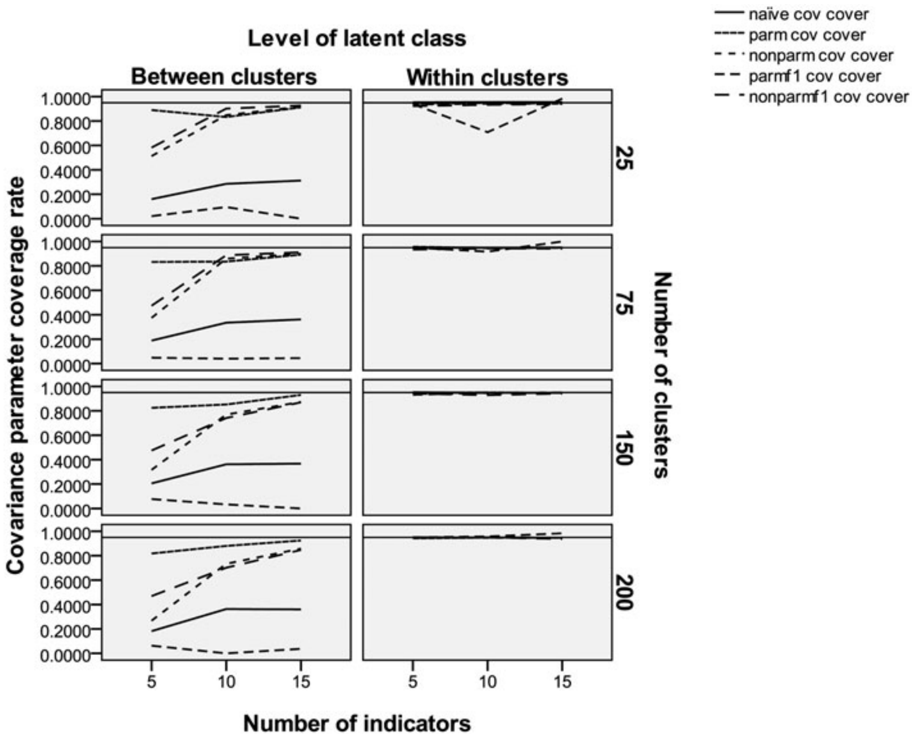


FIGURE 8 Mean covariate parameter coverage rates by model, number of indicators, number of clusters, and level of latent class.

As was the case for thresholds, researchers using within-cluster indicator variables can expect covariate parameter coverage rates near the nominal 0.95 regardless of the model they elect to use. However, when the indicators are between variables researchers should avoid the naive model, as it will perform poorly in terms of covariate parameter coverage. When working with between-cluster variables, researchers should expect the parametric model to provide optimal coverage for the covariate parameter, although the two nonparametric models will also perform well. In addition, each of these models will provide higher coverage rates when more indicators and more clusters are present in the data. Thus, researchers may consider trying to add indicators

TABLE 2
Mean Covariate Coverage Rates, by Method, Sample Size, and Level of Latent Class (Within/Between)

Sample size	*Naive*	*Parametric*	*Nonparametric*	*Parametric with one factor*	*Nonparametric with one factor*
10	0.95 / 0.38	0.95 / **0.87**	0.95 / 0.69	0.93 / **0.05**	0.94 / 0.73
20	0.95 / 0.28	0.95 / **0.87**	0.95 / 0.68	0.93 / **0.04**	0.93 / 0.74
35	0.95 / 0.21	0.95 /**0.87**	0.94 / 0.69	0.94 / **0.03**	0.94 / 0.73

Note. Especially notable results appear in boldface.

in particular to their models to achieve optimal covariate parameter coverage. Last, as was the case with the threshold parameter coverage values, researchers should expect lower than the nominal 0.95 rate when working with between-cluster indicators.

DISCUSSION

This study examined the performance of 5 approaches for fitting a multilevel latent class model given a correct specification of the number of latent classes. Given the complexity of the study, the results presented earlier are summarized in Table 3. These results indicated when only Level 1 latent classes and covariates are present, data fit, classification quality, and parameter coverage is comparable across the 5 models. This result, that all methods work comparably and reasonably well when only Level 1 variables are present, is consistent with other studies that have examined the performance of single and multilevel models (e.g., regression, logistic regression, or multiple indicators, multiple causes [MIMIC]) in the presence of multilevel data (Finch & French, 2011; French & Finch, 2010; Hox & Moss, 2001). In contrast, for Level 2 latent classes and covariates, the naive approach ignoring multilevel data consistently performed poorly across our outcome variables. This result is not particularly surprising, given similar findings with regard to ignoring multilevel data in other modeling contexts (e.g., Hox, 1998; Maas & Hox, 2004). However, results demonstrated that conducting analysis in a business-as-usual manner (e.g., assuming independence) will likely lead to false conclusions. Therefore, researchers are advised to avoid using the naive model in situations where multilevel data with Level 2 indicators and/or covariates are present. These findings extend past work in MLCA, which demonstrated that researchers should account for multilevel data structure when determining the number of latent classes that might be present at Levels 1 and 2 (Lukociene & Vermunt, 2002) by examining other aspects of model performance, including parameter estimation and model selection. The present work also shows that researchers must be cognizant of the multilevel structure when interpreting the fitted results of the model (i.e., classifications and model parameters).

With respect to the models, beyond the naive case, concluding that one approach will always provide optimal outcomes vis-à-vis the others is difficult. The parametric and nonparametric models performed well in a number of situations. In particular, the nonparametric model consistently provided the highest quality predictions of group membership, based on the entropy and posterior probabilities. The parametric and nonparametric models yielded the highest rates of threshold parameter coverage in the between-subjects indicator condition, whereas the parametric model had the highest coverage rates for the covariate in the between condition. In contrast, the parametric one-factor model had the lowest coverage rates for both thresholds and covariates in the between-subjects indicator condition, except when compared with the naive model. On the other hand, classification quality for this model was comparable to that of the other multilevel models. Thus, performance may depend on conditions being examined. Nevertheless, we use these results to make some recommendations for practitioners working with such data and models.

Practical Implications

A major goal of this study was to assist researchers in selecting an appropriate model in the presence of a multilevel data structure. Whereas past work has demonstrated the use of MLCA models

TABLE 3
Summary of Study Results

	Outcomes			
Manipulated factors	*Sample size adjusted Bayesian information criterion*	*Entropy*	*Threshold coverage*	*Covariate coverage*
Level of effect, within or between subjects	Naive: best model fit with within-cluster variables Parametric with one factor: best model fit with between-cluster variables	Naive: highest entropy and posterior probability values with within-cluster Nonparametric and nonparametric with one factor: highest entropy with between-cluster case All models: entropy and posterior probability higher in the within case	Within-cluster: all methods had rates at 0.95 All models: parameter threshold coverage lower in between-cluster vs. within-cluster Between-cluster: Naive had lowest coverage rates; nonparametric and parametric had highest	Within-cluster: all models had covariate coverage rates near 0.95 Between-cluster: naive and parametric with one factor had lowest coverage, parametric had the highest
Number of clusters	Within-cluster: with an increase in cluster number, naive method resulted in suboptimal fit; parametric with one factor resulted in optimal fit, regardless of cluster number	Between-cluster: more clusters resulted in lower entropy and posterior probability values for naive; nonparametric and nonparametric with one factor had highest entropy and posterior probability values regardless of cluster number	Between-cluster: large cluster number resulted in higher coverage for naive and parametric with one factor, lower coverage for nonparametric with one factor	Between-cluster: large cluster number resulted in the highest coverage rates for parametric
Sample size	Within-cluster: as the sample size increased, naive was increasingly resulted in optimal fit; parametric with one factor always resulted in optimal model fit	Sample size did not result in differences in entropy or posterior probability	Between-cluster: coverage rates for naive and parametric were lower for larger values of sample size Within-cluster: sample size did not result in different coverage rates	Within-cluster: sample size not associated with covariate coverage rates Between-cluster: larger sample size associated with lower coverage for naive Between-cluster: sample size not associated with coverage rates

Number of latent classes	Number of classes not associated with differences in selection of the optimal model	Within-cluster: more latent classes resulted in lower entropy and posterior probability values Between-cluster: number of classes not associated with entropy or posterior probability	Number of classes not associated with differences in selection of the optimal model	Positive association between number of indicators and coverage for nonparametric and nonparametric with one factor was somewhat stronger in the presence of more latent classes
Number of observed indicator variables	Within-cluster: more indicators resulted in nonparametric and naive providing optimal fit at the same rate; parametric with one factor always provided optimal fit	Within-cluster: more indicators led to greater classification accuracy Between-cluster: naive had the lowest entropy and posterior probability values, nonparametric and nonparametric with one factor had the highest values	Between-cluster: presence of more indicators resulted in better coverage, particularly for parametric with one factor and nonparametric with one factor	Between-cluster: presence of more indicators was associated with higher coverage rates, particularly for the nonparametric and nonparametric with one factor

(Henry & Muthen, 2010) and provided some insights into the appropriate statistics for model selection (Lukociene & Vermunt, 2002), the present study is an important step in systematically comparing these modeling approaches in terms of their ability to accurately fit the data, provide quality classification solutions, and estimate both threshold and covariate parameters under a variety of known conditions. On the basis of these results, it seems reasonable for researchers using data with between-subjects indicator variables to use the parametric or nonparametric models, when their primary interest is in obtaining quality classification, and/or accurate parameter estimation. Either of these models should provide comparable outcomes to one another, and optimal outcomes when compared with the other models examined in this study. When the indicators are all within subjects and the data structure is multilevel, it would appear that any of these methods will provide good parameter coverage and high-quality classification solutions, even the naive model. Furthermore, given the occasional estimation problems previously cited with the parametric approach, the nonparametric model may be a useful default approach for researchers. The parametric model did not present such estimation difficulties except for the smallest sample size condition with the largest number of indicators, suggesting that lack of convergence may not be particularly common for many applications.

Limitations and Future Work

There are limitations to these simulation conditions that should be addressed in future work. First, we did not vary the level of the dependence in the data. The ICC was set at a constant value of 0.15 across simulations. Although the value was realistic given past work and that the fixed value was desirable to keep the present study at a manageable size in terms of number of conditions and potential interactions, it may be worth examining the effect of varying levels of ICC on the performance of these models as results could change depending on the level of clustering. It is common for this correlation to differ depending on context (e.g., business organization compared with educational environment) and even with context depending on outcomes (e.g., math vs. reading) and the value used falls in the middle range. Other methods accounting for multilevel data structures display different patterns of behavior (e.g., Type I error, Power; bias) when the ICC differs (e.g., Finch & French, 2011; French & Finch, 2010; Hox & Moss, 2001). Future work will confirm the expectations of behavior on the basis of what has been documented in similar multilevel work with other methods.

Second, the models used were best fitting or properly specified models. That is, we worked under the assumption that the correct number of classes was modeled. However, this may not be the case with real data. In addition, other types of latent models do differ in terms of outcomes (Type I error and power) when misspecification is present (e.g., French & Finch, 2011; Yuan & Bentler, 2004). Thus, future work should focus on the outcomes studied here and various MLCA models when models are not properly specified. Third, the models were all based on a normally distributed latent variable, a primary assumption, which provided the parametric models a distinct advantage. Future research should examine the effect of nonnormality on the performance of each model, particularly the parametric approaches, which hold normality as a central assumption. Fourth, only a limited number of simulation conditions were included in this study. These conditions were selected to match applied research, as cited earlier, however the limited number of these conditions serve as a limiting factor that must be considered when interpreting these results. For example, the clusters were all simulated to be of the same size, which would not be typical in practice. In addition, only dichotomous indicators were included in the study. Much research is conducted using ordinal categorical variables, which may lead to somewhat different results than those found here. Future research building upon this study should include differing cluster sizes and ordinal indicators.

CONCLUSION

The use of LCA and particularly MLCA, given the common hierarchical data structures in the social, educational, and medical research environment, is likely to continue to expand. These methods are becoming particularly popular given the accessibility of the software to implement these analytic techniques and the user-friendly articles that provide excellent explanations and code (e.g., Henry & Muthén, 2010). Because these models are applied more frequently, methodological research on the conditions under which these various models function best to accurately capture individuals behavior is essential. Evidence needs to support the accuracy of these models as the results are used to make important decisions about individuals and courses of treatment or interventions in all types of environments where the stakes are high (e.g., education, health, employment outcomes) for persons in these environments.

AUTHOR NOTES

W. Holmes Finch is a professor of psychometrics and statistics at Ball State University in Muncie, Indiana. He works in the areas of latent variable modeling and measurement. His work has contributed to an increased understanding of factor analysis and dimensionality assessment, as well as group differences in latent variable structures. **Brian F. French** is a professor in psychometrics and Director of the Learning and Performance Research Center at Washington State University in Pullman, Washington. His program of research contributes to a growing body of literature related to test score validity. He has specific interest in measurement invariance and factor analysis.

NOTE

1. For the parametric model, see Vermunt 2003, Vermunt 2008, and Asparouhov & Muthén 2008; for the parametric one-factor model, see Vermunt 2008 and Asparouhov & Muthén 2008; for the nonparametric model, see Vermunt 2008, and Bijmolt, Paas, & Vermunt 2004; and for the nonparametric one-factor model, see Vermunt 2008.

REFERENCES

Asparouhov, T., & Muthen, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27–51). Charlotte, NC: Information Age.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, *9*, 3–29.

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC model parameters with multilevel data. *Structural Equation Modeling*. doi: 10.1080/10705511.2011.557338

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, *47*, 299–317. doi: 10.111/j.1745-3984.2010.00115.x

French, B. F., & Finch, W. H. (2011). Model misspecification and invariance testing via confirmatory factor analytic procedures. *Journal of Experimental Education*, *79*, 404–428. doi: 10.1080/00220973.2010.517811

Gilthorpe, M. S., Harrison, W. J., Downing, A., Forman, D., & West, R. M. (2011). Multilevel latent class casemix modelling: A novel approach to accommodate patient case mix. *BMC Health Services Research*, *11*, 1–7. doi: 10.1186/1472–6963–11–53

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

Halpin, P. F., & Maraun, M. D. (2010). Selection between linear factor models and latent profile models using conditional covariances. *Multivariate Behavioral Research*, *45*, 910–934. doi: 10.1080/00273171.2010.532361

Hedden, S L., Whitaker, D. E., & von Thomsen, S. (2011). Latent patterns of risk behavior in urban African American middle school students in Baltimore. *Journal of Child & Adolescent Substance Abuse*, *20*, 34–47. doi: 10.1080/1067828X.2011.534358

Henry, K. L., Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling*, *17*, 193–215. doi: 10.1080/107055110036 59342

Hernandez, M. B., Blavo, C., Hardigan, P. C., Perez, A. M., & Hage, K. (2010). Differences in perceived stress, depression, and medical symptoms among medical, nursing, and physician assistant student: A latent class analysis. *Annals of Behavioral Science and Medical Education*, *16*, 35–39.

Hox, J. J. (1998). Multilevel modeling: Whan and why? In R. Balderjahn & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). New York, NY: Springer-Verlag.

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groupsand small samples. *Structural Equation Modeling*, *8*, 157–174. doi: 10.1207/S15328007SEM0802_1

Huang, G.-H., Wang, S.-M., & Hsu, C.-C. (2011). Optimization-based model fitting for latent class and latent profile analyses. *Psychometrika*, *76*, 584–611. doi: 10.1007/s11336-011-9227-3

Johnson, J. E., Burlingame, G. M., Olsen, J. A., Davies, D. R., & Gleave, R. L. (2005). Group climate, cohesion, alliance, and empathy in group psychotherapy: Multilevel structural equation models. *Journal of Counseling Psychology*, *52*, 310–321. doi: 10.1037/0022-0167.52.3.310

Kaplan, D., & Keller, B. (2011). A note on cluster effects in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 525–536. doi: 10.1080/10705511.2011.607071

Komro, K. A., Tobler, A. L., Maldonado-Molina, M. M., & Perry, C. L. (2010). Effects of alcohol use initiation patterns on high-risk behaviors among urban, low-income, young adolescents. *Prevention Science*, *11*, 14–23. doi: 10.1007/s11121-009-0144-y

Kuntsche, E., Kuendig, H., & Gmel, G. (2008). Alcohol outlet density, perceived availability and adolescent alcohol use: A multilevel structural equation model. *Journal of Epidemiology and Community Health*, *62*, 811–816. doi: 10.1136/jech.2007.06537

Lazaesfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88, 767–778. doi: 10.1093/biomet/88.3.767

Lukociene, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, *40*, 247–283. doi: 10.1111/j.1467–9531.2010.01231.x.

Maas, C. & Hox, J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137.

Montanari, G. E., Ranalli, G., & Eusebi, P. (2010). Multilevel latent class models for evaluation of long-term care facilities. *Data Analysis and Classification*, *5*, 249–256. doi: 10.1007/978-3-642-03739-9_29

Muthén, L. K., & Muthén, B. O. (2008). *Mplus user's guide, version 5.2*. Los Angeles, CA: Authors.

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

Palardy, G. J. (2008). Differential school effects among low, middle, and high social class composition schools: A multiple group, multilevel latent growth curve analysis. *School Effectiveness and School Improvement*,*19*, 21–49. doi: 10.1080/09243450801936845

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

Ramaswamy, V., DeSarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticicties with PIMS data. *Marketing Science*, *12*, 103–124.

Rindskopf, D. (2006). Heavy alcohol use in the "Fighting Back" survey sample: Separating individual and community-level influences using multilevel latent class analysis. *Journal of Drug Issues*, *36*, 441–462.

Schrepp, M. (2003). A method for the analysis of hierarchical dependencies between items of a questionnaire. *Methods of Psychological Research*, *8*, 43–79.

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

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*, 213–239.

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*, 33–51. doi: 10.1177/0962280207081238

Vermunt, J. K., & Magidson, J. (2003). Latent class models for classification. *Computational Statistics and Data Analysis*, *41*, 531–537. doi: 10.1111/j.0081-1750.2003.t01-1-00131.x

Vuolo, M., Staff, J., & Mortimer, J. T. (2012). Weathering the Great Recession: Psychological and behavioral trajectories in the transition from school to work. *Developmental Psychology*, *48*, 1759–1773. doi: 10.1037/a0026047

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and *z* tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, *64*, 737–757. doi: 10.1177/0013164404264853