Research Report

# Hierarchical General Diagnostic Models

Matthias von Davier

Research & Development

May 2007
RR-07-19

May 2007

# Hierarchical General Diagnostic Models

Matthias von Davier
ETS, Princeton, NJ

May 2007

**Abstract**

This paper introduces multilevel extensions for the general diagnostic model (GDM) following recent developments on extensions of latent class analysis (LCA) to hierarchical models. The GDM is based on LCA as well as discrete latent trait models and may be viewed as a general modeling framework for confirmatory multidimensional item response models.

The multilevel extensions presented in this paper enable one to check the impact of clustered data, such as data for students within schools in large scale educational surveys, on the structural parameter estimates of the GDM. Moreover, the multilevel version of the GDM allows study of differences in skill distributions across these clusters.

Key words: Latent class analysis, multilevel extensions, item response models, diagnostic models, logistic models

**Acknowledgments**

This paper is the result of a study of literature on hierarchical latent class analysis and discussions of the issue with Shelby Haberman, Henry Braun, and Kentaro Yamamoto. I would like to express my thanks to my colleagues for their insights and their willingness to share ideas and discuss my initial attempts to develop this model extension. The opinions expressed are those of the author and not necessarily those of ETS.

## 1 Introduction

This paper introduces a hierarchical extension of the general diagnostic model (GDM; von Davier, in press-a) similar to extensions for latent class analysis (LCA; Lazarsfeld & Henry, 1968) to multilevel latent class models (Vermunt, 2003). Hierarchical extensions have also been developed for linear models (e.g., Bryk & Raudenbush, 1992; Goldstein, 1987) as well as for Rasch-type models (e.g., Kamata & Cheong, 2006) and more general IRT models (e.g., Fox & Glas, 2001).

The GDM is based on LCA as well as discrete latent trait models (Heinen, 1996) and may be viewed as a general modeling framework for confirmatory multidimensional item response models (see von Davier, in press-b; von Davier & Rost, 2006; von Davier & Yamamoto, 2007).

The multilevel extensions presented in this paper enable one to check the impact of the clustering of observed data, such as data for students within schools in large scale educational surveys, on the structural parameter estimates of the GDM. Moreover, the multilevel version of the GDM allows the study of differences in skill distributions across these clusters.

## 2 The General Diagnostic Model

Assume an $I$-dimensional categorical random variable $\vec{x} = (x_1, \ldots, x_I)$ with $x_i \in \{0, \ldots, m_i\}$for $i \in \{1, \ldots, I\}$, referred to as a response vector in the following. Further assume that there are $N$ independent and identically distributed (i.i.d.) realizations $\vec{x}_1, \ldots, \vec{x}_N$ of this random variable $\vec{x}$, so that $x_{ni}$ denotes the $i$-th component of the $n$-th realization $\vec{x}_n$. In addition, assume that there are $N$ unobserved realizations of a $K$-dimensional categorical variable, $\vec{a} = (a_1, \ldots, a_K)$, so that the vector

$$(\vec{x}_n, \vec{a}_n) = (x_{n1}, \ldots, x_{nI}, a_{n1}, \ldots, a_{nK})$$

exists for all $n \in \{1, \ldots, N\}$. The data structure

$$(X, A) = ((\vec{x}_n, \vec{a}_n))_{n=1,\ldots,N}$$

is referred to as the complete data, and $(\vec{x}_n)_{n=1,\ldots,N}$ is referred to as the observed data matrix. Denote $(\vec{a}_n)_{n=1,\ldots,N}$ as the latent skill or attribute patterns, which is the unobserved target of inference.

Let $P(\vec{a}) = P\left(\vec{A} = (a_1, \ldots, a_K)\right) > 0$ for all $\vec{a}$ denote the nonvanishing discrete count density of $\vec{a}$. Assume that the conditional discrete count density $P(x_1, \ldots, x_I \mid \vec{a})$ exists for all $\vec{a}$. Then the probability of a response vector $\vec{x}$ can be written as

$$P(\vec{x}) = \sum_{\vec{a}} P(\vec{a}) P(x_1, \ldots, x_I \mid \vec{a}).$$

### *2.1 Conditional Independence*

So far, no assumptions have been made about the specific form of the conditional distribution of $\vec{x}$ given $\vec{a}$, other than that $P(x_1, \ldots, x_I \mid \vec{a})$ exists. For the general diagnostic model, local independence (LI) of the components $x_i$ given $\vec{a}$ is assumed, which yields

$$P(x_1, \ldots, x_I \mid \vec{a}) = \prod_{i=1}^{I} p_i(x = x_i \mid \vec{a})$$

so that the probability $p_i(x = x_i \mid \vec{a})$ is the one component left to be specified to arrive at a model for $P(\vec{x})$.

### *2.2 Logistic Model Specification*

Logistic models have widespread applications and apart from early disputes about the merits of probit versus logit models (Berkson as cited in Cramer, 2003) have secured a prominent position among models for categorical data. The general diagnostic model is also specified as model with a logistic link between an argument, which depends on the random variables involved and some real valued parameters, and the probability of the observed response.

Using the above definitions, the GDM is defined as follows. Let

$$Q = (q_{ik}), \; i = 1, \ldots, I, \; k = 1, \ldots, K$$

be a binary $I \times K$ matrix, that is $q_{ik} \in \{0, 1\}$. Let

$$(\gamma_{ikx}), \; i = 1, \ldots, I, \; k = 1, \ldots, K, \; x = 1, \ldots, m_i$$

be a cube of real valued parameters, and let $\beta_{ix}$ for $i = 1, \ldots, I$ and $x \in \{0, \ldots, m_i\}$ be real valued parameters. Then define

$$p_i(x \mid \vec{a}) = \frac{\exp\left(\beta_{ix} + \sum_k \gamma_{ikx} h(q_{ik}, a_k)\right)}{1 + \sum_{y=1}^{m_i} \exp\left(\beta_{iy} + \sum_k \gamma_{iky} h(q_{ik}, a_k)\right)}.$$

It is often convenient to constrain the $\gamma_{ikx}$ somewhat and to specify the real valued function $h(q_{ik}, a_k)$ and the $a_k$ in a way that allows emulation of models frequently used in educational measurement and psychometrics. It is convenient to choose $h(q_{ik}, a_k) = q_{ik}a_k$, and $\gamma_{ikx} = x\gamma_{ik}$, which defines the general diagnostic model for partial credit data (Muraki, 1992).

Von Davier (2005) has shown that this model already contains several models from the areas of item response theory (IRT; Lord & Novick, 1968), latent-class analysis (Lazarsfeld & Henry, 1968), multiple classification latent-class models (Goodman, 1974; Haberman, 1979; Maris, 1999) and diagnostic models (see, for example, von Davier, DiBello, & Yamamoto, 2006).

## 3 Mixture General Diagnostic Models

Von Davier (in press-b) introduced the discrete mixture distribution version of the GDM, referred to as the MGDM. In discrete mixture models for item response data (Mislevy & Verhelst, 1990; Rost, 1990; for an overview, see von Davier & Rost, 2006), the probability of an observation $\vec{x}$ depends on the unobserved latent trait in the case of the GDMs, $\vec{a}$, and on a subpopulation indicator $g$, which is also unobserved. The rationale for mixture distribution models is that observations from different subpopulations may either differ in their distribution of skills or in their approach to the items (e.g., in terms of strategies employed) or both. A discrete mixture distribution in the setup of random variables as introduced above includes an unobserved grouping indicator $g_n$ for $n = 1, \ldots, N$. The complete data for examinee $n$ then becomes $(\vec{x_n}, \vec{a}_n, g_n)$, of which only $\vec{x}_n$ is observed in mixture distribution models. In multiple group models, $(\vec{x}_n, g_n)$ is observed.

The conditional independence assumption has to be modified to account for differences between groups, that is

$$P(\vec{x} \mid \vec{a}, g) = P(x_1, \ldots, x_I \mid \vec{a}, g) = \prod_{i=1}^{I} p_i(x = x_i \mid \vec{a}, g).$$

Moreover, assume that the conditional probability of the components $x_i$ of $\vec{x}$ depends on nothing but $\vec{a}$ and $g$, that is,

$$P(\vec{x} \mid \vec{a}, g, z) = \prod_{i=1}^{I} p_i(x = x_i \mid \vec{a}, g) = P(\vec{x} \mid \vec{a}, g) \tag{1}$$

for any random variable $z$. In mixture models, when the $g_n$ are not observed, the marginal probability of a response vector $\vec{x}$ needs to be found, that is,

$$P(\vec{x}) = \sum_{g} \pi_g P(\vec{x} \mid g), \tag{2}$$

where $P(\vec{x} \mid g) = \sum_{\vec{a}} p(\vec{a} \mid g) P(\vec{x} \mid \vec{a}, g)$. The $\pi_g = P(G = g)$ are referred to as mixing proportions, or class sizes. The class-specific probability of a response vector $\vec{x}$ given skill pattern $\vec{a}$ in class $g$ is then

$$P(\vec{x} \mid \vec{a}, g) = \prod_{i=1}^{I} P(x_i \mid \vec{a}, g) = \prod_{i=1}^{I} \left[ \frac{\exp\left(\beta_{ixg} + \sum_k x_i \gamma_{ikg} q_{ik} a_k\right)}{1 + \sum_y \exp\left(\beta_{iyg} + \sum_k y \gamma_{ikg} q_{ik} a_k\right)} \right]. \quad (3)$$

with class-specific item difficulties $\beta_{ixg}$. The $\gamma_{ikg}$ are the slope parameters relating skill $k$ to item $i$ in class $g$.

Note that mixture models and multiple group models are two extremes, for mixtures models no $g_n$ is observed, while for multiple group models all $g_n$ are observed. Von Davier and Yamamoto (2004) pointed this out and described an extension of the GPCM for mixture versions, multiple group versions, and partially observed grouping versions, where the $g_n$ information is missing only for a portion of the sample.

One important special case of the MGDM is a model that assumes measurement invariance across populations, which is expressed in the equality of $p(\vec{x} \mid \vec{a}, g)$ across groups, or, more formally:

$$P(x_i \mid \vec{a}, g) = p(x_i \mid \vec{a}, c) \text{ for all } i \in \{1, \ldots, I\} \text{ and all } g, c \in \{1, \ldots, G\}.$$

This assumption allows one to write the model equation without the group index $g$ in the conditional response probabilities, so that

$$P(\vec{x}) = \sum_g \pi_g P(\vec{x} \mid g) = \sum_g \pi_g \sum_{\vec{a}} p(\vec{a} \mid g) \prod_{i=1}^{I} P(x_i \mid \vec{a}). \quad (4)$$

Note that the differences between groups are only present in the $p(\vec{a} \mid g)$, so that the skill distribution is the only component with a condition on $g$ in the above equation. The next section introduces hierarchical GDM based on mixture distribution versions of the GDM.

## 4 Hierarchical General Diagnostic Models

Hierarchical models introduce an additional structure, often referred to as a cluster variable, in the modeling of observed variables to account for correlations in the data. These are attributed to the complex structure of the environment in which the data are observed. More concretely, one standard example for clustered data is the responses to educational assessments sampled from students within schools or classrooms. As a rather sloppy explanation, it seems plausible to assume that students within schools are more similar than students across schools (even though the amount to which this statement is true may depend on the educational system). Hierarchical models have been developed for linear models (e.g., Bryk & Raudenbush, 1992; Goldstein, 1987) as well as for Rasch-type models (e.g., Kamata & Cheong, 2006).

For the developments presented here, the extension of the LCA to a hierarchical model (e.g., Vermunt, 2003, 2004) is of importance. In addition to the latent class or grouping variable $g$, the hierarchical extension of the LCA assumes that each observation $n$ is characterized by an outcome $s_n$ on a clustering variable $s$. The clusters identified by this outcome may be schools, classrooms, or other sampling units representing the hierarchical structure of the data collection. As Vermunt outlined, the (unobserved) group membership $g_n$ is thought of as an individual classification variable; for two examinees $n \neq m$ there may be two different group memberships, that is, both $g_n = g_m$ and $g_n \neq g_m$ are permissible even if they belong to the same cluster (i.e., $s_n = s_m$).

Moreover, it is assumed that the skill distribution depends only on the group indicator $g$ and no other variable, that is,

$$P(\vec{a} \mid g, z) = P(\vec{a} \mid g) \tag{5}$$

for any random variable $z$. More specifically, for the clustering variable $s$,

$$P(g) = \sum_{s=1}^{S} p(s) P(g \mid s).$$

With Equation 5,

$$P(\vec{a} \mid s) = \sum_{g} P(g \mid s) P(\vec{a} \mid g),$$

for

$$P(g \mid s) P(\vec{a} \mid g) = p(g \mid s) P(\vec{a} \mid g, s) = P(\vec{a}, g \mid s).$$

As above for the MGDM, assume that the observed responses $\vec{x}$ depend on the skill pattern $\vec{a}$ and the group index $g$ only. Then

$$P(\vec{x} \mid g, s) = \sum_{\vec{a}} p(\vec{a} \mid g, s) P(\vec{x} \mid \vec{a}, g, s) = \sum_{\vec{a}} p(\vec{a} \mid g) P(\vec{x} \mid \vec{a}, g) = P(\vec{x} \mid g)$$

with Equations 1 and 5. Then the marginal distribution of a response pattern $\vec{x}$ in the hierarchical GDM (HGDM) is given by

$$P(\vec{x}) = \sum_{s} p(s) \sum_{g} P(g \mid s) \sum_{\vec{a}} P(\vec{a} \mid g) P(\vec{x} \mid \vec{a}, g), \tag{6}$$

where, as before in the MGDM, the $p(\vec{a} \mid g)$ denote the distribution of the skill patterns in group $g$, and the $p(\vec{x} \mid \vec{a}, g)$ denote the distribution of the response vector $\vec{x}$ conditional on skill pattern $\vec{a}$

and group $g$. A hierarchical GDM that assumes measurement invariance across clusters and across groups is defined by

$$P(\vec{x}) = \sum_{s} p(s) \sum_{g} P(g \mid s) \sum_{\vec{a}} P(\vec{a} \mid g) P(\vec{x} \mid \vec{a}), \tag{7}$$

with conditional response probabilities $p(\vec{x} \mid \vec{a}) = \prod_i p(x_i \mid \vec{a})$ that do not depend on cluster or group variables.

The increase in complexity of hierarchical GDMs over nonhierarchical versions lies in the fact that the group distribution $P(g \mid s)$ depends on the cluster variable $s$. This increases the number of group or class size parameters depending on the number of clusters $\#\{s : s \in S\}$. The estimation of item parameter $\beta_{ix(g)}$and $\gamma_{ik(g)}$ as well as the additional conditional probabilities of group sizes given clusters $p(g \mid s)$ and skill patterns given group $P(\vec{a} \mid g)$ is outlined in the next section.

## 5 Estimation of Hierarchical General Diagnostic Models

The case of fitting models with cluster-dependent response probabilities $P(\vec{x} \mid \vec{a}, s)$ will not be discussed here. The reason is that a model in which both the skill distributions and the probability of correct responses depend on the cluster variable does not allow attribution of the variation of observed responses across clusters to differences in skill distributions. Such a model would essentially assume that items have different difficulty in different clusters. Even though this is a very empathic view of the world, this does not allow drawing any conclusions involving cluster differences other than clusters are different. Apart from that, the fact that most applications of hierarchical models offer only moderate sample sizes within clusters makes the estimation of a multitude of cluster-specific parameters infeasible.

The estimation of GDMs and MGDMs has been outlined in von Davier (in press-a, in press-b). This approach is extended here to the estimation of HGDMs. The expectation-maximization (EM) algorithm has been shown to be a suitable one for this kind of estimation problems (Vermunt, 2003), so that other, more computationally costly methods are not necessary. For the most part, researchers will be concerned with fitting less highly parameterized versions of the HGDM, such as the models given in Equations 6 and 7.

*mdltm* software (von Davier, 2005) enables one to estimate MGDMs and GDMs with HGDMs according to Equations 6 and 7. The extensions to enable estimation of these models were recently implemented in *mdltm* based on the research presented in this paper.

Since the data are structured hierarchically, the first step is to define the complete data for the case of the HGDM. Let $S$ denote the number of clusters in the sample, and let $N_s$ denote the number of examinees in cluster $s$, for $s = 1, \ldots, S$. Then

- let $x_{ins}$ denote the $i$-th response of the $n$-th examinee in cluster $s$ and let $\vec{x}_{ns}$ denote the complete observed response vector of examinee $n$ in cluster $s$
- let $a_{kns}$ denote the $k$-th skill of examinee $n$ in cluster $s$ and let $\vec{a}_{ns}$ denote the skill pattern of examinee $n$ in cluster $s$
- let $g_{ns}$ denote the group membership of examinee $n$ in cluster $s$

Note that only the $x_{ins}$ are observed, as are the cluster sizes $N_s$ and the number of clusters $S$. The $a_{kns}$ and $g_{ns}$ are unobserved and have to be inferred by making model assumptions and calculating posterior probabilities such as $P(g \mid s)$ and $P(\vec{a}, g \mid \vec{x}, s)$.

## *5.1 Marginal Calculations in Hierarchical General Diagnostic Models*

For the complete data (i.e, the observed data $\vec{x}$ in conjunction with the unobserved skill profiles $\vec{a}$ and group membership $g$), the marginal likelihood is

$$L = \prod_{s=1}^{S} \prod_{n=1}^{N_s} P(\vec{x}_{ns}, \vec{a}_{ns}, g_{ns}; s),$$

that is, a sum over cluster-specific distributions of the complete data. With the above assumptions,

$$L = \prod_{s=1}^{S} \prod_{n=1}^{N_s} P(\vec{x}_{ns} \mid \vec{a}_{ns}, g_{ns}) p(\vec{a}_{ns} \mid g_{ns}) p(g_{ns} \mid s),$$

which equals

$$L = L_{\vec{x}} \times L_{\vec{a}} \times L_g,$$

with

$$L_{\vec{x}} \times L_{\vec{a}} \times L_g = \left( \prod_{s=1}^{S} \prod_{n=1}^{N_s} P(\vec{x}_{ns} \mid \vec{a}_{ns}, g_{ns}) \right) \left( \prod_{s=1}^{S} \prod_{n=1}^{N_s} p(\vec{a}_{ns} \mid g_{ns}) \right) \left( \prod_{s=1}^{S} \prod_{n=1}^{N_s} p(g_{ns} \mid s) \right).$$

Note that these components may be rearranged and rewritten as

$$L_{\vec{x}} = \prod_{s=1}^{S} \prod_{n=1}^{N_s} \prod_{i=1}^{I} P(x_{ins} \mid \vec{a}_{ns}, g_{ns}) = \prod_{g} \prod_{\vec{a}} \prod_{i} \prod_{x} P(X_i = x \mid \vec{a}, g)^{n_i(x, \vec{a}, g)},$$

with $n(x_i, i, \vec{a}, g) = \sum_s n(x_i, i, \vec{a}, g, s)$ is the frequency of category $x_i$ responses on item $i$ for examinees with skill pattern $\vec{a}$ in group $g$. Also,

$$L_{\vec{a}} = \prod_{s=1}^{S} \prod_{n=1}^{N_s} p(\vec{a}_{ns} \mid g_{ns}) = \prod_{\vec{a}} \prod_{g} p(\vec{a} \mid g)^{n(\vec{a};g)},$$

where $n(\vec{a}; g)$ is the frequency of skill pattern $\vec{a}$ in group $g$. Finally,

$$L_g = \prod_{s=1}^{S} \prod_{n=1}^{N_s} p(g_{ns} \mid s) = \prod_{s} \prod_{g} p(g \mid s)^{n(g;s)}$$

holds. The $n(g; s)$ represent the frequency of group membership in $g$ in cluster $s$.

### *5.2 Estimation of Cluster-Skill Distributions With the EM Algorithm*

Since unobserved latent variables are involved, the EM algorithm (Dempster, Laird, & Rubin, 1977) is a convenient choice for estimating GDMs (von Davier, in press-a) as well as MGDMs (von Davier, in press-b) and HGDMs. The EM algorithm cycles through the generation of expected values and the maximization of parameters given these preliminary expectations until convergence is reached. For details on this algorithm, refer to McLachlan and Krishnan (2000). For the HGDM, there are three different types of expected values to be generated in the E-step:

1. $\hat{n}_i(x, \vec{a}, g) = \sum_s \sum_n 1\{x_{ins} = x\} P(\vec{a}, g \mid \vec{x}_{ns}, s)$ is the expected frequency of response $x$ to item $i$ for examinees with skill pattern $\vec{a}$ in group $g$, estimated across clusters and across examinees within clusters

2. $\hat{n}(\vec{a}, g) = \sum_s \sum_n P(\vec{a}, g \mid \vec{x}_{ns}, s)$ is the expected frequency of skill pattern $\vec{a}$ and group $g$, estimated across clusters and across examinees within clusters

3. $\hat{n}(g; s) = \sum_n P(g \mid \vec{x}_{ns}, s)$ is the expected frequency of group $g$ in cluster $s$, estimated across examinees in that cluster

For the first and second type of the required expected counts, this involves estimating

$$P(\vec{a}, g \mid \vec{x}, s) = \frac{P(\vec{x}, s, \vec{a}, g)}{\sum_g P(\vec{x}, s, g)} = \frac{P(\vec{x} \mid \vec{a}, g) p(\vec{a} \mid g) p(g \mid s)}{\sum_g P(\vec{x}, s, g)},$$

with

$$P(\vec{x}, s, g) = \sum_{\vec{a}} P(\vec{x}, s, \vec{a}, g) = \sum_{\vec{a}} P(\vec{x} \mid \vec{a}, g) p(\vec{a} \mid g) p(g \mid s)$$

for each response pattern $\vec{x}_{ns}$, for $s = 1, \ldots, S$ and $n = 1, \ldots, N_s$. For the third type of expected count, use

$$p(g \mid \vec{x}, s) = \sum_{\vec{a}} P(\vec{a}, g \mid \vec{x}, s),$$

which is equivalent to

$$p(g \mid \vec{x}, s) = \frac{P(\vec{x}, s, g)}{\sum_g P(\vec{x}, s, g)} = \frac{\sum_{\vec{a}} P(\vec{x} \mid \vec{a}, g) p(\vec{a} \mid g) p(g \mid s)}{\sum_g \left[\sum_{\vec{a}} P(\vec{x} \mid \vec{a}, g) p(\vec{a} \mid g) p(g \mid s)\right]}.$$

This last probability then allows one to estimate the class membership $g$ given both the observed responses $\vec{x}$ and the known cluster membership $s$. The utility of the clustering variable may be evaluated in terms of increase of the maximum a posteriori probabilities $p(g \mid \vec{x}, s)$ over $p(g \mid \vec{x})$. If the clustering variable $s$ is informative for the classification $g$, a noticeable increase of the maximum posterior probabilities should be observed. The improvement should also be seen in terms of the marginal log-likelihood if $s$ in informative for $g$.

## 6 An Application to TOEFL® iBT Datasets

Simulated data have advantages, such as the truth (i.e., the set of generating values) is known and comparisons of different levels of model complexity and misspecification can be made on the basis of known deviations from the true model. The disadvantage is that simulated data are by origin artificial, so that the impact of model assumptions on model-data fit can only be studied under often less than realistic settings. The accuracy of parameter recovery using simulated data has been studied with quite satisfactory results for the GDM by von Davier (in press-a) using flat item response data with no missing values, and by Xu and von Davier (2006) for sparse matrix samples of item responses as collected in national and international surveys of educational outcomes.

The current exposition focuses on the comparison of results based on two administration of the TOEFL® iBT. The target of inference is the stability of estimates relating to clustering variables given by language group. The analyses carried out are independent scalings of two TOEFL iBT administrations for which Q-matrices were produced. Von Davier (press-a) pointed out that the GDM applied to TOEFL data resulted in highly correlated skill variables, and found that a two-dimensional, two-parameter logistic (2PL) IRT model across reading and listening domains provided a more parsimonious data description. However, the eight-skill model across reading and listening domains was the subject of further investigation by TOEFL experts, so that this model is adopted for the analyses with the hierarchical GDM.

In a first step, the HGDM was compared to the GDM without hierarchical extension, both adopting the same Q-matrix based on eight mastery/nonmastery skills for the February and November administrations of the TOEFL iBT. The HGDM was estimated according to Equation 7. In other words, measurement invariance was assumed across mixture components so that only the skill distribution could vary across clusters and the response probabilities $P(\vec{x} \mid \vec{a})$ depended on the skill profile only, not on cluster $s$ or mixture component $g$. Table 1 shows the skill correlations for the February administration, as well as the marginal skill mastery probabilities for the GDM. Table 2 shows the same information for the November administration.

**Table 1**

***Skill Correlations and Marginal Probabilities of Skill Mastery for the February Administration Based on the Nonhierarchical Eight-Skill General Diagnostic Model Across 76 Items Assuming Four Listening and Four Reading Skills***

| | Skill 1 | Skill 2 | Skill 3 | Skill 4 | Skill 5 | Skill 6 | Skill 7 | Skill 8 |
|---|---|---|---|---|---|---|---|---|
| Skill 1 | 1.00 | 0.76 | 0.73 | 0.80 | 0.75 | 0.60 | 0.69 | 0.57 |
| Skill 2 | | 1.00 | 0.83 | 0.81 | 0.65 | 0.64 | 0.67 | 0.58 |
| Skill 3 | | | 1.00 | 0.75 | 0.68 | 0.69 | 0.70 | 0.63 |
| Skill 4 | | | | 1.00 | 0.61 | 0.55 | 0.58 | 0.45 |
| Skill 5 | | | | | 1.00 | 0.79 | 0.76 | 0.66 |
| Skill 6 | | | | | | 1.00 | 0.86 | 0.80 |
| Skill 7 | | | | | | | 1.00 | 0.80 |
| Skill 8 | | | | | | | | 1.00 |
| P(master) | 0.63 | 0.61 | 0.57 | 0.69 | 0.54 | 0.46 | 0.49 | 0.39 |

The correlations range between 0.67 and 0.86 for skills of the same domain (i.e., among the four reading or four listening skills) and are slightly lower across the domains as expected. For correlations between one of the four reading skills and one of the four listening skills, the range is 0.56 to 0.77. These are still substantial correlations, which is due to the fact that overall reading and listening domains themselves are highly correlated. A two-dimensional 2PL IRT model

**Table 2**

***Skill Correlations and Marginal Probabilities of Skill Mastery for the November Administration Based on the Nonhierarchical Eight-Skill General Diagnostic Model Across 76 Items Assuming Four Listening and Four Reading Skills***

| | Skill 1 | Skill 2 | Skill 3 | Skill 4 | Skill 5 | Skill 6 | Skill 7 | Skill 8 |
|---|---|---|---|---|---|---|---|---|
| Skill 1 | 1.00 | 0.79 | 0.81 | 0.67 | 0.62 | 0.62 | 0.57 | 0.59 |
| Skill 2 | | 1.00 | 0.86 | 0.68 | 0.60 | 0.61 | 0.56 | 0.59 |
| Skill 3 | | | 1.00 | 0.70 | 0.64 | 0.63 | 0.58 | 0.60 |
| Skill 4 | | | | 1.00 | 0.71 | 0.67 | 0.77 | 0.67 |
| Skill 5 | | | | | 1.00 | 0.82 | 0.78 | 0.80 |
| Skill 6 | | | | | | 1.00 | 0.85 | 0.72 |
| Skill 7 | | | | | | | 1.00 | 0.86 |
| Skill 8 | | | | | | | | 1.00 |
| P(master) | 0.63 | 0.62 | 0.62 | 0.44 | 0.48 | 0.47 | 0.40 | 0.43 |

estimated with the *mdltm* software (von Davier, 2005) results in estimated correlations of between the reading and listening domains of 0.81 and 0.85 for the two administrations.

When estimating the HGDM for the two administrations, the resulting statistics differ from those from the GDM in two ways. First, there are two skill distributions $P(\vec{a} \mid c)$ estimated, one for each of two mixture components $c = 1$ and $c = 2$, representing the largest of the between-cluster differences (here language group) that can be expected. Then cluster-skill distributions are formed by a cluster-specific (a language group, $s$) proportion $P(c \mid s)$, or the probability of belonging to each of these skill distributions. Note that the TOEFL iBT datasets used here are composed of students from various language groups, and some of these languages are represented by only a few students. Since every cluster receives a different set of $P(c \mid s)$, one should consider collecting students representing very small language groups into larger clusters to avoid numerically unstable estimates of proportions for these small groups.

The log likelihood for the eight-skill GDM and HGDM are reported in Table 3 together with the number of estimated parameters and the average log likelihood per observation. Note that the November administration included a larger number of language groups, some of which were of rather small size. This led to a larger increase in the number of estimated parameters from GDM to HGDM for the November administration than for the February administration.

**Table 3**

***Log Likelihood and Number of Parameters for the Eight-Skill General Diagnostic Model and Hierarchical General Diagnostic Model for Both Administration***

| | Log likelihood | Parameters | Average likelihood |
|---|---|---|---|
| Independence | | | -43.24 |
| FEB GDM | -164435.20 | 194 | -38.83 |
| FEB HGDM | -163883.00 | 318 | -38.70 |
| FEB 2PL2 | -160799.26 | 160 | -37.97 |
| FEB H2PL2 | -160297.34 | 251 | -37.85 |
| Independence | | | -41.92 |
| NOV GDM | -196009.88 | 195 | -37.44 |
| NOV HGDM | -195480.19 | 337 | -37.34 |
| NOV 2PL2 | -191431.64 | 160 | -36.57 |
| NOV H2PL2 | -190905.63 | 269 | -36.47 |

The average likelihood per response pattern is improved by a small amount when including the language group as clustering variable. However, compared to the gain by assuming the GDM rather than independence of all observed variables, the gain in going from GDM to HGDM seems comparably small. For comparisons, the loglikelihood, parameters, and average-response pattern likelihoods are also presented for the two-dimensional 2PL/GPC model, which are estimated as a nonhierarchical model (2PL2) and a hierarchical model (H2PL2), and are also given in the table. As von Davier (in press-a) reported, the two-dimensional 2PL IRT model is a more parsimonious description of the TOEFL iBT pilot data than the eight-skill model, a result that holds up for both the February and the November administrations. The eight-skill model, however, is the focus of an ongoing methods comparison by TOEFL researchers, so it is adopted for subsequent comparisons between GDM and HGDM here without any comparisons to the two-dimensional 2PL/GPCM model.

Table 4 shows the two resulting marginal skill distributions for the February administration, and Table 5 shows the same information for the November administration. For both administrations, the mixture component $C1$ shows much lower mastery probabilities than component $C2$. The mixture component $C2$ is characterized by high probabilities of mastery of all eight skills for both administrations. The marginal sizes of the two components $\pi_{C2,Feb}$ and $\pi_{C2,Nov}$ for the two administrations differ somewhat; there is about 42% in the high proficiency class in November, whereas there is about 51% in February.

**Table 4**

***Marginal Skill Distributions for the Two Mixture Components*** $C1$ ***and*** $C2$ ***in the February Eight-Skill Hierarchical General Diagnostic Model With Skill Mastery Probabilities Given and Marginal Sizes of the Mixture Components Are*** $\pi_{C1} = 0.49$ ***and*** $\pi_{C2} = 0.51$

| | Skill 1 | Skill 2 | Skill 3 | Skill 4 | Skill 5 | Skill 6 | Skill 7 | Skill 8 |
|---|---|---|---|---|---|---|---|---|
| P(master\|C1) | 0.33 | 0.28 | 0.16 | 0.38 | 0.15 | 0.05 | 0.13 | 0.06 |
| P(master\|C2) | 0.92 | 0.93 | 0.96 | 0.97 | 0.94 | 0.85 | 0.85 | 0.72 |

**Table 5**

***Marginal Skill Distributions for the Two Mixture Components*** $C1$ ***and*** $C2$ ***in the November Eight-Skill Hierarchical General Diagnostic Model With Skill Mastery Probabilities Given and Marginal Sizes of the Mixture Components Are*** $\pi_{C1} = 0.58$ ***and*** $\pi_{C2} = 0.42$

| | Skill 1 | Skill 2 | Skill 3 | Skill 4 | Skill 5 | Skill 6 | Skill 7 | Skill 8 |
|---|---|---|---|---|---|---|---|---|
| P(master\|C1) | 0.40 | 0.38 | 0.38 | 0.16 | 0.16 | 0.11 | 0.04 | 0.09 |
| P(master\|C2) | 0.97 | 0.95 | 0.94 | 0.90 | 0.93 | 0.98 | 0.93 | 0.91 |

The two mixture components $C1$ and $C2$ represent the largest possible differences between clusters (language groups) in the sample, since each cluster receives an estimate of a proportion $P(C2 \mid s)$—and with that, implicitly, $P(C1 \mid s) = 1 - P(C2 \mid s)$—of members estimated to belong in the high versus low proficiency components $C2$ and $C1$. Since the mastery probabilities of all skills are much higher in $C2$ compared to $C1$ for both administrations, this proportion can be interpreted as the proportion of examinees in each language group who are highly proficient with respect to the assessment items reflected in the skill definitions. These proportions can be studied across administrations, so that the variation (or the lack thereof) of the proportion of highly proficient students in the language groups becomes a target of inference. This target delivers information about how well-aligned the TOEFL assessment is for the different language groups represented in the sample.

Figure 1 shows the proportion of students falling in the high performing class for the November and February administrations. The table contains only those language groups for which at least 10 students were observed for each administration of the TOEFL. It can be seen that the class sizes vary across administrations but are relatively stable when languages are compared. For example, the proportion of students with a Chinese (CHI) language background is smaller than the proportion of students with a French (FRE) language background (see the appendix for the language-specific class sizes). The correlation between the two high proficient class-size estimates across 37 countries is 0.787. When a weighted correlation (with weights defined as the geometric mean of the two language–group-specific sample sizes, one for each administration) across all 116 language groups is calculated, the correlation between the class-size estimates is 0.89.

The consistency of the language–group-proportion estimates and the substantial correlation of these estimates across the two administrations are evident from Figure 1. For estimates of the skill-mastery probabilities of language groups, the $P(C2 \mid s)$ and the mixture-component skill probabilities can be combined, resulting in

$$P(\vec{a} \mid s) = \sum_{c=C1}^{C2} P(\vec{a} \mid c) P(c \mid s)$$

for the language–group-specific skill distribution. As an illustration, the marginal skill mastery probabilities for the November and February administrations have been calculated for the CHI and Spanish (SPA) language groups. Table 6 shows the language–group-specific marginal skill mastery probabilities for CHI and SPA for the two administrations. It can be seen that the skill mastery probabilities range between 0.54 and 0.69 for the listening skills in the Spanish language sample and between 0.40 and 0.58 for the Chinese language sample for the November administration. For the reading skills, the mastery probabilities range between 0.32 and 0.41 for the Chinese language sample and between 0.49 and 0.55 for the Spanish language sample.

It is important to note that the language-group proportions as well as the estimates of

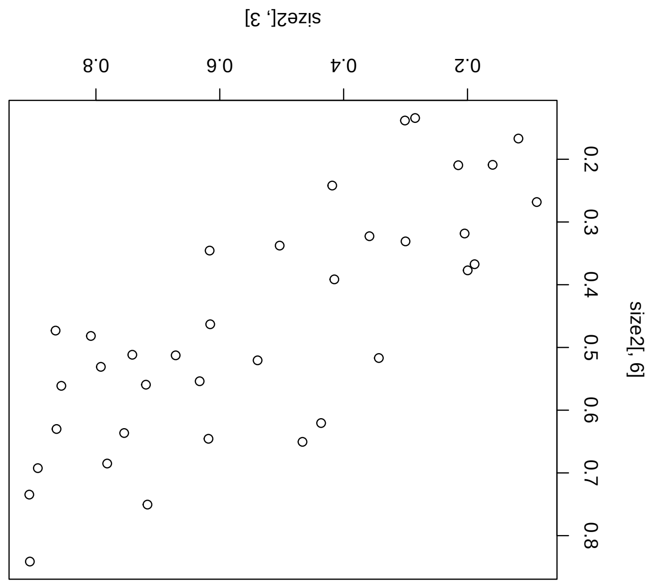


***Figure* 1 Plot of the high proficiency class-size correspondence across two administrations of the TOEFL iBT based on 37 language groups for which sample sizes exceeded 10 in both administrations.**

skill-mastery probabilities will vary somewhat over the administrations, even though the ordering of language–group-specific mastery estimates may stay stable. The estimates presented here are based on 4 + 4 skills with high correlations within the reading and listening domains as well as across. Therefore, a similar analysis may be tried with a model that joins the four postulated skills per domain into one overarching dimension by estimating a two-dimensional model instead. However, for the purpose of providing statistics on skill mastery for ongoing TOEFL research, it was necessary in the current study to use the expert-generated eight-skill matrix. As a result, the language–group-specific profiles of skill mastery will, due to the nature of the highly correlated skills, mostly reflect overall differences in the proficiency level of the applicant samples across language groups.

## 7 Conclusions

This paper introduces a hierarchical version of the GDM (von Davier, in press-a) and shows the effect of clustering through a comparison of results from two administrations of the TOEFL

**Table 6**

***Language–Group-Specific Mastery Probabilities Exemplified Using the November and February Administrations Based on Mixing Components and the Chinese and Spanish Language Groups***

| CHI and SPA in Nov | Skill 1 | Skill 2 | Skill 3 | Skill 4 | Skill 5 | Skill 6 | Skill 7 | Skill 8 |
|---|---|---|---|---|---|---|---|---|
| C1: 0.68 (CHI), 0.49 (SPA) | 0.40 | 0.38 | 0.38 | 0.16 | 0.16 | 0.11 | 0.04 | 0.09 |
| C2: 0.32 (CHI), 0.51 (SPA) | 0.97 | 0.95 | 0.94 | 0.90 | 0.93 | 0.98 | 0.93 | 0.91 |
| P(SKILL\|CHI) | 0.58 | 0.56 | 0.56 | 0.40 | 0.41 | 0.39 | 0.32 | 0.35 |
| P(SKILL\|SPA | 0.69 | 0.67 | 0.67 | 0.54 | 0.55 | 0.55 | 0.49 | 0.51 |

| CHI and SPA in Feb | Skill 1 | Skill 2 | Skill 3 | Skill 4 | Skill 5 | Skill 6 | Skill 7 | Skill 8 |
|---|---|---|---|---|---|---|---|---|
| C1: 0.79 (CHI), 0.33 (SPA) | 0.33 | 0.28 | 0.16 | 0.38 | 0.15 | 0.05 | 0.13 | 0.06 |
| C2: 0.21 (CHI), 0.67 (SPA) | 0.92 | 0.93 | 0.96 | 0.97 | 0.94 | 0.85 | 0.85 | 0.72 |
| P(SKILL\|CHI) | 0.45 | 0.42 | 0.33 | 0.50 | 0.32 | 0.22 | 0.28 | 0.20 |
| P(SKILL\|SPA) | 0.63 | 0.61 | 0.57 | 0.68 | 0.55 | 0.46 | 0.50 | 0.40 |

iBT assessment when estimating language–group-specific proficiencies. The HGDM provides reliable estimates of proportions of high proficiency across language groups. The correlation of the estimates is 0.78 for the 37 largest language groups not weighted by sample size, and it increases to 0.89 when all language groups that are present in both administrations are weighted according to their pooled sample size.

If the clustering is informative as it seems to be in the TOEFL case, the prediction of proficiency can potentially be improved, as seen in the slight increase of average log-likelihood (see Table 3). The clustering, or language-group membership in the analyses presented here, acts as ancillary information, so that the fit of the HGDM to the observed cognitive item responses can be compared to models without a clustering variable. The results presented here indicate that a mixture of different class-specific skill distributions is a useful tool in conjunction with cluster-specific mixing proportions to model the dependency of skill distribution on a clustering variable. The approach estimates conditional skill distributions across the whole sample, which, in the parameterization chosen, represent different expected skill profiles in unknown subpopulations of a mixture distribution. The cluster-specific mixing proportions then estimate the composition of the clusters—here language groups—based on the assumption that the mixture-distribution subpopulations are represented in varying levels across clusters. In this example, the mixture components turned out to be ordered proficiency classes, due to the nature of the eight skills applied, which are known to be substantially correlated.

The estimated proportions, more specifically the variance of these proportions across clusters, and the consistency of identified proportions across administrations can provide valuable information about the sources of proficiency variation in hierarchically organized data. The HGDM provides a tool to study such variations in the context of item response models, latent class models, and diagnostic models for profile scoring.

## References


Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods* (1st ed.). Newbury Park, CA: Sage Publications.

Cramer, J. S. (2003). *Logit models from economics and other fields.* Cambridge, UK: Cambridge University Press.

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

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

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

Haberman, S. J. (1979). *Qualitative data analysis: Vol. 2. New developments.* New York: Academic Press, 1979.

Heinen, T. (1996). *Latent class and discrete latent trait models, similarities and differences.* Thousand Oaks, CA: Sage Publications.

Kamata, A., & Cheong, Y. F. (2007). Hierarchical Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models—Extensions and applications* (pp. 217–232). New York: Springer.

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.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187–212.

McLachlan, G., & Krishnan, T. (2000) *The EM-algorithm and extensions.* New York: Wiley.

Mislevy, R. J., & Verhelst, N. D. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195–215.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16* (2), 159–176.

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

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

Vermunt, J. K. (2004). An EM-algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica, 58* (2), 220–233.

von Davier, M. (2005). mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]. Princeton, NJ: ETS.

von Davier, M. (in press-a). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology.*

von Davier, M. (in press-b). *Mixtures of general diagnostic models.* Princeton, NJ: ETS.

von Davier, M., DiBello, L., & Yamamoto, K. Y. (2006). *Reporting test outcomes with models for cognitive diagnosis* (ETS Research Rep. No. RR-06-28). Princeton, NJ: ETS.

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 643–768). Amsterdam: Elsevier.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement, 28*(6), 389–406.

von Davier, M., & Yamamoto, K. (2007). Mixture distribution Rasch models and hybrid Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models—Extensions and Applications* (pp. 99–116). New York: Springer.

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data* (ETS Research Rep. No. RR-06-08). Princeton: ETS.

## Appendix

**Proportions of High Proficiency Class Membership by Country for the February and November Administration of the TOEFL iBT for 37 Language Groups for Which Sample Sizes Exceeded 10 in Both Administrations**

| Lang. | N(FEB) | P(C2\|FEB) | N(NOV) | P(C2\|NOV) |
|---|---|---|---|---|
| CHI | 609 | 0.2046 | 657 | 0.3185 |
| VIE | 33 | 0.0883 | 92 | 0.2681 |
| KOR | 604 | 0.2148 | 832 | 0.2094 |
| RUM | 32 | 0.8560 | 35 | 0.5611 |
| FRE | 467 | 0.7818 | 357 | 0.6850 |
| URD | 29 | 0.3433 | 43 | 0.5167 |
| GER | 458 | 0.9067 | 433 | 0.8412 |
| POL | 27 | 0.8082 | 58 | 0.4816 |
| ITA | 378 | 0.6154 | 331 | 0.4629 |
| IND | 27 | 0.1886 | 45 | 0.3673 |
| SPA | 294 | 0.6712 | 483 | 0.5125 |
| TAM | 21 | 0.6182 | 26 | 0.6456 |
| JPN | 245 | 0.2847 | 410 | 0.1344 |
| BEN | 19 | 0.4665 | 19 | 0.6505 |
| ARA | 119 | 0.3010 | 187 | 0.1382 |
| BUL | 19 | 0.7412 | 19 | 0.5115 |
| TGL | 82 | 0.5033 | 111 | 0.3377 |
| HEB | 19 | 0.8938 | 28 | 0.6925 |
| RUS | 74 | 0.7192 | 136 | 0.5592 |
| MAL | 18 | 0.8636 | 25 | 0.6301 |

(table continues)

Table (continued)

| Lang. | N(FEB) | P(C2\|FEB) | N(NOV) | P(C2\|NOV) |
|---|---|---|---|---|
| TEL | 60 | 0.6326 | 43 | 0.5539 |
| UKR | 15 | 0.4151 | 15 | 0.3913 |
| POR | 59 | 0.7922 | 73 | 0.5308 |
| ALB | 14 | 0.4363 | 18 | 0.6206 |
| ENG | 58 | 0.5389 | 66 | 0.5206 |
| CZE | 13 | 0.6163 | 13 | 0.3456 |
| THA | 48 | 0.1178 | 91 | 0.1669 |
| IBO | 12 | 0.8651 | 13 | 0.4731 |
| HIN | 48 | 0.7168 | 70 | 0.7504 |
| PAN | 11 | 0.3585 | 15 | 0.3225 |
| TUR | 48 | 0.3000 | 76 | 0.3310 |
| N/A | 11 | 0.7544 | 20 | 0.6363 |
| FAS | 43 | 0.4183 | 58 | 0.2420 |
| YOR | 10 | 0.9075 | 11 | 0.7347 |
| GUJ | 37 | 0.1997 | 40 | 0.3770 |
| AMH | 10 | 0.1595 | 23 | 0.2089 |