Ignoring a Multilevel Structure in Mixture Item Response Models: Impact on Parameter Recovery and Model Selection

Applied Psychological Measurement
2018, Vol. 42(2) 136–154
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621617711999
journals.sagepub.com/home/apm

Woo-yeol Lee¹, Sun-Joo Cho¹, and Sonya K. Sterba¹

Abstract

The current study investigated the consequences of ignoring a multilevel structure for a mixture item response model to show when a multilevel mixture item response model is needed. Study 1 focused on examining the consequence of ignoring dependency for within-level latent classes. Simulation conditions that may affect model selection and parameter recovery in the context of a multilevel data structure were manipulated: class-specific ICC, cluster size, and number of clusters. The accuracy of model selection (based on information criteria) and quality of parameter recovery were used to evaluate the impact of ignoring a multilevel structure. Simulation results indicated that, for the range of class-specific ICCs examined here (.1 to .3), mixture item response models which ignored a higher level nesting structure resulted in less accurate estimates and standard errors (SEs) of item discrimination parameters when the number of clusters was larger than 24 and the cluster size was larger than six. Class-varying ICCs can have compensatory effects on bias. Also, the results suggested that a mixture item response model which ignored multilevel structure was not selected over the multilevel mixture item response model based on Bayesian information criterion (BIC) if the number of clusters and cluster size was at least 50, respectively. In Study 2, the consequences of unnecessarily fitting a multilevel mixture item response model to single-level data were examined. Reassuringly, in the context of single-level data, a multilevel mixture item response model was not selected by BIC, and its use would not distort the within-level item parameter estimates or SEs when the cluster size was at least 20. Based on these findings, it is concluded that, for class-specific ICC conditions examined here, a multilevel mixture item response model is recommended over a single-level item response model for a clustered dataset having cluster size >20 and the number of clusters >50 .

Keywords

mixture item response model, model selection, multilevel data

Single-level mixture item response models or categorical-item factor mixture models were proposed to account for both categorical latent variables (i.e., latent classes) and continuous latent variables (i.e., factors) in a population (e.g., Rost, 1990). The single-level mixture item response

¹Vanderbilt University, Nashville, TN, USA

Corresponding Author:

Sun-Joo Cho, Vanderbilt University, Peabody Hobbs 213A, 230 Appleton Place, Nashville, TN 37203, USA.
Email: sj.cho@vanderbilt.edu

model is similar to a multigroup item response model (Bock & Zimowski, 1997), except that the group of interest is a latent class or categorical latent variable. The two-parameter single-level mixture item response model as an extension of mixture Rasch model (Rost, 1990) can be written as

$$\text{logit}[P(y_{ji} = 1 | \theta_{jg}, C_j)] = \alpha_{ig}\theta_{jg} - \beta_{ig}, \quad (1)$$

where j is an index for a person ($j = 1, \dots, J$), i is an index for an item ($i = 1, \dots, I$), C_j is a categorical latent variable at the within level (called a within-level latent classification variable; $C_j = 1, \dots, g, \dots, G$) for a person j , θ_{jg} is a class-specific continuous latent variable (e.g., ability; assumed to follow $N(0, \sigma_g^2)$), α_{ig} is a class-specific item discrimination parameter, and β_{ig} is a class-specific item location parameter (e.g., difficulty parameter). The mixture item response models have interested researchers because they are useful for measuring individual differences when distinct subpopulations are assumed in the overall population (see Cho, Suh, & Lee, 2016, for reviews of applications of mixture item response models).

Cluster or multistage sampling is common in educational and psychological research. When persons (e.g., students) are nested within a higher level structure (e.g., schools), the persons' scores in the same higher level unit are likely to be more highly correlated with one another than those from different higher level units. If the dependency due to clustering is not accounted for when it exists, results from a latent variable model will be less accurate because of a violation of the (local) independence assumption. In this regard, multilevel mixture item response models (Vermunt, 2007) were developed to account for possible dependency due to cluster- or multistage sampling. In the model, dependency is taken into account by incorporating continuous and/or categorical latent variables at the higher level. The multilevel mixture item response models were applied to detect differential item functioning across latent classes in multilevel data (Bennink, Croon, Keuning, & Vermunt, 2014; Cho & Cohen, 2010; Finch & Finch, 2013) and to investigate the individual differences in math ability growth within latent classes over time in multilevel longitudinal data (Cho, Cohen, & Bottge, 2013; von Davier, Xu, & Carstensen, 2011). Although many data structures in educational and psychological research involve clustering, multilevel mixture item response models are seldom used over single-level mixture item response models. As an example, in a survey of Cho et al. (2016), five multilevel mixture item response model applications were found, whereas 19 single-level mixture item response model applications were found.

To examine the necessity of multilevel mixture item response models, additional research is needed to investigate the impact of inappropriately modeling multilevel data. Recent simulation studies showed that misspecified models which ignored a higher level nesting structure resulted in less accurate estimates and lower classification accuracy in a growth mixture model (Chen, Kwok, Luo, & Wilson, 2010) and a latent class model (Kaplan & Keller, 2011; Park & Yu, 2016). Chen et al. (2010) found that, compared with a fitted multilevel growth mixture model, the classification of persons using a growth mixture model was less accurate, and the classification accuracy was affected by the intraclass correlation (ICC), mixing proportions (i.e., the number of persons within a class), and within-class variances and covariances of random effects. Kaplan and Keller (2011) showed that a larger ICC and smaller sample size at the higher level resulted in misclassified persons in a latent class model, controlling for the total sample size. Park and Yu (2016) found inflated standard errors (hereafter, *SEs*) for parameter estimates of a latent class model when a higher level nesting structure was ignored, whereas *SE* deflation is often found when ignoring nesting in other modeling contexts (e.g., linear regression; Raudenbush & Bryk, 2002). The consequences were more severe as the between-level latent classes became more separated from one another in terms of the mixing proportion of within-level latent classes.

However, the impact of ignoring dependency in item responses (due to clustering) on parameter accuracy has not been investigated in mixture item response modeling. Dependency in item responses is explained by both continuous and categorical latent variables in mixture item response models, whereas it is explained only by categorical latent variables in latent class models. Because of this difference in complexity between the two models, the sample sizes and ICCs where the impact of ignoring dependency becomes consequential in mixture item response models may not be directly inferred from prior simulation results that used latent class models as in Park and Yu (2016) and Kaplan and Keller (2011). Different kinds of the latent variables are expected to yield different results regarding parameter recovery and model selection. The same design conditions can have very different effects on, for instance, mixture model selection and class enumeration depending on the complexity of the within-class model (e.g., Lubke & Muthén, 2007, vs. Tofghi & Enders, 2007).

Furthermore, in conventional mixture item response model applications, the number of latent classes is not a model parameter and is typically chosen using information criteria (IC) when comparing models having different numbers of latent classes. Thus, a researcher deciding whether to fit a single-level versus multilevel mixture item response model to multilevel data must consider not only the risk of biased estimates and *SEs* when fitting a misspecified single-level model but also whether the correct number of classes could be recovered when fitting a multilevel model. There have been studies investigating IC such as Akaike's information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978) for class enumeration in (single-level) mixture item response models (e.g., Li, Cohen, Kim, & Cho, 2009; Preinerstorfer & Formann, 2012). These studies found that BIC performed best among model selection indices they considered. BIC includes the number of parameters and sample size in the penalty term. For multilevel data, Yu and Park (2014) compared different kinds of IC for multilevel latent class models. They reported that using the number of clusters leads to a slightly better performance than total sample size in BIC and consistent AIC (CAIC; Bozdogan, 1987). Akaike's BIC (ABIC; Akaike, 1980) performed much better when total sample size is used. However, it has not been shown whether these findings from the multilevel latent class models can be generalized to multilevel mixture item response models having continuous latent variables.

Thus, the purpose of the current study is to investigate via simulation the consequences of fitting a single-level mixture item response model (a) in the presence of multilevel structure and (b) when it is not needed. Specifically, the accuracy of parameter recovery and model selection was compared when fitting single-level versus multilevel mixture item response models. For this comparison purpose, the performance of IC was investigated for selecting the "correct" number of latent classes in (multilevel) mixture item response models.

In the following, multilevel mixture item response models are described. Next, evaluation measures are shown. Subsequently, two simulation studies are presented to achieve the purpose of the study. Finally, simulation results are summarized and discussed.

Multilevel Mixture Item Response Models

Multilevel mixture item response models account for possible dependency due to clustering by incorporating continuous and/or categorical latent variables at the between level. Vermunt (2007) suggested eight possible versions of two-level (e.g., students nested within schools) mixture item response models. Because mixture models posit categorical latent variables and item response models posit continuous latent variables, latent variables at each level may be categorical, continuous, or both categorical and continuous. In the current study, the population-generating multilevel item response model has both categorical and continuous latent variables

at the within level and a continuous latent variable at the between level for the following two reasons: First, it has been common for empirical applications using cross-sectional nested data to theorize and specify continuous latent variables rather than categorical latent variables (i.e., between-level latent classes) at the between level (see Asparouhov & Muthén, 2006). Second, when categorical latent variables at the between level are specified, they are often used simply to nonparametrically approximate continuous latent variable(s) at the between level (e.g., Rights & Sterba, 2016; Vermunt, 2008).

Online Appendix A depicts a two-level mixture item response model with within-level latent classes. In the figure, the squares represent item responses, and the ellipses represent latent variables. The 1 inside the triangle is given to represent a vector of 1s. As shown in the figure, dependency in item responses $[y_{jk1}, \dots, y_{jki}, \dots, y_{jkl}]'$ is explained by both categorical and continuous latent variables (C_{jk} and θ_{jkg} , respectively) at the within level and by a continuous latent variable (θ_k) at the between level. The arrows from the continuous latent variables to the item responses represent item discrimination parameters ($\alpha_{ig,W}$ and $\alpha_{i,B}$ at Levels 2 and 3, respectively), and the arrows from the triangle to the item responses represent item location parameters. The dotted arrows from the categorical latent variable to the other arrows indicate that all item parameters are class-specific.

The population-generating multilevel mixture item response model described in Online Appendix A can be specified as follows:

$$\text{logit}[P(y_{jki} = 1 | \theta_{jkg}, \theta_k, C_{jk})] = \alpha_{ig,W} \theta_{jkg} + \alpha_{i,B} \theta_k - \beta_{ig}, \quad (2)$$

where k is an index for a cluster ($k = 1, \dots, K$), C_{jk} is a categorical latent variable at the within level (called a within-level latent classification variable; $C_j = 1, \dots, g, \dots, G$) for a person j nested within a cluster k , $\alpha_{ig,W}$ is a class-specific within-level item discrimination parameter, $\alpha_{i,B}$ is a between-level item discrimination parameter, β_{ig} is a class-specific item location parameter, θ_{jkg} is a class-specific within-level continuous latent variable (assumed to follow $N(0, \sigma_g^2)$), and θ_k is a between-level continuous latent variable (assumed to follow $N(0, \tau^2)$).

For model selection purposes, candidate-fitted models include not only Equation 1 and Equation 2 models but also the most complicated multilevel item response model having both categorical and continuous latent variables at the within level and at the between level. The latter model can be specified by adding a subscript (e.g., h) to all parameters of Equation 2 to indicate that these parameters can vary across levels $h = 1, \dots, H$ of a between-level latent categorical latent variable.

Class-Specific ICC

Dependency in item responses due to clustering at the between level can be characterized with ICC. In multilevel mixture item response models, a class-specific ICC can be specified for each item. It can be interpreted as the proportion of variance which is accounted for at the between level for that item. The class-specific ICC, ICC_{ig} , can be derived as the correlation coefficient among probabilities of item responses on the logit scale for the same cluster k but different persons j and j' , as shown in Equation 3. Note that in Equation 3, a shorthand is used, where $P'(y_{jki}) = \text{logit}[P(y_{jki} = 1 | \theta_{jkg}, \theta_k, C_{jk})]$:

$$ICC_{ig} = \frac{\text{Cov}(P'(y_{jki}), P'(y'_{jki}))}{\sqrt{\text{Var}(P'(y_{jki}))} \cdot \sqrt{\text{Var}(P'(y'_{jki}))}} = \frac{\alpha_{i,B}^2 \tau^2}{\sqrt{\alpha_{ig,W}^2 \sigma_g^2 + \alpha_{i,B}^2 \tau^2} \cdot \sqrt{\alpha_{ig,W}^2 \sigma_g^2 + \alpha_{i,B}^2 \tau^2}}. \quad (3)$$

As shown in Equation 3, ICC_{ig} is a function of item discrimination parameters or variances of continuous latent variables. This fact has two important implications: The first implication is that increasing ICC_{ig} implies larger differences in the between- versus within-discrimination parameters ($\alpha_{i,B}$ and $\alpha_{i,W}$ in Equation 3), controlling for the variances of the latent variables. The second implication is that different degrees of ICC can be found across latent classes when $\alpha_{ig,W}$ or σ_g^2 differs across latent classes. These two implications will be made use later in the simulation; specifically, the variance of the continuous latent variable will be fixed for identification purposes, and then ICC_{ig} will be manipulated across conditions exclusively by manipulating the sizes of discrimination parameters. The derivation of Equation 3 is presented in Online Appendix B.

Simulation studies were designed to investigate (a) the consequences of fitting a single-level mixture item response model in the presence of multilevel structure (characterized by ICC) in Simulation Study 1 and (b) the consequences of fitting a multilevel mixture item response model when it is not needed in Simulation Study 2. Below, evaluation measures were first presented.

Evaluation Measures

In Simulation Study 1, the impact of ignoring a multilevel structure is evaluated based on parameter recovery and model selection accuracy. The true model in Simulation Study 1 is a multilevel mixture item response model with two within-level latent classes (called *multilevel model* hereafter; that is, Equation 2), whereas the misspecified model is a single-level mixture item response model with two latent classes (called *single-level model* hereafter; that is, Equation 1). In Simulation Study 2, parameter recovery and model selection accuracy were also considered to examine the consequences of fitting multilevel models when they are not needed. The true model in Simulation Study 2 is a single-level model (Equation 1), whereas the misspecified model is the multilevel model (Equation 2). The accuracy of parameter recovery was compared between the true and misspecified models. In both simulation studies, several candidate models with different numbers of within-level and between-level classes were compared based on IC.

Parameter Recovery

Bias and root mean square error (RMSE) of item parameter estimates were compared between the true and misspecified models. Bias was calculated by $\hat{\alpha}_i - \alpha_i$ for an item discrimination as an example. RMSE was computed using $\sqrt{\sum_{r=1}^R (\hat{\alpha}_i - \alpha_i)^2 / R}$, where r indicates the r th replication ($r = 1, \dots, R$). RMSE combines bias and the sampling variance of the parameter estimate. In addition, the accuracy of *SEs* was evaluated by comparing (a) the means of *SEs* of item parameter estimates of fitted single- and multilevel models with (b) the standard deviation of item parameter estimates across replications in a true model.

Model Selection

IC were used for model selection. The proportion of samples in which each information criterion selects the true model was compared with the proportion of samples where it selects the misspecified model.

A general form of IC is as follows:

$$IC = -2LL + C, \quad (4)$$

where LL is a log likelihood and C indicates the model complexity. The first term, $-2LL$, decreases as a model becomes complex. The second term, C , penalizes the complexity of the model. Therefore, a smaller value of IC means that the model has relatively a better model fit compared with other competing models.

Several kinds of IC have been suggested that differ in the penalty for complexity. AIC considers the number of model parameters in the penalty term as follows:

$$AIC = -2LL + 2P, \quad (5)$$

where P is the number of model parameters.

BIC uses the number of model parameters and sample size in the penalty term. For the sample size in BIC, it is common to use the number of persons in multilevel modeling (e.g., Hamaker, van Hattum, Kuiper, & Hoijtink, 2011) and in multilevel item response modeling (e.g., Cohen & Cho, 2016), as specified below:

$$BIC = -2LL + \log(n)P, \quad (6)$$

where n the number of persons.

Simulation Study I

Simulation Conditions

Three simulation conditions were theorized to affect model selection performance and parameter accuracy in the context of a multilevel data structure because they can lead to between-cluster responses that are more heterogeneous and within-cluster responses that are more homogeneous (e.g., Chen et al., 2010; Kaplan & Keller, 2011; Preacher, Zhang, & Zyphur, 2011): the number of clusters (two levels), cluster size (three levels), and class-specific ICC (three levels). The total number of simulation conditions is 18 ($=2 \times 3 \times 3$).

Other conditions that may be less relevant specifically to ignoring multilevel data structure were considered fixed conditions in the simulation study. That is, factors that simply generically affect classification accuracy for mixture models—such as the number of items, the number of latent classes, mixing proportions, and different item profiles—were considered as fixed conditions. Twenty items were used, which is a reasonable size for mixture item response model studies (e.g., Finch & French, 2012). Eighty percent of the total items had class-specific item parameters, and 20% were class-invariant items. The two within-level latent classes (called Class 1 and Class 2 hereafter) in the true model had equal mixing proportions. Every cluster had the same proportion of within-level latent classes (because variation in these proportions could otherwise lead to detection of between-level latent classes). Differences in class-specific item discrimination parameters were manipulated by ICC (see Equation 3). The item profile of the two (within-level) latent classes was sufficiently distinct in terms of item difficulty parameters for there to be good parameter recovery if a single-level mixture item response model was fit to single-level data (e.g., Li et al., 2009). Below, each varying simulation condition is described.

Number of clusters. The number of clusters was set to $K = 24$ or 50. A sample of 24 and 50 clusters is common in experimental intervention research.

Cluster size. Balanced cluster sizes were selected as $n_k = 6, 20$, or 50, as used in previous multilevel modeling studies (e.g., Preacher et al., 2011). A cluster size of six is found in small group designs (e.g., Kenny, Mannetti, Pierro, Livi, & Kashy, 2002).

The alternative numbers of clusters and cluster sizes imply six different total numbers of persons: $J = 144, 300, 480, 1,000, 1,200$, or $2,500$.

ICC. To investigate the consequences of fitting a single-level mixture item response model when multilevel structure is needed, class-specific ICCs were manipulated as .1 and .3, .2 and .2, and .3 and .1 (for Class 1 and Class 2, respectively). For the level of .2 and .2, item discrimination parameters were identical between the two latent classes, and item difficulty parameters were different between the classes. For the levels of .1 and .3 and .3 and .1, both item discrimination and difficulty parameters differed between the two latent classes. The ICC was manipulated via differences between item discrimination parameters at the within level and at the between level. The ICC in a condition was the same across items to control for the effect of different degrees of ICC.

Data Generation

The true model is a multilevel model (i.e., Equation 2; a multilevel mixture item response model with two within-level latent classes). Binary responses were generated based on the latent variables and item parameter values. R (R Core Team, 2015) was used to generate item responses.

Class 1 and Class 2 had one common continuous latent variable at the between level (θ_k) and separate latent variables at the within level (θ_{jkg}). These continuous latent variables were generated from standard normal distributions. There were K generated θ_k and $J/2 (= (n_k \times K)/2)$ generated θ_{jkg} for each of the two within-level latent classes. The item parameter values were obtained by taking the following steps:

Step 1. Twenty within-level item discrimination parameters of Class 1 ($\alpha_{i1.W}$) were generated from a normal distribution with a mean of 1.13 and a variance of 0.36, which is a default prior on item discriminations in BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). The within-level item discrimination parameters of Class 2 ($\alpha_{i2.W}$) were calculated based on the generated $\alpha_{i1.W}$ using ICC. For example, for the level of .1 and .3, the between-level item discrimination parameter for the first item was 0.484. The within-level item discrimination parameter was 1.452 ($0.1 = 0.484^2 / (0.484^2 + 1.452^2)$) for Class 1 and 0.739 ($0.3 = 0.484^2 / (0.484^2 + 0.739^2)$) for Class 2, respectively. Within-level item discrimination parameters of Items 1 to 16 (i.e., items having class-specific item parameters) were calculated. Items 17 to 20 were class invariant, so the same within-level discrimination parameters were used for Class 1 and Class 2 regardless of ICC. For the .2 and .2 condition, Class 1 and Class 2 had the same within-level item discrimination parameters. For the .3 and .1 condition, $\alpha_{i1.W}$ was the same as $\alpha_{i2.W}$ in the .1 and .3 conditions and vice versa. Item difficulty parameters of Class 1 (β_{i1}) were generated from a standard normal distribution.

Step 2. Between-level item discrimination parameters ($\alpha_{i.B}$) were calculated based on ICC. The between-level item discrimination parameters of Class 1 were identical to Class 2 to model a continuous latent variable only at the between level.

Step 3. Item difficulty parameters of Class 2 (β_{i2}) were calculated based on differences in item difficulty between the within-level latent classes. Specifically, the values of 1.0, 2.0, -1.0, and -2.0 were added to the item difficulty parameters of Items 1 to 4, Items 5 to 8, Items 9 to 12, and Items 13 to 16, respectively. The manipulation resulted in a distinct item profile because Items 1 to 8 are harder to persons in Class 1 than persons in Class 2, whereas Items 9 to 16 are easier for persons in Class 1 than for persons in Class 2. The item parameters of Items 17 to 20 were identical between the latent classes. Online Appendix D shows an example of generated item parameters for the .1 and .3 condition.

Analysis

Model identification and scale comparability constraints. If there are class-invariant items in single-level and multilevel mixture item response models, equality constraints can be set on item parameters across the latent classes ($\alpha_{i1} = \dots = \alpha_G$; $\beta_{i1} = \dots = \beta_G$). The mean and variance of the continuous latent variable in the first latent class ($\sigma_1^2 = 1$) are set to 0 and 1, respectively, to identify the model. The means and variances of the continuous latent variable in the other latent classes ($[\theta_{j2}, \dots, \theta_{jG}]'$) are estimated because of the equality constraints on the item parameters. If the class-invariant items are not present, a standard normal distribution can be set for each continuous latent variable in the other classes ($[\theta_{j2}, \dots, \theta_{jG}]'$) for model identification.

Fitting models. As mentioned previously, the true model was Equation 2, a multilevel model with an across-class equality constraint (Items 17-20) for the levels of .1 and .3, .2 and .2, and .3 and .1 in the ICC simulation condition. When fitting the Equation 2 model, one true and eight other candidate specifications were compared (i.e., models with one or two between-level latent classes/one, two, three, or four within-level latent classes). When fitting the Equation 1 model, four candidate specifications were compared (i.e., models with one, two, three, or four latent classes). Therefore, 13 models were used for estimation per replication of each condition.

The equality constraint in the generating model was not imposed when fitting 12 of the candidate models, as often assumed in model selection (e.g., Li et al., 2009). Fifty replications were used for each condition. Accordingly, the number of total runs was 11,700 (= 50 replications $\times$ 13 models $\times$ 18 conditions).

Parameter estimation and prediction. Item parameter estimation and scoring were implemented using Mplus 7.4 (Muthén & Muthén, 1998-2015). Marginal maximum-likelihood estimation method with the MLR estimator option was used for parameter estimation. The MLR option provides a test statistic and *SEs* using the Huber–White sandwich estimator that are robust against nonnormality.¹ Expected a posteriori (EAP) scoring was used for latent variable predictions. An example Mplus code for the true model is provided in Online Appendix D.

For classification of persons to a within-level latent class g , persons are assigned to the latent class for which they have the highest posterior probability of group membership. Specifically, given the estimated item parameters ($\hat{\alpha}_{ig,W}$, $\hat{\alpha}_{i,B}$, and $\hat{\beta}_{ig}$), predicted scores ($\tilde{\theta}_{jkg}$ and $\tilde{\theta}_k$), and item responses ($y = [y_{jk1}, \dots, y_{jki}, \dots, y_{jkl}]'$) for each person j nested with a cluster k , the posterior probability of belonging to each latent class, P_{jkg} , is calculated as follows:

$$P_{jkg} = \frac{\hat{\pi}_g \cdot \prod_{i=1}^I (P(y_{jki} = 1 | \tilde{\theta}_{jkg}, \tilde{\theta}_k, C_{jk}))^{y_{jki}} [1 - P(y_{jki} = 1 | \tilde{\theta}_{jkg}, \tilde{\theta}_k, C_{jk})]^{1-y_{jki}}}{\sum_{g=1}^G \hat{\pi}_g \cdot \prod_{i=1}^I (P(y_{jki} = 1 | \tilde{\theta}_{jkg}, \tilde{\theta}_k, C_{jk}))^{y_{jki}} [1 - P(y_{jki} = 1 | \tilde{\theta}_{jkg}, \tilde{\theta}_k, C_{jk})]^{1-y_{jki}}}, \quad (7)$$

where $\hat{\pi}_g$ is an estimated mixing proportion. For each person, $\sum_{g=1}^G P_{jkg} = 1$. A similar procedure is applied to a single-level mixture item response model for classification.

There are several issues to be considered specific to mixture modeling. One issue is label switching (see McLachlan & Peel, 2000, for details). To monitor the label switching problem, mean square error (MSE) was calculated based on the parameter for Class 1 and Class 2, respectively, prior to calculating bias and RMSE. At each replication, the MSEs of the item difficulty estimates with two kinds of item parameters were compared, and the latent classes were reassigned according to the smaller MSE. Because the item profile of the two latent classes in terms of item difficulty parameters was distinct, the reassignment of the latent classes was generally not problematic. When there was label switching, the measures for parameter recovery (i.e.,

bias, RMSE, and the ratio of SE) were calculated after the latent classes were renamed. To monitor local maxima for the mixture model likelihood, 200 random starting values were employed for initial iterations; the best 20 were iterated to completion (STARTS = 200 20 in Mplus).

Results

Out of 11,700 replications, 613 replications (5.2%) had convergence problems. The convergence problems occurred mainly when the cluster size was the smallest ($n_k = 6$) and classes were overextracted (three or four latent classes). The converged results from 11,087 replications were included in the analysis. Because factors that affect model selection for mixture models were fixed, the classification accuracy, posterior probabilities, and entropy were similar between the single-level model and the multilevel model as presented in Online Appendix E.

Parameter recovery. Figure 1 presents the item parameter recovery results under the multilevel model (i.e., true model with equality constraints on item parameters for class-invariant items) and the single-level model with two latent classes (i.e., a misspecified model). The smallest sample size condition ($K = 24$ and $n_k = 6$) often produced extreme values of estimates. Item parameter estimates (i.e., $\hat{\alpha}_{i1.W}$, $\hat{\alpha}_{i2.W}$, $\hat{\alpha}_{i.B}$, $\hat{\beta}_{i1}$, $\hat{\beta}_{i2}$) were identified as outliers if the absolute value of the bias was greater than 5. By this criterion, 1,558 estimates were deleted out of 262,000 (18 conditions $\times$ 50 replications $\times$ 100 [multilevel model] or 80 [single-level model] item parameters) total estimates. Most of the outliers (1,454 estimates; 93.3%) occurred in $K = 24$ and $n_k = 6$ conditions. The multilevel model produced 699 (0.78%) and the single-level model produced 365 (0.51%) outliers.

Bias. For each class-specific parameter in a model, a three-way ANOVA was conducted with simulation conditions (K , n_k , and ICC) as between factors and the bias as the dependent variable to examine the source of bias. The single-level model revealed main effects of and an interaction between K and n_k for bias of $\hat{\alpha}_{i1}$ and $\hat{\alpha}_{i2}$ (partial $\eta^2 \geq .009$). However, the main effect of K was the only significant factor in the bias of $\hat{\alpha}_{i1}$ of the multilevel model, explaining 0.8% of the total variance. The explained variance by class-specific ICC was small under the multilevel and single-level models. However, this result does not imply that the class-specific ICC is not an important factor in explaining variability in bias. Because Class 1 may have a lower ICC when Class 2 has a higher ICC (i.e., .1 and .3) and Class 1 may have a higher ICC when Class 2 has a lower ICC (i.e., .3 and .1), and the class-specific item parameters are estimated simultaneously, bias of item parameter estimates from the two latent classes can be compensatory.

The top row in Figure 1 presents the bias results. Within-level item discrimination parameter estimates ($\hat{\alpha}_{ig.W}$) of the multilevel model were compared with the item discrimination parameter estimates ($\hat{\alpha}_{ig}$) of the single-level model to examine the consequences of ignoring the multilevel structure. The discrimination parameter estimates of the single-level model which ignored a multilevel structure were more biased than those of the multilevel model except for the following conditions: $n_k = 6$ and $K = 24$, also $n_k = 6$, $K = 50$, and ICC = .1 and .3 for $\hat{\alpha}_{i1}$; $K = 50$ and ICC = .2 and .2 for $\hat{\alpha}_{i2}$. The degree of bias of $\hat{\alpha}_{ig}$ in the single-level model increased with increasing n_k in $K = 24$, but the pattern was reversed in $K = 50$. The degree of the bias in the multilevel model largely decreased with increasing n_k and K . The bias ranged from 0.066 to 0.223 for the multilevel model and from -0.063 to 0.179 for the single-level model.

Regarding item difficulty parameters, the bias from the single-level model was larger than the bias from the multilevel model with the following exceptions: $n_k = 6$ and $K = 24$; also $n_k = 20$ and $K = 24$ for $\hat{\beta}_{i2}$. For both models, the degree of bias was less than or equal to 0.212 across all conditions and revealed a decreasing pattern as the number of clusters and cluster size increased.

Figure I. Simulation Study I: Bias, RMSE, and SE ratio of item parameter estimates.
Note. For the SE ratio, outliers were not included in the figure. Outliers included 4 points (62.720, 9.605, 60.312, 32.896) of $\alpha_{i1,w}$, 4 points (101.545, 46.548, 87.701, 11.150) of $\alpha_{i2,w}$, 3 points (14.410, 22.830, 76.124) of β_{i1} , and 4 points (16.502, 72.402, 66.676, 4.562) of β_{i2} . RMSE = root mean square error; SE = standard error; ICC = intraclass correlation.

RMSE. ANOVA results were explained the most by the main effects of and interaction between K and n_k under the multilevel and single-level model, except for $\hat{\beta}_{i1}$. The RMSE of $\hat{\beta}_{i1}$ was not affected by simulation conditions.

Figure 1 (middle) shows the RMSE of item parameter estimates. The RMSE from the single-level model was higher than that from the multilevel model when $n_k = 6$ with the following exceptions: the RMSE of $\hat{\alpha}_{i2}$ in ICC = .1 and .3 conditions and $\hat{\alpha}_{i1}$ in ICC = .3 and .1 conditions. The latent class having the higher ICC (i.e., having the higher item discriminations as shown in Equation 3) revealed a higher RMSE under the single-level model than under the multilevel model. The RMSE of the within-level discrimination estimates did not differ between the models when $n_k = 20$ and 50. Concerning the within-level discrimination parameters, RMSE ranged from 0.119 to 1.118 for the multilevel model. For the single-level model, the range of RMSE was from 0.166 to 1.036. The interaction effect between n_k and K resulted from the relatively high RMSE in $n_k = 6$ and $K = 24$ conditions. The effect of the number of clusters at the level of $n_k = 20$ and 50 was not as large as at $n_k = 6$.

Regarding the difficulty parameters, the RMSE ranged from 0.087 to 1.008 for the multilevel model and from 0.081 to 0.801 for the single-level model. The RMSE from the single-level model was higher than that from the multilevel model only in $n_k = 6$ and $K = 24$ conditions.

The ratio of SE. The ratio of the mean of estimated *SE* across replications in the multilevel model to the standard deviation of estimates across replications in the multilevel model (indicated by multi/*SD* in Figure 1) and the ratio of the mean of estimated *SE* across replications in the single-level model to the standard deviation of estimates across replications in the multilevel model (indicated by single/*SD* in Figure 1) were calculated to evaluate the *SE* of item parameter estimates (see Figure 1, bottom). The ratio is greater than 1 if the estimated *SE* is overestimated compared with the *SE* of estimates in the multilevel model. Below, ANOVA results are first presented and then Figure 1 results are interpreted.

ANOVA result confirmed that cluster size explained the greatest variability in the ratio under the multilevel model ($\eta^2 = .010, .009$, and $.008$ for $\alpha_{i1.w}$, $\alpha_{i2.w}$, and β_{i2} , respectively). For β_{i1} , no simulation factors were significant ($p > .284$, $\eta^2 \leq .006$). Under the single-level model, the interaction effect between cluster size and the number of clusters ($\eta^2 = .019$ and $.008$ for α_{i1} and β_{i1} , respectively) and the main effect of cluster size ($\eta^2 = .021$ and $.011$ for α_{i1} and β_{i1} , respectively) were two most influential factors of the variability in the ratio for $\alpha_{i1.w}$ and β_{i1} . The interaction effect can be explained from the underestimated *SEs* in $n_k = 6/K = 24$ condition. No simulation factors revealed any significant result for α_{i2} and β_{i2} , respectively.

The *SEs* of item parameter estimates tended to be overestimated when multilevel structures are ignored. As shown in Figure 1, the multilevel model had extremely large mean *SE* (>4.5) in $n_k = 6$ and $K = 24$ conditions, also $n_k = 6$, $K = 50$, and ICC = .1 and .3 condition. These extreme *SEs* are not presented in Figure 1 (bottom). On the contrary, the *SE* under the single-level model was often underestimated (<0.8) in the same conditions. Except for the $n_k = 6$ and $K = 24$ at all ICC levels, and $n_k = 6$ and $K = 50$ at the level of ICC = .1 and .3 conditions, the ratio of the single-level model was larger than that of the multilevel model. The ratio ranged from 0.832 to 1.181 in the multilevel model and from 1.008 to 1.541 in the single-level model for discrimination parameters, and ranged from 0.949 to 1.316 in the multilevel model and from 1.027 to 1.519 in the single-level model for difficulty parameters. The patterns in the ratio were as follows except for the condition of $n_k = 6$. As n_k increased, the ratio for discrimination parameter estimates decreased, and the ratio for difficulty parameter estimates increased under the multilevel model. In the single-level model, the ratio for discrimination and difficulty parameter estimates decreased as n_k and K increased.

Model selection. Table 1 (top) presents the model selection results by AIC and BIC. The performance of AIC and BIC for the multilevel model (i.e., true model; the multilevel model with two within-level classes and equality constraint on item parameters for class-invariant items denoted by **2a** in Table 1) is first described. BIC selected the multilevel model at least 88% of the time when n_k is 50. The single-level models with two classes were more frequently selected by BIC in $n_k = 20$, also $n_k = 6$ and $K = 50$ conditions. The single-level model with one latent class (i.e., two-parameter item response model) was dominantly selected by BIC in $n_k = 6$ and $K = 24$. The multilevel model was not often selected by AIC (i.e., 4%-24%). The single-level models with three or four classes were more frequently selected by AIC than the multilevel models when $n_k = 6$, and the multilevel mixture item response model with three or four classes was selected by AIC over the multilevel model when $n_k = 20$ or 50.

Simulation Study 2

Simulation Study 1 found that, even when a nested data structure does exist, it is often but not always beneficial in terms of bias, RMSE, and/or *SE* ratio to fit a multilevel rather than single-level, mixture item response model. Hence, it is relevant in Simulation Study 2 to investigate the consequences of fitting a multilevel mixture item response model when there is actually no nested data structure.

Simulation Conditions

As in Simulation Study 1, two levels ($K = 24, 50$) of the number of clusters and three levels ($n_k = 6, 20, 50$) of cluster sizes were included in the simulation condition. Class-specific ICC was not manipulated as the simulation condition as the true model was a single-level model. The total number of simulation conditions was six.

Data Generation

Item parameters for 20 items were generated. Item discrimination parameters were identical to Class 1 of .1 and .3 conditions in Simulation Study 1. The mean and variance of the item discrimination parameters are 1.13 and 1, respectively. The two latent classes had the same item discrimination parameters. Class-specific item difficulty parameters were the same as in Simulation Study 1. Person parameters (i.e., continuous latent variables) were also identical to the within-level latent variables in Simulation Study 1. Two class-specific person parameters were generated from a standard normal distribution, having the sample sizes of 72, 240, 600, 300, 500, and 1,250 for each condition.

Analysis

Model identification and scale comparability constraints. For model identification, the mean and variance of the person parameters of Class 1 were constrained to 0 and 1, respectively. An equality constraint on class-invariant item parameters between latent classes was imposed for scale comparability. The means and the variances of the person parameters of the other classes were estimated.

Fitting models. As Simulation Study 1, the true model (two latent classes with equality constraint on item parameters for class-invariant items) and eight candidate multilevel mixture item response models (i.e., models with one or two between-level latent classes/one, two, three, or

Table 1. Proportion of Model Selection for Simulation Study 1 (top) and Simulation Study 2 (bottom).

IC	K	n_k	ICC	$\frac{l}{g}$	0				1				2					
				1	2	3	4	1	2a	2b	3	4	1	2	3	4		
AIC	24	6	.1/3		.00	.12	.44	.08	.00	.04	.00	.20	.12	.00	.00	.00	.00	
			.2/2		.00	.04	.46	.12	.00	.04	.00	.22	.12	.00	.00	.00	.00	
			.3/1		.00	.18	.48	.08	.00	.04	.00	.18	.04	.00	.00	.00	.00	
		20	.1/3		.00	.00	.26	.16	.00	.12	.02	.24	.20	.00	.00	.00	.00	
			.2/2		.00	.02	.08	.18	.00	.14	.00	.32	.26	.00	.00	.00	.00	
			.3/1		.00	.00	.16	.12	.00	.10	.02	.32	.28	.00	.00	.00	.00	
		50	.1/3		.00	.00	.00	.00	.00	.22	.00	.40	.38	.00	.00	.00	.00	
			.2/2		.00	.00	.00	.00	.00	.08	.00	.44	.48	.00	.00	.00	.00	
			.3/1		.00	.00	.00	.00	.00	.20	.02	.36	.42	.00	.00	.00	.00	
		50	6	.1/3		.00	.02	.38	.20	.00	.02	.00	.16	.20	.00	.02	.00	.00
				.2/2		.00	.06	.48	.22	.00	.00	.00	.10	.14	.00	.00	.00	.00
				.3/1		.00	.02	.28	.26	.00	.06	.00	.24	.14	.00	.00	.00	.00
	20		.1/3		.00	.00	.00	.00	.00	.16	.00	.44	.40	.00	.00	.00	.00	
			.2/2		.00	.00	.00	.00	.00	.12	.00	.44	.44	.00	.00	.00	.00	
			.3/1		.00	.00	.00	.00	.00	.22	.00	.50	.28	.00	.00	.00	.00	
	50		.1/3		.00	.00	.00	.00	.00	.24	.00	.56	.20	.00	.00	.00	.00	
			.2/2		.00	.00	.00	.00	.00	.10	.00	.54	.36	.00	.00	.00	.00	
			.3/1		.00	.00	.00	.00	.00	.08	.00	.44	.48	.00	.00	.00	.00	
	BIC	24	6	.1/3		.92	.08	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
				.2/2		.94	.06	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
				.3/1		.92	.08	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
			20	.1/3		.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
				.2/2		.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
				.3/1		.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
50			.1/3		.00	.12	.00	.00	.00	.88	.00	.00	.00	.00	.00	.00	.00	
			.2/2		.00	.00	.00	.00	.00	1.00	.00	.00	.00	.00	.00	.00	.00	
			.3/1		.00	.04	.00	.00	.00	.96	.00	.00	.00	.00	.00	.00	.00	
50			6	.1/3		.04	.96	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
				.2/2		.26	.74	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
				.3/1		.26	.74	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	20	.1/3		.00	.82	.00	.00	.00	.18	.00	.00	.00	.00	.00	.00	.00		
		.2/2		.00	.54	.00	.00	.00	.46	.00	.00	.00	.00	.00	.00	.00		
		.3/1		.00	.74	.00	.00	.00	.26	.00	.00	.00	.00	.00	.00	.00		
	50	.1/3		.00	.00	.00	.00	.00	1.00	.00	.00	.00	.00	.00	.00	.00		
		.2/2		.00	.00	.00	.00	.00	1.00	.00	.00	.00	.00	.00	.00	.00		
		.3/1		.00	.00	.00	.00	.00	1.00	.00	.00	.00	.00	.00	.00	.00		
IC	K	n_k		$\frac{l}{g}$	0				1				2					
				1	2	3	4	1	2a	2b	3	4	1	2	3	4		
AIC	24	6			.00	.22	.46	.08	.00	.02	.00	.18	.04	.00	.00	.00	.00	
		20			.00	.08	.52	.34	.00	.02	.00	.00	.04	.00	.00	.00	.00	
		50			.00	.08	.62	.22	.00	.00	.00	.02	.06	.00	.00	.00	.00	
	50	6			.00	.12	.48	.24	.00	.04	.00	.08	.04	.00	.00	.00	.00	
		20			.00	.00	.64	.28	.00	.00	.00	.04	.04	.00	.00	.00	.00	
		50			.00	.04	.60	.30	.00	.02	.00	.02	.02	.00	.00	.00	.00	
BIC	24	6		1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	
		20			.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	
		50			.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	
	50	6			.42	.58	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	
		20			.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	
		50			.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	

Note. Bold column corresponds to the true model; l = the number of between-level latent classes; g = the number of within-level latent classes; 2a = the model with equality constraint; 2b = the model without equality constraint. IC = information criteria; ICC = intraclass correlation; AIC = Akaike's information criterion; BIC = Bayesian information criterion.

four within-level latent classes) were compared. For a single-level mixture item response model, four candidate models (i.e., models with one, two, three, or four latent classes) were compared. Thus, 13 models were considered for estimation per replication of each condition. For each condition, 50 replications were used. Consequently, the number of total runs was 3,900 ($= 50 \text{ replications} \times 13 \text{ models} \times \text{six conditions}$).

Results

Two hundred twenty-two replications (5.7%) out of 3,900 did not produce a converged solution when the data were generated from the single-level mixture item response model. This rate was comparable with the conditions where the multilevel model was the true model. As in the multilevel model conditions, the convergence problems occurred mostly when the number of latent classes was set to 3 and 4. As shown in Online Appendix F, the classification accuracy, posterior probability, and entropy were similar between the single-level model and the multilevel model because factors that affect model selection for mixture models were fixed in simulation conditions.

Parameter recovery. Parameter estimation was not stable in the smallest sample size condition ($K = 24$ and $n_k = 6$). Although the model converged, extremely large estimates and *SE* occasionally occurred. Because these outliers in one out of six conditions can distort the results, the ANOVA was not conducted. Below, bias and RMSE were compared between the true and multilevel (with two within-level latent classes and one between-level latent class) models.

Bias. Figure 2 (top) presents the bias of the estimates under the single-level and multilevel models. When a between-level continuous latent variable did not exist, bias of within-level item discrimination estimates was close to 0 under the multilevel and single-level models except for one condition: The estimates of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ from the multilevel model were more positively biased when the $n_k = 6$ and $K = 50$. The difficulty parameter estimates were generally unbiased under both models.

RMSE. Figure 2 (middle) depicts the RMSE of the estimates. The RMSE of item discrimination and difficulty parameter estimates under the multilevel model were larger than the single-level model in $n_k = 6$ and $K = 50$, also in $n_k = 20$ and $K = 24$ conditions. However, as the sample size increased, the difference in RMSE between the two models decreased.

The ratio of *SE*. As in Simulation Study 1, the ratio of the estimated *SE* (from the fitted model) and the standard deviation of estimates across replications (from the true—that is, single-level model) was calculated (see Figure 2, bottom). The *SE* from the multilevel model was extremely overestimated in $n_k = 6$ condition. The *SE*s in single-level model were underestimated in $n_k = 6$ and $K = 24$ conditions. Except for $n_k = 6$ condition, the ratio for discrimination parameter estimates from the multilevel model increased along with n_k and K , whereas that from the single-level model decreased. The ratio for difficulty parameter estimates showed a decreasing pattern with increasing n_k and K under the multilevel and single-level models.

Model selection. The model selection results are presented in Table 1 (bottom). When a between-level continuous latent variable did not exist, BIC always selected the correct model when $K > 24$ and $n_k > 6$. A single-level model with one latent class was the most often selected by BIC when the total sample size was 144 and 300. However, the true model was rarely (0%-22%) successfully selected by AIC, consistent with Preinerstorfer and Formann (2012) and Li et al. (2009), which showed that BIC outperformed in selecting the number of latent classes for single-level mixture item response models.

Figure 2. Simulation Study 2: Bias, RMSE, and SE ratio of item parameter estimates.

Note. For bias, 5 points at the smallest sample size are not presented in the figure because the values are out of range ($>.25$ or $<-.25$; $\hat{\alpha}_2 = -.560$, $\hat{\beta}_1 = -.304$, and $\hat{\beta}_2 = -1.103$ for the single-level model; $\hat{\alpha}_{2w} = -.741$ and $\hat{\beta}_2 = -2.403$ for the multilevel model). For RMSE, 7 data points in $n_k = 6$ and $K = 24$ conditions are not presented in the figure because the values are greater than 1.25 ($\hat{\alpha}_1 = 2.522$, $\hat{\alpha}_2 = 4.506$, $\hat{\beta}_1 = 1.731$, and $\hat{\beta}_2 = 10.358$ for the single-level model; $\hat{\alpha}_1 = 1.772$, $\hat{\alpha}_2 = 4.379$, and $\hat{\beta}_2 = 6.151$ for the multilevel model). For the SE ratio, the outliers from the multilevel model were not presented. The outliers include $\alpha_1 = 26.10$, $\alpha_2 = 34.97$, $\beta_1 = 4.42$, $\beta_2 = 55.59$ in $n_k = 6$ and $K = 24$ conditions; $\alpha_1 = 49.88$, $\alpha_2 = 44.90$, $\beta_2 = 4.95$ in $n_k = 6$ and $K = 50$ conditions. RMSE = root mean square error. SE = standard error.

Summary and Conclusion

The first purpose of the current study was to examine the consequences of ignoring multilevel structure when fitting a single-level mixture item response model. When the number of clusters is larger than 24 and the cluster size is larger than six, ignoring multilevel structure in the applications of single-level mixture item response models can be problematic regarding the accuracy of item discrimination estimates and *SEs* of item discrimination estimates. As long as the cluster size is greater than six, the bias of item discrimination parameter estimates from the single-level model was higher than those from multilevel model even when the number of latent classes is correctly specified; bias increased with increasing cluster size and number of clusters. The *SEs* of the item discrimination and difficulty parameter estimates were relatively overestimated in the single-level model compared with the multilevel model. A single-level model which ignored multilevel structure was not selected over the multilevel model based on BIC if the number of clusters and cluster size were at least 50, respectively. AIC always selected a model having incorrect number of latent classes. AIC is not recommended for detecting the within-level latent classes in the applications of multilevel mixture item response models. Note that, in small sample sizes ($n_k = 6$ and $K = 24$), the estimates from the multilevel model were more biased than the single-level model. In addition, the *SE* of item parameter estimates was also extremely overestimated under these conditions. This result implies that the multilevel mixture item response model does not perform well in small sample sizes (e.g., $n_k = 6$ and $K = 24$), and a single-level mixture item response model may be advisable to employ in such a condition.

The second purpose of this study was to examine the consequences of fitting a multilevel mixture item response model when there is no multilevel structure. It was expected that the multilevel mixture item response model as well as the single-level model would perform similarly in estimating within-level item parameters because the single-level mixture item response model is a special case of the multilevel model (when $ICC = 0$). As expected, when there is no continuous latent variable at the between level, using multilevel item response models did not distort the within-level item parameter estimates when the cluster size is as large as 20. The *SE* of item parameter estimates from the multilevel model was inflated in small cluster size, but they were comparable with the *SE* from the single-level model if the cluster size is larger than six. Reassuring, the multilevel model was not detected over the single-level model by AIC and BIC when a between-level continuous latent variable does not exist.

There are methodological limitations in the present study. First, the item parameter estimates were often extremely large for small cluster size ($n_k = 6$) for both the true and misspecified models. Those extreme values made it difficult to interpret the main effect of other simulation factors such as the number of clusters and cluster sizes. When the number of clusters was six, the number of clusters in each latent class was 3 in the simulation design. The small sample size within cluster might make difficult to estimate the within-level item discrimination parameters. However, this level of cluster size was included because small cluster size design is frequently used in practice. Also, the current study showed that the parameter recovery improved as the total sample size increased despite small cluster size.

Second, the simulation conditions affecting the classification accuracy and quality were not varied, and the latent classes were well separated across conditions because the investigation of the performance of IC in various item profiles between latent classes was not the primary goal of this study. Further studies are needed to investigate to what extent classification accuracy and quality depend on the number of items, the number of latent classes, mixing proportions, and different item profiles in the applications of multilevel mixture item response models.

Third, the population-generating multilevel mixture item response model chosen had categorical latent variables at the within level only, although there are continuous latent variables at

the within level and at the between level. This generating model is chosen because categorical latent variables can be an approximation to a continuous latent variable at the between level. Thus, simulation results of the current study are limited to the case in which there are no between-level latent classes. Additional simulation studies are needed to generalize the simulation results to the other possible multilevel mixture item response models having categorical latent variables (i.e., between-level latent classes), or categorical and continuous latent variables at the between level.

To conclude, the findings in this study have implications for researchers. Ignoring a multilevel structure in mixture item response models results in less accurate model selection results, item discrimination estimates, and their *SEs*, particularly when there is a large number of clusters and cluster sizes. In multilevel linear modeling, the necessity of the multilevel modeling is typically justified with ICC (e.g., Raudenbush & Bryk, 2002). In this study, class-specific ICC derivations are presented in terms of discrimination parameters, and detailed steps are presented for generating a complex data structure relevant to multilevel mixture item response models, which can be useful information for researchers working with these models. However, as shown in this study, it is difficult to provide general guidelines in terms of class-specific ICC necessitating multilevel mixture modeling when ICC differs between (within-level) latent classes. Instead, the following guideline applies to the range of class-specific ICCs examined here: A multilevel mixture item response model is recommended for a dataset having $n_k > 20$ and $K > 50$.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. The standard error (*SE*) adjustment with the MLR option is not for nonindependent variables because we did not use it with the TYPE = COMPLEX option.

Supplemental Material

Supplementary material is available for this article online.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267-281). Budapest, Hungary: Akademiai Kiado.
- Akaike, H. (1980). Likelihood and the Bayes procedure. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & M. Smith (Eds.), *Bayesian statistics* (pp. 143-166). Valencia, Spain: University Press.
- Asparouhov, T., & Muthén, B. O. (2006). *Multilevel mixture models* (Unpublished webnote). Available from www.statmodel.com
- Bennink, M., Croon, M. A., Keuning, J., & Vermunt, J. K. (2014). Measuring student ability, classifying schools, and detecting item bias at school level based on student-level dichotomous items. *Journal of Educational and Behavioral Statistics*, 39, 180-201.

- Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433-448). New York, NY: Springer-Verlag.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.
- Chen, Q., Kwok, O.-M., Luo, W., & Wilson, V. L. (2010). The impact of ignoring a level of nesting structure in multilevel growth mixture models: A Monte Carlo study. *Structural Equation Modeling*, 17, 570-589.
- Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, 35, 336-370.
- Cho, S.-J., Cohen, A. S., & Bottge, B. A. (2013). Detecting intervention effects using a multilevel latent transition analysis with a mixture IRT model. *Psychometrika*, 78, 576-600.
- Cho, S.-J., Suh, Y., & Lee, W.-y. (2016). An NCME instructional module on latent DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice*, 35, 48-61.
- Cohen, A. S., & Cho, S.-J. (2016). Information criteria. In W. J. van der Linden (Ed.), *Handbook of item response theory, models, statistical tools, and applications* (Vol. 2, pp. 363-378). Boca Raton, FL: Chapman & Hall/CRC Press.
- Finch, W. H., & Finch, M. E. H. (2013). Investigation of specific learning disability and testing accommodations based differential item functioning using a multilevel multidimensional mixture item response theory model. *Educational and Psychological Measurement*, 73, 973-993.
- Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods*, 11, 167-178.
- Hamaker, E. L., van Hattum, P., Kuiper, R. M., & Hoijsink, H. (2011). Model selection based on information criteria in multilevel modeling. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 231-255). New York, NY: Taylor & Francis.
- Kaplan, D., & Keller, B. (2011). A note on cluster effects in latent class analysis. *Structural Equation Modeling*, 18, 525-536.
- Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology*, 83, 126-137.
- Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, 33, 353-373.
- Lubke, G., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling*, 14, 27-47.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.
- Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus users guide* (7th ed.). Los Angeles, CA: Author.
- Park, J., & Yu, H.-T. (2016). The impact of ignoring the level of nesting structure in nonparametric multilevel latent class models. *Educational and Psychological Measurement*, 76, 824-847.
- Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling*, 18, 161-182.
- Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, 65, 251-262.
- R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from <https://www.R-project.org/>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.
- 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.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271-282.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

- 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). Greenwich, CT: Information Age.
- Vermunt, J. K. (2007, August). *Multilevel mixture item response theory models: An application in education testing*. Paper presented at Bulletin of the International Statistical Institute 56th Session, Lisbon, Portugal.
- Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, 17, 33-51.
- von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, 76, 318-336.
- 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.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software.