PSYCHOMETRIKA—VOL. 76, NO. 2, 318–336
APRIL 2011
DOI: 10.1007/S11336-011-9202-Z

MEASURING GROWTH IN A LONGITUDINAL LARGE-SCALE ASSESSMENT WITH A GENERAL LATENT VARIABLE MODEL

MATTHIAS VON DAVIER AND XUELI XU

ETS

CLAUS H. CARSTENSEN

BAMBERG UNIVERSITY

The aim of the research presented here is the use of extensions of longitudinal item response theory (IRT) models in the analysis and comparison of group-specific growth in large-scale assessments of educational outcomes.

A general discrete latent variable model was used to specify and compare two types of multidimensional item-response-theory (MIRT) models for longitudinal data: (a) a model that handles repeated measurements as multiple, correlated variables over time and (b) a model that assumes one common variable over time and additional variables that quantify the change. Using extensions of these MIRT models, we approach the issue of modeling and comparing group-specific growth in observed and unobserved subpopulations. The analyses presented in this paper aim at answering the question whether academic growth is homogeneous across types of schools defined by academic demands and curricular differences. In order to facilitate answering this research question, (a) a model with a single two-dimensional ability distribution was compared to (b) a model assuming multiple populations with potentially different two-dimensional ability distributions based on type of school and to (c) a model that assumes that the observations are sampled from a discrete mixture of (unobserved) populations, allowing for differences across schools with respect to mixing proportions. For this purpose, we specified a hierarchical-mixture distribution variant of the two MIRT models. The latter model, (c), is a growth-mixture MIRT model that allows for variation of the mixing proportions across clusters in a hierarchically organized sample. We applied the proposed models to the PISA-I-Plus data for assessing learning and change across multiple subpopulations. The results of this study support the hypothesis of differential growth.

Key words: item response theory, growth models, multidimensional IRT, longitudinal models, diagnostic models, large-scale assessments.

1. Introduction

The aim of the research presented here is the use of extensions of longitudinal item-response-theory (IRT) models in the analysis and comparison of group-specific growth in large-scale assessments of educational outcomes. Measurement of change in student performance between testing occasions is a central topic in educational research and assessment (Fischer, 1995). Measurement of differential growth in proficiency in groups that are subjected to different types of curricular and learning expectations is of central importance for large-scale national and international programs aimed at monitoring educational progress. It may be conjectured that students who operate at a higher level of learning expectations are in a better position to gain more within a given period of schooling. Differences between high- and low-performing students tend to accentuate over time. This so-called Matthew effect ("he who has will be given more ...") has been found to apply in education in such areas as science knowledge (Walberg & Tsai, 1983) and reading literacy (Stanovich, 1986). In our study, we will compare group-specific growth measures to

Any opinions expressed in this paper are those of the author(s) and not necessarily of Educational Testing Service. Requests for reprints should be sent to Matthias von Davier, ETS, Princeton, NJ, USA. E-mail: mvondavier@ets.org

examine whether a similar effect can be found in a representative sample of students tested in a large-scale study and followed up in a longitudinal design.

Most research on the measurement of growth has been conducted using small-scale data collections in fields such as developmental, educational, clinical, and applied psychology. Change across occasions can be meaningfully measured by focusing either on the group level (Fischer, 1973, 1976; Wilson, 1989) or by focusing on the individual (Andersen, 1985; Andrade & Tavares, 2005; Embretson, 1991; Fischer, 1995). The research presented here provides extensions of item-response-theory (IRT) approaches for measuring growth in representative samples of student populations. While most large-scale educational assessments, such as the National Assessment of Educational Progress (NAEP) or the Programme for International Student Assessment (PISA), are cross-sectional studies, there are some studies that monitor student growth in longitudinal designs. NAEP and PISA, as well as other national or international survey assessments of educational outcomes, aim at monitoring educational outcomes at the system level by repeated cycles of assessments using representative samples in different age or grade cohorts of the same target population (for example, eighth-graders in NAEP, 15-year-olds in PISA) over time. Longitudinal survey assessments, such as the Early Childhood Longitudinal Study (ECLS) and the PISA-L used in our study, aim at following a representative sample of the target population over time. PISA-L was implemented as a national option that conducted follow-up measurement on the operational PISA 2003 sample. The necessary details on the PISA-L study will be given in a section describing the data below.

1.1. Measuring Group Differences in Growth

Fischer (1973, 1976) proposed a linear logistic test model (LLTM) based on the dichotomous Rasch model (Rasch, 1980). The Rasch model assumes that the probability of a correct response by person j on item i can be written as

$$P(X_{ij} = 1) = \frac{\exp(\theta_j - \beta_i)}{1 + \exp(\theta_j - \beta_i)}, \tag{1}$$

where X_{ij} is the response variable with values in $\{0, 1\}$, θ_j is person j's ability, and β_i is item i's difficulty. The LLTM entails linear constraints across item parameters β_i for the purpose of representing a structural relationship between the difficulties of different item sets (here: items given at different points in time). The LLTM is a constrained Rasch model and thus a unidimensional model, it does not involve multiple skills but rather a single ability dimension. However, it can be used to model growth (Fischer, 1995; Glück & Spiel, 1997) by specifying linear constraints that represent time point effects, group effects, and other item features. For a set of I items given at T time points in G treatment groups, a group-specific model for growth can be specified in the LLTM using

$$\beta_i = \sum_{l=1}^{p} w_{il}\alpha_l + c, \tag{2}$$

in which the effects from α_1 to α_I are the baseline item difficulties, α_{I+1} is the effect of Time Point 2, α_{I+T-1} is the effect of Time Point T, $\alpha_{I+T-1+1}$ is the effect of Group 2, $\alpha_{I+T-1+G-1}$ is the effect of Group G, and $p = I + T - 1 + G - 1$. This example assumes only main effects for base item difficulties, time points, and groups; Time Point 1 and Group 1 are the reference groups. A model with group-specific time point effects is also easily specified within this framework. Note that the LLTM model for growth does not measure change at the individual level because the α effects do not depend on individuals.

Wilson (1989) presented the Saltus model, which assumes student progression through developmental stages. As in the LLTM, in the Saltus model an additive constant that modifies item difficulty represents the effect of belonging to one of several developmental stages. However, the

LLTM breaks item difficulties into known components, whereas the Saltus model does not assume that the student's current stage is known. In the Saltus approach, the student's current stage and the student's ability measured within this stage are latent variables that must be inferred by using the model's assumptions and plugging in the observed responses. Examinees are assigned an ability parameter θ_j and a class membership c_j representing their developmental stage. The classes $c = 1, \ldots, G$ (developmental stages) enter the model through stage parameters τ_{ck} for subsets of items belonging to the same group (item type) and indexed with the same $k(i)$ and same developmental stage $c(j)$ of examinee j. The equation for the Saltus model is

$$P(X_{ij} = 1|\theta_j, \beta_i) = \frac{\exp(\theta_j - \beta_i + \tau_{c(j)k(i)})}{1 + \exp(\theta_j - \beta_i + \tau_{c(j)k(i)})}. \tag{3}$$

The Saltus model parameters are the ability θ_j of respondent j, the baseline difficulty β_i of item i, and the Saltus parameters $\tau_{c(j)k(i)}$ as defined above. The Saltus model can also be specified for polytomous items (Draney & Wilson, 2007; Wilson & Draney, 1997). It is a constrained version of the mixture-distribution Rasch model (Rost, 1990; von Davier & Rost, 1995). Like the LLTM, the Saltus model is an approach to modeling growth by structuring or constraining item difficulties. Unlike the LLTM, it includes a latent class variable that determines which structural parameter applies to the examinee, depending on his or her class membership $c(j)$. Models whose population structure consists of an unobserved mixture of subpopulations can be used to model different trajectories of growth for different subpopulations. In such models, growth is a group-specific trajectory.

1.2. Measuring Individual Differences in Growth

The multidimensional Rasch model allows for the modeling of individual growth. Indeed, Andersen (1985) proposed that it be used for the repeated administration of the same items over time points. The following equation expresses Andersen's model:

$$P(X_{ijk} = 1|\theta_{jk}, \beta_i) = \frac{\exp(\theta_{jk} - \beta_i)}{1 + \exp(\theta_{jk} - \beta_i)}, \tag{4}$$

where θ_{jk} represents the ability of person j at occasion k, and β_i is the difficulty of item i. Note that item difficulties remain constant across time points (occasions), but the ability associated with each occasion may differ. Thus, measurement occasions are represented by multiple correlated ability variables. In Andersen's model, abilities are specific to occasions; they do not quantify change but represent ability level at each occasion (Embretson, 1991). Therefore, deriving measures of change across occasions based on the model requires calculation of differences between occasion-specific abilities.

The model proposed by Andrade and Tavares (2005) can be viewed as an extension of Andersen's (1985) model. It describes latent ability changes within an item-response-theory (IRT) framework and assumes known fixed values of item parameters. The latent ability structure describes the changes over occasions. This model can be written as

$$\begin{aligned} P(X_{ijk} = 1|\theta_{jk}, \alpha_i, \beta_i, c_i) &= c_i + (1 - c_i)\frac{\exp[\alpha_i(\theta_{jk} - \beta_i)]}{1 + \exp[\alpha_i(\theta_{jk} - \beta_i)]}, \\ \theta_j = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{jK})^T &\sim \text{MVN}_K(\mu, \Sigma), \end{aligned} \tag{5}$$

in which θ_{jk} and β_i are defined as in Equation (3), α_i and c_i are the discrimination and guessing parameters defined as in the two- and three-parameter IRT models (2PL and 3PL IRT; Lord & Novick, 1968), and $\text{MVN}_k(\mu, \Sigma)$ is the k-dimensional multivariate normal distribution with mean vector μ and covariance matrix Σ.

Embretson (1991) proposed a multidimensional Rasch model for learning and change (MRMLC) to provide parameters for individual differences in change. She postulated the involvement of K abilities in item responses within K occasions. Specifically, the MRMLC assumes that on the first occasion ($k = 1$) only an initial ability is involved in the item responses and that on later occasions ($k > 1$) ability plus $k - 1$ additional abilities are involved in the performance. Thus, the number of abilities increases at each time point (occasion). The MRMLC can be written as

$$P\left(X_{ijk} = 1|(\theta_{j1}, \ldots, \theta_{jk}), \beta_i\right) = \frac{\exp\left(\sum_{m=1}^{k} \theta_{jm} - \beta_i\right)}{1 + \exp\left(\sum_{m=1}^{k} \theta_{jm} - \beta_i\right)}, \tag{6}$$

where θ_{jm} and β_i are defined as in Equation (4). Embretson (1997) presented a version of the MRMLC that contains slope parameters and could be referred to as the 2PL version of the MRMLC. Note that the same items are repeated over occasions in Andersen's (1985) model while Embretson (1991) developed the MRMLC for situations in which items are not necessarily repeated. Repeated item presentation to the same test taker may lead to practice effects and/or memory effects, and local dependency among item responses may result. Equation (6) indicates that for an item i observed at time k, the abilities up to time k are involved, including initial ability (θ_{j1}) and $k - 1$ time-point specific abilities ($\theta_{j2}, \ldots, \theta_{jk}$), termed "modifiabilities" in Embretson's model. The change between condition $k - 1$ and k equals the kth modifiability (θ_{jk}). Using the partial credit model (PCM, Masters, 1982), Fischer (2001) extended MRMLC to polytomous items.

2. An Item Response Model for Individual Change and Average Growth

In this study, we specified extensions of the Andersen (1985) and Embretson (1991) models using a framework of general latent-variable modeling. We chose these two approaches because they represent two alternatives of MIRT models that enjoy continued attention of researchers using item-response models for operational analysis of large-scale assessment data. In addition, we chose these two approaches since they lead to identical results (modulo parameter transformations) when using them in their original formulation based on the Rasch measurement model, but are different when utilizing a more general IRT measurement model. Details on the relationship between these models will be given below. We utilized the general diagnostic model (GDM; von Davier, 2005) framework to implement these approaches with extensions that allow for (a) the use of more general IRT measurement models and (b) more complex population structures:

(a) The GDM allowed a multidimensional generalization of the two-parameter logistic (2PL) model and the generalized partial credit model (GPCM; Muraki, 1992), rather than the Rasch model. The 2PL model and GPCM enable estimation of multiple slope parameters for items assumed to be multidimensional. In the case of simple-structure models such as those found in large-scale surveys, multidimensional GDMs can be used to simultaneously estimate the different IRT scales and the multidimensional ability distribution.

(b) The GDM also allowed the use of multiple-group (Xu & von Davier, 2006) and mixture-distribution versions of IRT and MIRT models (von Davier & Rost, 2006) and mixture distribution diagnostic models (von Davier, 2007a, 2007b, 2010). Multiple-group extensions of IRT (Bock & Zimowski, 1997) should be used whenever a sample is drawn from a population that comprises multiple subpopulations. In survey assessments, students often are sampled from composite populations in which subpopulations are defined by variables such as geographical region, socioeconomic status, curriculum, instructional track, or type of school.

Altogether, 2 (types: Embretson vs. Andersen) $\times 3$ (population model: single group vs. multiple groups vs. hierarchical mixture) $\times 2$ (scale parameterization: 1PL vs. 2PL) $= 12$ different models were compared. We will refer to models with occasion specific factors as generalized Andersen models even for the cases where we talk about models that are generalized to include slope parameters or multiple populations. We will refer to models that include the same baseline dimension for all occasions plus occasion-specific secondary dimensions (modifiabilities) as (generalized) Embretson-type models. The 12 models compared in this study were estimated using the EM algorithm that was implemented in the software *mdltm* (von Davier, 2005). Standard errors of item parameters were estimated using a jackknife resampling approach. Details on the procedure used for variance estimation can be found in Hsieh, Xu, and von Davier (2009). When applying the GDM to longitudinal data, the probability of a response depends on item-difficulty and occasion-specific parameters. By representing the latter in a design matrix, we were able to specify the original Andersen (1985) approach, the Rasch-type and 2PL-type Embretson (1991) approach, and our generalizations (described below) of these within a single framework. Group differences in skills distribution may exist and are represented as model parameters. Therefore, to allow for differences in proficiency distributions and in amount of change across student groups, we compared single-group Andersen and Embretson approaches to Andersen and Embretson approaches with multiple populations. Details about these approaches will be presented in the next section.

For an application of the GDM to longitudinal data, let time points be denoted by $T \in \{1, 2\}$, and let the D represent the total number of different items presented at time points 1 and 2. Let item index be $i = 1, \ldots, 2D$. While we will assume that only some subset of item responses is observed per time point, we will assume without loss of generality that any item is represented in the vector of observed variables twice, that is, an item that is indexed by j at Time Point 1 will have index $j + D$ at Time Point 2. We will assume that the item index is arranged so that $i \in \{1, \ldots, D\}$ is the index set unique to Time Point 1 while $i \in \{D + 1, \ldots, 2D\}$ is the index set unique to Time Point 2. Let x_{ij} denote the response of examinee $j = 1, \ldots, N$ to items $i = 1, \ldots, D, D + 1, \ldots, 2D$ at time points $T = 1, 2$. Note that we only observe some of these responses per person j. Let $P_{ij} = P(X_{ij} = 1)$ denote the probability of a correct response for examinee $j = 1, \ldots, N$ to item $i = 1, \ldots, D, D + 1, \ldots, 2D$ at time point $T = 1, 2$.

Then, both the Andersen and the Embretson approaches follow the general equation

$$P_{ij} = \frac{1}{1 + \exp(\beta_i - \alpha_{i1}\theta_1 - \alpha_{i2}\theta_2)} \tag{7}$$

with suitably chosen α_{i1} and α_{i2}. The above model is a two-dimensional IRT model at both time points, but the Andersen and Embretson approaches differ in how they assume time points $T = 1, 2$ and dimensions θ_1 and θ_2 are related. The Andersen model does not assume that there is a first dimension θ_1 at Time Point 2 or a second dimension θ_2 at Time Point 1, while the Embretson model assumes that there is a first (interpretation = baseline) dimension θ_1 at time points $T = 1$ and $T = 2$, while there is a second dimension θ_2 unique to $T = 2$.

2.1. The Andersen Longitudinal Model

The Andersen case constrains the general case (7) by assuming for time point one, that is, for items $i \in \{1, \ldots, D\}$,

$$\alpha_{i1} = \alpha_{(i+D)2}, \qquad \alpha_{i2} = 0 \tag{8}$$

for $i \in \{D + 1, \ldots, 2D\}$, that is, Time Point 2

$$\alpha_{i1} = 0, \quad \alpha_{i2} = \alpha_{(i-D)1}, \quad \text{and} \quad \beta_i = \beta_{i+D} \quad \text{for } i \in \{1, \ldots, D\}. \tag{9}$$

TABLE 1.
Andersen model within the General Diagnostic Model (GDM) framework.

Items	First dimension	Second dimension
Items unique to Time Point 1	X	
Items unique to Time Point 2		X
Common item i at Time Point 1	X	
Common item i at Time Point 2		X

In the GDM, the Andersen model is represented by a simple-structure design matrix, with nonzero entries $q_{i1} = 1$ for items representing Time Point 1, $i \in \{1, \ldots, D\}$, in the first column representing Dimension 1, and zero entries for Dimension 2, $q_{i2} = 0$ for these items. For Time Point 2, $i \in \{D + 1, \ldots, 2D\}$, there are nonzero entries in the second column $q_{i2} = 1$ representing Dimension 2 and zero entries in column 1, $q_{i1} = 0$. In addition, equality constraints are imposed across items i and $i + D$. The original Andersen model only requires equality constraints on item difficulties. The 2PL extension of the Andersen approach requires constraints that ensure the same items have the same discrimination parameters over time but for different ability (time-point specific) dimensions. This constraint enables construction of a common scale for comparisons between the two dimensions defined by Time Points 1 and 2. Table 1 shows the structure of the Andersen-type model within the GDM framework.

With these constraints in place, we use the convention often found in Rasch modeling to normalize item difficulties and slopes (see below). We estimate the mean and variance for each of the time-specific ability dimensions. In addition, a correlation $\rho(\theta_1, \theta_2)$ is estimated.

2.2. *The Embretson Longitudinal Model*

The Embretson model is characterized by the following set of constraints imposed on the general model given in Equation (7). Time point 1 is defined by constraints for items in $i \in \{1, \ldots, D\}$ of the form

$$\alpha_{i1} = \alpha_{(i+D)1}, \qquad \alpha_{i2} = 0. \tag{10}$$

For Time Point 2, that is, items in $i \in \{D + 1, \ldots, 2D\}$, there are two nonzero loadings, both α_{i1} and α_{i2} are estimated with a constraint

$$\alpha_{i1} = \alpha_{(i-D)1} \tag{11}$$

imposed on the first dimension.

As before, the assumption that the same item i given at two time points $T = 1, 2$ has the same item difficulty is implemented by the constraint $\beta_i = \beta_{i+D}$ for $i \in \{1, \ldots, D\}$. While the Embretson model started out as a model that does not assume repeated items within individuals, it assumes that all items may appear at Time Point 1 and Time Point 2. This can be achieved with balanced incomplete block (BIB) designs that present different sets of items to each test taker, while all item blocks are given at both time points. Our application deviates from a complete balance over time by having a few items that are unique to a time point, while the great majority of items are administered at both time points. Through the second ability variable θ_2, the Embretson approach models average growth as the mean of this ability variable $M(\theta_2) = \mu$. Most importantly, we now have a unidimensional model at Time Point 1

$$\text{logit}(P_{ij}) = \alpha_{i1}\theta_{1j} - \beta_i \tag{12}$$

TABLE 2.
Embretson model within the General Diagnostic Model (GDM) framework.

Items	First dimension	Second dimension
Items unique to Time Point 1	X	
Items unique to Time Point 2	X	X
Items common to both time points	X	X

and a multidimensional model at Time Point 2, namely

$$\text{logit}(P_{ij}) = \alpha_{(i-D)1}\theta_{1j} + \alpha_{i2}\theta_{2j} - \beta_{(i-D)}. \tag{13}$$

The design matrix utilized to specify the Embretson model in the GDM differs from the Andersen case in the following way: Items presented at Time Point 1 load only on Dimension 1, while items presented at Time Point 2 are loading on both Dimensions 1 and 2. Table 2 shows the structure of the Embretson-type model under the linking design within a GDM framework.

For identification, we use the convention customarily used in Rasch modeling to set the average of item difficulty parameters zero. It is well known that the Rasch model does not contain a slope parameter. This is equivalent to using the more general 2PL model with the constraint that all slopes equal 1.0. We will use this convention to estimate Rasch models; and in order to estimate 2PL extensions of the Andersen and the Embretson model, we will use the convention to normalize the slope using an average of 1.0. In addition, a correlation $\rho(\theta_1, \theta_2)$ can be estimated. This correlation should be lower in the Embretson compared to the Andersen approach, since the baseline θ_1 and the average expected growth μ already take care of a significant proportion of the variance in educational settings where students go through more or less the same amount of teaching until retested.

An examination of the two model specifications reveals that the Andersen model is a special case of the Embretson model: By means of the constraint $\alpha_{i2} = \alpha_{i1}$ we get

$$\text{logit}\big(P(x_{ij})\big) = \alpha_{i1}\theta_1 + \alpha_{i1}\theta_2 - \beta_i = \alpha_{i1}(\theta_1 + \theta_2) - \beta_i = \alpha_{i1}\theta_{2(\text{Andersen})} - \beta_i \tag{14}$$

for items $i \in \{D+1, \ldots, 2D\}$ presented at Time Point 2. Clearly, the Andersen model being a re-parameterized special case of the Embretson model with fewer parameters will necessarily exhibit a higher deviance than the Embretson model. Also, going back from the Andersen model to a constrained Embretson model would be possible via $\theta_{2(\text{Embretson})} = \theta_{2(\text{Andersen})} - \theta_1$. This gives an indication of how the ability parameters in the two approaches can be interpreted: The mean and variance associated with the Embretson change dimension $\theta_{2(\text{Embretson})}$ give an indication of what the baseline dimension θ_1 cannot explain. In some sense, $\theta_{2(\text{Embretson})}$ can be said to more directly represent the actual growth rather than $\theta_1 + \theta_{2(\text{Embretson})} = \theta_{2(\text{Andersen})}$, which already contains the baseline ability level: A student who sets out high, say $\theta_1 = 2.2$, and stays at the same level, $\theta_{2(\text{Andersen})} = 2.2$, did not grow at all, while a student that started with $\theta_1 = 0.0$ but has $\theta_{2(\text{Andersen})} = 2.2$ as well can be considered to have grown substantially. This is to say that the same $\theta_2 = 2.2$ may mean very different things in the Andersen approach given different levels of initial ability.

3. Multiple Group IRT Growth Models—Measuring Change in Multiple Populations

Extending the Anderson and Embretson approaches to multiple populations requires the introduction of an additional variable: The indicator function for the group membership variable plays a crucial role in these models. We will introduce these extensions in the framework of

the GDM (von Davier, 2005). The multiple group variable is represented by $1_c[g(j)]$, where $g(j)$ denotes the group membership of student j. If $c = g(j)$, let $1_c[g(j)] = 1$; otherwise let $1_c[g(j)] = 0$. If the group membership $g(j)$ is unknown, the multiple-population GDM becomes a discrete mixture distribution GDM (von Davier & Rost, 2006; von Davier & Yamamoto, 2007). The general model for measuring change in the GDM for multiple observed populations or indirectly observed mixture components can be written as

$$P\left(X_{ij} = 1 | \underline{q}_i, \underline{\theta}_j, \underline{\underline{\alpha}}_i, \underline{\beta}_i\right) = \sum_{c=1}^{G} \pi_{cj} \frac{\exp\left(\sum_{l=1}^{k} q_{il}\alpha_{icl}\theta_{jl} - \beta_{ic}\right)}{1 + \exp\left(\sum_{l=1}^{k} q_{il}\alpha_{icl}\theta_{jl} - \beta_{ic}\right)}. \tag{15}$$

The parameters and design variables are defined as follows: The variable q_{il} stands for the (i, l)th entry of an $M(I, k)$ design matrix with binary (0/1) entries, which will be referred to as the Q-matrix here, and $\underline{q}_i$ denotes the ith row in this matrix. These entries determine which ability dimensions l are required for which item i. The parameter θ_{jl} is lth component of the ability $\underline{\theta}_j$ of examinee j, $\underline{\beta}_i = (\beta_{i1}, \ldots, \beta_{iG})$ is the vector-valued item difficulty parameter of item i in group c. The parameter α_{icl} is the discrimination parameter of ability component l for item i in group c, and $\underline{\underline{\alpha}}_i$ denotes the matrix-valued collection of these slope parameters for item i across dimensions and groups. The parameter π_{cj} is defined as follows: For a multiple group model, $\pi_{cj} = 1_c[g(j)]$; for a mixture model with unknown group membership, $\pi_{cj} = \pi_c$ for all j, where π_c denotes class size or mixing proportion of a discrete-mixture model (McLachlan & Peel, 2000); and for a hierarchical-mixture model, $\pi_{cj} = \pi_{cs(j)}$, with $s(j)$ representing an index variable that represents the membership of student j to cluster s (Vermunt, 2003; von Davier, 2007b). In Equation (15), the model is defined for binary responses only. Note that this is a simplification of the original GDM, which is defined both for dichotomous and polytomous response variables (von Davier and Yamamoto, 2004; von Davier, 2005, 2008).

If groups of students go through differentially paced curricula, there may be differences in average growth that can be modeled as group-level differences. We utilize the multiple-group general diagnostic models (MG-GDMs) to specify Andersen's and Embretson's model for samples that are based on compositions of multiple populations such as students from multiple school types following different curricula. The item-level constraints over time points are used as described above and are additionally imposed on difficulty and slope parameters across all groups to ensure that the same constraints hold in each of these. However, in order to remove the indeterminacy of the IRT scale, additional constraints are needed. Either the mean and variance of the ability distribution in a population can be constrained, or the average item difficulties and slope parameters can be constrained for this purpose (von Davier & von Davier, 2007). We chose to constrain item parameters. Item difficulties were constrained to a mean of 0.0, and slope parameters to an average of 1.0 (when estimated) in order to maintain comparability of group specific parameters of initial state and growth as represented in the mean and variance estimates of the latent variables across Rasch-type and 2PL-type models. As before, the correlations $\rho_g(\theta_1, \theta_2)$ are estimated, but separate for each group. The group-specific Time Point 2 means that μ_{g2} are parameters of interest, since these represent the average expected growth in the Embretson model, understood as the growth over and above the average of the baseline proficiency μ_{g1}. In the Andersen model, a growth measure for each group is given by the difference $\varepsilon_g = \mu_{g2} - \mu_{g1}$, i.e., the average proficiency at Time Point 2 minus the average proficiency at Time Point 1. This is a derived quantity, not a model parameter, in the Andersen model.

Growth may follow different trajectories in different subpopulations g. Schools pace their curricula differently, and different types of schools may have dramatically different curricula. Even within seemingly homogeneous groups of learners, different trajectories may emerge, based on differences between students regarding how they acquire knowledge. Wilson's (1989) Saltus model addresses these different growth rates in different populations (Draney & Wilson, 2007;

TABLE 3.
Different levels of complexity of population models used as extensions of the Embretson- and Andersen-type longitudinal item-response-theory models.

Type of population model	Assumption about schools and types of schools
Single-group	All schools and types of schools have the same ability distribution and gain
Multiple-group	Each type of school has a potentially different distribution of student proficiencies and gain
Hierarchical-mixture	Different profiles of student proficiencies and gain exist, and different schools may have different prevalence for each profile (e.g., some schools have a larger proportion of fast learners than others)

Wilson & Draney, 1997), multidimensional IRT (MIRT) models may address them when allowing for multiple groups (Xu & von Davier, 2006), and latent growth-curve models address them in an IRT mixture (Meiser, Hein-Eggers, Rompe, & Rudinger, 1995; Rijmen, de Boeck, & Maas, 2005).

Different groups may be needed to represent differences in initial proficiency distributions and differences in the amount of change over time. A sample may contain a clustered sample of students from different school types, as in the case of the data used below. Each of these school types is characterized by potentially different (a) distributions of initial proficiency and (b) levels of change in proficiency, because schools differ in terms of curriculum and in the proficiency level of students entering these schools. A model for group-specific growth must be able to account for these potential differences. An exploration of whether the Matthew effect can be found in the PISA L data is facilitated by utilizing (a) multiple-group models using school-type as grouping variable and (b) mixture models in which populations with different growth rates are identified from observed patterns of student responses. We use a hierarchical-mixture model (Vermunt, 2003; von Davier, 2007b) in the application, since the data collection design of PISA L is a two-stage cluster sample with schools sampled within the country and students sampled within schools. The hierarchical GDM assumes that students within schools fall into one of several latent proficiency distributions while allowing for school-specific proportions. For example, in one school, 80% of students may fall into a class with high average ability and high gain, and the other 20% may fall into a class with low average ability and moderate gain. In another school, 50% of students may fall into a high-average, high-gain class, and the other 50% may fall into a low-average, moderate-gain class. The hierarchical GDM simultaneously estimates school-based cluster proportions and class-specific profiles and item parameters.

Table 3 shows the succession of model variants that we estimated and their main assumptions about the population structure. We estimated the model variants for longitudinal IRT models of both the Andersen (1985) and Embretson (1991) type of measuring growth.

The three types of population models were estimated using both the Andersen (1985) and the Embretson (1991) approaches. As previously noted, Andersen and Embretson originally developed their models as extensions of the Rasch model, and the operational model used in PISA is also based on the Rasch model. Therefore, we considered it necessary to evaluate the appropriateness of the Rasch model's assumption of constant discrimination across items. As a consequence, all model variants in Table 3 were estimated in two versions: (a) using a Rasch-model/PCM measurement model with constant slope parameter across items and (b) using a 2PL measurement model that allowed assessing whether different items should receive different discrimination parameters. Results were compared between the two versions (Rasch versus 2PL) and across the Andersen and Embretson models, as well as the three population-model variants. All together, six models (three population models, each in two longitudinal IRT approaches) were compared separately for the Rasch and the 2PL version in terms of model-data fit.

TABLE 4.
The number of schools, classrooms, and students in the study sample.

Type of school by instructional track	2003 assessment			2004 assessment		
	Schools	Classrooms	Students	Schools	Classrooms	Students
Lower secondary track (*Hauptschule*)	43	81	1,348	–	–	–
Lower and intermediate secondary track (*Realschule*)	23	46	932	22	33	653
Intermediate secondary track	51	101	2,535	50	98	2,199
Integrative school (*Gesamtschule*)	20	39	743	19	28	504
Higher secondary track (*Gymnasium*)	61	120	3,001	61	116	2,664
Total	198	387	8,559	152	275	6,020

Note. A *Hauptschule* (literally "general school") is basically a vocational school for Grades 5 through 9 or 10, a *Realschule* is a school for students ages 10–11 to 16–17, a *Gesamtschule* is a comprehensive school for students ages 11–16+, and a *Gymnasium* is a college-preparatory school.

4. Data and Analysis

4.1. Data Description

Through its Programme for International Student Assessment (PISA), the Organization for Economic Cooperation and Development (OECD) conducts international surveys of 15-year-olds to assess their academic skills. Since the first surveys began in 2000, the number of participating countries and the surveys' impact has increased. In Germany the 2003 assessment (OECD, 2003, 2004) was expanded to address several additional research questions, including student gains in proficiency over a school year (Prenzel, Carstensen, Schöps, & Maurischat, 2006). In addition to including a sample of 15-year-old students for international comparisons, the survey included a sample of ninth graders who were reassessed in 2004 in a study called PISA-I-Plus. This paper focuses on a longitudinal analysis of these students' "mathematical literacy" performance. Items for math literacy were developed in accordance with PISA's framework. The 2003 assessment used 77 items; the 2004 assessment used the same items plus 22 items unique to Time Point 2.

The sample used in this study is representative of ninth graders in Germany and includes all types of schools. The sample used in our study included all students promoted from Grade 9 to Grade 10. The project aimed at finding students who had moved to a different school, whenever possible, so these students could be included in the study. Table 4 gives the number of schools, classrooms, and students in each assessment. The analyses presented here are based on a sample of 6,020 students from 152 schools tested in both 2003 and 2004. The sampling design of PISA-I-Plus is a two-stage cluster: schools were selected in the first stage of the sampling process, and students within schools were selected in the second stage.

Previous PISA surveys, as well as other assessments, have shown that school type is the main source of between-cluster school differences. Germany's educational system places students into high, medium, and low academic tracks and, in some states, additional integrative schools with more heterogeneous student populations. As a result, different types of schools considerably

differ in students' average proficiency. As outlined above, our analysis took this fact into account by incorporating multiple-group models that reflect these differences.

The data were collected in the PISA study using a test-booklet design of four 30-minute blocks from different domains and a questionnaire administered after the test. In 2003, 13 different booklets were used, and in 2004, repeated and new items were recombined into 6 different booklets. Test questions were multiple-choice or required a short constructed response. All item responses were dichotomously scored. The PISA-I-Plus data set included items repeated over time and items unique to different time points. We carried out our analyses using the survey weights that were provided by the sampling organization for the assessments.

4.2. Analysis Plan

We chose the models developed by Andersen (1985) and Embretson (1991) as the basis for our analysis of the PISA-I-Plus math data in our study. By including items repeated over time, we were able to apply models that use these items as the anchor set; we therefore were able to link scales over time points. In addition, each student also received a small number of the 22 items unique to Time Point 2. Although there were 77 items in common across the two time points, each student was administered only a small number (slightly less than one third) of the common items due to the sparse matrix design of the test booklets.

The PISA-I-Plus math assessment design shares some features with designs for which Andersen (1985) and Embretson (1991) developed their models. As previously mentioned, Andersen developed his model for situations in which the same items are repeatedly administered over time, whereas Embretson developed her model for situations in which different item sets are administered on different occasions. However, by using a partial balanced incomplete block (pBIB) design (sometimes referred to as a "multimatrix design") and by assuming that item characteristics stay the same over time points, we can achieve a link between Time Points 1 and 2 that is based on 77 items out of a total of 99. Our model combines the features of not repeating items within a respondent, while repeating items over time to facilitate linkage enabled by an assessment design that administers different items to students at the different time points while a substantial portion of items have been repeatedly administered over time. The main dimension in our version of the Embretson model is not a time point-dependent dimension, but a dimension that "explains" the common part of the response variance across all time points. Therefore, items unique to time points 2 (and higher) can also load on the common dimension. Only the modifiabilities (change dimensions) are the ones unique to the time points 2 (and higher). Our analysis entails a multidimensional generalization of the 2PL/GPCM (Muraki, 1992) in addition to the Rasch versions and an extension of the growth model proposed by Andersen and Embretson.

We applied the three approaches given in Table 3 when modeling the proficiency distributions' dependence on type of school. As a baseline, we assumed no differences between school types leading to models with one (common) proficiency distribution. This baseline model was compared to a multiple-group version of the extended Andersen and Embretson models, where the groups represent potentially different proficiency distributions (over time) for each of the school types. The item parameters, however, are assumed to be the same across school types, so that the measurement model is the same, while the population distributions might be different for different school types. These models are then compared to a hierarchical-mixture distribution longitudinal IRT model.

5. Results

5.1. Model Fit

For comparison, we conducted our analysis under both a single-group and a multiple-group assumption. Under the former assumption, all students are assumed to come from a single pop-

TABLE 5.
The two-parameter logistic/generalized partial credit models: Akaike Information Criterion (AIC) and log likelihood.

Model	Item parameters	Skill distribution parameters	AIC	BIC	Log likelihood
Andersen					
Single-group	196 (198-2)	5	308,522.80	309,870.1	−154,060.40
Multiple school type	196 (198-2)	20	306,511.84	307,959.7	−153,039.92
Hierarchical-mixture	196 (198-2)	12	306,053.54	307,447.7	−152,818.77
Embretson					
Single-group	294 (297-3)	5	307,238.48	309,242.6	−153,320.24
Multiple school-type	294 (297-3)	20	305,230.58	307,335.3	−152,301.29
Hierarchical-mixture	294 (297-3)	12	304,187.88	306,238.9	−151,787.94

TABLE 6.
The Rasch-type models: Akaike Information Criterion (AIC) and log likelihood.

Model	Item parameters	Skill distribution parameters	AIC	BIC	Log likelihood
Andersen					
Single-group	98 (99-1)	5	310,838.94	311,529.37	−155,316.47
Multiple school type	98 (99-1)	20	308,853.82	309,644.85	−154,308.91
Hierarchical-mixture	98 (99-1)	12	308,399.82	309,137.13	−154,089.91
Embretson					
Single-group	98 (99-1)	5	310,839.12	311,529.53	−155,316.56
Multiple school type	98 (99-1)	20	308,855.08	309,646.27	−154,309.54
Hierarchical-mixture	98 (99-1)	12	308,398.78	309,136.15	−154,089.39

ulation with the same ability distribution. Under the latter assumption, students from different groups are assumed to come from different populations with potentially different ability distributions. For the data set used in this study, the groups are defined by (a) school types or (b) latent classes based on a hierarchical-mixture model (Vermunt, 2003; von Davier, 2007b, 2010). In the multiple-group and hierarchical-mixture analysis, we set the item parameters to be equal across groups. Therefore, we did not consider possible differential item functioning.

We conducted two sets of analysis, one using 2PL variants (Table 5) and the other using the Rasch model (Table 6). Each set included six models: (a) Andersen single-group, (b) Andersen school-type, (c) Andersen hierarchical-mixture, (d) Embretson single-group, (e) Embretson school-type, and (f) Embretson hierarchical-mixture. Tables 5 and 6 present the Akaike information criterion (AIC; Akaike, 1974), the BIC (Schwarz, 1978), and the log-likelihood for each of the models. The models will be evaluated mainly on the basis of the AIC and the log-likelihood, since the utility of the BIC has been questioned (Gilula & Haberman, 2001), and the two-stage cluster nature of the sample does not necessarily support the choice of $\ln(N)$ in the penalty term of the BIC.

Both tables also contain the number of parameters, separated into item parameters and parameters needed to fit the two-dimensional ability distribution(s). One two-dimensional distribution is fitted in the case of the single-group models, four bivariate distributions are fitted for the 4-population school-type models, and two bivariate distributions in the case of the 2-population hierarchical-mixture models. We fitted two moments for each ability dimension in each population and one parameter for the covariance between the two dimensions, which results in 5 parameters fitted for each of the bivariate ability distributions in the estimated models—higher-order moments (see Xu & von Davier, 2008) were not estimated. The hierarchical discrete mixture

required two additional parameters for estimation of the Dirichlet distribution of the class sizes (see von Davier, 2007b, 2010).

Of the models shown in Table 5, the hierarchical-mixture model (with two mixture components and School as the clustering variable to account for the two-stage cluster sampling) using the Embretson-type measurement of growth, assuming two different latent growth trajectories, has the smallest AIC and the largest log-likelihood.[1] This model shows a slightly better fit in terms of AIC and log-likelihood than the Embretson-type multiple-group model using school type as the grouping variable. Note that the single-group models that do not allow for group differences in average initial ability and growth do not perform well compared to the models with multiple (observed or latent) ability distributions.

Table 6 shows goodness-of-fit information for the Rasch-type measurement models that were estimated in conjunction with the same set of single-group, school-type multiple group, and hierarchical-mixture population models.

A comparison of the data in Tables 5 and 6 shows that the Rasch-type models fit worse than their 2PL counterparts do. Note that the fit of the Embretson and Andersen models is almost identical when looking at the same population model specification for the Rasch-type models. This is an expected result, as it was shown in the section introducing a common modeling framework above that Andersen and Embretson models are equivalent when estimated as Rasch-type models. The small differences observed may result from slight differences with regard to when the algorithm terminated iterations.

Note that the models using the 2PL as the basis for analysis show improved model fit, judging based on AIC, over their Rasch model counterparts. However, within the group of Rasch-type models, the hierarchical-mixture model again shows the relatively best model-data fit. The reason for the hierarchical-mixture model's improvement of fit over the multiple-group model might be that the observed classification into school types does not completely reflect the actual differences in average proficiency and growth across schools. The hierarchical-mixture model allows students within schools to be unmixed into the components of a mixture distribution, and this might reflect more appropriately than the observed type of school that students from different schools grow at different rates.

The reason for the superior performance of the 2PL type models may be that some, but not all, of the tasks readministered at Time Point 2 were affected by the growth of student skills modeled using the second dimension. If that is the case, some items that show reasonable discrimination parameters for Dimension 1 (overall proficiency) may lack a significant loading on Dimension 2 (in the Embretson approach, the change dimension). If so, those items would be poorly represented by the Rasch-type approach, in which all items receive the same discrimination parameter.

5.2. Latent Growth Measure

By design, the Embretson model describes a base ability for each person by defining the main dimension to involve all items across time. Also, the items at Time Point 2 are related to a second dimension unique to growth at this time point. However, this dimension measures only what cannot be explained by the baseline ability, so that it has inherently lower reliability than the first (main) dimension; therefore, the change dimension does not seem suitable for reporting growth for individual students. In contrast, measures of group-level distributions can reliably

[1]The fit can be slightly improved by using more than two mixture components. We ran the hierarchical mixture model for the 2PL Embretson case with 2, 3, and 4 mixture components. The results are only marginally better in terms of the AIC per response. When comparing the hierarchical mixture model to a nonhierarchical version, no more than two mixture components can be found. Therefore, and for the sake of (relative) simplicity of presentations, we restrict our exposition to the results of the two-component mixture, which already fits slightly better than the school-type model.

TABLE 7.
Means and standard deviations of multiple groups for the Embretson-type model.

Model		Multiple group model: School type				Hierarchical mixture	
		Low	Medium	Integr.	High	Class 1	Class 2
Main dimension	Mean	−0.896	−0.451	−1.063	0.443	−1.041	0.417
	s.e.	0.035	0.022	0.043	0.020	0.034	0.026
	SD	0.792	0.822	0.968	0.814	0.637	0.690
	s.e.	0.031	0.016	0.041	0.014	0.017	0.016
Change dimension	Mean	0.529	0.491	0.461	0.542	0.523	0.637
	s.e.	0.028	0.018	0.030	0.023	0.021	0.020
	SD	0.368	0.324	0.318	0.324	0.275	0.306
	s.e.	0.025	0.016	0.032	0.014	0.015	0.010
Correl.	Estimate	0.028	0.075	0.172	−0.167	0.197	−0.081
	s.e.	0.080	0.044	0.097	0.051	0.065	0.049

Note. s.e. = standard error.

indicate average growth even if individual measures are noisy due to considerable measurement error (Mislevy, 1991). In the Embretson model, growth at the group level can be identified by the group mean on the growth dimension θ_2, as shown in Table 7. To set the scale in the 2PL variant of the Embretson MIRT model, we constrained the average difficulty to a mean of zero and the average slope to 1.0 for both dimensions, and constrained the item parameters to be the same across groups.

The Embretson model yields the estimates for the school-type-based multiple-group model and hierarchical-mixture model shown in Table 7. Shown in the table are the estimates for the mean and standard deviation of the main and change dimensions, as well as the estimated correlations between these by school type and latent class. All estimates are accompanied by standard errors. The standard errors are based on the jackknife resampling procedure described in Hsieh, Xu, and von Davier (2009). Both the hierarchical-mixture and the multiple-group models are presented in Table 7 as they are the two relatively best fitting models, and because school-type based comparisons are of importance for policy makers. In addition, the fit of the school-type multiple-group model is only marginally inferior to the hierarchical-mixture model when looking at the Embretson 2PL-based models.

For the multiple-group model, the students in the high-track school are high performing in the main dimension, and they show the largest improvement at Time Point 2 (Table 7). The students in the integrative school are lowest performing in the main dimension; they also show the least improvement at Time Point 2. The students at the medium and low academic track schools show similar improvement at Time Point 2 compared to the "high" academic track school. Across school types, correlations between common dimension and time-point-specific dimension are negligible when evaluating these in light of their estimated standard error.

For the hierarchical-mixture model, the results show one class with high-average baseline ability (Class 2) that is similar to the baseline ability of the "high" school-type in the multiple-group model, and one class with lower average baseline ability (Class 1) close to the baseline average ability of the "integrative" school type. The standard deviations of both latent classes in the hierarchical-mixture model are somewhat lower than the conditional standard deviations in the school-type multiple-group model. This is due to the fact that the mixture model partitions the data into more homogeneous groups compared to an observed variable (school-type) that may not completely coincide with differences between status and growth trajectories. Recall that the resulting somewhat different profiles are based on the hierarchical-mixture model that shows

TABLE 8.
Means and standard deviations of multiple groups for the Andersen-type model.

Model		Multiple group model: School types				Hierarchical mixture	
		Low	Medium	Integr.	High	Class 1	Class 2
First dimension	Mean	−0.836	−0.370	−1.005	0.563	−0.934	0.569
	s.e.	0.042	0.021	0.046	0.020	0.034	0.034
	SD	0.845	0.837	0.977	0.866	0.635	0.747
	s.e.	0.030	0.020	0.048	0.018	0.027	0.021
Second dimension	Mean	−0.339	0.048	−0.590	0.982	−0.534	1.008
	s.e.	0.039	0.023	0.047	0.021	0.043	0.028
	SD	0.873	0.930	1.084	0.801	0.738	0.671
	s.e.	0.030	0.019	0.044	0.016	0.031	0.020
Corr.	Estimate	0.862	0.885	0.960	0.891	0.862	0.835
	s.e.	0.028	0.012	0.009	0.011	0.015	0.020
Growth	Mean	0.497	0.419	0.415	0.419	0.400	0.440
	s.e.	0.030	0.018	0.039	0.020	0.028	0.020

Note. s.e. = standard error.

the relatively best fit among the models compared here. This model divides students into two groups: Class 1 is low-performing with moderate growth between Time Points 1 and 2 and Class 2 is high-performing with larger growth between Time Points 1 and 2. The class size of the high performing Class 2 across schools is 54.1%. For the low performing school, 14.3% are expected in the high performing class, for the medium track school type, the value is 37.7%, and for the integrated schools, it is 22.2%, while for the high track schools, it is 91.4%.

Table 8 shows the average ability estimates, standard deviations, estimated of growth and correlation between scales for the Andersen-type 2PL multiple-group and hierarchical-mixture models. Even though these two models do not fit as well as their Embretson-type counterparts do, their results are of some interest since the constraints used in the Andersen approach may make this model seem the more parsimonious way of measuring growth: Recall that the Andersen model can be viewed as a special case of the Embretson model that is obtained when the slope parameters of the second dimension are constrained to equal those of the first dimension. For the Andersen model, the difference between the averages of Dimensions 1 and 2 can be viewed as a measure of group-level growth. Table 8 shows the means by school type and class, the correlations between scales by school type and class, and the average growth measures under the Andersen-type 2PL model.

Note that the direction of growth is consistent between the Embretson and Andersen models, but the ranking of school types by average growth is not. For example, under the Andersen school type model, low-track schools show the largest growth, whereas under the Embretson model, high academic track schools do. For the hierarchical-mixture model, the results are more similar to the corresponding Embretson model. The latent class that performs higher on Dimension 1 (Class 2 in the Embretson and in the Andersen hierarchical mixtures) shows a larger growth parameter than the initially lower performing class (Class 1). Correlations between time point specific factors are ranging from about 0.84 to 0.96 across observed and latent populations.

Recall that the Embretson 2PL hierarchical mixture shows slightly better fit than the multiple-group model, and among the Andersen-based models, the hierarchical-mixture model also shows the relatively best fit. Both hierarchical models show larger growth for the latent class

with higher baseline (initial) ability, compared to the other class with lower average baseline (initial) ability. An inspection of the resulting parameters reveals that the way the Embretson model is specified allows some items to receive slope parameters close to zero for the loadings on the change dimension, whereas other slope parameters quantify how much the conditional-response probabilities of the items assessed at Time Point 2 depend on the second (growth) dimension in our model. As mentioned above, the average of slope parameters was constrained to be 1.0 for both dimensions to make the estimates comparable to the Rasch model case and comparable across dimensions. It turns out that this resulted in differences between the variance for baseline dimension and change dimension. More specifically, for the hierarchical-mixture model the estimated standard deviation for the baseline dimension is about 0.6–0.7, while it is about 0.3 for the change dimension. At the same time, while the average slopes for the two dimensions were constrained to be the same, the standard deviation of slope parameters varies substantially between the two dimensions. That is, the standard deviation of slope parameters is 0.31 for the baseline dimension and 0.77 for the change dimension, indicating that the improved fit is due to allowing the slopes for the growth dimension to vary, while the ability variance of the growth dimension is smaller than for the baseline dimension. Several slope parameter estimates are indeed close to 0.0, and a few are even slightly negative for the growth dimension indicating that the responses on these items are unaffected by growth, while other slope parameters are as large as 2.5 and above, indicating that the responses to these items are substantially affected by growth. The extended Andersen model, on the other hand, cannot fit these differential patterns, since in this approach the same items given at different time points have exactly the same parameters.

6. Discussion

In this study, we examined and extended models for measuring change in longitudinal designs within the context of IRT: (a) the Andersen model that assumes a unique dimension per testing occasion and (b) the Embretson model that assumes a baseline dimension across all testing occasions, starting with the first, and additional change dimensions unique to subsequent occasions. We extended both models via the GDM framework (von Davier, 2005) for multiple populations (Xu & von Davier, 2006) and for clustered data (von Davier, 2007b). These extensions give researchers tools to examine growth in multiple observed and latent populations. In the multiple-group longitudinal IRT models, different observed populations are assumed to have potentially different patterns of growth. In the hierarchical-mixture version of this model, groups with differential growth may exist within each cluster, while every cluster may contain a different proportion of these groups characterized by a specific growth trajectory. An important example of differential growth is the observation that led to the description of the Matthew effect: Learners starting out at a higher level gain more on average than learners starting at a lower level of proficiency.

The Matthew effect was studied by applying the developments presented in this paper in an application to a longitudinal large-scale survey assessment indicated that the Embretson-type 2PL model, extended to a hierarchical-mixture MIRT model to account for variance between school types, fits the data best. Therefore, this paper's main findings are based on this model. They are supported by corresponding findings that we estimated from other models.

The two identified growth trajectories on the basis of the best fitting model can be described as a high-performance, high-gain trajectory and a low-performance, moderate-gain trajectory. This finding is in agreement with the Matthew effect in education (Stanovich, 1986; Walberg and Tsai, 1983). The same pattern was found for the relatively best fitting model among

the Andersen-type models. However, this finding could only be partially confirmed for the Embretson approach when estimating a multiple-group model using the observed classification of schools into types of varying academic performance. Therefore, these results have to be taken with a grain of salt: If we allow an algorithm to "unmix" students within schools into homogeneous groups that are conforming more to a model while assuming that the same measurement model holds across groups, an increase in homogeneity will potentially be achieved by reducing variance within groups. This will likely lead to latent classes consisting of quite similar observations within groups and will at the same time increase differences across groups. This may have contributed to a homogenization of groups with respect to both initial level and expected growth. Note however, that the Embretson model also showed distinct differences in growth between the academically lowest performing school type at the baseline and the highest performing school type at the baseline. At least for these two extreme school types, the Matthew effect was also found for the multiple-group model that showed slightly inferior model data fit compared to the discrete mixture model.

Another issue that requires some further thought is the question of what the two different growth measures mean from a content perspective. The extended Andersen approach measures change by assuming that the items at Time Point 2 and higher require exactly the same ability dimension; but, if growth occurred, the students would move up in terms of average achievement. Note that the correlation between the two time-point specific dimensions in the Andersen-type models was 0.86 and higher (within school-types and within latent classes). This shows that there is a substantial amount of common variance across time points. This high consistency over time suggests that a single dimension could be assumed across occasions. The Embretson model does exactly that by assuming a common dimension across time points plus a specific dimension for all but the baseline. The time-point-specific dimension explains what is not common to all occasions. This approach allows estimation of loadings (slopes) on the time-point-specific dimensions and is more general than the Andersen approach. The unique dimension in the Embretson approach explains response variance that cannot be explained by means of the underlying common factor. In that sense the extended Embretson approach explains what is "new" (unique) to the response behavior of students up and above the results expected when assuming homogeneous growth.

This paper presented analytical tools that allow stakeholders and policymakers to quantify changes in different groups assessed in longitudinal large-scale surveys. At the same time, the multiple-group Embretson-type GDM involves a design matrix and parameters that can be constrained to be the same across groups (as in the example presented here) or specific to the groups assessed so that the items that are more sensitive to growth can be identified. Future assessment cycles can target specific areas of the proficiency domain that are of interest in assessing change in proficiency over time.

The question of what the right model for growth might be is, to some extent, futile. Models will fit a given dataset more or less well, and more general models will (generally) fit data better than more restrictive models. The Andersen and the Embretson models address similar questions; and, indeed, the Andersen model can be transformed into a reparameterized, constrained version of the Embretson model. The Andersen parameterization allows assessments of how much a student grows only relative to the initial state, while the Embretson approach directly parameterizes an average growth measure. The evaluation of the differences between models showed that the implied relative magnitude of growth agreed with the Matthew effect for both the Andersen and Embretson models when looking at the relatively best fitting population model, the hierarchical-mixture model.

Finally, the proposed hierarchical-mixture and multiple-group general diagnostic models used in this study to implement Embretson-type (and Andersen-type) growth models can be with ease extended to three or more time points. Beyond 4 or 5 time points, the computational burden should be reduced through the use of efficient estimation methods such as the ones described by Rijmen (2009), von Davier and Sinharay (2007, 2010), or Cai (2010).

References

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

Andersen, E.B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16.

Andrade, D.F., & Tavares, H.R. (2005). Item response theory for longitudinal data: population parameter estimation. *Journal of Multivariate Analysis*, *95*, 1–22.

Bock, R.D., & Zimowski, M.F. (1997). Multiple group IRT. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York: Springer.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*, 33–57.

Draney, K., & Wilson, M. (2007). Application of the Saltus model to stage-like data: some applications and current developments. In M. von Davier, & C.H. Carstensen (Eds.), *Multivariate and mixture distribution rasch models* (pp. 119–130). New York: Springer.

Embretson, S.E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515.

Embretson, S.E. (1997). Structured ability models in tests designed from cognitive theory. In M. Wilson, G. Engelhard, Jr., & K. Draney (Eds.), *Objective measurement: theory into practice* (Vol. 4, pp. 223–236). Greenwich: Ablex.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Fischer, G.H. (1976). Some probabilistic models for measuring change. In D.N.M. de Gruijter, & L.J.T. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97–110). New York: Wiley.

Fischer, G.H. (1995). Some neglected problems in IRT. *Psychometrika*, *60*, 459–487.

Fischer, G.H. (2001). Gain scores revisited under an IRT perspective. In A. Boomsma, M.A.J. Van Duijn, & T.A.B. Snijders (Eds.), *Essays on item response theory* (pp. 43–68). New York: Springer.

Gilula, Z., & Haberman, S.J. (2001). Analysis of categorical response profiles by informative summaries. *Sociological Methodology*, *31*, 193–211.

Glück, J., & Spiel, C. (1997). Item response models for repeated measures designs: application and limitations of four different approaches. *Methods of Psychological Research Online*, *2*(1), 1–18. Retrieved March 12, 2009, from http://www.dgps.de/fachgruppen/methoden/mpr-online/issue2/art6/article.html.

Hsieh, C., Xu, X., & von Davier, M. (2009). Variance estimation for NAEP data using a resampling-based approach: an application of cognitive diagnostic models. In M. von Davier, & D. Hastedt (Eds.), *IERI monograph series: Vol. 2. Issues and methodologies in large scale assessments* (pp. 161–174). Hamburg/Princeton: IEA-ETS Research Institute.

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

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

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

Meiser, T., Hein-Eggers, M., Rompe, P., & Rudinger, G. (1995). Analyzing homogeneity and heterogeneity of change using Rasch and latent class models: a comparative and integrative approach. *Applied Psychological Measurement*, *19*(4), 377–391.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–177.

Organisation for Economic Co-operation and Development (2003). *The PISA 2003 assessment framework: mathematics, reading, science and problem solving knowledge and skills*. Paris: Author.

Organisation for Economic Co-operation and Development (2004). *Learning for tomorrow's world: first results from PISA 2003*. Paris: Author.

Prenzel, M., Carstensen, C.H., Schöps, K., & Maurischat, C. (2006). Die Anlage des Längsschnitts bei PISA 2003 [The design of the longitudinal PISA assessment]. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand et al. (Eds.), *PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres [Studies on the development of competencies over the course of a school year]* (pp. 29–63). Münster: Waxmann.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Rijmen, F. (2009). *Efficient full information maximum likelihood estimation for multidimensional IRT models* (ETS Research Report No. RR-09-03). Princeton, NJ: ETS.

Rijmen, F., de Boeck, P., & Maas, H. (2005). An IRT model with a parameter-driven process for change. *Psychometrika*, *70*, 651–669.

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

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

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407.

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

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report No. RR-05-16). Princeton, NJ: ETS.

von Davier, M. (2007a). *Mixture general diagnostic models* (ETS Research Report No. RR-07-32). ETS: Princeton, NJ.

von Davier, M. (2007b). *Hierarchical mixtures of diagnostic models* (ETS Research Report No. RR-07-19). ETS: Princeton, NJ.
von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287–307.
von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling*, *52*(1), 8–28. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2010/02_vonDavier.pdf.
von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G.H. Fischer, & I.W. Molenaar (Eds.), *Rasch models: foundations, recent developments, and applications* (pp. 371–379). New York: Springer.
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. 643–661). Amsterdam: Elsevier.
von Davier, M., & Sinharay, S. (2007). An importance sampling EM algorithm for latent regression models. *Journal of Educational and Behavioral Statistics*, *32*(3), 233–251.
von Davier, M., & Sinharay, S. (2010). Stochastic approximation for latent regression item response models. *Journal of Educational and Behavioral Statistics*, *35*(2), 174–193.
von Davier, M., & von Davier, A. (2007). A unified approach to IRT scale linkage and scale transformations. *Methodology*, *3*(3), 115–124.
von Davier, M., & Yamamoto, K. (2004). A class of models for cognitive diagnosis. Paper presented at the *4th Spearman invitational conference*, October, Philadelphia, PA.
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* (pp. 99–115). New York: Springer.
Walberg, H.J., & Tsai, S.-L. (1983). Matthew effects in education. *American Educational Research Journal*, *20*, 359–373.
Wilson, M. (1989). Saltus: a psychometric model for discontinuity in cognitive development. *Psychological Bulletin*, *105*, 276–289.
Wilson, M., & Draney, K. (1997). Partial credit in a developmental context: the case for adopting a mixture model approach. In M. Wilson, G. Engelhard, Jr., & K. Draney (Eds.), *Objective measurement: theory into practice* (Vol. 4, pp. 333–350). Greenwich: Ablex.
Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data* (ETS Research Report No. RR-06-08). Princeton, NJ: ETS.
Xu, X., & von Davier, M. (2008). *Comparing multiple-group multinomial loglinear models for multidimensional skill distributions in the general diagnostic model* (ETS Research Report No. RR-08-35). Princeton, NJ: ETS.

Manuscript Received: 30 MAR 2009
Final Version Received: 20 AUG 2010
Published Online Date: 2 FEB 2011