Research Report

Comparing Multiple-Group Multinomial Log-Linear Models for Multidimensional Skill Distributions in the General Diagnostic Model

Xueli Xu

Matthias von Davier

June 2008

ETS RR-08-35

Comparing Multiple-Group Multinomial Log-Linear Models for Multidimensional Skill Distributions in the General Diagnostic Model

Xueli Xu and Matthias von Davier

ETS, Princeton, NJ

June 2008

Abstract

The general diagnostic model (GDM) utilizes located latent classes for modeling a multidimensional proficiency variable. In this paper, the GDM is extended by employing a log-linear model for multiple populations that assumes constraints on parameters across multiple groups. This constrained model is compared to log-linear models that assume separate sets of parameters to fit the distribution of latent variables in each group of a multiple-group model. Estimation of these constrained log-linear models using iterative weighted least squares (IWLS) methods is outlined and an application to NAEP data exemplifies the differences between constrained and unconstrained models in the presence of larger numbers of group-specific proficiency distributions.

The use of log-linear models for the latent skill space distributions using constraints across populations allows for efficient computations in models that include many proficiency distributions.

Key words: General diagnostic model, multiple-group models, log-linear models, IWLS

1. Introduction

Located latent class (LLC) models (McCutcheon, 1987) relate a set of observed discrete multivariate variables to a set of discrete latent variables. LLC models are alternatives to item response theory (IRT) models (Lord & Novick, 1968) in analyzing item responses. Though it is convenient and parsimonious to assume latent ability be a continuous random variable with a limited number of parameters as in IRT models, some research (Follmann, 1988; Haberman, 2005) has demonstrated that only a finite number of points along the hypothetical scale of the latent ability can be identified. LLC models assume discrete latent random variables. In LLC models, latent abilities are conceptualized as an ordered or unordered set of a finite number of fixed classes (Haberman, 1979; Heinen, 1996; Lazarsfeld & Henry, 1968). Latent classes are defined by the values of this random variable if there is only one dimension in the latent space. If the latent ability space is multidimensional, the latent classes are defined by the combinations of the values of this latent ability vector (Goodman, 1974; Haberman, 1979).

In the LLC modeling framework, the probability of obtaining score x for item j, conditional on a latent class c, is denoted by $p(Y_j = x|c)$. Here Y_j is the response variable for item j. Often the constraint $\sum_{k=1}^{C} P(c_k) = 1$ is imposed to make the parameters identifiable, where C is the total number of latent classes. A latent class c is usually a realization of a discrete latent vector. So the conditioning probability can also be written as $p(Y_j = x|c) = p(Y_j = x|\theta_1, \theta_2, \ldots, \theta_M)$, where M is the dimension of the latent vector, and $\theta_1, \ldots, \theta_M$ are M latent random variables. The latent class space can be expanded geometrically as the dimensionality of the vector and levels for each component in the vector increase. If a latent vector contains M discrete variables with K real values for each, there will be $K^M - 1$ latent classes in total. With the increases in K and M, the total number of latent classes increases so quickly that they cannot even be identified in most data sets. For instance, when $K = 4$ and $M = 5$, there will be 1,024 latent classes, or 1,023 independent parameters in the unconstrained case, which results in problems with identifiability in most data sets. To address this issue, Xu and von Davier (2008) applied a log-linear model for the latent class space (Nerlove & Press, 1973) to capture basic features of the latent class distribution without loosing model fit. Specifically, they modeled the latent class distribution $p(\theta_1, \theta_2, \ldots, \theta_M)$

as:

$$log[p(\theta_1, \theta_2, \ldots, \theta_M)] = \alpha + \sum_{m=1}^{M} \beta_{1(m)}\theta_m + \sum_{m=1}^{M} \beta_{2(m)}\theta_m^2 + \sum_{m=1}^{M} \beta_{3(m)}\theta_m^3 + \sum_{m_i=1}^{M-1} \sum_{m_j=m_i}^{M} \beta_{4(m_i m_j)}\theta_{m_i}\theta_{m_j}, \tag{1}$$

where m is the index for the latent random variable, M is the total number of dimensions for this latent random vector, and α is a normalizing constant. Note that the fourth component is included only when there are at least four levels for a latent random variable.

By implementing this log-linear model, Xu and von Davier (2008) successfully reduced the parameters in the latent class space, hence increasing estimation efficiency. However, (1) does not allow for differences between subgroups. To evaluate the differences between groups, Xu and von Davier (2006) used a multiple-group assumption to analyze data. Under this assumption, all subgroups are calibrated concurrently with the item parameters constrained to be the same across these subgroups. In the meantime, the latent class distributions of different subgroups are estimated separately. For a test with four latent variables, four levels in each of these variables, and four prespecified subgroups, there will be $4(groups) \times 18 = 72$ distributional parameters in the latent class space if (1) is used for each subgroup. Since α is a normalizing constant, it is completely determined by the other parameters in (1). Compared to (1), this multiple-group assumption indeed puts a subgroup indicator in every term of the model:

$$log[p_g(\theta_1, \theta_2, \ldots, \theta_M)] = \alpha_g + \sum_{m=1}^{M} \beta_{1g(m)}\theta_m + \sum_{m=1}^{M} \beta_{2g(m)}\theta_m^2 + \sum_{m=1}^{M} \beta_{3g(m)}\theta_m^3 + \sum_{m_i=1}^{M-1} \sum_{m_j=m_i}^{M} \beta_{4g(m_i m_j)}\theta_{m_i}\theta_{m_j}, \tag{2}$$

where g is an indicator for subgroup membership, and $g = 1, \ldots, G$. Immediately this raises a question: Has the difference between subgroups been overparameterized?

To address this, another model with fewer parameters is proposed:

$$log[p_g(\theta_1, \theta_2, \ldots, \theta_M)] = \alpha_g + \sum_{m=1}^{M} \beta_{1g(m)}\theta_m + \sum_{m=1}^{M} \beta_{2(m)}\theta_m^2 + \sum_{m=1}^{M} \beta_{3(m)}\theta_m^3 + \sum_{m_i=1}^{M-1} \sum_{m_j=m_i}^{M} \beta_{4(m_i m_j)}\theta_{m_i}\theta_{m_j}. \tag{3}$$

Group differences are present only in the first moment of the latent variables, while higher-order moments and linear-by-linear interactions are assumed to be the same across groups. For a test with four latent variables, four levels in each variable, and four prespecified subgroups, there will be $14 + 4 * 4(groups) = 30$ distributional parameters in the latent class space if (3) is used. This model reduces the required parameters to model the latent class space to a large extent, which enables one to carry out subgroup analysis even when the number of subgroups is large.

The primary goal of this paper is to conduct model comparisons with varying numbers of groups (here $G = 1, 2, 4, 8$) using the constrained model in (3) versus unconstrained multiple-group models, such as (2). A general diagnostic model (GDM) is used to connect the observed item responses to the latent classes, and (2) and (3) are used to capture the characteristics of the latent classes distributions. The remainder of this paper is organized as follows. Section 2 provides a brief introduction to the GDM, while section 3 details the estimation of (3) using iterative weighted least squares (IWLS) methods. Section 4 describes our real analysis plan and the data itself. Section 5 presents the results, and section 6 provides a discussion and summary.

2. An Extension of General Diagnostic Model

A compensatory GDM suitable for dichotomous and polytomous ordinal items (von Davier, 2005) is given by:

$$\log P(Y_j = x|\theta_1, \theta_2, \ldots, \theta_M) = a_j(b_{jx}, \gamma_{jm}) + b_{jx} + \sum_{m}^{M} x\gamma_{jm}\theta_m q_{jm}, \tag{4}$$

where Y_j is the response variable for item j, $a_j(\cdot)$ is a normalizing term, and b_{jx} is a location parameter for score x on item j. Furthermore, m is an indicator for the m-th latent variable, and M is the total number of dimensions of the latent vector. Also, γ_{jm} is the index for the slope parameter for item j associated with mth latent variable, and q_{jm} is the entry in a Q-matrix for item j and latent variable m. The Q-matrix specifies the correspondence between items and the latent variables. Specifically, $q_{jm} = 1$ if item j requires the m-th latent skill, otherwise $q_{jm} = 0$. Finally, K real values are assigned to each discrete latent variable θ_m to represent performance levels along each variable.

Using the model in (1), Xu and von Davier (2008) extended GDM model by structuring the latent class distribution. The observed log-likelihood of response vector $Y_1, Y_2, \ldots, Y_J$ is

$$logp(Y_1, Y_2, \ldots, Y_J) = log \sum_{\theta_1,\theta_2,\ldots,\theta_M} \prod_{j=1}^{J} p(Y_j|\theta_1, \theta_2, \ldots, \theta_M)p(\theta_1, \theta_2, \ldots, \theta_M), \tag{5}$$

where $p(Y_j|\theta_1, \theta_2, \ldots, \theta_M)$ comes from (4).

The observed log-likelihood of response vector $Y_1, Y_2, \ldots, Y_J$ combined with the constrained latent space model in (3) is

$$logp_g(Y_1, Y_2, \ldots, Y_J) = log \sum_{\theta_1,\theta_2,\ldots,\theta_M} \prod_{j=1}^{J} p(Y_j|\theta_1, \theta_2, \ldots, \theta_M) p_g(\theta_1, \theta_2, \ldots, \theta_M), \tag{6}$$

where $p(Y_j|\theta_1, \theta_2, \ldots, \theta_M)$ uses the form of (4), and $p_g(\theta_1, \theta_2, \ldots, \theta_M)$ utilizes the form of (3). There is a difficulty in solving this log-likelihood since a summation is included in the logarithm. Instead, a complete log-likelihood is derived that treats both item responses and latent variables as observable, and an expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is usually used to solve this type of optimization problem. The likelihood of the complete data may be written as

$$logp_g(Y_1, Y_2, \ldots, Y_J, \theta_1, \theta_2, \ldots, \theta_M) = \sum_{j=1}^{J} logp(Y_j|\theta_1, \theta_2, \ldots, \theta_M) + logp_g(\theta_1, \theta_2, \ldots, \theta_M). \tag{7}$$

Note that the first component in the right side contains only the parameters related to the GDM model, and the parameters of latent space structure are included only in the second component if the value of these latent skill variables is known. So in this complete log-likelihood, the values of $p_g(\theta_1, \theta_2, \ldots, \theta_M)$ need to be imputted, and then the maximum likelihood estimates (MLE) of the parameters can be obtained. Impute $p_g(\theta_1, \theta_2, \ldots, \theta_M)$ by employing the posterior distribution of $p_g(\theta_1, \theta_2, \ldots, \theta_M|Y_1, Y_2, \ldots, Y_J)$. The parameters related to GDM, such as (4), in the first component of (7) can be estimated by methods in von Davier (2005). The estimates of parameters related to (3) in the second component of (7) are outlined in next section.

3. Computational Formula for the Parameters in (3)

The model in (3) is also referred to as the product multinomial log-linear model (Lang, 1996) in the statistical literature. This paper uses iterative weighted least squares methods (IWLS) to derive the estimates and their estimation errors. In our experience, this method has proven stable and quick, and it is one common choice to estimate this type of model.

3.1 Estimation of the Constrained Multiple-Group Model

To enable estimation using IWLS methods, the design matrix for (3) has to be specified. The design matrix that enters the IWLS algorithm puts group dependent parameters in different columns and group independent parameters in the same column. This yields

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{D} \end{pmatrix}, \tag{8}$$

where

$$\boldsymbol{A} = \begin{pmatrix} \mathbf{0}_C & \mathbf{0}_C & \dots & \mathbf{0}_C & \boldsymbol{\theta_1} & \mathbf{0}_C & \dots & \mathbf{0}_C & \dots & \boldsymbol{\theta_M} & \mathbf{0}_C & \dots & \mathbf{0}_C \\ \mathbf{1}_C & \mathbf{0}_C & \dots & \mathbf{0}_C & \mathbf{0}_C & \boldsymbol{\theta_1} & \dots & \mathbf{0}_C & \dots & \mathbf{0}_C & \boldsymbol{\theta_M} & \dots & \mathbf{0}_C \\ \mathbf{0}_C & \mathbf{1}_C & \dots & \mathbf{0}_C & \mathbf{0}_C & \mathbf{0}_C & \dots & \mathbf{0}_C & \dots & \mathbf{0}_C & \mathbf{0}_C & \dots & \mathbf{0}_C \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_C & \mathbf{0}_C & \dots & \mathbf{1}_C & \mathbf{0}_C & \mathbf{0}_C & \dots & \boldsymbol{\theta_1} & \dots & \mathbf{0}_C & \mathbf{0}_C & \dots & \boldsymbol{\theta_M} \end{pmatrix},$$

where $C = K^M$ is the total number of latent classes, and M is the total number of latent variables. Each entry in this matrix represents a vector. For example, $\mathbf{0}_C = \{0, 0, \dots, 0\}^T$ with a total of C elements.

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\theta_1}^2 & \boldsymbol{\theta_2}^2 & \dots & \boldsymbol{\theta_M}^2 & \boldsymbol{\theta_1}^3 & \boldsymbol{\theta_2}^3 & \dots & \boldsymbol{\theta_M}^3 \\ \boldsymbol{\theta_1}^2 & \boldsymbol{\theta_2}^2 & \dots & \boldsymbol{\theta_M}^2 & \boldsymbol{\theta_1}^3 & \boldsymbol{\theta_2}^3 & \dots & \boldsymbol{\theta_M}^3 \\ \boldsymbol{\theta_1}^2 & \boldsymbol{\theta_2}^2 & \dots & \boldsymbol{\theta_M}^2 & \boldsymbol{\theta_1}^3 & \boldsymbol{\theta_2}^3 & \dots & \boldsymbol{\theta_M}^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\theta_1}^2 & \boldsymbol{\theta_2}^2 & \dots & \boldsymbol{\theta_M}^2 & \boldsymbol{\theta_1}^3 & \boldsymbol{\theta_2}^3 & \dots & \boldsymbol{\theta_M}^3 \end{pmatrix},$$

and

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{\theta_1\theta_2} & \boldsymbol{\theta_1\theta_3} & \dots\boldsymbol{\theta_{M-1}\theta_M} & \\ \boldsymbol{\theta_1\theta_2} & \boldsymbol{\theta_1\theta_3} & \dots\boldsymbol{\theta_{M-1}\theta_M} & \\ \boldsymbol{\theta_1\theta_2} & \boldsymbol{\theta_1\theta_3} & \dots\boldsymbol{\theta_{M-1}\theta_M} & \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\theta_1\theta_2} & \boldsymbol{\theta_1\theta_3} & \dots\boldsymbol{\theta_{M-1}\theta_M} & \end{pmatrix},$$

where $\boldsymbol{\theta_m} = (\theta_{m,1}, \theta_{m,2}, \dots, \theta_{m,C})^T$.

3.2 Parameter Vector and Response Vector for (3)

The parameter vector $\boldsymbol{\beta}$ to be estimated by IWLS for (3) consists of the following concatenation of components:

$$\boldsymbol{\beta} = (\alpha_g, \beta_{1g(m)}, \beta_{2(m)}, \beta_{3(m)}, \beta_{4(m_i m_j)})^T,$$

for $g = 1, \dots, G$, and the latent variable indices $m = 1, \dots, M$, $m_i = 1, \dots, M-1$ as well as $m_j = 2, \dots, M$.

The response vector is

$$\boldsymbol{Y} = (log[p_{g=1}(\theta_1, \ldots, \theta_M)]^T, \ldots, log[p_{g=G}(\theta_1, \ldots, \theta_M)]^T)^T,$$

which are imputed by the posterior distribution.

3.3 Parameter Estimates for (3)

With design-matrix $\boldsymbol{Z}$ and parameter vector $\boldsymbol{\beta}$ set up for (3), the estimation equation can be written in matrix form:

$$\boldsymbol{Y} = \boldsymbol{Z\beta}.$$

The steps to be cycled through in the IWLS method include:

1. $\boldsymbol{\mu^{(0)}} = \boldsymbol{n} = (\boldsymbol{n_1'}, \ldots, \boldsymbol{n_g'}, \ldots, \boldsymbol{n_G'})^T$, where $\boldsymbol{n_g} = (n_{g,1}, \ldots, n_{g,C})^T$ is a vector of elements in the latent classes for subgroup g. Here $n_{g,C}$ stands for the number of students in group g who belong to attribute pattern c (composed by $\theta_1, \ldots, \theta_M$). Once the probability of each attribute pattern is known, this number can be easily calculated.
2. $\boldsymbol{\mu^{(t)}} = Z\boldsymbol{\beta^{(t)}}$
3. $X^{(t)} = log(\boldsymbol{\mu^{(t)}}) + (\boldsymbol{n} - \boldsymbol{\mu^{(t)}})/\boldsymbol{\mu^{(t)}}$
4. $\boldsymbol{v^{(t)}} = diag(\boldsymbol{\mu^{(t)}})$
5. $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z^T v^{(t)} Z})^{-1} \boldsymbol{Z^T v^{(t)} X^{(t)}}$

Iteration t will continue until a prespecified convergence criterion is met. Then the covariance of the estimates $\hat{\boldsymbol{\beta}}$ is given by:

$$cov(\hat{\boldsymbol{\beta}}) = (\boldsymbol{Z^T v Z})^{-1}.$$

4. Analysis Plan and Data

The models in (2) and (3) with $G = 1, 2, 4, 8$ will be used to analyze two reading data sets from the National Assessment of Educational Progress (NAEP). When $G = 1$, this model is called a single-group model. In this model, no subgroup differences are accounted for. When $G = 2$, the two groups are identified by the gender variable. For the case of $G = 4$, the race or ethnicity variable is used as the indicator of the subgroups. Specifically, they are White, Black, Hispanic, and Asian-American groups. When $G = 8$, a complete factorial design of race and gender is defined and used as the indicator variable for the subgroups. These four models are compared using the following: model fit, item parameter estimates, as well the marginal distribution of each latent random variable.

The two data sets are from the Grade 4 reading assessment in 2003 and the Grade 8 reading assessment in 2005, respectively. Each data set contains a representative sample of the student population in the target grade. The Grade 4 data contains the responses of 191,300 students to 102 items, while the Grade 8 data consists of the responses of 159,500 students to 142 items. The item sets for both grades include both multiple-choice items and constructed-response items. A partially balanced incomplete block (pBIB; Allen, Donoghue, & Schoeps, 1998) was utilized in these two assessments. This means that each student took only approximately one sixth of the entire test, and different students may have taken different subsets of the assessment.

The Grade 4 assessment was designed to measure two content areas, reading for literary experience and reading to get information. The Grade 8 assessment was designed to measure three content areas, reading for literary experience, to gain information, and to perform a task. A Q-matrix is an integral part of the GDM. In these two assessments, specifically, the content domains defined in the NAEP reading framework serve as the skill dimensions in the Q-matrix, and the correspondence between the content area and items serves as the Q-matrix. Since one item measures only one skill in the Q-matrix, this Q-matrix is also referred to as a simple-structure Q-matrix. Though the NAEP assessments were not originally developed for the purpose of skills diagnosis, the application of the GDM enables a multidimensional analysis.

5. Results

Tables 1 and 2 give an overview of the results in terms of indices of model-data fit. The Akaike information criterion (AIC; Akaike, 1974) is one of the indices used to assess model fit. A model is said to be better when it results in a smaller fit indices and a larger log-likelihood. Compared to the single-group analysis (when $G = 1$), all multiple-group models improve the log-likelihood as well as the AIC index. Among the models examined in this study, the subgroup analysis being defined by the complete factorial of race and gender has the smallest AIC and the largest log-likelihood. Conditional on eight-group analysis, the unrestricted group analysis using

(2) gives a better fit and a larger likelihood than the restricted model in (3). The second best fit is provided by the subgroup analysis using four racial groups.

Table 1.
Model Fit for 2003 Grade 4 Reading Data

	(3)	# of parameters	Log-likelihood	AIC
G=1	Single group	230	-2,123,709.38	4,247,879
G=2	Gender	244	-2,122,960.92	4,246,410
G=4	Race	250	-2,112,678.39	4,225,857
G=8	Race * Gender	262	-2,111,761.79	4,224,048
	(2)	# of parameters	Log-likelihood	AIC
G=2	Gender	247	-2,122,875.37	4,246,245
G=4	Race	261	-2,112,606.41	4,225,735
G=8	Race * Gender	289	-2,111,585.27	4,223,749

Note. AIC = Akaike information criterion.

Table 2.
Model Fit for 2005 Grade 8 Reading Data

	(3)	# of parameters	Log-likelihood	AIC
G=1	Single group	344	-1,866,356.62	3,733,401
G=2	Gender	348	-1,865,083.91	3,730,864
G=4	Race	356	-1,858,987.40	3,718,687
G=8	Race * Gender	372	-1,857,542.85	3,715,830
	(2)	# of parameters	Log-likelihood	AIC
G=2	Gender	353	-1,865,062.97	3,730,832
G=4	Race	375	-1,858,767.67	3,718,285
G=8	Race * Gender	419	-1,857,135.02	3,715,108

Note. AIC = Akaike information criterion.

Next, the item parameter estimates obtained from these models are compared. There are 233 item parameters in the Grade 4 data, and 331 item parameters for the Grade 8 data. Due to limitations of space, this paper lists only the 10 items with the largest differences between the models. A list of these items and their parameter estimates are given in Tables 3 and 4. An inspection of the results shows that these item parameters estimates are similar across different models. The largest difference is equal to 0.1, which should be evaluated in comparison to the range of parameters within models, which is about (-4,4).

The marginal distributions of the latent classes from the eight-group analysis are shown in

Table 3.
Ten Items With the Largest Difference in Parameter Estimates for 2003 Grade 4 Data

	Estimates			
Item ID	$G = 1$	$G = 2$	$G = 3$	$G = 4$
190	1.3924	1.3954	1.3732	1.3720
26	1.6230	1.6362	1.6120	1.6249
105	1.1713	1.1719	1.1953	1.1955
192	1.0023	1.0048	0.9805	0.9802
63	-3.7367	-3.7529	-3.7273	-3.7277
100	2.8840	2.8912	2.8650	2.8673
45	-4.8598	-4.8697	-4.8846	-4.8916
107	2.0844	2.0768	2.1119	2.1042
99	1.9779	1.9876	1.9481	1.9570
188	1.6616	1.6641	1.6222	1.6236

Table 4.
Ten Items With the Largest Difference in Parameter Estimates for 2005 Grade 8 Data

	Estimates			
Item ID	$G = 1$	$G = 2$	$G = 3$	$G = 4$
202	1.1173	1.1601	1.1286	1.1542
189	-3.5310	-3.5530	-3.5100	-3.5310
195	0.8017	0.8194	0.8411	0.8450
270	1.1000	1.1075	1.1434	1.1417
260	-1.0940	-1.0794	-1.0500	-1.0473
197	-1.0830	-1.0937	-1.1260	-1.1301
276	1.0508	1.0529	1.0989	1.0898
20	-3.6116	-3.5983	-3.6303	-3.5810
205	-3.8372	-3.9006	-3.8489	-3.8922
315	1.6713	1.6706	1.7414	1.7234

Tables 5 to 6. Among the 48 probabilities in the 2003 data for Grade 4 , the largest difference between the unrestricted and restricted models is 0.018 (i.e., 1.18 %), and 5 of the probabilities are larger than 0.01. The differences for the rest are less than 0.01. For Grade 8 in 2005, the largest discrepancy between these two models is less than 0.03 (i.e., 3%), and the differences in most probabilities are less than 0.01 (i.e., 1%). In short, the results show that the marginal distribution resulting from these two models are similar.

Table 5.
The Marginal Distribution From Eight-Group Analysis for 2003 Grade 4 Data

		$G = 8$ of (3)				$G = 8$ of (2)			
Groups	Skills	Levels				Levels			
		-1.73205	-0.57735	0.57735	1.73205	-1.73205	-0.57735	0.57735	1.73205
White	Skill 0	0.041	0.176	0.440	0.345	0.042	0.174	0.441	0.344
	Skill 1	0.047	0.239	0.377	0.337	0.049	0.239	0.378	0.334
Black	Skill 0	0.178	0.364	0.362	0.098	0.179	0.368	0.357	0.096
	Skill 1	0.218	0.439	0.260	0.085	0.207	0.454	0.258	0.082
Hispanic	Skill 0	0.161	0.340	0.384	0.116	0.165	0.334	0.388	0.113
	Skill 1	0.204	0.417	0.280	0.099	0.211	0.408	0.290	0.092
Asian-	Skill 0	0.052	0.197	0.425	0.327	0.056	0.197	0.407	0.341
American	Skill 1	0.063	0.263	0.360	0.315	0.067	0.268	0.345	0.321
Male	Skill 0	0.102	0.262	0.415	0.221	0.103	0.261	0.424	0.213
	Skill 1	0.115	0.314	0.340	0.232	0.116	0.313	0.344	0.228
Female	Skill 0	0.066	0.209	0.418	0.306	0.069	0.208	0.410	0.313
	Skill 1	0.091	0.293	0.340	0.275	0.091	0.296	0.340	0.273

Table 6.
The Marginal Distribution From Eight-Group Analysis for 2005 Grade 8 Data

		$G = 8$ of (3)				$G = 8$ of (2)			
Groups	Skills	Levels				Levels			
		-1.73205	-0.57735	0.57735	1.73205	-1.73205	-0.57735	0.57735	1.73205
White	Skill 0	0.055	0.202	0.376	0.367	0.063	0.195	0.378	0.363
	Skill 1	0.015	0.115	0.390	0.481	0.017	0.117	0.389	0.477
	Skill 2	0.006	0.088	0.393	0.514	0.008	0.090	0.387	0.514
Black	Skill 0	0.210	0.378	0.301	0.111	0.197	0.386	0.317	0.101
	Skill 1	0.055	0.296	0.460	0.189	0.056	0.301	0.479	0.163
	Skill 2	0.029	0.269	0.501	0.201	0.031	0.267	0.518	0.184
Hispanic	Skill 0	0.178	0.344	0.332	0.145	0.177	0.335	0.352	0.133
	Skill 1	0.054	0.276	0.455	0.214	0.052	0.280	0.480	0.185
	Skill 2	0.035	0.267	0.488	0.209	0.039	0.275	0.471	0.213
Asian-	Skill 0	0.059	0.200	0.360	0.379	0.065	0.209	0.342	0.383
American	Skill 1	0.018	0.121	0.371	0.488	0.023	0.125	0.347	0.504
	Skill 2	0.007	0.094	0.378	0.518	0.012	0.106	0.352	0.529
Male	Skill 0	0.126	0.280	0.352	0.240	0.137	0.263	0.368	0.231
	Skill 1	0.036	0.198	0.427	0.338	0.039	0.197	0.435	0.327
	Skill 2	0.020	0.179	0.451	0.348	0.024	0.182	0.447	0.345
Female	Skill 0	0.074	0.227	0.359	0.340	0.069	0.235	0.357	0.337
	Skill 1	0.020	0.142	0.395	0.443	0.019	0.150	0.398	0.432
	Skill 2	0.008	0.113	0.399	0.479	0.010	0.117	0.393	0.479

6. Discussion

In this study, the unrestricted multiple-group model in (2) and restricted multiple-group model in (3) with different group skill distribution constraints were compared in terms of model fit, item parameter estimates, as well as the marginal distributions for these groups. Four different group definitions were used in this analysis. These are a single-group analysis, a two-group analysis defined by gender, a four-group analysis defined by race, and an eight-group analysis defined by a factorial design of gender and race.

For overall model fit indices such as the log-likelihood and AIC, the eight-group analysis is the best compared to other group analyses. For the eight-group model, the unrestricted model in (2) has better fit than the restricted model in (3). The results are within expectation since (2) relaxes the homogeneity constraints on group effects of higher moments in the log-linear model of the class distribution. This may imply that the parameter reduction in (3) has overconstrained the group effects. However, the results in this paper do not mean that (3) is inadmissible. This model might outperform (2) when there are large numbers of subgroups to take into account. For example, in the current NAEP analysis procedure, hundreds of background variables are included in the latent regression model (Mislevy, 1991). If one wants to mimic the procedure in the GDM framework, the parameter set of (2) will be much larger than that of (3) so that the parameters might not be identified. Hence, (3) is a substitute for (2) when the latter is not permissible in analysis.

Although the model fit of the restricted and unrestricted models differ substantially, a comparison of the marginal distributions of the latent classes under different models show that the differences are small in this regard. For the two data sets analyzed in this paper, the maximum differences are 1.8% and 2.7%, respectively. A comparison of item parameter estimates across these models provides information on the potential of introducing undesirable effects on the item parameters by using a constrained model for the proficiency variables. Effects that are evident can be interpreted as differential item functioning (DIF), that is, distortions of item parameters due to constrained shapes of conditional proficiency distributions. In particular, if significant differences are found in these estimates between using a single-group assumption and using a multiple-group assumption, one may argue that DIF could be introduced via use of a constrained multiple-group model. However, the analysis in this paper has shown that there is little effect as evidenced in Tables 3 and 4, and we are confident that no DIF is introduced by using these models to fit the data.

Finally, the selected variables race and gender are just two out of several hundred of available background variables in NAEP. Since most variables might not have a major effect on the conditional proficiency distributions, it is neither practical nor useful to include every background variable in the latent space model. In addition, while (3) is instrumental in reducing the number of parameters in the latent space without sacrificing accuracy, the number of model parameters will increase too much if many more variables are used in the definition of subgroups. The challenge lies in how to choose group predictors and where to limit the addition of variables to the latent space model.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Allen, N., Donohue, J., & Schoeps, T. (1998). *The NAEP 1998 technical report* (NCES 2001-509). Washington, DC: National Center for Education Statistics.

Dempster, A., Laird, N., & Rubin, R. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B. 39*(1), 1-38.

Follmann, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika, 53,* 553–562.

Goodman L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215–231.

Heinen, T. (1996). *Discrete latent variables models: Similarities and differences.* Thousand Oaks, CA: Sage.

Haberman, S. J. (1979). *Analysis of qualitative data: Vol. 2. New developments.* New York: Academic Press.

Haberman, S. (2005). *Latent class item response models*(ETS Research Rep. No. RR-05-28). Princeton, NJ: ETS.

Lang, J. (1996). On the comparison of multinomial and Poisson loglinear models. *Journal of Royal Statistical Society Series B, 58*, 253–266.

Lang, J. (2004). Multinomial-Poisson homogeneous models for contingency tables. *Annals of Statistics, 32*, 340–383 .

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McCutcheon, A. L. (1987). *Latent class analysis.* Thousand Oaks, CA: Sage.

Mislevy, R. (1991). Randomization-based inference about latent varaibles from complex samples. *Psychometrika, 56,* 177–196.

Nerlove, M., & Press, S. J. (1973). *Univariate and multivariate log-linear and logistic models* (Technical Rep. No. R-1306-EDA/NIH). Santa Monica, CA: Rand Corporation.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS

Research Rep. No. RR-05-16). Princeton, NJ: ETS.

Xu, X., & von Davier, M. (2006). *Applying the general diagnostic model to data from large scale educational surveys* (ETS Research Rep. No. RR-06-08). Princeton, NJ: ETS.

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model for NAEP data* (ETS Research Rep. No. RR-08-27). Princeton, NJ: ETS.