A Multilevel Mixture IRT Model With an Application to DIF

Sun-Joo Cho and Allan S. Cohen

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2010 35: 336
DOI: 10.3102/1076998609353111

Journal of Educational and Behavioral Statistics
June 2010, Vol. 35, No. 3, pp. 336–370
DOI: 10.3102/1076998609353111

A Multilevel Mixture IRT Model With an Application to DIF

Sun-Joo Cho
Vanderbilt University

Allan S. Cohen
University of Georgia

Mixture item response theory models have been suggested as a potentially useful methodology for identifying latent groups formed along secondary, possibly nuisance dimensions. In this article, we describe a multilevel mixture item response theory (IRT) model (MMixIRTM) that allows for the possibility that this nuisance dimensionality may function differently at different levels. A MMixIRT model is described that enables simultaneous detection of differences in latent class composition at both examinee and school levels. The MMixIRTM can be viewed as a combination of an IRT model, an unrestricted latent class model, and a multilevel model. A Bayesian estimation of the MMixIRTM is described including analysis of label switching, use of priors, and model selection strategies. Results of a simulation study indicated that the generated parameters were recovered very well for the conditions considered. Use of MMixIRTM also was illustrated with the standardized mathematics test.

Keywords: *finite mixture modeling; item response theory; multilevel modeling; MCMC*

Introduction

Mixture item response theory (MixIRT) models have been studied for use in describing a number of important effects in test data including differential use of response strategies (Bolt, Cohen, & Wollack, 2001; Mislevy & Verhelst, 1990; Rost, 1990), speededness (Bolt, Cohen, & Wollack, 2002; Yamamoto &

The research for this study was funded by the 2006-2007 College Board Research Grant Program. The authors would like to especially thank Dr. Wayne Camara and Dr. Vytas Laitusis for their support. We also thank an anonymous reviewer and the editor for their insightful comments and suggestions. The authors also thank Dr. Paul De Boeck and Dr. Sophia Rabe-Hesketh for their valuable comments. Any opinions, findings, or conclusions are those of the authors and do not necessarily reflect the views of the supporting agency.

Everson, 1997), the impact of testing accommodations (Cohen, Gregg, & Deng, 2005), and detection of differential item functioning (DIF; Cohen & Bolt, 2005; Samuelsen, 2005). The utility of MixIRT models is that they provide a means of detecting groups formed by dimensionality arising directly from the test data. To the extent these groups are substantively meaningful, they provide a potentially important means of understanding how and why examinees respond the way they do.

In this regard, Rost (1990, 1997) described a mixture Rasch model (MRM) in which an examinee population is assumed to be composed of a fixed number of discrete latent classes. In each latent class, a Rasch model is assumed to hold, but each class may have different item difficulty parameters. The probability of a correct response in the MRM can be given as

$$P(y_{ijg} = 1|g, \theta_{jg}) = \frac{1}{1 + \exp[-(\theta_{jg} - \beta_{ig})]}, \quad (1)$$

where $g = 1, \ldots, G$ is an index for latent class, $j = 1, \ldots, J$ is an index for examinees, θ_{jg} is the latent ability of examinee j within class g, and β_{ig} is the Rasch difficulty parameter of item i for class g. The structure of ability in the MRM is

$$\theta_{jg} \sim N\left(\mu_g, \sigma_g^2\right), \quad (2)$$

where μ_g is a class-specific mean of ability and σ_g^2 is a class-specific variance of ability.

Rost (1990) suggested that the primary diagnostic potential of the MRM is in its use in accounting for qualitative differences among examinees and simultaneously to quantify ability with respect to the same items. The MRM is used to identify latent classes of examinees that are homogeneous with respect to item response patterns. The members of each latent class vary in ability, and the response strategies differ among classes. Based on this rationale, in this study, we use an MRM to investigate DIF among latent classes. An important limitation of MRM is that it essentially ignores the basic multilevel structure that is present beyond the student level in much of educational test data.

Multilevel models, also known as hierarchical linear models (HLM), allow the natural multilevel structure of educational and psychological data to be represented formally in the analysis (Bryk & Raudenbush, 1992; Goldstein, 1987; Longford, 1993). The combination of HLM with IRT is advantageous, as it provides more accurate estimates of the standard errors of the model parameters (Fox, 2005; Maier, 2001, 2002). This combination has led to the development of psychometric models for item response data that contain hierarchical structure, thus enabling a researcher to study the impact of different predictors such as schools and curriculum on the lower level units such as students (e.g., Fox & Glas, 2001; Kamata, 2001; Maier, 2001, 2002; Rabe-Hesketh, Skrondal, & Pickles, 2004). A limitation of multilevel IRT models, from the perspective of

mixture IRT modeling, is that they do not provide information about group membership beyond that given by manifest predictors included in the model.

In this study, we incorporate that multilevel structure into a mixture IRT model and extend the model into a multilevel mixture IRT model (MMixIRTM). This MMixIRTM is then used to detect and compare latent groups in the data that have different measurement characteristics. The utility of this approach lies in the fact that latent groups, although not immediately observable, are defined by certain shared response propensities that can be used to help explain item level performance about how members of one latent group differ from another. It is these differences in response propensities that help explain the potential causes of these differential measurement characteristics. The approach proposed in this study provides this information at the student and school levels along with information describing the composition of the different latent groups. We begin below by illustrating how the MMixIRTM can be viewed as incorporating features from multilevel models to form a general MMixIRTM.

Multilevel Mixture IRT Model

The proposed MMixIRTM has mixtures of latent classes at two levels, a student level and a school level. Student-level latent classes capture the association between the responses at the student-level unit. The MixIRT model assumes that there may be heterogeneity in response patterns at the student level, which should not be ignored (Mislevy & Verhelst, 1990; Rost, 1990). The MMixIRTM does not exclude the possibility, however, that there may be no student-level latent classes. It is interesting to note that, if no student-level latent classes exist, this would indicate that there would also be no school-level latent classes. The reason is that school-level units are clustered based on the likelihood of their students belonging to one of the latent classes. In a MMixIRTM as presented in this study, in other words, it is not meaningful to have school-level classes if no student-level latent classes are present.

Viewed in this way, school-level latent classes capture the association between the students within school-level units (Vermunt, 2003). Latent classes at the school level, however, may differ in the probability that students belong to particular latent classes. This is accommodated in the MMixIRTM by allowing for the possibility that school-level latent classes may differ in the proportions of students in each student-level latent class contained in a school-level latent class.

The probability of getting a correct response in the MMixIRTM can be given as follows:

$$\mathrm{P}\left(y_{ijt}=1|g,k,\theta_{jtgk}\right)=\frac{1}{1+\exp\left[-\left(\theta_{jtgk}-\beta_{igk}\right)\right]}, \tag{3}$$

where $g=1,\ldots,G$ is an index for student-level latent classes, $k=1,\ldots,K$ is an index for school-level latent classes, $j=1,\ldots,J$ is an index for examinees,

TABLE 1
Structure of Mixing Proportions in the MMixIRTM

	$K = 1$	$K = 2$	.	.	$K = K$
$G = 1$	$\pi_{1\|1}$	$\pi_{1\|2}$	.	.	$\pi_{1\|K}$
$G = 2$	$\pi_{2\|1}$	$\pi_{2\|2}$	.	.	$\pi_{2\|K}$
.	.	.	.	.	.
.	.	.	.	.	.
$G = G$	$\pi_{G\|1}$	$\pi_{G\|2}$	.	.	$\pi_{G\|K}$
Sum	$\sum_{g=1}^{G} \pi_{g\|1} = 1$	$\sum_{g=1}^{G} \pi_{g\|2} = 1$	.	.	$\sum_{g=1}^{G} \pi_{g\|K} = 1$

$t = 1, \ldots, T$ is an index for schools, for a test of $i = 1, \ldots, I$ items, θ_{jtgk} is the ability of examinee j in school t and in latent classes g and k, and β_{igk} is the difficulty of item i for latent classes g and k.

Mixture Proportion Structure of MMixIRTM

There are two mixture proportions in MMixIRTM, $\pi_{g|k}$ and π_k. The $\pi_{g|k}$ indicate the relative sizes of latent classes at the student-level conditional on latent class membership at the school level, and π_k is the proportion of schools for each class. As shown in Table 1, there are K probability arrays, $\pi_{1:G|k}$, $k = 1, \ldots, K$, where G is the dimension of each array.

Item Difficulty Structure of MMixIRTM

In the general MMixIRTM, item difficulty parameters have both student- and school-level class-specific values. These can be represented as β_{igk}, which is an interaction effect of the student-level latent class (g) and school-level latent class (k). The meaning of an interaction effect is that the characteristics of school-level latent classes change as the number of school-level DIF items increases.

Ability Structure

As indicated in Equation 4, abilities θ_{jtgk} have mean μ_{gk} and variance σ^2_{gk} as follows:

$$\theta_{jtgk} \sim N\left(\mu_{gk}, \sigma^2_{gk}\right). \tag{4}$$

θ_{jtgk} is reparameterized into $\sigma^2_{gk} \cdot \eta_{jtgk}$ where $\eta_{jtgk} \sim \text{Normal}(\mu_{gk}, 1)$.

Priors and Models on Parameters

The following priors were used for the MMixIRTM:

$$\begin{array}{lcl} g & \sim & \text{Multinomial}\left(1, \pi_{g|k}[1:G]\right) \\ k & \sim & \text{Multinomial}\left(1, \pi_{k}[1:K]\right) \\ \eta_{jtgk} & \sim & \text{Normal}\left(\mu_{gk}, 1\right), \\ \mu_{gk} & \sim & \text{Normal}\left(0, 1\right),\ \mu_{11} = 1 \\ \sigma_{gk} & \sim & \text{Normal}\left(0, 1\right) I(0,), \\ \beta_{igk} & \sim & \text{Normal}\left(0, 1\right), \end{array}$$

where $I(0,)$ indicates that observations of σ were sampled above 0. A mildly informative prior on item difficulty was set at $\beta_{igk} \sim N(0,1)$ for items across classes as the use of diffuse priors failed to provide enough bound on the item difficulty and standard deviation of ability parameters for the MMixIRTM. The use of such priors provided rough bounds on the parameters of the model and made fitting procedures more stable (Bolt et al., 2001, 2002; Cohen & Bolt, 2005; Cohen, Cho, & Kim, 2005; Samuelsen, 2005; Wollack, Cohen, & Wells, 2003).

The probabilities of mixtures were modeled using two approaches. In the first method, priors were incorporated into the probabilities of mixtures. For the prior of $\pi_{g|k}$, a Dirichlet distribution can be used as the conjugate prior of the parameters of the multinomial distribution:

$$\frac{\Gamma\left(\sum_{g} \alpha_g\right)}{\prod_{g} \Gamma(\alpha_g)} \cdot \prod_{g}^{G} \pi_{g|k}^{\alpha_g - 1}, \tag{5}$$

where $\sum_{g=1}^{G} \pi_{g|k} = 1$ for all school-level latent class ks with the proportion of π_k, and G indicates the number of student-level latent classes. One way to sample $\pi_{g|k}$ from the Dirichlet distribution is to sample G independent random variables $\pi^*_{g|k}$ from the Gamma distribution, $\text{Gamma}(\alpha_g, 1)$, $g = 1, \ldots, G$ normalizing

$$\pi_{g|k} = \frac{\pi^*_{g|k}}{\sum_{g=1}^{G} \pi^*_{g|k}}, \tag{6}$$

for each k. In a similar way, the Dirichlet distribution with Gamma sampling was used as a prior of π_k (Gelman, Carlin, Stern, & Rubin, 2003).

In the second method, a multinomial logistic regression with a covariate model (Dayton & Macready, 1988; Vermunt & Magidson, 2005) was used for representing student-level mixtures conditional on a particular school-level mixture (i.e., $\pi_{g|k}$). The following model with covariates was used:

$$\pi_{g|k,W_j} = \frac{\exp\left(\gamma_{0gk} + \sum_{p=1}^{P} \gamma_{pg} W_{jp}\right)}{\sum_{g=1}^{G} \exp\left(\gamma_{0gk} + \sum_{p=1}^{P} \gamma_{pg} W_{jp}\right)}. \quad (7)$$

Priors of γ_{0gk} and γ_{pg} were set to $N(0,1)$. For identifiability, $\gamma_{01} = 0$ and $\gamma_{p1} = 0$.

The probability of a school belonging to latent class k, π_k, can be written as

$$\pi_{k|W_t} = \frac{\exp\left(\gamma_{0k} + \sum_{p=1}^{P} \gamma_{pk} W_{tp}\right)}{\sum_{k=1}^{K} \exp\left(\gamma_{0k} + \sum_{p=1}^{P} \gamma_{pk} W_{tp}\right)}. \quad (8)$$

Priors of γ_{0k} and γ_{pk} were set to $N(0,1)$. For identifiability, $\gamma_{01} = 0$ and $\gamma_{p1} = 0$.

Special Cases of the MMixIRTM

Below, we introduce three special cases of the MMixIRTM, each of which has utility for estimation of differences in item parameters in the multilevel model. These are each obtained through the use of different sets of constraints. The general MMixIRTM described earlier and the three special cases of the general model are described in Table 2 for each ability distribution and for item difficulty parameters.

Special Case I

The first special case of a MMixIRTM is that in which the proportions of student-level latent classes are same among school-level latent classes. When this assumption holds, item (i.e., β_{igk}) and ability (i.e., θ_{jtgk}) parameters can be split into student-level and school-level parameters. We illustrate this special case as follows:

$$P\left(y_{ijtgk} = 1 | g, k, \theta_{jtg}, \theta_{jtk}\right) = \frac{1}{1 + \exp\left[-\left(\theta_{jtg} + \theta_{jtk}\right) - \left(\beta_{ig} + \beta_{ik}\right)\right]}. \quad (9)$$

Equation 9 shows that the item difficulty parameters β_{ig} and β_{ik} are estimated separately for the student-level and for the school-level latent groups. If this assumption holds, then the formulation indicates that differences in item response characteristics can be analyzed separately at the student level and at the school level.

TABLE 2
Comparisons of the Multilevel Mixture IRT Model

Model	Model Rationale	Level	Ability Distribution	Proportions	Item Difficulty
General model	k has different proportions of gs	Both student and school levels	$\theta_{jtgk} \sim N\left(\mu_{gk}, \sigma^2_{gk}\right)$	$\pi_{g\|k}, \pi_k$	β_{igk}
Special Case I	k has the same proportions of gs	Student level	$\theta_{jtg} \sim N\left(\mu_g, \sigma^2_g\right)$	π_g	β_{ig}
		School level	$\theta_{jtk} \sim N\left(\mu_k, \sigma^2_k\right)$	π_k	β_{ik}
Special Case II (Asparouhov & Muthén, 2008)	gs are clustered with respect to same student-level ability distribution	Student level	$\theta_{jtg} \sim N\left(\mu_g, \sigma^2_g\right)$	π_g	β_{ig}
		School level	$\theta_{jt} \sim N(0, 1)$	NA	NA
Special Case III (Vermunt, 2007a)	ks are clustered with respect to same school-level ability distribution	Student level	$\theta_{jt} \sim N(0, 1)$	NA	NA
		School level	$\theta_{jtk} \sim N\left(\mu_k, \sigma^2_k\right)$	π_k	β_{ik}

Note: IRT = item response theory; NA = not applicable.

Special Case II

The second special case we consider is that for which item and ability parameters do not vary across school-level classes. This model can be useful when the purpose of analysis is to identify different students' strategies with incorporating multilevel data structure. We illustrate this special case as follows:

$$\mathrm{P}\left(y_{ijtg}=1|g,\theta_{jtg},\theta_{jt}\right)=\frac{1}{1+\exp\left[-\left(\theta_{jtg}+\theta_{jt}-\beta_{ig}\right)\right]}. \quad (10)$$

It can be seen in Equation 10 that the item difficulty parameter β_{ig} differs among student-level latent classes. It does not contain a k subscript indicating that the same parameters hold for each school-level latent class. If this model holds, then differences in item characteristics are present only at the student level. The j subscript in θ_{jt} of Equation 10 indicates students in school t of class g. A similar model was described in Asparouhov and Muthén (2008).

Special Case III

The third special case of the MMixIRTM is one in which item and ability parameters do not vary among student-level classes. This special model is of interest in case we seek to obtain only school-level DIF information. In fact, this model contains information aggregated across student-level latent classes for each school-level latent class. We illustrate this special case as follows:

$$\mathrm{P}\left(y_{ijtk}=1|k,\theta_{jt},\theta_{jtk}\right)=\frac{1}{1+\exp\left[-\left(\theta_{jt}+\theta_{jtk}-\beta_{ik}\right)\right]}. \quad (11)$$

The item difficulty estimates, β_{ik}, contain a k subscript but not a g subscript indicating that they differ only by school-level latent class. This formulation was illustrated in Vermunt (2007a) and is of interest when we seek to examine DIF only among the school-level latent classes.

Parameter Estimation

The MMixIRTM parameters were estimated using a Markov chain Monte Carlo (MCMC) algorithm written in WinBUGS 1.4 (Spiegelhalter, Thomas, & Best, 2003). Below, we examine issues that need to be considered during estimation.

Label Switching

Label switching occurs when latent classes change meaning over the estimation chain. This also occurs in other types of estimation as well (e.g., maximum likelihood estimation) but is of particular interest here, in the context of MCMC. There are three possible types of label switching that can arise in a MMixIRTM. The three types are similar in that the meaning of the label has changed but they

differ in their causes and consequences. Label switching within a MCMC chain is the first type, and label switching across chains is the second type. The third type of label switching is a variant of the second type in that student-level latent classes switch within a school-level latent class. Each type of label switching is described below.

The first type of label switching, switching within a chain on subsequent iterations, can be a serious problem in Bayesian estimation. This type of label switching occurs essentially because there is a lack of sufficient information available to the algorithm to discriminate between latent groups of mixture models belonging to the same parametric family (McLachlan & Peel, 2000). This problem is harder to deal with for a MMixIRTM because of the multilevel structure of mixtures. We use the notion of the hierarchical mixtures of experts model (Jordan & Jacobs, 1992) and view the MMixIRTM as having a two-layer mixture structure. The higher level is the school level and mixing proportions in the latent classes are π_k. At the lower level, the student level, latent class has mixing proportions $\pi_{g|k}$. A necessary condition for identification of a multilevel mixture model is that the higher level model has the structure of an identifiable latent class model (Vermunt, 2007b). Separate identifiability of the lower part of the model is a sufficient condition but not always necessary when the number of higher level classes is larger than 1. A necessary condition for identification is that the π_k for the K latent classes be identified. As noted earlier, label switching can be observed, when distinct jumps occur in the traces of a parameter and when the density for the parameter has multiple modes (Stephens, 2000). If multiple modes do not exist for the π_k, the first-type label of switching is not present and a necessary condition for identification has been satisfied.

The second type of label switching arises in a MCMC chain such as for different replications as in a simulation study or for different initial values. A variant of this kind of label switching also can happen within each school-level mixture in the MMixIRTM, because the student-level proportions are modeled within a school-level mixture. This latter variant of label switching is the third type of label switching and occurs when the labeling of student-level membership is different for each school-level mixture. If this kind of label switching occurs in a simulation study, the parameter estimates can be compared with the generating parameters to determine which labels should be applied to each of the latent classes for each school-level mixture. With real data, group memberships can be matched across chains and across school-level mixtures by looking at the patterns of the means of ability, mixture proportions, and difficulty.

Posterior Distribution

The probability of getting a correct response is a function of g, k, and θ_{jtgk}. The class-specific probability, $P(y_{ijt} = 1|g, k, \theta_{jtgk})$, is composed of mixing

proportions, π_k and $\pi_{g|k}$. Unobserved indicator variables, ζ^l_{jtgk}, are introduced, because an individual j in group t is assigned to both a student-level latent class g and a school-level latent class k in iteration l. The likelihood function of the MMixIRTM is as follows:

$$\begin{aligned} L(g,k,\theta_{jtgk}) \quad = \quad & \prod_{i=1}^{I}\prod_{j=1}^{J}\left[\left\{\sum_{k=1}^{K}\sum_{g=1}^{G}\pi_k\cdot\pi_{g|k}P\left(y_{ijt}=1|g,k,\theta_{jtgk}\right)\right\}^{u_{ij}}\right. \\ & \left.\cdot\left\{1-\sum_{k=1}^{K}\sum_{g=1}^{G}\pi_k\cdot\pi_{g|k}P\left(y_{ijt}=1|g,k,\theta_{jtgk}\right)\right\}^{1-u_{ij}}\right]^{\zeta^l_{jtgk}}, \end{aligned} \tag{12}$$

where u_{ij} are dichotomously scored responses as 0 and 1, ζ^l_{jtgk} is 1 if the examinee j is from mixtures g and k and $\zeta^l_{jtgk}=0$ otherwise at an iteration l.

Reparameterizing $\theta_{jtgk}=\sqrt{\sigma^2_{gk}}\cdot\eta_{jtgk}$, the joint posterior distribution for the use of priors on $\pi_{g|k}$ and π_k,

$$S=\left\{g,k,\eta_{jtgk},\mu_{gk},\sigma_{gk},\beta_{igk},\pi_k,\pi_{g|k}\right\},$$

can be written as

$$\begin{aligned} P(S|U) \quad & \propto L\left(g,k,\eta_{jtgk}\right)P\left(\eta_{jtgk}|\mu_{gk}\right)P\left(\mu_{gk}\right)P\left(g|\pi_{g|k}\right) \\ & \cdot P\left(\pi_{g|k}\right)P(k|\pi_k)P(\pi_k)P\left(\sigma_{gk}\right)P\left(\beta_{igk}\right). \end{aligned} \tag{13}$$

The joint posterior distribution for the use of a multinomial logistic regression model on $\pi_{g|k}$ and π_k,

$$S=\left\{g,k,\eta_{jtgk},\mu_{gk},\sigma_{gk},\beta_{igk},\pi_k,\pi_{g|k},\gamma_{0gk},\gamma_{pg},\gamma_{0k},\gamma_{pk}\right\},$$

can be written as

$$\begin{aligned} P(S|U) \quad & \propto L\left(g,k,\eta_{jtgk}\right)P\left(\eta_{jtgk}|\mu_{gk}\right)P\left(\mu_{gk}\right)P\left(g|\pi_{g|k}\right)P\left(\pi_{g|k}|\gamma_{0gk},\gamma_{pg}\right)P\left(\gamma_{0gk}\right)P\left(\gamma_{pg}\right) \\ & \cdot P(k|\pi_k)P\left(\pi_k|\gamma_{0k},\gamma_{pk}\right)P\left(\gamma_{0k}\right)P\left(\gamma_{pk}\right)P\left(\sigma_{gk}\right)P\left(\beta_{igk}\right). \end{aligned} \tag{14}$$

Sampling in WinBUGS

Figure 1 presents a graphical representation showing the sequencing of MCMC sampling used by the WinBUGS code given in the Appendix. The processing in WinBUGS proceeds by sampling all nodes starting at the outer edge of the diagram beginning with the hyperparameters and working inward to the p[j,i]. As an example, pi is the variable name used in the program code for $\pi_{g|k}$, pi1 is the variable name used in the program code for π_k, g[j] is the index for group membership for student j, gg[group[j]] is the index for group membership for school k, eta[j] is η_j, a[g[j],gg[group[j]] is σ_{gk}, beta[i,g[j],gg[group[j]]] is β_{igk} and mutg[g[j],gg[group[j]]] is μ_{gk}, and the r[j,i] are the response data. A solid

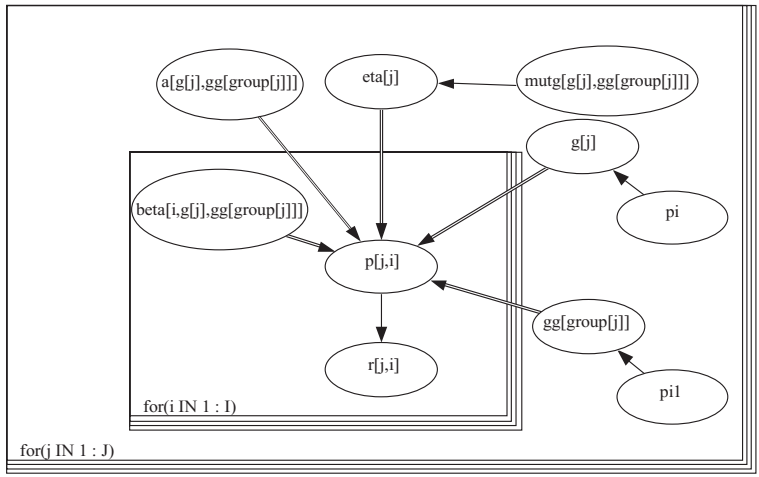

FIGURE 1. *Graphical representation of MMixIRTM with priors.*

arrow indicates a stochastic dependence and a hollow arrow indicates a logical function. From the diagram, it can be seen that eta[j] depends on mutg[g[j], gg[group[j]]], g[j] depends on pi, and gg[group[j]] depends on pi1. p[j,i] (which is the code in the program for $p(P(y_{ijt} = 1|g, k, \theta_{jgk}))$ is a logical function of a[g[j],gg[group[j]], beta[i,g[j],gg[group[j]]], eta[j], mutg[g[j],gg[group[j]]], g[j], and gg[group[j]].

Once the model is fully specified, WinBUGS then determines the necessary sampling methods directly from the structure in the diagram. The form of the full conditional distribution of μ_{gk}, $\pi_{g|k}$, and π_k is a conjugate distribution of the parameters (i.e., normal and Dirichlet distributions), so that in this study, direct sampling was conducted using standard algorithms. The form of the full conditional distribution of g and k is a discrete distribution with a finite upper bound so that the inversion method is implemented. The form of the full conditional distribution of η_j, a_{gk}, and β_{igk} is log-concave, so that WinBUGS uses derivative-free adaptive rejection sampling. The truncated version of the normal distribution of σ_{gk} is log-concave as well.

Starting values are needed for each parameter being sampled to define the first state of the Markov chain. The starting values for the remaining model parameters were randomly generated in the WinBUGS software except that the school-level group membership was randomly set.

Model Selection

When the number of latent classes in a model is unknown, the parameter space is ill defined and of infinite dimension (McLachlan & Peel, 2000). One approach

in Bayesian estimation of mixture models is to consider the number of latent classes, g and k, to be an unknown parameter with a prior distribution. Richardson and Green (1997) describe an approach in which the number of latent classes is an unknown parameter that is to be estimated. The usual approach in mixture modeling, and the one taken in this article, however, is to fit a range of mixture models each with a different number of latent classes and then to consider the selection of a model according to some theoretical rationale often including use of some appropriate statistical criteria.

Akaike's (1974) information criterion (AIC) and Schwarz's (1978) Bayesian information criterion (BIC) were used in this study for selection of the model with the correct number of mixtures. Because the ζ_{jtgk}^{l} may be different at each iteration in sampling, it is necessary to monitor the likelihood at each iteration. The definitions of AIC and BIC in this study are as follows:

$$\text{AIC} = -2\log(L)^{l} + 2m, \tag{15}$$

$$\text{BIC} = -2log(L)^{l} + m\log n, \tag{16}$$

where $\log(L)^{l}$ is a log likelihood at an iteration l, m is the number of parameters, and n is the number of observations. The model selection strategy taken in this study is one in which the candidate models, each describing different numbers of latent classes, are run in parallel. Information is then accumulated over iterations to provide a probability that a specific model is selected by AIC and BIC (Congdon, 2003). This approach compares the averages of the AIC and BIC fit measures over iterations in a MCMC run following burn-in.

A Multilevel Analysis of Latent Groups DIF

In this section, we provide a simulation study and a real data example illustrating how the MMixIRTM can be used. We provide this in the context of a multilevel analysis of DIF and motivate the example as follows: The usual approach to detection of DIF is to compare responses from manifest groups conditional on some measure of ability, oftentimes simply the score on the test being studied. Unfortunately, comparisons among members of manifest groups cannot accurately identify those examinees in each group who do or do not respond differentially to an item (Cohen & Bolt, 2005; Cohen et al., 2005; DeAyala, Kim, Stapleton, & Dayton, 2002; Samuelsen, 2005). The use of MixIRT models for DIF comparisons has been suggested as an alternative to manifest groups because they help provide better identification and more accurate explanations of the causes of DIF. In this example, we show how the MMixIRTM can be used to identify and describe characteristics of latent groups in the context of a multilevel DIF detection analysis.

Scale Anchoring and Linking for the MMixIRTM

Linking of scales is necessary to make comparisons of item parameter estimates among the different latent groups in the model. In this study, we anchor the metric with respect to the ability distribution. For metric anchoring, the mean of η_{gk} is set to 0 for the reference group (in this study, $\mu_{11} = 0$). The means for the remaining latent classes are set to μ_{gk}; that is, they are to be estimated. Thus, the mean and variance of the distributions of the other groups are estimated relative to the $N(0, 1)$ scale of the reference group in terms of η_{jtgk}. The procedure can be described as follows:

$$\text{logit}\left[P\left(y_{ijtgk} = 1 | g, k, \eta_{jtgk}\right)\right] = \sqrt{\sigma^2_{gk}} \cdot \eta_{jtgk} - \beta_{igk}. \tag{17}$$

DIF analysis with the MMixIRTM is done over the same set of items across latent classes, g and k. That is, each latent class of examinees responded to the same set of items. In this way, one can think of every item on the scale as being a potential anchor item to be used in estimating an appropriate link. This is similar to a common-item internal anchor nonequivalent groups linking design, although, in the MMixIRTM, class-specific item difficulties as well as group memberships g and k are estimated simultaneously. For comparisons of the item difficulties across latent classes, g and k, the $\widehat{\beta}_{igk}$ are transformed for each classes g and k with the $\sum_i^I \widehat{\beta}_{igk} = 0$. The result of this transformation is that the mean of item difficulties is 0 for each class for a DIF analysis (Lord, 1980; Samuelsen, 2005, 2008; Wright & Stone, 1979; Zimowski, Muraki, Mislevy, & Bock, 1996).

DIF Detection Procedure

DIF at the student level is defined based on differences in item parameters within a school-level latent class. For studied item i of a school level in latent class k, the null hypothesis can be stated as of interest:

$$\text{H}_0 : \beta_{igk} - \beta_{ig'k} = 0, \tag{18}$$

indicating no difference in item difficulties between two student-level groups.

DIF at the school level can be determined by comparing the differences between the item difficulties among school-level latent classes. Let group $k = 1$ be the focal group for the focal school and the remaining $K - 1$ groups be the reference groups for that school. For the studied item i and for student-level latent class g, the null hypothesis can be stated as

$$\text{H}_0 : \beta_{igk} - \beta_{igK} = 0, \tag{19}$$

indicating no difference in item difficulties between two school-level groups.

The difference between each pair of item difficulties can be tested with a highest posterior density (HPD) interval (Box & Tiao, 1973). Assuming a nominal α

TABLE 3
Proportions Simulated in Each Latent Class

	$P(K = 1) = .5$	$P(K = 2) = .5$
$P(G = 1)$	.92	.32
$P(G = 2)$	.08	.68

of .05 for rejection of the null hypothesis in which a parameter value or a function of parameter values is 0, the HPD interval can be used to test, if the value differs significantly from 0 (Box & Tiao, 1973). Using the HPD method, Samuelsen and Bradshaw (2008) found the power to detect DIF was a function of the magnitude of the DIF and whether the ability distributions of the focal and reference groups were matched. Type I error rates using this method were generally within acceptable levels. In this study, if the HPD interval of the DIF measure for each item did not include 0, the item was considered DIF (Samuelsen, 2005; Samuelsen & Bradshaw, 2008).

Simulation Study

We first present the simulation study to examine the performance of the MMixIRTM under some practical DIF testing conditions. The data in this simulation were simulated to possibly have differentially functioning items at each of two different levels.

Design of Study

The following conditions were examined to reflect practical testing situations for a single 40-item test: five effect sizes of DIF (0.4, 0.6, 0.8, 1, and 1.2), two percentages of DIF items at the school level (10% and 30% DIF items), two student and school sample size combinations (25 students per school with 320 schools and 100 students per school with 80 schools), and two models of probabilities of latent classes (a Dirichlet distribution with Gamma sampling and a multinomial logistic regression model). Finally, two percentages of the overlap between covariates and latent classes (100% and 60%) were examined for multinomial regression modeling of the mixing proportions of latent classes. (Examples of typical covariates considered in a DIF analysis include manifest characteristics such as gender or ethnicity.)

Two latent classes were generated at both the student level and the school level using the mixing proportions given in Table 3. Both school-level latent classes were simulated to have mixing proportions of .5. Within school-level latent Class 1 (i.e., $k = 1$), student-level latent Class 1 (i.e., $g = 1$) was simulated to be larger, containing 92% of the examinees. For school-level latent Class 2, the

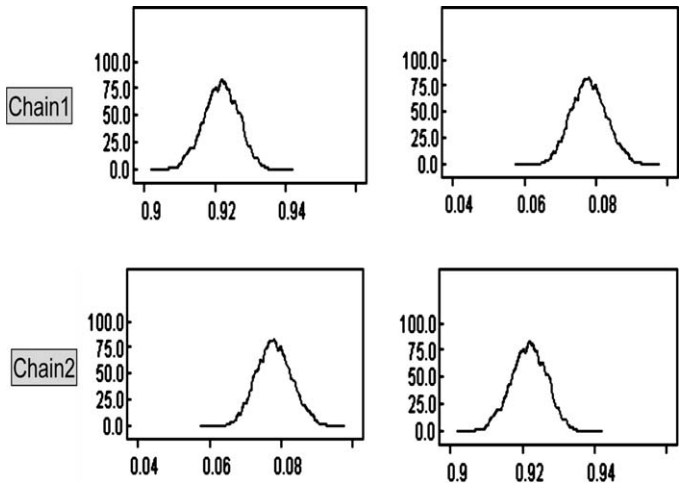

FIGURE 2. *The marginalized posterior density of* $\pi_{g|k}$: *for Type 2 label switching at the school level*.

second student-level class (i.e., $g = 2$) was simulated as the larger group with 68% of the cases. This reflected a less dominant group in the second school-level latent class. Five replications were generated for each condition.

Detection of Label Switching

Because the estimated marginal posterior densities for π_k in the MMixIRTM were unimodal, it was appropriate to conclude that the first type of label switching, label switching within the MCMC chains, did not occur for models with and without covariates. In this case, it was possible to infer that a necessary condition for identification was satisfied. However, the second type of label switching, label switching across chains, was observed for a few of the simulation conditions for both the models with and without covariates. These conditions included six simulation conditions with 80 schools and 100 students per school. Figure 2 shows the marginalized posterior density of π_k from the two different chains to illustrate the second type of label switching at the school level. Chain 1 has the same latent labeling as generated (i.e., $\pi_{1|1}$ and $\pi_{2|1}$ as shown in Table 3); however, the labeling switched in Chain 2 such that $\pi_{1|2}$ and $\pi_{2|2}$. By comparing parameter estimates with the generating parameters, it was possible to determine which labels should be applied to each of the latent classes.

The third type of label switching was not detected because labeling of student-level group membership was observed to be consistent across school-level mixtures for models with and without covariates.

Monitoring Convergence

Three convergence methods were used, the Gelman and Rubin (1992) method as implemented in WinBUGS, the Geweke (1992) method, and the Raftery and Lewis (1992) method as implemented in the computer program Bayesian Output Analysis (Smith, 2004). In addition, autocorrelation plots from WinBUGS were examined. A conservative burn-in of 7,000 iterations was used in this study followed by 8,000 post–burn-in iterations for all conditions. Thinning was set at 40, meaning that 32,000 iterations were required after burn-in to obtain the 8,000 iterations.

Model Selection

The recovery of the generating model was assessed using two commonly used information-based indices, AIC and BIC, to determine model fit. For the 10% DIF condition, the correct number of latent classes (i.e., $G = 2$ and $K = 2$) was not selected. In the 30% DIF condition, however, the BIC accurately selected the correct model for both the 25 and 100 students per school conditions. Consistent with previous research, AIC tended to select the more complex model. BIC correctly identified the generating MMixIRTM for the 2-group solution in the 30% DIF condition. This result suggests that more than 10% DIF was needed at the school level to detect the correct number of latent classes at student and school levels for the given conditions.

Recovery of Generating Parameters

The recovery study was conducted to determine the success of the algorithm in detecting the correct numbers of latent classes. The recovery results for the 30% DIF condition are shown in Table 4.

The recovery of group membership for both student- and school-level mixtures was good for both 25 students/320 schools and 100 students/80 schools. With respect to the correctly selected number of mixtures and identified model, the accuracy of detection of student-level group membership was 98.5 to 99.6% and the accuracy of detection of school-level group membership was 100%.

The marginalized posterior density for item difficulty of all items was nearly symmetric. Thus, posterior means were considered for the parameter recovery of item difficulties. The density plots given in Figure 3 are similar to those for the rest of the items on the test. Table 5 shows item difficulty estimates and their HPDs only for 13 DIF items from one replication of one simulation condition (i.e., for 30% DIF items, 25 students per school with 320 schools, and use of a Dirichlet distribution with Gamma sampling on the probabilities of mixtures)

TABLE 4
Model Parameter Recovery Results

A. Group Membership Recovery With 30% DIF Items

		25 Students/320 Schools		100 Students/80 Schools	
		Student Level	School Level	Student Level	School Level
Prior		98.7	100	98.6	100
Covariate	60%	98.7	100	98.6	100
	100%	99.4	100	99.9	100

B. Item Difficulty Recovery With 30% DIF Items

		25 Students/320 Schools			100 Students/80 Schools		
		RMSE	Bias	Correlation	RMSE	Bias	Correlation
Prior		0.096	0.033	.997	0.101	0.000	.997
Covariate	60%	0.097	−0.001	.997	0.100	0.000	.997
	100%	0.094	0.000	.997	0.115	0.001	.997

Note: DIF = differential item functioning; RMSE = root mean squared error.

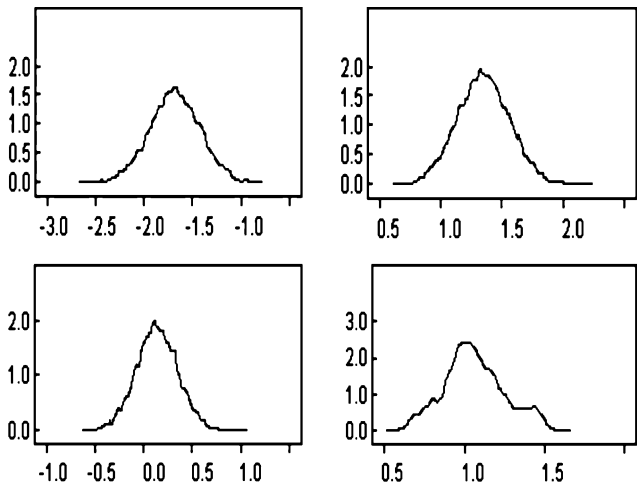

FIGURE 3. *Marginalized posterior densities of difficulties for four items.*

as an example. The root mean squared errors (RMSEs) for item difficulty were .094 to .115, biases were −.001 to .033, and Pearson correlations were .997 for all 30% DIF items conditions. This result indicates that the item difficulty parameter was recovered well.

TABLE 5
Item Difficulty Estimates for 13 Items[a]

	$G = 1, K = 1$		$G = 2, K = 1$		$G = 1, K = 2$		$G = 2, K = 2$	
Item	Parameters	Estimates	Parameters	Estimates	Parameters	Estimates	Parameters	Estimates
1	−2.000	−1.957 (−2.090, −1.827)	0.000	0.028 (−0.329, 0.461)	−2.400	−2.302 (−2.588, −1.937)	0.400	0.402 (0.049, 0.733)
2	−2.000	−2.022 (−2.153, −1.890)	0.000	−0.066 (−0.411, 0.380)	−3.200	−3.181 (−3.467, −2.746)	1.200	1.262 (0.913, 1.594)
3	−1.750	−1.759 (−1.881, −1.640)	0.000	0.047 (−0.100, 0.482)	−2.550	−2.500 (−2.786, −2.115)	0.800	0.758 (0.409, 1.093)
4	−1.750	−1.717 (−1.837, −1.600)	0.250	0.279 (0.071, 0.269)	−2.750	−2.650 (−2.936, −2.248)	1.250	1.319 (0.969, 1.651)
5	−1.500	−1.524 (−1.642, −1.410)	0.250	0.229 (−0.168, 0.614)	−2.100	−2.186 (−2.472, −1.814)	0.850	0.863 (0.518, 1.197)
6	−1.500	−1.457 (−1.570, −1.345)	0.250	0.254 (−0.139, 0.647)	−2.700	−2.514 (−2.800, −2.133)	1.450	1.429 (1.081, 1.756)
7	−1.250	−1.260 (−1.367, −1.151)	0.250	0.258 (−1.029, 0.656)	−1.650	−1.673 (−1.959, −1.304)	0.650	0.631 (0.281, 0.961)
8	−1.250	−1.265 (−1.373, −1.156)	0.500	0.544 (0.146, 0.938)	−2.450	−2.364 (−2.650, −1.989)	1.700	1.659 (1.313, 1.985)
9	−1.000	−0.962 (−1.066, −0.862)	0.500	0.410 (0.011, 0.804)	−1.800	−1.913 (−2.199, −1.550)	1.300	1.287 (0.933, 1.620)
10	−1.000	−1.026 (−1.128, −0.923)	0.500	0.417 (0.016, 0.808)	−2.000	−1.940 (−2.226, −1.572)	1.500	1.449 (1.100, 1.781)
11	−0.500	−0.524 (−0.621, −0.432)	1.000	0.944 (0.560, 1.337)	−1.100	−1.012 (−1.298, −0.670)	1.600	1.648 (1.298, 1.977)
12	−0.500	−0.461 (−0.553, −0.366)	1.000	0.870 (0.482, 1.265)	−1.700	−1.626 (−1.912, −1.266)	2.200	2.064 (1.720, 2.395)
13	−0.500	−0.537 (−0.632, −0.444)	1.000	1.085 (0.688, 1.485)	−0.500	−0.510 (−0.797, −0.171)	1.000	0.998 (0.171, 0.171)

a. Simulation condition with 30% differential item functioning (DIF) items, 25 students per school with 320 schools, and use of a Dirichlet distribution with Gamma sampling on the probabilities of mixtures. Estimates: posterior mean. Credibility interval in parentheses.

Multilevel DIF: An Empirical Illustration on a Standardized Mathematics Test

In this section, the MMixIRTM is illustrated with data from the mathematics section of the 2006 form of a large-scale standardized test. The test is intended to provide high school students with practice for taking a college-level admissions test and to give students an opportunity to enter college scholarship programs.

Data

The sample was selected using only students in the 10th or 11th grades for the current school year, who had converted scores that fell within the limits of the score scale. Nonstandard students were excluded from the sample. The sample included 987 schools and $39,614$ students. An approximately 20% random sample of 206 schools and $8,362$ students was drawn for this illustration.

Intraclass correlations (ICCs, Raudenbush & Bryk, 2002) were used to determine the hierarchical structure of the data. The ICC for the multilevel IRT model was .289, indicating 28.9% of the total variance was explained at the school level. An ICC with a linear mixed effects model fitted using the *lmer* function (Bates & Debroy, 2004) was .219 based on the total scores, indicating 21.9% of the total variance was explained at the school level.

Results

The same priors on probabilities of mixtures, label switching analyses, convergence checking, and model selection analyses were used for the empirical study as described above for the simulation study. As previously noted, it was possible to use either of the priors on the probabilities of mixtures; that is, either a model without covariates or a multinomial logistic regression model (i.e., with covariates). Results from both are described below.

Table 6 presents averaged AIC and BIC values across iterations following burn-in. The probabilities that a model had the minimum AIC and the minimum BIC over the MCMC chain are shown in parentheses. Based on the BIC values shown in Table 6, $G = 4$ and $K = 2$ were chosen based on the model with and without covariates.

Table 7 presents student-level latent class proportions for each school-level latent class, k. There was a similar pattern in proportions with the use of covariates and without. As indicated values in bold in Table 7, school-level Class 1 had student-level Class 4 as a dominant group, whereas school-level Class 2 had student-level Class 1 as a dominant group.

Using models with and without covariates, student-level latent Classes 1, 2, 3, and 4 were essentially "Average," "Low," "High," and "Very Low" ability groups, respectively. This is shown in Table 8. That is, it appears that

TABLE 6
Model Selection Result for Mathematics Section

Number of Mixtures	Model	Npar[b]	BIC (Prop)	AIC (Prop)
A. Model Without Covariates[a]				
$G = 1\ K = 1$	Multilevel Rasch model	40	310200 (0)	309900 (0)
$G = 2\ K = 2$	MMixIRTM	162	300500 (0)	299300 (0)
$G = 3\ K = 2$	MMixIRTM	244	299000 (0.12)	297300 (0)
$G = 4\ K = 2$	MMixIRTM	326	**298800 (0.88)**	296500 (0.25)
$G = 5\ K = 2$	MMixIRTM	408	299500 (0)	**296400 (0.67)**
$G = 3\ K = 3$	MMixIRTM	367	300200 (0)	297200 (0.02)
$G = 4\ K = 3$	MMixIRTM	490	300800 (0)	297300 (0.01)
$G = 5\ K = 3$	MMixIRTM	613	301200 (0)	296900 (0.05)
B. Model With Covariates[a]				
$G = 1\ K = 1$	Multilevel Rasch model	40	310200 (0)	309900 (0)
$G = 2\ K = 2$	MMixIRTM	182	300800 (0)	299500 (0)
$G = 3\ K = 2$	MMixIRTM	266	299700 (0.05)	297800 (0)
$G = 4\ K = 2$	MMixIRTM	370	**299400 (0.64)**	296800 (0.03)
$G = 5\ K = 2$	MMixIRTM	438	299500 (0.31)	**296400 (0.93)**
$G = 3\ K = 3$	MMixIRTM	409	301900 (0)	299100 (0)
$G = 4\ K = 3$	MMixIRTM	545	301400 (0)	297600 (0)
$G = 5\ K = 3$	MMixIRTM	681	301700 (0)	296900 (0.04)

a. Proportion (prop) model selected shown in parentheses.
b. The number of parameters.

TABLE 7
Class Proportions for Mathematics Section Within School-Level Group Membership

		$\hat{\pi}_k$	$G = 1$	$G = 2$	$G = 3$	$G = 4$
Without covariates	$K = 1$	.462	.187	**.386**	.058	**.369**
	$K = 2$	.538	**.404**	.272	.194	.130
With covariates	$K = 1$	.465	.246	.272	.130	**.351**
	$K = 2$	.535	**.396**	.279	.100	.224

school-level Class 1 was a low-ability group, whereas school-level Class 2 was a high-ability group.

Table 9 presents the DIF items detected for the model without covariates. DIF items at the school level were detected by doing comparisons of student-level item difficulties across school-level classes. Values inside parentheses in Table 9 indicate the lower and upper bounds of the HPD for each item. Items with boldface entries were detected as DIF items. At the school level (see Table 9), one

TABLE 8
Distribution of Ability for Mathematics Section: With and Without Covariates

		$G = 1$	$G = 2$	$G = 3$	$G = 4$
Without covariate	$K = 1$	$N(0^{a}, 0.572^2)$	$N(-1.965, 0.431^2)$	$N(1.344, 0.893^2)$	$N(-4.557, 0.299^2)$
	$K = 2$	$N(0.562, 0.575^2)$	$N(-1.768, 0.412^2)$	$N(1.474, 1.067^2)$	$N(-4.001, 0.319^2)$
With covariate	$K = 1$	$N(0^{a}, 1.131^2)$	$N(-1.842, 0.425^2)$	$N(0.535, 0.564^2)$	$N(-4.524, 0.311^2)$
	$K = 2$	$N(0.570, 0.563^2)$	$N(-1.730, 0.442^2)$	$N(1.621, 1.032^2)$	$N(-4.250, 0.307^2)$

a. Fixed for identification.

TABLE 9
Magnitude of School-Level Differential Item Functioning (DIF) Values for Mathematics Section

Item	Across School-Level Latent Classes			
	Student-Level Class 1	Student-Level Class 2	Student-Level Class 3	Student-Level Class 4
1	−0.400 (−0.944, 0.096)	−0.135 (−0.409, 0.140)	0.710 (−0.160, 1.617)	−0.101 (−0.3644, 0.1543)
2	0.033 (−0.401, 0.456)	−0.012 (−0.235, 0.217)	0.165 (−0.791, 1.124)	−0.186 (−0.448, 0.08163)
3	0.234 (−0.430, 0.849)	0.141 (−0.100, 0.384)	1.104 (−0.041, 2.304)	−0.077 (−0.4136, 0.2386)
4	−0.018 (−0.364, 0.316)	0.079 (−0.134, 0.296)	−0.376 (−1.067, 0.324)	**−0.318 (−0.6132, −0.0276)**
5	−0.142 (−0.564, 0.251)	−0.096 (−0.328, 0.134)	−0.063 (−0.845, 0.685)	0.049 (−0.1984, 0.2948)
6	−0.012 (−0.366, 0.334)	**−0.296 (−0.514, −0.071)**	0.610 (−0.138, 1.361)	−0.089 (−0.3819, 0.1951)
7	−0.276 (−0.607, 0.042)	−0.099 (−0.340, 0.140)	−0.045 (−0.694, 0.588)	−0.034 (−0.3402, 0.2432)
8	0.184 (−0.131, 0.514)	−0.196 (−0.405, 0.021)	0.115 (−0.576, 0.780)	−0.139 (−0.4146, 0.1394)
9	0.301 (−0.007, 0.612)	0.120 (−0.105, 0.341)	0.077 (−0.715, 0.769)	−0.144 (−0.4627, 0.1608)
10	0.133 (−0.131, 0.395)	0.037 (−0.173, 0.242)	**−0.636 (−1.345, −0.040)**	0.008 (−0.2778, 0.2994)
11	−0.114 (−0.380, 0.149)	−0.038 (−0.254, 0.183)	−0.319 (−1.202, 0.411)	−0.238 (−0.5167, 0.0329)
12	−0.082 (−0.342, 0.176)	−0.123 (−0.356, 0.110)	−0.145 (−0.706, 0.387)	−0.285 (−0.5922, 0.01118)
13	0.171 (−0.092, 0.434)	**0.334 (0.096, 0.578)**	−0.138 (−0.940, 0.451)	−0.170 (−0.4547, 0.1145)
14	0.039 (−0.220, 0.312)	**−0.270 (−0.505, −0.041)**	−0.099 (−0.579, 0.355)	0.051 (−0.2555, 0.3563)
15	0.104 (−0.186, 0.386)	0.110 (−0.208, 0.416)	−0.310 (−0.833, 0.190)	**−0.359 (−0.6692, −0.0523)**
16	−0.342 (−0.711, 0.046)	0.033 (−0.330, 0.395)	−0.208 (−0.655, 0.218)	−0.096 (−0.4, 0.2039)
17	0.101 (−0.241, 0.437)	−0.001 (−0.299, 0.287)	0.249 (−0.163, 0.677)	0.041 (−0.2594, 0.339)
18	0.422 (−0.127, 0.981)	−0.025 (−0.322, 0.261)	0.167 (−0.274, 0.597)	−0.196 (−0.5175, 0.1112)
19	0.022 (−0.408, 0.468)	−0.171 (−0.543, 0.198)	0.061 (−0.379, 0.476)	**−0.494 (−0.8123, −0.1651)**
20	0.087 (–0.277, 0.460)	–0.029 (–0.405, 0.334)	0.062 (–0.393, 0.478)	–0.108 (–0.458, 0.2501)
21	–0.254 (–0.635, 0.105)	0.069 (–0.168, 0.309)	–0.120 (–0.833, 0.564)	–0.105 (–0.4793, 0.2502)

(continued)

TABLE 9 *(continued)*

Item	Across School-Level Latent Classes			
	Student-Level Class 1	Student-Level Class 2	Student-Level Class 3	Student-Level Class 4
22	–0.474 (–1.051, 0.052)	–0.026 (–0.289, 0.239)	–0.184 (–1.262, 0.836)	0.009 (–0.2963, 0.3037)
23	0.015 (–0.267, 0.299)	–0.105 (–0.363, 0.173)	–0.248 (–0.823, 0.326)	–0.032 (–0.3722, 0.2982)
24	0.081 (–0.243, 0.393)	0.027 (–0.194, 0.251)	**0.915 (0.182, 1.652)**	0.029 (–0.305, 0.3517)
25	0.188 (–0.079, 0.457)	–0.001 (–0.236, 0.228)	–0.329 (–1.108, 0.263)	**–0.467 (–0.9077, –0.062)**
26	0.066 (–0.257, 0.390)	–0.165 (–0.466, 0.116)	–0.249 (–0.662, 0.179)	**–0.389 (–0.7019, –0.0794)**
27	–0.003 (–0.344, 0.401)	0.018 (–0.300, 0.331)	–0.392 (–0.883, 0.060)	0.298 (–0.0201, 0.6174)
28	–0.239 (–0.565, 0.110)	–0.290 (–0.609, 0.027)	0.046 (–0.366, 0.465)	0.233 (–0.0407, 0.5137)
29	–0.001 (–0.299, 0.310)	**0.287 (0.048, 0.518)**	–0.402 (–1.298, 0.378)	**–0.275 (–0.5471, –0.0118)**
30	–0.232 (–0.516, 0.044)	0.077 (–0.170, 0.331)	–0.059 (–0.564, 0.461)	0.407 (–0.2039, 0.9806)
31	**–0.419 (–0.764, –0.096)**	**–0.224 (–0.445, –0.012)**	0.384 (–0.274, 1.072)	0.013 (–0.433, 0.4189)
32	0.178 (–0.139, 0.493)	0.188 (–0.069, 0.447)	0.000 (–0.710, 0.697)	0.323 (–0.2898, 0.8889)
33	–0.033 (–0.335, 0.251)	–0.018 (–0.253, 0.216)	0.049 (–0.581, 0.586)	–0.075 (–0.7754, 0.5585)
34	–0.216 (–0.576, 0.132)	0.079 (–0.185, 0.343)	–0.149 (–0.557, 0.253)	0.077 (–0.2354, 0.3758)
35	0.376 (–0.218, 1.039)	0.029 (–0.880, 0.928)	0.204 (–0.368, 0.707)	0.408 (–0.7822, 1.584)
36	0.006 (–0.300, 0.309)	–0.119 (–0.436, 0.183)	–0.275 (–0.704, 0.100)	0.861 (–0.0611, 1.795)
37	0.030 (–0.321, 0.396)	–0.172 (–0.507, 0.166)	–0.290 (–0.715, 0.109)	**0.768 (0.2857, 1.269)**
38	0.487 (–0.006, 1.106)	**0.983 (0.031, 1.964)**	0.116 (–0.416, 0.562)	0.803 (–0.4337, 2.025)
Number of DIF items	1	6	2	7

item (Item 31) in student-level Class 1, six items (Items 6, 13, 14, 29, 31, and 38) in student-level Class 2, two items (Items 10 and 23) in student-level Class 3, and seven items (Items 4, 15, 19, 25, 26, 29, and 37) in student-level Class 4 were detected as DIF items. In addition, student-level DIF items can be detected by pairwise comparisons of item difficulties within each school-level class. In school-level Class 1 (i.e., $K = 1$), the number of DIF items varied from 10 to 30 across pairwise comparisons of difficulties between student-level classes. At school-level Class 2 (i.e., $K = 2$), the number of DIF items varied from 16 to 30 across pairwise comparisons of difficulties between student-level classes. Similar DIF patterns were found with the covariate model.

The number of DIF items detected was larger than what one would expect from the usual DIF analysis using manifest groups. It occurs because the latent class approach maximizes differences among latent classes, resulting in larger numbers of DIF items and larger differences in item difficulties among latent classes (Samuelsen, 2005). This result was also consistent with previous research based on the use of MixIRT models for DIF analysis (Cohen & Bolt, 2005; Cohen et al., 2005; Samuelsen, 2005).

For the model without covariates, the association analysis between latent group membership and manifest group membership could be done at both student and school levels. Significant associations were observed between student-level latent group membership and ethnicity and gender. Significant school-level associations were found between school-level latent group membership and Title I schoolwide program, household income, and poverty level. Eighty-eight percent of students in a Title I program were categorized into school-level latent Class 1. Students from higher household income families were more likely to be in school-level latent Class 2, and all schools with at least a 30% poverty level were classed into school-level latent Class 1.

Tables 10A and B show student- and school-level covariate effects, respectively, based on the multinomial logistic regression covariate model (see Equations 7 and 8). Values in Table 10 are estimated regression coefficients. As shown in Table 10, panel A, males were more likely than females to belong to Class 1. American Indians or Alaskan Natives were more likely than Whites to belong to Class 4; Asians, Asian Americans, or Pacific Islanders were less likely than Whites to belong to Class 2; and African Americans, Mexicans, or Mexican Americans were more likely than Whites to belong to Classes 2 and 4. Puerto Ricans and other Latinos or Latin Americans were more likely than Whites to belong to Classes 2 and 4. At the school level (Table 10, panel B), only household income and poverty levels (the higher value indicates higher poverty) were significant among school-level covariates considered. Schools that have higher household incomes and lower poverty levels were more likely to belong to school Class 2.

Finally, the last five items on the test were examined as a group to determine whether particular response patterns might emerge conditional on group

TABLE 10
Covariate Effects for Mathematics Section

A. Student Level		$G = 1$	$G = 2$	$G = 3$	$G = 4$
Intercept	$K = 1$	0.380 (0.002, 1.017)*	0.133 (–0.166, 0.612)	0[a]	–0.587 (–0.900, –0.158)*
	$K = 2$	–0.590 (–0.929, –0.250)*	–0.359 (–0.571, –0.127)*	0[a]	–1.431 (–1.682, –1.188)*
Gender	Female	–0.277 (–0.475, –0.0328)*	0.124 (–0.019, 0.279)	0[a]	0.124 (–0.019, 0.279)
Ethnicity	No response	–0.207 (–1.092, 0.546)	0.702 (0.077, 1.273)*	0[a]	2.165 (1.616, 2.656)*
	American Indian or Alaska Native	–0.429 (–1.617, 0.757)	0.343 (–0.587, 1,253)	0[a]	1.040 (0.262, 1.856)*
	Asian, Asian American, or Pacific Islander	–0.303 (–0.608, 0.059)	–0.735 (–1.016, –0.399)*	0[a]	–0.327 (–0.667, 0.008)
	Black or African American	–0.123 (–0.608, 0.330)	1.502 (1.218, 1.775)*	0[a]	2.620 (2.353, 2.902)*
	Mexican or Mexican American	–0.612 (–1.555, 0.217)	1.384 (1.015, 1.773)*	0[a]	1.912 (1.597, 2.302)*
	Puerto Rican	–0.496 (–1.492, 0.449)	1.172 (0.576, 1.852)*	0[a]	2.187 (1.622, 2.768)*
	Other Hispanic, Latino, or Latin American	–0.443 (–1.061, 0.131)	1.256 (0.894, 1.584)*	0[a]	1.884 (1.544, 2.26)*
	Other	–0.163 (–0.833, 0.427)	0.418 (–0.038, 0.843)	0[a]	1.075 (0.657, 1.473)*

*significant at $p = 0.05$.

TABLE 10 *(continued)*

B. School Level		$K = 1$	$K = 2$
Intercept	Intercept	0[a]	–1.192 (–2.859, 0.295)
Metropolitan code	Suburban	0[a]	0.088 (–0.578, 0.844)
	Rural	0[a]	0.287 (–0.560, 1.187)
	Unclassified	0[a]	–0.484 (–2.091, 1.089)
School enrollment size code		0[a]	0.276 (–0.295, 0.847)
Title I schoolwide program	No Title I	0[a]	0.730 (–0.120, 0.810)
Household income		0[a]	0.145 (0.080, 0.180)*
Poverty level code		0[a]	–0.805 (–1.001, –0.712)*

Note: Type II label switching across without and with covariate models. 95% credibility interval in parentheses.

a. Fixed for model identification.

* Significant at p.

membership. This was done for models with and without covariates. These five items were gridded-in items and were classified as high difficulty in the item descriptions. Several patterns of omitted responses were noted: 99900, 99909, 99990, and 99999 (where 0, 1, and 9 indicate incorrect, correct, and omitted responses, respectively). No students assigned to Class 3, the high-ability group, had any omitted responses. Students with 0s and 1s, that is, students who tried to answer the question, were classified mostly into the average-and high-ability groups.

Discussion

A MMixIRTM was described for modeling multilevel item response data at both the student level and school level. The model developed in this study used features of an IRT model, an unrestricted latent class model, and a multilevel model. The student level of the model provides an opportunity to determine whether latent classes exist that differ in their response strategies to answering questions. Information at the school level can be used to reveal possible differences that might be due to curricular or pedagogical differences among latent classes. In addition, a Bayesian solution was described for estimation of the model parameters as implemented using the freely available software, WinBUGS. A simulation study was presented to investigate the performance of the model under some practical DIF testing conditions. Generated parameters were recovered very well for the conditions considered. Use of MMixIRTM was also illustrated with a standardized mathematics test.

The MMixIRTM makes it possible to describe differential item performance of a target school using descriptions of student- and school-level characteristics associated with the given school compared to characteristics associated with other schools not in the same latent class as the target school. This description can then be used to provide schools with a framework within which to compare the results of their school with other schools in their latent class and in the other latent classes. As an example, schools classified into the same latent class as the target school (i.e., school-level Class 2) were characterized by lower Title I enrollment, higher household income, lower poverty levels, and a predominance of students in student-level latent Class 1. Students in each of the latent classes were also characterized by differences in ability, as well as by differences in response strategies, particularly at the end of the test. In the DIF example, providing a description of all schools that are members of the same school-level latent class facilitated comparisons among latent classes and allowed for the possibility that there may be more than one comparison group.

In this study, the pairwise standardized item difficulty differences between groups were used for DIF detection (Lord, 1980; Wright & Stone, 1979; Zimowski et al., 1996) in the context of the HPD interval (Samuelsen, 2005, 2008; Samuelsen & Bradshaw, 2008). This procedure is only one of the ways

of implementing a DIF analysis in a Rasch model. This method can yield DIF inflation when DIF is asymmetrical (De Boeck, 2008; Wang, 2004). DIF is asymmetrical if it is restricted to a subset of the items showing a mean difference in difficulty between the two groups (De Boeck, 2008). The likelihood ratio test (Thissen, Steinberg, & Wainer, 1988, 1993) is a more general model-based approach. One issue that needs to be addressed in a MixIRTM DIF study is finding anchor items for scale comparability across latent classes.

Information in Table 9 was used to characterize latent classes with respect to item difficulty at both student and school levels. Information regarding item-level content and surface characteristics of items would also be helpful for characterizing patterns of responses in latent classes. Because of the secure nature of the test, however, test content was not available for analysis in this study. Some success has been shown, for example, using cognitive characteristics of mathematics test items to explain differences between gender groups (Gallagher, 1998; Gallagher & DeLisi, 1994; Gallagher, Morley, & Levin, 1999). Similarly, surface characteristics of items have been used to explain some gender DIF in mathematics test items (Li, Cohen, & Ibarra, 2004). Cognitive skills have been modeled to reflect basic features of items in the linear logistic test model (Fisher, 1983). Tatsuoka (1983) described a Q matrix to contain this information, the entries of which indicate whether a particular cognitive skill is required by attribute h in item i. Elements of the Q matrix, q_{ih}, are either 1, if attribute h is required by item i, or 0, if it is not. One way this matrix can be incorporated into the MMixIRTM is with the following linear structure:

$$\beta_{igk} = \sum_{h} b_{gkh} q_{ih}, \tag{20}$$

where b_{gkh} is the contribution to item difficulty required by attribute h for each class, g and k.

The substantive usefulness of the MMixIRTM is based on the assumption that the resulting latent classes represent discrete subpopulations and are not just statistical artifacts of nonnormality that may incidentally exist in the data (Bauer & Curran, 2003). At the present time, little evidence exists indicating what schools look like based on latent classes. Results of the simulation study suggest that the findings of this study were not artifacts in that the algorithm was able to correctly detect generated groups in the data. The resulting student- and school-level mixtures in the data examined in the empirical example, likewise, were clearly distinguishable in terms of ability levels, item difficulty profiles, student and school demographic characteristics, and response patterns. When several factors determine a class, however, finding those factors that cause the DIF potentially may be difficult.

In the MMixIRTM, as the number of school-level DIF items increases, the characteristics of school-level latent classes should also change. That is, the model rationale for constructing school-level latent classes is dependent on the

proportion of school-level DIF items. It may be that the number of non-DIF items that are needed to anchor the construct across school-level mixtures be considered a substantive issue as well as a statistical one. In the empirical illustration, we demonstrated that it was possible to obtain class-specific item difficulties for each *g* and *k* and to express these on the same scale. No constraints were set that required identifying non-DIF items across groups. The analysis needed to ensure the meaning of *g* was the same or similar across classes after class-specific item difficulties were obtained. In the example, item profiles were similar between the same student-level latent classes across school-level latent classes. It appears, in other words, that the construct characterized by class-specific item difficulties is similar across school-level classes.

The second type of label switching, that is, switching among different replications of the same simulation conditions, was observed in the simulation study. This type of label switching can be problematic (a) for checking convergence using more than two chains, (b) for comparing student-level latent group membership across school-level latent classes (because the student-level probability of mixtures is modeled for each school-level class), and (c) for comparing results with and without covariates. The WinBUGS code developed for this study was not designed to prevent the second type of label switching. Although this type of label switching can be easily detected and taken care of in a simulation study, it can be a serious problem with real data. The strategy used in this study to check for this type of label switching was to investigate item difficulty profiles and ability patterns for each school latent class to see whether the representation was similar across school-level mixtures and to cross-tabulate group memberships to find the dominant group.

Noninformative prior distributions for parameters were not considered for two reasons. First of all, improper priors cannot be implemented in WinBUGS, and second, attempts using diffuse priors resulted in a substantial number of traps. The focus here was with the analysis of large-scale achievement data having a hierarchical data structure, so that the effect of prior information was likely negligible anyway. In this regard, the result with the WinBUGS code has been shown to be comparable with the result with LatentGOLD syntax module based on marginal maximum likelihood estimation (Vermunt & Magidson, 2007).

von Davier and Yamamoto (2007) have noted that MCMC typically requires substantial computing time to obtain usable results. This is due, in part, to the use of multiple starting points necessary with empirical data to determine whether the algorithm has converged. For this reason, the amount of computing required can sometimes be very large. In the case of the example in this study, 90 hours were required for one condition in the simulation study and 121.5 hours were required for the empirical study on a 3.0-GHz computer with 1 GB of RAM to complete a single replication. Results such as this are not uncommon, and operational estimation of model parameters for long tests and very large samples using

MCMC is clearly going to require either substantial computing resources or development of speedier algorithms.

Appendix

WinBUGS Code Used for MMixIRTM

```
# 2 student-level class (G) with prior
# 2 school-level class (K)with prior
# J: the number of students
# I: the number of items
# T: the number of schools
# g: group (i.e. student) membership at the student-level
# gg (k): group (i.e., school) membership at the school-level
# a: the SD of ability
# b: item difficulty
# eta: ability
# mutg: the mean of ability
model
{
for (j in 1:J) {
    for (i in 1:I) {
    r[j,i] <- resp[j,i]
}}
# G=2
for (j in 1:J) {
    for (i in 1:I) {
    logit(p[j,i]) <- a[g[j], gg[group[j]]] *eta[j]
    - b[i,g[j],gg[group[j]]]
    r[j,i]~dbern(p[j,i])
}}
# Ability
for (j in 1:J) {
    eta[j]~dnorm(mutg[g[j],gg[group[j]]], 1)
}
mutg[1,1] <- 0
mutg[2,1] ~ dnorm(0,1)
mutg[1,2] ~ dnorm(0,1)
mutg[2,2] ~ dnorm(0,1)

# SD of Ability
for (g in 1:G2) {
    for (k in 1:K2){
    a[g, k] ~ dnorm(0,1) I(0,)
}}
# Student Level
for (j in 1:N) {
    g[j] ~ dcat(pi[gg[group[j]],1:G2])
}
for (k in 1:K2) {
   for (g in 1:G2) {
    pi[k,g] <- delta[k,g] /sum(delta[k,])
    delta[k,g] ~ dgamma(alpha[g],1)
   }
```

(*continued*)

Appendix (*continued*)

```
}
# School Level
for (t in 1:T){
    gg[t] ~ dcat(pi1[1:K2])
}
for (k in 1:K2) {
    pi1[k] <- delta1[k]/sum(delta1[1:K2])
    delta1[k] ~ dgamma(alpha1[k],1)
}
# Item Difficulty
for (i in 1:T) {
    for (g in 1:G2) {
    for (k in 1:K2){
    b[i,g,k]~dnorm(0,1)
}}}
# Log-Likelihood
for (j in 1:J) {
    for (i in 1:I) {
    l[j,i]<-log(p[j,i])*r[j,i]+log(1-p[j,i])*(1-r[j,i])
}}
loglik <-sum(l[1:J,1:I])
AIC <- -2*(loglik - np)
BIC <- -2*loglik + np*log(N)
}
# Initial Value of School-Level Group Membership
list(gg=c(1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
…
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2))

# Input Data
# 25 students
list(J=8000, I=40, T=320, G2=2, K2=2, np=169, alpha=c(1,1),
alpha1=c(1,1),
group=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,
3,3,3,3...),
resp=structure(.Data=c(
1.0,1.0,1.0,1.0,1.0,
…
1.0,0.0,0.0,0.0,0.0),.Dim = c(8000,40)))
```

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Asparouhov, T., & Muthén, B. (2008). Multilevel mixture model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 25–51). Greenwich, CT: Information Age Publishing, Inc.

Bates, D., & Debroy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, *91*, 1–17.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*, 338–363.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response for multiple-choice data. *Journal of Educational and Behavioral Statistics*, *26*, 381–409.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a Mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331–348.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, *97*, 65–108.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*, 133–148.

Cohen, A. S., Cho, S.-J., & Kim, S.-H. (2005, April). *A mixture testlet model for educational tests.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice*, *20*, 225–233.

Congdon, P. (2003). *Applied Bayesian modeling.* New York: Wiley.

Dayton, C. M., & Macready, G. B. (1988). Concomitant variable latent-class models. *Journal of the American Statistical Association*, *83*, 173–178.

DeAyala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, *2*, 243–276.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

Fisher, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, *48*, 3–26.

Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, *58*, 145–172.

Fox, J.-P., & Glas, C. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271–288.

Gallagher, A. M. (1998). Gender and antecedents of performance in mathematics testing. *Teachers College Record*, *100*, 297–314.

Gallagher, A. M., & DeLisi, R. (1994). Gender differences in scholastic aptitude test-mathematics problem solving among high ability students. *Journal of Educational Psychology*, *86*, 204–211.

Gallagher, A. M., Morley, M. E., & Levin, J. (1999). Cognitive patterns of gender differences on mathematics admissions tests. In *Graduate Record Examinations FAME Report* (pp. 4–11). Princeton, NJ: Education Testing Service.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (Vol. 4, pp. 169–194). UK: Oxford University Press.

Goldstein, H. (1987). *Multilevel models in education and social research.* London: Charles Griffin and Co.

Jordan, M. I., & Jacobs, R. A. (1992). Hierarchies of adaptive experts. In J. Moody, S. Hanson, & R. Lippmann (Eds.), *Advances in neural information processing systems* (Vol. 4, pp. 985–993). San Mateo, CA: Morgan Kaufmann.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79–93.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing*, *4*, 115–136.

Longford, N. T. (1993). *Random coefficient models.* New York: Oxford University Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, *26*, 307–330.

Maier, K. S. (2002). Modeling incomplete scaled questionnaire data with a partial credit hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, *27*, 271–289.

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

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195–215.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190.

Raftery, A. L., & Lewis, S. (1992). How many iterations in the Gibbs sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (Vol. 4, pp. 763–773). UK: Oxford University Press.

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

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B, 59*, 731–792. Correction (1998). *Journal of the Royal Statistical Society Series, Series B, 60*, 661.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Rost, J. (1997). Logistic mixture models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449–463). New York: Springer.

Samuelsen, K. M. (2005). *Examining differential item functioning from a latent class perspective.* Unpublished doctoral dissertation, University of Maryland, College Park.

Samuelsen, K. M. (2008). Examining differential item functioning from a latent class perspective. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 67–113). Charlotte, NC: Information Age Publishing.

Samuelsen, K. M., & Bradshaw, L. (2008). *The credibility interval method for the detection of DIF within a Bayesian framework.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York City.

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

Smith, B. (2004, April). Bayesian output analysis program (BOA) (Version 1.1.2 for R and S-PLUS) [Computer program]. Iowa City: University of Iowa, Department of Biostatistic.

Spiegelhalter, D., Thomas, A., & Best, N. (2003). WinBUGS (Version 1.4) [Computer program]. Cambridge UK: MRC Biostatistics Unit, Institute of Public Health.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, *62*, 795–809.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Vermunt, J. K. (2007a). Multilevel mixture item response theory models: An application in education testing. ISI 2007 Proceedings.

Vermunt, J. K. (2007b). A hierarchical model for clustering three-way data sets. *Computational Statistics & Data Analysis*, *51*, 5368–5376.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced.* Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2007). Latent GOLD 4.5 syntax module [Computer program]. Belmont, MA: Statistical Innovations Inc.

von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 99–115). New York: Springer.

Wang, W.-C. (2004). Effect of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*, *72*, 221–261.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement*, *40*, 307–330.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: Measurement, Evaluation, Statistics, and Assessment Press.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langenheine (Eds.),

Applications of latent trait and latent class models in the social sciences. New York, NY: Waxmann.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software.

Authors

SUN-JOO CHO is an assistant professor at Peabody College of Vanderbilt University, Hobbs 213A 230 Appleton Place, Nashville, TN 37203; e-mail: sj.cho@vanderbilt.edu. Her research interests are the modeling of item response data using item response theory and its parameter estimation.

ALLAN S. COHEN is a professor at University of Georgia College of Education, 570 Aderhold Hall, Athens, GA30602; email: acohen@uga.edu. His research interests are applied statistics and psychometric theory.

Manuscript received April 29, 2008
Revision revised May 16, 2009
Accepted August 18, 2009