Article

Recommendations on the Sample Sizes for Multilevel Latent Class Models

Educational and Psychological Measurement
1–25

DOI: 10.1177/0013164417719111
journals.sagepub.com/home/epm
SAGE

Jungkyu Park[1] and Hsiu-Ting Yu[2]

Abstract

A multilevel latent class model (MLCM) is a useful tool for analyzing data arising from hierarchically nested structures. One important issue for MLCMs is determining the minimum sample sizes needed to obtain reliable and unbiased results. In this simulation study, the sample sizes required for MLCMs were investigated under various conditions. A series of design factors, including sample sizes at two levels, the distinctness and the complexity of the latent structure, and the number of indicators were manipulated. The results revealed that larger samples are required when the latent classes are less distinct and more complex with fewer indicators. This study also provides recommendations about the minimum required sample sizes that satisfied all four criteria—model selection accuracy, parameter estimation bias, standard error bias, and coverage rate—as well as rules of thumb for sample size requirements when applying MLCMs in data analysis.

Keywords

latent class models, multilevel modeling, sample size

Latent class models (LCMs; Goodman, 1974; Lazarsfeld & Henry, 1968) are a standard research tool in many fields—such as sociology, psychology, education, marketing, and medicine—in which observations are often nested within higher level units. Over the past decade, researchers have extended LCMs to analyze data arising from hierarchically nested structures (e.g., Asparouhov & Muthén, 2008; Di & Bandeen-Roche, 2008; Henry & Muthén, 2010; Varriale & Vermunt, 2012; Vermunt, 2003,

[1]McGill University, Montreal, Quebec, Canada
[2]National Chengchi University, Taipei, Taiwan

Corresponding Author:
Hsiu-Ting Yu, National Chengchi University, No. 64, Sec. 2, Zhinan Road, Taipei, 11605, Taiwan.
Email: hsiutingyu@gmail.com

2004; Vermunt & Magidson, 2008). The multilevel latent class model (MLCM; Vermunt, 2003, 2004) is an extension of the LCM that incorporates possible dependency on data because of the multilevel structure where individuals are nested within a group. A number of LCMs also have been introduced to deal with other multilevel data structures, such as the measurement occasions nested within individuals (Chung, Anthony, & Schafer, 2011) or the three-level hierarchical structure (Bartolucci, Pennoni, & Vittadini, 2011; Palardy & Vermunt, 2010).

One common feature of these extensions is the use of random effects to capture the dependency due to multilevel structures. The incorporation of random effects into LCMs enables researchers to evaluate the effects of clustering or repeated sampling and to disentangle them from the lower level. Several researchers have discussed the various specifications of random effects. For example, Vermunt (2003, 2004) proposed continuous random effects originating from a normal distribution (i.e., a parametric MLCM), while Hedeker (2003) and Asparouhov and Muthén (2008) introduced the factor analytic approach, which includes a common factor in the parametric MLCM to reduce the dimensionality of continuous random effects. Furthermore, Di and Bandeen-Roche (2011) proposed a random effect following the Dirichlet distribution. The specification of random effects in MLCMs was discussed in more detail by Finch and French (2013) and Henry and Muthén (2010).

Despite such various extensions of the LCM, the focus of this study is the MLCM with a nonparametric specification (Aitkin, 1999; Laird, 1978). The nonparametric version of the MLCM includes a discrete random effect at the higher level. The discrete higher level random effect in this model is represented by a finite number of latent classes, relying on the assumption that each higher level unit is assigned to one of the higher level latent classes. Thus, the key feature of the nonparametric MLCM is providing classification information for both higher level and lower level units (e.g., Bassi, 2009; Bijmolt, Paas, & Vermunt, 2004; da Costa & Dias, 2014; Finch & Marchant, 2013; Onwezen et al., 2012; Pirani, 2011; Rindskopf, 2006; Rüdiger & Hans-Dieter, 2013). Moreover, the nonparametric MLCM does not involve unverifiable distributional assumptions regarding the random effects (Aitkin, 1999) and is computationally less intensive than the parametric approach (Rights & Sterba, 2016; Vermunt, 2004). Several measures can be used to decide whether a higher level class structure should be incorporated in analyses using a nonparametric MLCM, such as the pseudo–intraclass correlation coefficient (ICC)—the ratio of the higher level class variance to the sum of class variances at both higher and lower levels (Rights & Sterba, 2016)—and *R*-square entropy—the index representing the degree of latent class separation for a given mixture model (Ramaswamy, DeSarbo, Reibstein, & Robinson, 1993).

When designing a study, one common question is ''What sample size do I need?'' Unfortunately, there is no standard or simple answer to this question. The answer depends on many factors, including the purpose of the study, the type of model used in the analyses, data characteristics, expected power, population size, and cost and resource considerations. For researchers using highly complex mixture models such

as the MLCM, this question is even more important, because the maximum likelihood estimation methods used to estimate parameters are asymptotic, meaning properties of the estimates rely on the assumption that the sample size is sufficiently large. Therefore, the minimum sample size requirement becomes a crucial question in empirical applications to obtain reliable and unbiased results.

Prior studies related to the statistical inference regarding the number of latent classes often provide some insights about sample size requirements for single-level LCMs (e.g., Henson, Reise, & Kim, 2007; Nylund, Asparouhov, & Muthén, 2007; Yang & Yang, 2007). These studies indicated that the required sample sizes could vary depending on many factors, such as the number of indicators, the structure of latent classes (i.e., the number of classes and structure of conditional response probabilities), and the existence of covariates (Wurpts & Geiser, 2014; Yang, 2006). Nonetheless, there is no consensus so far regarding the minimum required sample sizes to achieve satisfactory performance in correctly identifying the number of latent classes. For example, a simulation study by Nylund et al. (2007) showed that sample sizes of 500 are sufficient for detecting the correct number of latent classes using the bootstrapped likelihood ratio test (BLRT) and adjusted BIC statistics under most simulation conditions. Conversely, Henson et al. (2007) showed that a sample size of 500 was not enough under some conditions to accurately identify the true number of latent classes due to a lack of statistical power and high estimation error.

More recently, several studies have discussed sample size determinations to attain a specific power level regarding the model selection test for the single-level LCM. For example, Dziak, Lanza, and Tan (2014) studied the minimum sample sizes required to avoid underestimating the number of latent classes when using the BLRT. The study proposed new effect size measures using Cohen's w (Cohen, 1998) and the Kullback–Leibler discrepancy, and provided empirical power curves to illustrate the required sample sizes in LCMs. Tekle, Gudicha, and Vermunt (2016) also performed a power analysis for the BLRT with a new computationally efficient approach to determine required sample sizes. In addition, Gudicha, Tekle, and Vermunt (2016) presented power and sample size computations using the Wald test for LCM parameters.

Several studies have investigated similar topics with other mixture models. For example, Kim (2012) conducted a simulation study to investigate the sample size requirement for the growth mixture model, and Tein, Coxe, and Cham (2013) explored the statistical power of various statistical tests used for determining the number of latent classes in latent profile analysis. Moreover, Tueller and Lubke (2011) examined a minimum sample size for structural equation mixture models, whereas Gudicha, Schmittmann, and Vermunt (2016) conducted a power analysis on latent Markov models to provide guidance on the required sample size and number of measurement occasions needed to attain acceptable levels of power.

Determining sample sizes in models with multiple levels is more complicated than it is for single-level models, mainly because the sample sizes at the higher level (number of groups) and the lower level (number of individuals per group) should be

considered simultaneously during a study's design stage. Despite this difficulty, sample size requirements have been extensively studied in the context of linear mixed models, also referred to as random effects models, hierarchical models, and multilevel models (Bryk & Raudenbush, 1992; Goldstein, 1995; Hox, 2010; Skrondal & Rabe-Hesketh, 2004; Snijders & Bosker, 2012). These studies have demonstrated that the sample size requirements for obtaining unbiased estimates are influenced by several elements, including the magnitude of the ICC, the scale of the dependent variables (i.e., binary or continuous), and the estimation methods (i.e., full maximum likelihood, restricted maximum likelihood or Bayesian estimation methods; Browne & Draper, 2006; Moineddin, Matheson, & Glazier, 2007; Maas & Hox, 2005; McNeish & Stapleton, 2016). Although there is no widely accepted minimum sample size to ensure unbiased estimates, findings from the results of the simulation studies provided some general guidelines. For example, Kreft (1996) suggested the ''30/30 rule'': If researchers are interested in the fixed effects, they should strive for a sample of at least 30 groups with at least 30 individuals per group. Hox (2010) advocated for 50 groups and 20 individuals per group—or 100 groups and 10 individuals per cluster—for situations in which the model of interest involves random effects. More recently, Snijders and Bosker (2012) indicated that a minimum of 20 groups is sufficient to run linear mixed models.

Simulation studies also have found that when the dependent variable is binary, small sample sizes are more problematic than in the conditions with a continuous dependent variable, and the sample sizes at the lower level are more important than the sample sizes at the higher level (Austin, 2010; McNeish & Harring, 2017; Moineddin et al., 2007). Moreover, Bayesian Markov chain Monte Carlo estimation methods have been suggested as alternatives to overcome the problems of likelihood-based estimation methods when the sample sizes are small (Baldwin & Fellingham, 2013; Browne & Draper, 2006; Stegmueller, 2013).

The MLCM has been demonstrated to be a useful tool in many empirical applications, such as the evaluation of interventions in group randomized trials (e.g., Kaplan & Keller, 2011; Kaplan, Kim, & Kim, 2009; Van Horn et al., 2008) and, more broadly, the investigation of national-level heterogeneity (e.g., Bijmolt et al., 2004; da Costa & Dias, 2014; Pirani, 2011). Despite this wide range of applications, no previous study has investigated the required and adequate sample sizes for fitting MLCMs. In this study, we systematically investigate the minimum sample sizes for fitting the MLCM within a situation in which true latent class solutions were either known or unknown. The simulation design covered a broad range of MLCM structures by manipulating design factors (class distinctness, model complexity, and number of indicators) that are known to influence separations among classes at two levels (Lukočiené, Varriale, & Vermunt, 2010). The sample size requirements under various simulation conditions were then determined by using several evaluation criteria.

Through this study, we aim to (a) evaluate the impact of varying sample sizes according to different evaluation criteria at both the group and individual levels; (b)

explore how design factors such as the complexity of the latent structure, distinctness among latent classes, and number of indicators affect the sample size requirements in an MLCM; and (c) establish recommendations and guidelines for adequate sample size (at the individual level) when analyzing data using an MLCM.

The remainder of this article is organized as follows: First, we will introduce the specification of the MLCM and then discuss the criteria used to evaluate appropriate sample sizes, including model selection accuracy, parameter bias, standard error bias, and coverage rates. Simulation studies designed to investigate appropriate sample sizes are then provided, followed by a summary of the simulation results. Finally, the ''Discussion and Conclusion'' section provides our recommendations, examines the limitations of this study, and offers suggestions for future research on sample size requirements for an MLCM.

Nonparametric Multilevel Latent Class Model

The MLCM with nonparametric specification can be specified as follows: X_{gi} denotes the discrete latent variables at the lower level (individuals) with M categories, and H_g is the discrete latent variable at the higher level (groups level) with L categories. The categories or levels of discrete latent variables can be conceptualized as latent classes at each level; these classes are internally homogenous and each class is characterized with distinct response patterns. The terms *latent clusters* and *latent classes* are used to denote the latent classes at the higher level and lower level in this study, respectively.

Let Y_{gij} be the response to the jth item of an individual i in a group g, where $g = 1, \ldots, G$, $i = 1, \ldots, n_g$, and $j = 1, \ldots, J$. The vector $\mathbf{Y}_{gi}$ denotes J responses for a subject i nested in group g, and $\mathbf{Y}_g$ represents the full vector of responses for all individuals in group g.

The standard LCM could be defined by a latent variable at individual level X_i without consideration of group level; therefore, the density of the response of subject i on item j is

$$P(Y_{ij}) = \sum_{m=1}^{M} P(X_i = m) \prod_{j=1}^{J} P(Y_{ij} | X_i = m). \tag{1}$$

An MLCM is defined by two separate equations for the higher and lower levels. For the higher level, the subscript i in Equation (1) is replaced by g. With these specifications, the probability of observing a response pattern for all subjects nested in group g is

$$P(\mathbf{Y}_g) = \sum_{l=1}^{L} P(H_g = l) \prod_{i=1}^{n_g} P(\mathbf{Y}_{gi} | H_g = l). \tag{2}$$

Equation (2) assumes that each group belongs to only one l (latent cluster), and the conditional densities for each of the individuals (n_g) within the group (g) are independent of each other given the latent cluster membership. The first term, $P(H_g = l)$ (called *latent cluster probability*), is represented by a vector, with each element representing the probabilities of g being present in the cluster l ($l = 1, \ldots, L$). Since the latent clusters are assumed to be mutually exclusive and exhaustive, the elements of this vector can be conceptualized as cluster sizes and, thus, the sum of this vector equals 1.

At the individual level, the probability of obtaining a certain response pattern for each subject is

$$P(\mathbf{Y}_{gi}|H_g = l) = \sum_{m=1}^{M} P(X_{gi} = m|H_g = l) \prod_{j=1}^{J} f(Y_{gij}|X_{gi} = m). \tag{3}$$

By combining Equations (2) and (3) with the assumptions of local independence, the MLCM becomes

$$P(\mathbf{Y}_g) = \sum_{l=1}^{L} P(H_g = l) \left(\prod_{i=1}^{n_g} \sum_{m=1}^{M} P(X_{gi} = m|H_g = l) \prod_{j=1}^{J} f(Y_{gij}|X_{gi} = m) \right). \tag{4}$$

The term $P(X_{gi} = m|H_g = l)$ is the *conditional latent class probability*, which represents the distribution of latent class probabilities within a particular latent cluster. The conditional latent class probability can be parameterized as follows:

$$P(X_{gi} = m|H_g = l) = \frac{\exp(\gamma_{lm})}{\sum_{m=1}^{M} \exp(\gamma_{lm})}. \tag{5}$$

Note that that γ_{0ml} can be rewritten as $\gamma_m = \gamma_m + u_{lm}$, where discrete random variable u_{lm} does not assume any distributional assumption, but it varies across lower level classes while capturing the differences between the L classes at level 2 (Bijmolt et al., 2004; Henry & Muthén, 2010). Additional covariates at both the higher and lower levels can be included to facilitate the identification of the lower level classes and to predict class memberships (Henry & Muthén, 2010; Vermunt, 2003).

The term, $f(Y_{gij}|X_{gi} = m)$, is called the *conditional response density*, it is the probability of observing a particular response on variable j for individual i in group g given latent class membership (m). The conditional response density can have the various forms of distributions depending on the assumed characteristics of responses. If the response vector $\mathbf{Y}_{gi} = (Y_{gi1}, Y_{gi2}, \ldots, Y_{giJ})^T$ consists of J binary indicators, then the mth latent class density is given by $f(Y_{gij}|X_{gi} = m) = \rho_{mj}^{Y_{gij}}(1 - \rho_{mj})^{1-Y_{gij}}$, where ρ_{mj} presents the probability of endorsing item j for an individual belonging to latent class m.

The parameters of the MLCM can be estimated using the maximum likelihood (ML) method. The likelihood function is based on the probability density for the data of the higher level unit; and the log-likelihood to be maximized equals

$$\log L = \sum_{g=1}^{G} \log P(\mathbf{Y_g}|\mathbf{Z_g}). \quad (6)$$

A modified expectation–maximization algorithm obtains the ML estimates of the model parameters. The E-step modified by Vermunt (2003) is called the *upward–downward procedure*, which makes use of the conditional independence assumptions. Specifically, the latent variables are summed out by moving from higher to lower level units, and the marginal posteriors are then obtained by going from lower to higher level units. Vermunt (2003) gave the details of this algorithm.

Evaluation Criteria

The model selection accuracy and three evaluation criteria were considered to determine the minimum sample size requirements for MLCMs: parameter bias, standard error bias, and coverage rates. The MLCMs with different combinations of sample sizes were evaluated under the assumptions that the true numbers of latent classes are either unknown or known; the first criterion is based on the assumption that the true latent structure is unknown, whereas the other three criteria presume that the true latent structure is known in advance.

Although power can be another consideration in determining sample sizes, we did not include it as a criterion in this study. The main reason for this is that there is no consensus regarding the components of power calculation in MLCMs, such as effect size, null and alternative hypotheses, and test statistics. We also used the simulation approach, which provides power estimates for the individual effects of interest (Muthén & Muthén, 2002), but any specific parameter effect (whether a parameter is significantly different from 0) was not of interest in the current study. Finally, power cannot be calculated for parameters with population values equal to 0 (Kim, 2012); however, some population parameters in our simulation study were very close to 0.

Model Selection Accuracy

Model selection accuracy was used as the criterion for the sample size requirement. The model selection task in the MLCM is to determine the number of latent classes at both the higher and lower levels. Two approaches were suggested to choose the optimal number of latent classes at multiple levels: the stepwise approach (Lukočienė et al., 2010; Lukočienė & Vermunt, 2010) and the simultaneous approach (Bijmolt et al., 2004). The first approach sequentially identifies the number of classes at each level in an iterative fashion, while the latter chooses the number of discrete components at two levels simultaneously. Previous studies have shown that the stepwise

approach and the simultaneous approach performed equally well in identifying the correct number of latent classes at two levels (Lukočiené et al., 2010; Yu & Park, 2014).

The most common and well-known methods for model selection in an MLCM are likelihood-based information criteria (IC). The IC measure the relative goodness of fit by finding a balance between model fit (the log-likelihood value) and model complexity (the number of parameters). Thus, the lowest value of a given IC indicates the best-fitting model (Nylund et al., 2007). Among many IC, the Bayesian information criterion (BIC; Schwarz, 1978) has been suggested as a standard criterion for identifying the number of latent classes in the MLCM with both stepwise and simultaneous approaches even with small sample sizes (Lukočiené et al., 2010; Yu & Park, 2014). The results of prior simulation studies also showed that using the number of groups as the penalty term leads to better performance than total sample size in the BIC (Lukočiené et al., 2010; Yu & Park, 2014). This is because the penalty using total sample size is too harsh for the BIC and leads to poorer performance when deciding the number of latent classes at both levels.

In the present study, the model selection accuracy was assessed by the percentages of simulation replications in which both the number of lower and higher level classes were correctly recovered simultaneously by the BIC with the number of groups as the sample size in the penalty term. For each replication, the model with the lowest BIC value was chosen as the best-fitting model. The percentage of replications correctly identifying the true model by the BIC was calculated.

The obtained percentages were then observed to determine the minimum sample sizes for correctly recovering the true number of latent classes when the true latent structure is unknown. The acceptable recovery rate was set at 80%, which corresponds to the acceptable power level in sample size studies; however, the percentages of correctly recovered latent classes by the IC cannot be exactly interpreted as ''power'' because the IC are not intended to conduct statistical tests with null and alternative hypotheses (Dziak et al., 2014).

Parameter Estimate Bias

The bias of parameter estimates is used to assess estimated parameter accuracy in simulation studies. In practice, the true population parameters can rarely be obtained from a sample. Therefore, one important task in a simulation study is to evaluate the precision of estimated model parameters based on samples obtained under ideal sampling conditions and with knowledge of the true parameters.

The relative percentage bias (RPB) of a parameter estimate was used to quantify the size of the parameter bias. The RPBs of three parameters—$P(H_g)$, $P(X_{gi}|H_g)$, and $P(Y_{gij}|X_{gi})$—were calculated by subtracting the population value from averaged parameter estimates over replications and dividing by the population value ($[(\hat{\theta} - \theta/\theta)]\times 100$) (Muthén & Muthén, 2002). The obtained RPBs were averaged

over all three parameters to present the overall level of precision for parameter estimates.

Standard Error Bias

The second criterion is standard error bias. This criterion is important because it directly relates to the accuracy of statistical decisions, that is, an underestimated standard error increases the risk of a Type I error, while an inflated Type II error is likely to occur when the standard error is overestimated.

The standard error bias is based on the central limit theorem, which proposes that when a population is repeatedly sampled the average sample statistic gradually approaches the true population value, and the standard deviation of the samples approximates the standard error associated with population parameters (Muthén & Muthén, 2002). Based on this theorem, the *standard deviation* of each parameter estimate over simulation replications is considered the population standard error.

The RPB of standard error was calculated in a way similar to the aforementioned parameter bias by subtracting the true value (the calculated *standard deviation* of each parameter estimate) from the estimate (the estimated standard error of each parameter) and dividing by the true value. The RPBs of three parameters were then averaged over simulation replications. In this study, RPB values of 10% or less were considered acceptable, following recommendations by Muthén and Muthén (2002).

Coverage Rates

The third evaluation criterion is coverage rates, which serve as another assessment of parameter estimates and their standard error. This criterion examines whether the 95% confidence interval for each parameter contains the true parameter value. When the MLCM fits well with sufficient sample sizes, the confidence interval will include the true population values in most simulation replications. The coverage rates refer to the observed proportion of those replications in which the population value is contained within confidence intervals.

In this study, the coverage rates were calculated by assuming the indicator variable for each parameter over simulation replications. The indicator variable was coded as 1 when an estimated parameter was within the 95% confidence interval (coveraged); however, when the estimated parameter was outside the interval (noncoveraged), the indicator variable was set to 0. The acceptable coverage rates include the population value of more than 90% of the simulated data.

Simulation Study

Design Factors

A simulation study was designed to investigate sample size requirements when fitting the MLCM. Because the primary interest of the present study is placed on sample

sizes at two levels, the number of units at the higher (group) and lower (individual) levels were varied systematically. Six levels of higher level sample sizes (the number of groups, G)—10, 20, 25, 30, 50, and 100—and eight levels of lower level sample sizes (the number of individuals per group, n_g) including 5, 10, 20, 30, 40, 50, 75, and 100 were considered. The specification of two factors resulted in 48 different sample size patterns with total sample sizes ranging from 50 to 100,000. These values were chosen to cover a wide range of empirical research settings from experiments of small group designs (e.g., Van Horn et al., 2008) to large-scale survey studies (e.g., Bijmolt et al., 2004; da Costa & Dias, 2014).

Three factors that are expected to affect required sample sizes in the context of an MLCM (Finch & French, 2013; Lukočienė et al., 2010) were also manipulated: (a) the complexity of the latent structure, (b) the distinctness of latent clusters and classes, and (c) the number of indicators. The complexity of the latent structure was defined by the number of latent clusters and classes at the higher and lower levels. Two levels of latent structure complexity were considered; the latent structure with two clusters and two classes represented the less complex scenario, whereas the latent structure with three clusters and three classes was chosen to represent the more complex scenario. These two scenarios are referred to as high and low complexity in terms of the latent structure and are denoted as C_H and C_L.

The second design factor was the extent to which the clusters and classes were distinct from others. Two levels of cluster/class distinctness were considered: high-distinct conditions and low-distinct conditions, which are denoted as D_H and D_L, respectively. These two conditions differed in the values of the conditional latent class probabilities, $P(X_{ig} = m|H_g = l)$, and the conditional response probabilities, $P(Y_{ij}|X_i = m)$. For high-distinct conditions, the population values of parameters differed greatly among the clusters as well as classes. The larger difference in population values led to more distinguishable clusters and classes, therefore inducing more separated latent clusters and classes. For example, in the case of C_L (two clusters and two classes), the conditional latent class probabilities were .8 and .2 under the high-distinct condition, while the probabilities were .65 and .35 under the low-distinct condition. Similarly, the conditional response probabilities for all indicators in classes with high-distinct conditions were .8 in one class and .2 in another, while the conditional response probabilities for all indicators were .7 and .3 with low-distinct conditions. The number of items in the simulation was set at either 6 or 12 (denoted as J_6 and J_{12}).

Note that the separation among latent clusters and classes was manipulated by the aforementioned factors (Lukočienė et al., 2010). Other factors that may influence cluster and class separation were fixed in the simulation design, according to other simulation studies for MLCMs (Lukočienė et al., 2010; Yu & Park, 2014). The latent cluster probabilities (i.e., the size of the latent clusters) were assumed to be homogeneous among latent clusters, and the number of categories of the indicators was fixed to two (i.e., binary indicators).

Table 1. Multilevel Latent Class Model Parameter Specifications for Simulation Design.

		Parameters		
Complexity	Distinctness	$P(H_g)$	$P(X_{gi}\|H_g)$	$P(Y_{gij}\|X_{gi})$
C_H	D_H	$\begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$	$\begin{bmatrix} .80 & .20 \\ .20 & .80 \end{bmatrix}$	$\begin{bmatrix} .80 & .80 & .80 & .80 & .80 & .80 \\ .20 & .20 & .20 & .20 & .20 & .20 \end{bmatrix}$
	D_L	$\begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$	$\begin{bmatrix} .65 & .35 \\ .35 & .65 \end{bmatrix}$	$\begin{bmatrix} .70 & .70 & .70 & .70 & .70 & .70 \\ .30 & .30 & .30 & .30 & .30 & .30 \end{bmatrix}$
C_L	D_H	$\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$	$\begin{bmatrix} .60 & .20 & .20 \\ .20 & .60 & .20 \\ .20 & .20 & .60 \end{bmatrix}$	$\begin{bmatrix} .80 & .80 & .80 & .80 & .80 & .80 \\ .80 & .80 & .80 & .20 & .20 & .20 \\ .20 & .20 & .20 & .20 & .20 & .20 \end{bmatrix}$
	D_L	$\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$	$\begin{bmatrix} .50 & .25 & .25 \\ .25 & .50 & .25 \\ .25 & .25 & .50 \end{bmatrix}$	$\begin{bmatrix} .70 & .70 & .70 & .70 & .70 & .70 \\ .70 & .70 & .70 & .30 & .30 & .30 \\ .30 & .30 & .30 & .30 & .30 & .30 \end{bmatrix}$

Note. The pattern for $P(Y_{gij}|X_{gi})$is doubled when $J = 12$.

The specification of the five manipulated factors and population values for the parameters are presented in Table 1. The specification of these factors has been chosen to cover a wide array of MLCM structures, ranging from highly separated to poorly separated clusters and classes. To ensure that our simulation design covers various MLCM structures, the levels of separation among clusters and classes were evaluated using the R^2 entropy index[1] (Ramaswamy et al., 1993). This index measures the uniqueness of discrete latent components based on their posterior probabilities; values close to 1 indicate that they are well-separated, whereas values close to 0 suggest that they are not distinguishable.

The specifications of the design resulted in the $R^2_{entropy.high}$ ranging from 0.01 to 1 and the $R^2_{entropy.low}$ ranging from 0.01 to 1. The average $R^2_{entropy.high}$ and $R^2_{entropy.low}$ across all conditions were 0.70 ($SD = 0.31$) and 0.66 ($SD = 0.24$), respectively. The average values of the two indices were slightly lower than those in the previous simulation study by Lukočienė et al. (2010).

Data Generation and Analysis

The design yielded 384 ($6 \times 8 \times 2 \times 2 \times 2$) simulation conditions by crossing the five manipulated factors as previously described. Based on the parameter specification in Table 1, 500 data sets were generated using the routine of the random number generator in R 3.1.1 software (R Development Core Team, 2010). The simulated data set was then fitted to the five models with different latent class structures (single-level LCMs with two and three classes; MLCMs with two clusters/two classes, two clusters/three classes, and three clusters/three classes) using the Latent GOLD 5.0 syntax module (Vermunt & Magidson, 2013). We introduced single-level LCMs as alternative models in addition to the MLCMs to consider the misspecified model in which

higher level structures were not incorporated in the models. Among the five models, one of them was the ''true model'' from which the data were generated. The log-likelihood values of each data set fitted to the five models were collected, and then we compute the BIC values. The model with the lowest BIC value was considered to be the best-fitting model, the percentage of correctly identifying the true model was also calculated among the replications.

The parameter estimates and their standard error were also recorded for each replication to calculate the RPBs and to determine whether the 95% confidence interval contained the specified population value or not. The obtained RPB and coverage rates for each parameter were averaged across the replications to determine whether they were within the acceptable range.

One potential issue to be considered while fitting the MLCM is label switching (McLachlan & Peel, 2000). This problem occurs because the ordering of latent classes and clusters is arbitrary in different iterations. Therefore, the ordering of the estimated parameters associated with latent clusters and classes may not be identical across data sets. To avoid the label switching, the estimated parameters were rotated back to a standard parameter order. This procedure was done by calculating the means for each parameter and rounding them to the first decimal point; according to the means, they were then rematched and rotated back to the same order. This procedure ensured that the parameter estimates across different data sets were in the same order for comparisons.

Simulation Results

Among the 192,000 ($6 \times 8 \times 2 \times 2 \times 2 \times 500$) replications, 164 replications (0.001%) produced a convergence problem. These problems were found mostly under the conditions of a highly complex latent structure (C_H) with the smallest sample sizes ($n_g = 5$ and $G = 10$). Any replication with the convergence problem was excluded from the analysis; thus, a total of 191,836 replications were included in the following analyses.

The averaged RPB and coverage rates of three MLCM parameters (i.e., latent cluster probability, conditional latent class probability, and conditional response probability) were evaluated to understand the effects of design factors in terms of the evaluation criteria.

Model Selection Accuracy

The percentage of accurately identifying the true model using the BIC is used to evaluate the model selection accuracy. The percentage of correctly identifying the true model was 62.5% across all simulation conditions. The detailed patterns showed that the BIC performed well in determining the correct number of latent classes and clusters under D_H and C_L conditions. Under D_L and C_H conditions, in contrast, the BIC was likely to underestimate the number of clusters and classes, particularly,

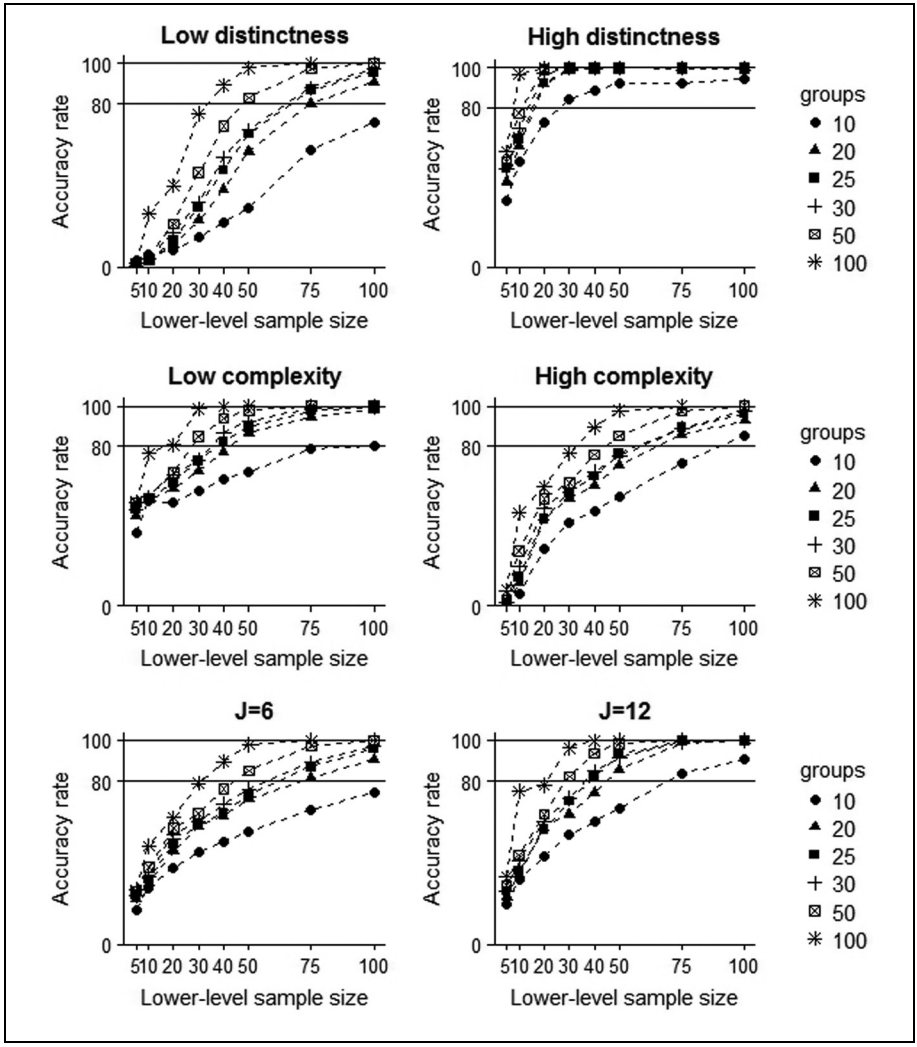

Figure 1. Average model selection accuracy rates across the 48 patterns of sample sizes and different levels of three design factors.

when sample sizes were small. The reason such underestimation occurred was that the BIC penalized the complexity of the model (additional estimated model parameters) severely in situations where small sample sizes were combined with low-distinct conditions and more complex latent structures.

Figure 1 exhibits the average model selection accuracy rates under the different levels of the three design factors. All figures feature the value of the average accuracy

rates on the y-axis and the change of n_g on the x-axis, which is spaced according to its size. As shown in the top panel of Figure 1, the average accuracy rates under the D_H condition (M = 87.64, SD = 19.28) were much higher than those under the D_L condition (M = 45.31, SD = 36.05). This pattern indicates that a greater degree of distinctness among clusters/classes provides better accuracy rates over poorer distinctness under most sample size conditions. The accuracy rate under the D_L condition with the smallest n_g (n_g = 5) was particularly low, showing nearly 0% of recovery rates (M = 0.75), while the average accuracy rate with the same n_g was 33.5% under the D_H condition.

The middle panel of Figure 1 presents the model selection accuracy rates under the two levels of model complexity. The results showed that the accuracy rate under the C_L condition was, on average, 75.97% (SD = 19.70), whereas the rate under the C_H condition was 56.98% (SD = 32.32). The accuracy rates clearly increased as the G and n_g increased under both conditions, but this pattern was more prominent under the C_H condition. One notable pattern was that when there was a sufficient n_g (n_g = 100), the accuracy rates under the C_L condition with small G (G = 10) were relatively lower than those under the C_H condition.

The bottom panel of Figure 1 exhibits the accuracy rates under the two different numbers of indicators. The results showed that slightly lower accuracy rates were observed under the condition of fewer indictors (J_6) (M = 61.82, SD = 25.08) than under the condition of more indicators (J_{12}) (M = 71.13, SD = 26.82). In particular, the accuracy rates reached almost 100% under the J_{12} condition when n_g was larger than 75 ($n_g > 75$) except for the smallest G (G = 10), but a similar level of accuracy rates could not be achieved under the J_6 condition even with the largest n_g (n_g = 100).

Parameter Estimate Bias

The average RPBs of parameter estimates under the two conditions of cluster/class distinctness are presented in the top panel of Figure 2. As the figure shows, the parameter estimates under the D_H condition (M = 9.26, SD = 5.15) were slightly less biased than the D_L condition (M = 11.02, SD = 4.92) in general.

The results also showed that the RPBs under the D_H condition were considerably lower than those under the D_L condition, particularly when n_g was smaller than 40 ($n_g < 40$). For instance, when n_g was 20 (n_g = 20), the average RPB over all of G was 9.18 (SD = 4.45) under the D_H condition, while the average value under the D_L condition was 12.60 (SD = 5.54). Furthermore, the RPBs clearly decreased as G increased under both conditions; however, the decreasing pattern was particularly prominent under the D_H condition with small n_g ($n_g < 30$).

The middle panel of Figure 2 illustrates the RPBs of parameter estimates under the two conditions of model complexity. In general, higher RPBs were consistently found under the C_H condition (M = 12.94, SD = 6.21) rather than the C_L condition (M = 7.35, SD = 3.81) across all the sample size conditions. The results also revealed that a drastically larger bias was found under the C_H condition, particularly when n_g

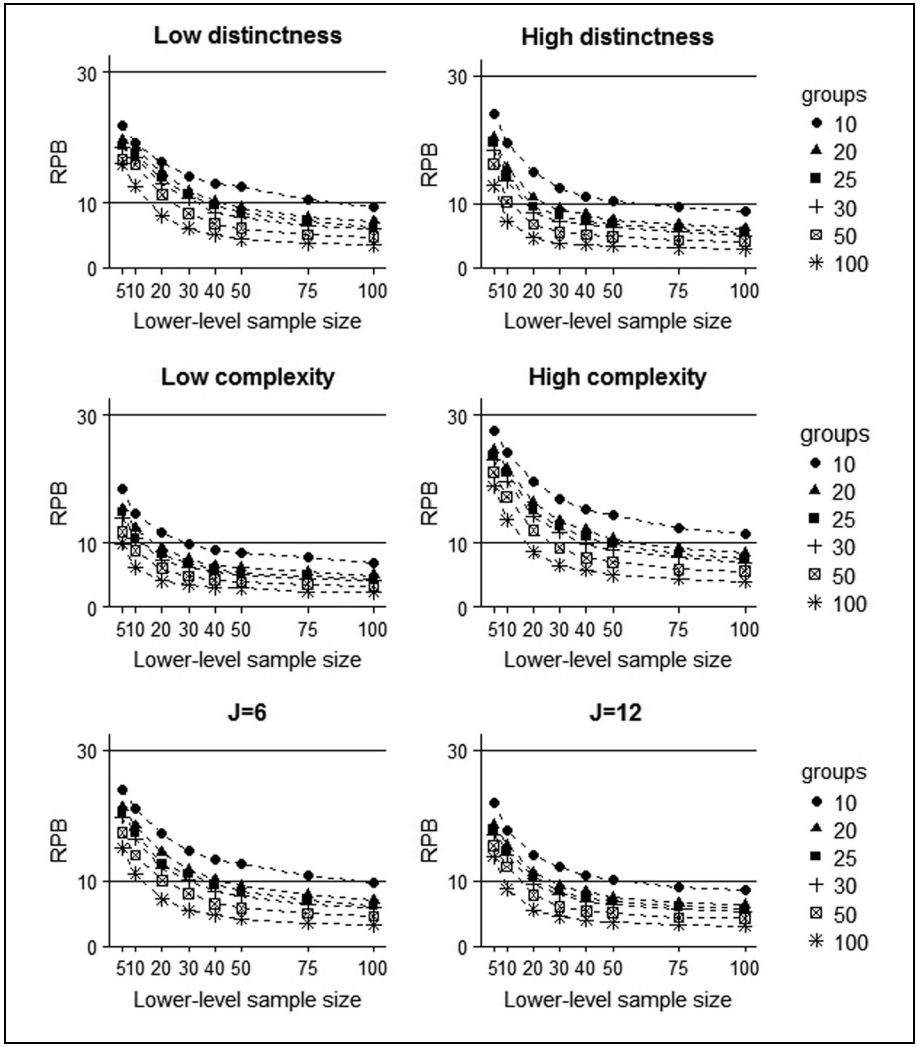

Figure 2. Average relative percentage bias (RPB) of parameter bias across the 48 patterns of sample sizes and different levels of three design factors.

was smaller than 20; the average RPB in such cases was 21.13 (*SD* = 5.24) under the C_H condition, while that under the C_L condition was 12.26 (*SD* = 3.42).

The bottom panel of Figure 2 displays the RPBs of parameter estimates under the two different conditions of the number of indicators. Overall, a similar pattern was found between the two conditions; the average RPB under the J_6 condition was 11.02 (*SD* = 5.32), and that under the J_{12} condition was 10.26 (*SD* = 4.63). Even with similar RPB patterns between two conditions, the J_6 condition required slightly larger

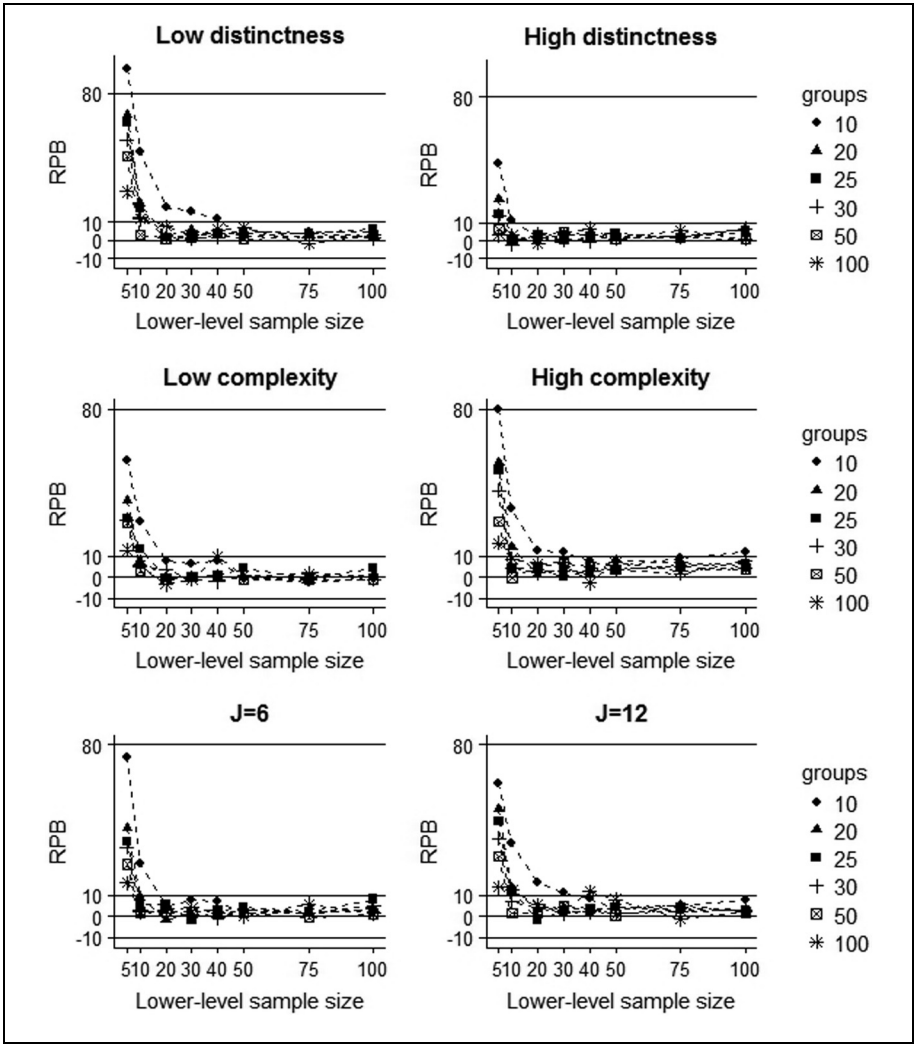

Figure 3. Average relative percentage bias (RPB) of standard error across the 48 patterns of sample sizes and different levels of three design factors.

samples than the J_{12} condition; the results suggest that an n_g of 75 was sufficient to reach the criterion under the J_{12} condition with any group size, while the J_6 condition needed the largest n_g ($n_g = 100$) to satisfy the criterion.

Standard Error Bias

The top panel of Figure 3 shows the RPBs of the standard error under the two levels of cluster and class distinctness. Largely biased standard errors were found under

both conditions when the sample sizes were smallest (n_g = 5 and G = 10); the average RPB was 42.90 under the D_H condition and 93.20 under the D_L condition. The RPBs decreased as G and (or) n_g increased, with the decreasing trend being relatively stronger under the D_L condition. Moreover, in the cases of small n_g ($n_g < 20$), the RPBs decreased rapidly as the sample sizes at both levels increased; however, after ensuring sufficient samples at both levels ($n_g > 20$ and $G > 20$), adding more samples did not influence the bias of standard errors for both conditions.

The RPBs of standard error under different complexity levels are shown in the middle panel of Figure 3. A similar overall pattern was held for the two conditions, except when the sample sizes were smallest (n_g = 5 and G = 10). In such a case, the average RPB under the C_H condition (M = 55.87) was particularly higher than that under the C_L condition (M = 55.87), and the RPBs dropped rapidly as the number of samples increased at both levels.

The bottom panel of Figure 3 presents the RPBs of standard error under different numbers of indicators. The results showed that there was a rapid decrease in the RPBs as the sample sizes at both levels increased, particularly when n_g was less than 20 ($n_g < 20$), while the decreasing pattern was less pronounced under the J_{12} condition.

These patterns suggest that the standard error was largely biased when n_g was small ($n_g < 30$), and such bias was more pronounced under less distinctive and more complex latent cluster and class structures. However, when there was sufficient n_g ($n_g > 30$), the standard error became stable under most conditions.

Coverage Rates

The average coverage rates under the two conditions of cluster and class distinctness are presented in the top panel of Figure 4. The overall coverage rates under the two conditions were within the acceptable range; however, the coverage rates under the D_H condition (M = 92.89, SD = 2.07) were slightly higher than those under the D_L condition (M = 91.03, SD = 2.06).

The results also showed that the coverage rates tended to increase as n_g increased under the D_H condition. However, reversed u-shaped patterns were observed under the D_L condition, particularly when G was larger than 25. This pattern was observed mainly because the parameter was well recovered in such conditions with an extremely small standard error. This resulted in very narrow confidence intervals around the sampling estimate distribution, thereby reducing the convergence rates because of rounding errors.

Coverage rates varied according to the levels of model complexity, as shown in the middle panel of Figure 4. Higher coverage rates were found under the C_L condition (M = 93.53, SD = 1.53) than under the C_H condition (M = 89.48, SD = 2.29), and the difference between the two conditions was more prominent when n_g was relatively small ($n_g < 30$). Moreover, there was a trend of increasing coverage rates for

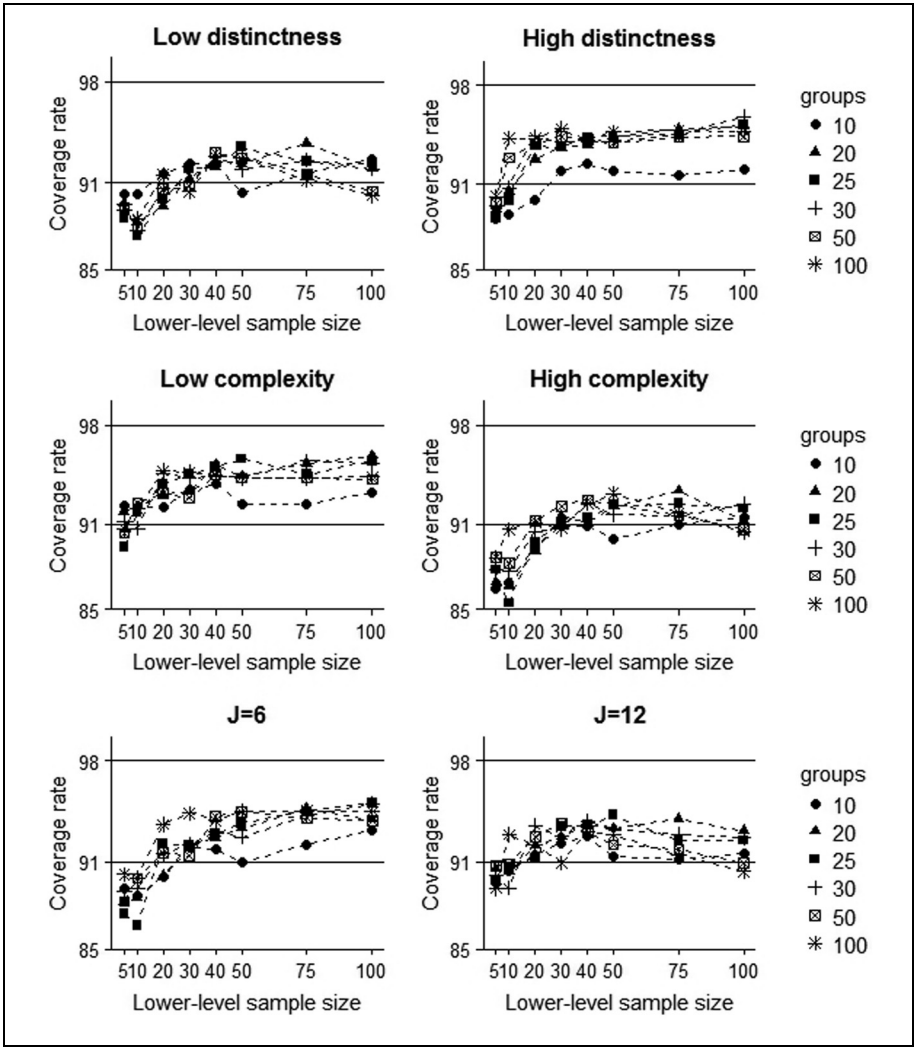

Figure 4. Average coverage rates across the 48 patterns of sample sizes and the different levels of three design factors.

both conditions as n_g increased, while increased numbers of G did not result in a dramatic increase in the coverage rates.

As shown in the bottom panel of Figure 4, the coverage rates were lower under the J_6 condition ($M = 90.48$, $SD = 1.97$) than under the J_{12} condition ($M = 93.53$, $SD = 1.52$), especially when n_g was small ($n_g < 20$). It is notable that reversed U-shaped patterns were also found under the J_{12} condition because of the small standard error estimate.

Table 2. Minimum Sample Size Requirements per Group (n_g) Under the Simulation Design.

Distinctness	Indicator	Number of parameters	G	Minimum n_g (C_L)	Minimum n_g (C_H)
D_H	6	15/26	10	40	>100
			20	20	50
			25	20	40
			30	10	40
			50	10	20
			100	10	20
	12	27/44	10	30	>100
			20	20	30
			25	20	30
			30	10	20
			50	5	20
			100	5	10
D_L	6	15/26	10	>100	>100
			20	75	100
			25	75	100
			30	50	100
			50	40	75
			100	30	50
	12	27/44	10	>100	>100
			20	50	75
			25	40	50
			30	40	50
			50	30	40
			100	10	30

Recommended Sample Sizes for Applying the MLCM

In multilevel studies, the problems related to sample sizes often occur at the group level because the higher level samples are usually smaller than the lower level samples, and increasing the number of groups may be more difficult than increasing the number of individuals due to the costs associated with data collection (Maas & Hox, 2005). Therefore, we suggest the minimum required number of lower level samples (n_g) satisfying the four criteria in each simulation condition.

Table 2 presents the minimum n_g required for each of the simulation conditions. The results revealed that the required n_g varied greatly depending on the size of G. In general, the required n_g was negatively associated to G. That is, as the number of G increased, the required n_g clearly decreased across all simulation conditions. For example, as G increased from 10 to 100, the required n_g under the C_L condition decreased from 75 to 5, whereas n_g under the C_H condition varied from 100 to 10. This pattern implies that a lower number of groups can be compensated for by an increased number of individuals within each group.

Three design factors (cluster and class distinctness, latent structure complexity, and the number of indicators) had substantial impacts on the required sample size.

The results revealed that the D_L condition required a considerably larger n_g than the D_H condition. That is, the required n_g to meet the priori criteria under the D_H condition ranged from 10 to 50, while the required n_g ranged from 30 to 100 under the D_L condition. Under the D_L condition with the smallest G ($G = 10$), even the largest n_g ($n_g = 100$) could not satisfy all the criteria regardless of other factors, which implies that n_g needed to be more than 100 under this condition.

The sample size requirements for the C_L condition satisfying all criteria were generally less than those under the C_H condition across all levels of G. The range of the minimum required n_g was 5 to 75 under the C_L condition, whereas the C_H condition necessitated an n_g ranging from 10 to 100. For the C_H condition with the smallest G ($G = 10$), the criteria could not be met even with largest n_g ($n_g = 100$), while an n_g of only 40 was sufficient to reach the criteria under the C_L condition with the same group size ($G = 10$) as long as the clusters and classes were distinctive enough (D_H).

Table 2 also suggests that the J_6 condition required larger n_g to meet the criteria than the J_{12} condition. According to the results, n_g needed to be at least 40 under the J_6 condition with high distinctness (D_H) and low-complex structure (C_L), but an n_g of 30 was sufficient to meet the criteria for the J_{12} condition with the same levels of distinctness and complexity.

Discussion and Conclusion

The parameters of the MLCM were estimated using the ML methods. The ML estimates would converge to their population values only when the sample size was sufficiently large. Therefore, knowing the minimum sample size requirement is important to obtain unbiased and reliable parameter estimates when applying the MLCM to empirical data analysis.

In this study, we conducted a simulation to investigate the sample size requirements for fitting MLCMs. We examined the effects of a series of design factors on the sample size requirements. The results revealed that the sample size requirements at the lower level depend heavily on the design factors, that is, smaller lower level samples (ranging from 5 to 30) were required for conditions with high distinctness, low-complex latent structure, and more indicators, whereas larger samples (ranging from 50 to 100) were needed in cases with low distinctness, high-complex latent class structure, and fewer indicators.

To determine the minimum sample size requirements for the MLCM, we considered four evaluation criteria: the accuracy of model selection, parameter estimate bias, standard error bias, and coverage rates. The findings from the current study related to model selection accuracy are consistent with previous studies. The results showed that the distinctness and complexity among latent clusters and classes, as well as the number of indicators, play a significant role in recovering the true latent structure (Lukočienė et al., 2010; Yu & Park, 2014). The findings of standard error bias and coverage rates are also partially in line with previous studies within linear mixed models, that is, the estimated standard error was downwardly biased (i.e., positive

values in the RPB) when higher level sample sizes were small ($G < 20$) (Maas & Hox, 2005; McNeish & Stapleton, 2016). Moreover, coverage rates improved as sample sizes at both levels increased, but the rates were particularly low under the condition in which the lower level sample was very small ($n_g = 5$) (McNeish & Harring, 2017).

We observed a clear trade-off between required sample sizes at two levels, which also is found in linear mixed models (Maas & Hox, 2005; Scherbaum & Ferreter, 2009). Specifically, as fewer groups are available, additional lower level samples are required to meet the evaluation criteria. This indicates that having sufficient numbers of lower level samples partially compensates for the problems caused by a small number of groups (e.g., biases in model selection and inaccurate parameter estimations). Therefore, if researchers are faced with a situation in which the number of available groups is limited, increasing the number of individuals per group is beneficial.

However, some of the findings are not consistent with previous sample size works. For example, as the number of indicators increased, the required number of sample sizes in the MLCM decreased to some extent; however, the decreasing trends are not as dramatic as those of other models, such as the structural equation model or growth mixture model (Kim, 2012; Wolf, Harrington, Clark, & Miller, 2013). This result may be due to the MLCM specification that indicators have a partial effect on only the lower level classes. Specifically, having a greater number of indicators provides additional information and results in better separation among the classes (Lukočienė et al., 2010), while the higher level clusters are built based directly on classes, not indicators. Furthermore, previous studies reported that the bias in fixed effect parameters was not affected by factors such as the ICC values and the number of groups (McNeish & Harring, 2017), but the bias in the MLCM parameters is heavily associated with factors related to class structure (model complexity and class distinctness) as well as sample sizes at both levels.

The main contribution of this study is providing rules of thumb for sample size requirements when applying the MLCM in data analysis. Specifically, the recommendations are as follows: (a) at least 20 groups are needed to meet all four criteria; (b) at least 10 individuals are needed per group to obtain reliable results, unless the number of groups and indicators are large enough; and (c) when fitting the complex model, each group needs to have at least 30 individuals if the number of indicators is limited. We believe that these guidelines will significantly help researchers who are in the planning stages.

This study provided some general guidelines for sample sizes in the MLCM; however, there is one limitation. In theory, various levels of cluster and class complexity may exist. In this study, only two representative cases were chosen: simple structure (two clusters, each with two classes) and complex structure (three clusters, each with three classes). Although the chosen structures reflect two different levels of complexity, they may not cover all possible latent structures in practice. Future direction in this line of research includes examining a broad array of latent structures, evaluating the impact of insufficient sample sizes when using MLCMs, and developing indices to quantify the dependency between higher and lower latent classes.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Note

1. The R^2 entropy index is denoted as $R^2_{entropy.high}$ at the higher level and $R^2_{entropy.low}$ at the lower level.

Supplementary Material

Supplementary material for this article is available online.

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*, 117-128.

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

Austin, P. C. (2010). Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures. *International Journal of Biostatistics, 6*, Article 16.

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, *18*, 151-164.

Bartolucci, F., Pennoni, F., & Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics*, *36*, 491-522.

Bassi, F. (2009). Latent class models for marketing strategies: An application to the Italian pharmaceutical market. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *5*, 40-45.

Bijmolt, T. H. A., Paas, L. J., & Vermunt, J. K. (2004). Country and consumer segmentation: Multi-level latent class analysis of financial product ownership. *International Journal of Research in Marketing*, *21*, 323-334.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473-514.

Bryk, A. S., & Raudenbush, S. W. (1992). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*, 147-158.

Chung, H., Anthony, J. C., & Schafer, J. L. (2011). Latent class profile analysis: An application to stage sequential processes in early onset drinking behaviours. *Journal of the Royal Statistical Society, Series A*, *174*, 689-712.

Cohen, M. P. (1998). Determining sample sizes for surveys with data analyzed by hierarchical linear models. *Journal of Official Statistics*, *14*, 267-275.

da Costa, L. P., & Dias, J. G. (2014). What do Europeans believe to be the causes of poverty? A multilevel analysis of heterogeneity within and between countries. *Social Indicators Research*, *122*, 1-20.

Di, C. Z., & Bandeen-Roche, K. (2011). Multilevel latent class models with dirichlet mixing distribution. *Biometrics*, *67*, 86-96.

Dziak, J. J., Lanza, S. T., & Tan, X. (2014). Effect size, statistical power and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. *Structural Equation Modeling*, *21*, 534-552. doi:10.1080/10705511.2014.919819

Finch, W. H., & French, B. F. (2013). Multilevel latent class analysis: Parametric and nonparametric models. *Journal of Experimental Education*, *82*, 307-333.

Finch, W. H., & Marchant, G. J. (2013). Application of multilevel latent class analysis to identify achievement and socio-economic typologies in the 20 wealthiest countries. *Journal of Educational and Developmental Psychology*, *3*, 201-221.

Goldstein, H. (1995). *Multilevel statistical models*. New York, NY: Halsted.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology*, *79*, 1179-1259.

Gudicha, D. W., Schmittmann, V. D., & Vermunt, J. K. (2016). Power computation for likelihood ratio tests for the transition parameters in latent Marko models. *Structural Equation Modeling*, *23*, 234-245.

Gudicha, D. W., Tekle, F. B., & Vermunt, J. K. (2016). Power and sample size computation for Wald tests in latent class models. *Journal of Classification*, *33*, 30-51.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, *22*, 1433-1446.

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.

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications* (2nd ed.). New York, NY: Routledge.

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling*, *14*, 202-226.

Kaplan, D., & Keller, B. (2011). A note on cluster effects in latent class analysis. *Structural Equation Modeling*, *18*, 525-536.

Kaplan, D., Kim, J.-S., & Kim, S.-Y. (2009). Multilevel latent variable modeling: Current research and recent developments. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *Handbook of quantitative methods in psychology* (pp. 592-612). Thousand Oaks, CA: Sage.

Kim, S.-Y. (2012). Sample size requirements in single- and multiphase growth mixture models: A Monte Carlo simulation study. *Structural Equation Modeling*, *19*, 457-476. doi: 10.1080/10705511.2012.687672

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Unpublished manuscript, California State University, Los Angeles.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*, 805-811.

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

Lukočiené, 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.

Lukočiené, O., & Vermunt, J. K. (2010). Determining the number of components in mixture models for hierarchical data. In A. Fink, L. Berthold, W. Seidel & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 241-249). Berlin-Heidelberg, Germany: Springer.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 85-91.

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

McNeish, D., & Harring, J. R. (2017). Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communications in Statistics: Simulation and Computation*, *46*, 855-869.

McNeish, D., & Stapleton, L. M. (2016). The effect of small sample size on two level model estimates: A review and illustration. *Educational Psychology Review*, *28*, 295-314.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study on the sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*, 1-10.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599-620.

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

Onwezen, M. C., Reinders, M. J., Lans, V. D. I., Sijtsema, S. J., Jasiulewicz, A., Guardia, M. D., & Guerrero, L. (2012). A cross-national consumer segmentation based on contextual differences in food choice benefits. *Food Quality and Preference*, *24*, 276-286.

Palardy, G., & Vermunt, J. K. (2010). Multilevel growth mixture models for classifying groups. *Journal of Educational and Behavioral Statistics*, *35*, 532-565.

Pirani, E. (2011). Evaluating contemporary social exclusion in Europe: A hierarchical latent class approach. *Quality and Quantity*, *47*, 923-941.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

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

Rights, J. D., & Sterba, S. K. (2016). The relationship between multilevel models and nonparametric multilevel mixture models: Discrete approximation of intraclass correlation, random coefficient distributions, and residual heteroscedasticity. *British Journal of Mathematical and Statistical Psychology*, *69*, 316-343.

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.

Rüdiger, M., & Hans-Dieter, D. (2013). University and student segmentation: Multilevel latent class analysis of students' attitudes toward research methods and statistics. *British Journal of Educational Psychology, 83*. 280-304.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, *12*, 347-367.

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

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC Press.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, *57*, 748-761.

Tein, J.-Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling*, *20*, 640-657.

Tekle, F. B., Gudicha, D. W., & Vermunt, J. K. (2016). Power analysis for the bootstrap likelihood ratio test for the number of classes in latent class models. *Advances in Data Analysis and Classification*, *10*, 209-224.

Tueller, S. J., & Lubke, G. H. (2011). Evaluation of structural equation mixture models: Parameter estimates and correct class assignment. *Structural Equation Modeling*, *17*, 165-192.

Van Horn, M. L., Fagan, A. A., Jaki, T., Brown, E. C., Hawkins, J. D., Arthur, M. W., & … Catalano, R. F. (2008). Using multilevel mixtures to evaluate intervention effects in group randomized trials. *Multivariate Behavioral Research*, *43*, 289-326.

Varriale, R., & Vermunt, J. K. (2012). Multilevel mixture factor models. *Multivariate Behavioral Research*, *47*, 247-275.

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

Vermunt, J. K. (2004). An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica*, *58*, 220-233.

Vermunt, J. K., & Magidson, J. (2008). *LG-Syntax user's guide: Manual for Latent GOLD 4.5 syntax module*. Belmont, MA: Statistical Innovations.

Vermunt, J. K., & Magidson, J. (2013). *Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*, 913-934.

Wurpts, I. C., & Geiser, C. (2014). Is adding more indicators to a latent class analysis beneficial or detrimental? Results of a Monte-Carlo study. *Frontiers in Psychology*, *5*, 1-15.

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

Yang, C. C., & Yang, C. C. (2007). Separating latent classes by information criteria. *Journal of Classification*, *24*, 183-203.

Yu, H.-T., & Park, J. (2014). Simultaneous decision on the number of latent clusters and classes for multilevel latent class models. *Multivariate Behavioral Research*, *49*, 232-244.