Research Report

Using the General Diagnostic Model to Measure Learning and Change in a Longitudinal Large-Scale Assessment

Matthias von Davier

Xueli Xu

Claus H. Carstensen

July 2009

ETS RR-09-28

Using the General Diagnostic Model to Measure Learning and Change in a Longitudinal Large-Scale Assessment

Matthias von Davier and Xueli Xu
ETS, Princeton, New Jersey

Claus H. Carstensen
University of Bamberg, Germany

July 2009

Abstract

A general diagnostic model was used to specify and compare two multidimensional item-response-theory (MIRT) models for longitudinal data: (a) a model that handles repeated measurements as multiple, correlated variables over time (Andersen, 1985) and (b) a model that assumes one common variable over time and additional orthogonal variables that quantify the change (Embretson, 1991). Using MIRT-model ability distributions that we allowed to vary across subpopulations defined by type of school, we also compared (a) a model with a single two-dimensional ability distribution to (b) extensions of the Andersen and Embretson approaches, assuming multiple populations. In addition, we specified a hierarchical-mixture distribution variant of the (Andersen and Embretson) MIRT models and compared it to all four of the above alternatives. These four types of models are growth-mixture models that allow for variation of the mixing proportions across clusters in a hierarchically organized sample. To illustrate the models presented in this paper, we applied the models to the PISA-I-Plus data for assessing learning and change across multiple subpopulations. The results indicate that (a) the Embretson-type model with multiple-group assumptions fits the data better than the other models investigated, and (b) the higher performing group shows larger improvement at Time Point 2 than the lower performing group.

Key words: Item response theory, growth models, multidimensional IRT, longitudinal models, diagnostic models, large scale assessments

Introduction

Measurement of change in student performance between testing occasions is a central topic in educational research and assessment (Fischer, 1995). Most research on such measurement 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 on either the group (Andersen, 1985; Andrade & Tavares, 2005; Fischer, 1973, 1976) or the individual (Embretson, 1991; Fischer, 1995).

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 *v* on item *i* can be written as

$$P(x_{vi}=1)=\frac{\exp(\theta_v-\beta_i)}{1+\exp(\theta_v-\beta_i)},$$

in which θ_v is the person's ability and β_i is the item'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 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 *J* items given at *T* time points in *G* treatment groups, a group-specific model for growth can be specified in the LLTM using

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

in which the effects from α_1 to α_J are the baseline item difficulties, α_{J+1} is the effect of Time Point 2, α_{J+T-1} is the effect of time point *T*, $\alpha_{J+T-1+1}$ is the effect of Group 2, and $\alpha_{J+T-1+G-1}$ is the effect of Group *G*. 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 (Fischer, 1973). 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_v and a class membership c_v that represent their developmental stage. The classes (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(i)* of examinee *i*. The equation for the Saltus model is

$$P(X_{ij}=1\,|\,\theta_j,b_j)=\frac{\exp(\theta_j-b_i+\tau_{c(j)k(i)})}{1+\exp(\theta_j-b_i+\tau_{c(j)k(i)})} \qquad (2)$$

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 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.

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:

$$P(X_{i(k)j} = 1 | \theta_{jk}, b_j) = \frac{\exp(\theta_{jk} - b_i)}{1 + \exp(\theta_{jk} - b_i)} \tag{3}$$

where θ_{jk} is the ability of person *j* at occasion *k*, and b_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 ability variables that might be correlated. In Andersen's model, abilities are specific to occasions; they do not quantify change but 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.

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

$$\begin{aligned} &P(X_{ijk} = 1 | \theta_{jk}, \varsigma_i) = c_i + (1 - c_i)\frac{\exp[a_i(\theta_{jk} - b_i)]}{1 + \exp[a_i(\theta_{jk} - b_i)]}, \\ &\theta_j = (\theta_{j1}, \theta_{j2}, ..., \theta_{jK})^T \sim MVN_K(\mu, \Sigma) \end{aligned} \tag{4}$$

in which θ_{jk} and b_i are defined as in Equation 3, a_i and c_i are the discrimination and guessing parameters in traditional three-parameter logistic models (Lord & Novick, 1968), and $MVN_k(\mu, \Sigma)$ 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 *M* abilities in item responses within *K* occasions. Specifically, the MRMLC assumes that (a) on the first occasion (*k* = 1) only an initial ability is involved in the item responses and (b) on later occasions (*k* > 1), that 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

$$P(X_{i(k)j} = 1 \mid (\theta_{j1},..,\theta_{jk}), b_j) = \frac{\exp(\sum_{m=1}^{k} \theta_{jm} - b_i)}{1 + \exp(\sum_{m=1}^{k} \theta_{jm} - b_i)}, \qquad (5)$$

in which θ_{jm} and b_i are defined as in Equation 3.

Note that same items are repeated over occasions in Andersen's (1985) model. In contrast, Embretson (1991) developed MRMLC for situations in which items are not repeated, to avoid well-known effects (e.g., practice effects and memory effects) of repeated item presentation and local dependency among item responses. Equation 5 indicates that in such situations, for an item *i* observed at time *k*, the abilities up to time *k* are involved. Therefore, an item observed at time *k* measures *k* abilities, including initial ability (θ_{j1}) and *k* – 1 time-point-specific abilities ($\theta_{j2},...,\theta_{jk}$), termed “modifiabilities” in Embretson's model. The change between condition *k* – 1 and *k* equals the *k*th modifiability (θ_{jk}). Using the partial credit model (PCM), Fischer (2001) extended MRMLC to polytomous items (Masters, 1982).

Measuring Change in Multiple Populations

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

In this study, we specified the Andersen (1985) and Embretson (1991) approaches using a framework of general latent-variable modeling. We used von Davier's (2005) general diagnostic model (GDM) to implement these approaches with extensions that allow for (a) the use of more-general IRT measurement models and (b) more-complex population structures.

The GDM allowed a multidimensional generalization of (a) the two-parameter logistic (2PL) model and (b) 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.

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). 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.

Within the framework of 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 Andersen (1985) approach, the Embretson (1991) approach, and our adaptations and generalizations (described below) within one framework. Group differences in skills distribution may exist. Therefore, to allow for differences in proficiency distributions and in amount of change across student groups, we used a single-group Andersen-Embretson type of model for change as well as other Andersen-Embretson approaches to represent multiple populations.

The multiple-population approach requires the introduction of an additional variable, the indicator function for the group membership variable, 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, probabilities of group membership $\pi_c(j)=P(c|X)$ may be used instead and 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

$$P(X_{jik} = x \mid \underline{q}_i, \underline{\theta}_j, \underline{\underline{\gamma}}_i, \underline{\underline{\beta}}_i) = \sum_{c=1}^{G} \pi_c[j] \frac{\exp\left(\sum_{l=1}^{k} x q_{il} \gamma_{icl} \theta_{jl} - \beta_{xic} \right)}{1 + \sum_{y=1..m_i} \exp\left(\sum_{l=1}^{k} y q_{il} \gamma_{icl} \theta_{jl} - \beta_{yic} \right)}. \tag{6}$$

Different groups may be needed to represent differences in initial proficiency distributions and differences in the amount of change over time. A data set may contain students from different school types, as in the case of the actual sample data 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 the proficiency level of students entering these schools. That implies that a model must be able to account for these potential differences. We used (a) multiple-group models in which the school of student *j* determined the student's membership in one school-type group as well as (b) hierarchical-mixture models in which membership in populations with different growth rates and potentially different initial-ability distributions was not determined by an observed variable but was inferred from observed patterns of student responses and from students' school (cluster) membership.

The most complex approach used in this study was a hierarchical extension of a longitudinal IRT model. This approach is based on the hierarchical GDM (von Davier, 2007) and assumes that students within schools fall into one of several proficiency distributions with school-specific proportions. It takes the hierarchical structure into account based on the identification of schools as clusters, and it attributes existing between-school differences to a mixture of students from several different proficiency distributions being present in each cluster.

This model, the hierarchical GDM, extends the hierarchical latent-class model (Vermunt, 2003) and allows clusters to vary in the proportions of mixture components (different ability distributions) represented in each class. This hierarchical approach allows for (a) differences between classes of proficiency distributions and (b) within-class variance of proficiencies. It attributes cluster-level variation to between-school differences in proportions of students belonging to the several proficiency distributions. 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 (a) these school-based cluster proportions, (b) the class-specific profiles, and (c) the item parameters, which are assumed to be the same across the latent classes defined by different ability and gain distributions.

Table 1 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.

Table 1

Different Complexity Levels 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.
Mixture	Different profiles of student proficiencies and gain (e.g., fast learners and slow learners) exist independently of schools and types of schools.
Hierarchical-mixture	Different profiles of student proficiencies and gain exist, and different schools have different prevalences for each profile (e.g., some schools have a larger proportion of fast learners than others).

Of the four types of population model, we estimated all but the mixture model for both the Andersen (1985) and the Embretson (1991) approach because there was no reason to assume that the different profiles of proficiency and gain were equally distributed across schools. As previously noted, Andersen and Embretson originally developed their models as extensions of the Rasch model. We considered it necessary to evaluate the appropriateness of the Rasch model's assumption of constant discrimination across items. Therefore, we estimated all approaches in two versions: (a) as a Rasch-model/PCM extension with constant slope parameter across items and (b) as a 2PL GPCM extension that allowed us to assess whether different items should receive different discrimination parameters. We compared results (a) between the two versions (Rasch versus 2PL), (b) across the Andersen and Embretson models, and (c) between the three population-model variants. We compared the six models (three population models, each in two longitudinal IRT approaches) separate for the Rasch and the 2PL version in terms of model-data fit based on a longitudinal data set collected as part of a study conducted in conjunction with an international survey of student skills.

Data and Analysis

Data Description

Through its Programme for International Student Assessment (PISA), the Organisation for Economic Co-operation and Development (OECD) conducts annual 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 an longitudinal analysis of these students' "mathematical literacy" performance. Items for math literacy were developed in accordance with PISA's framework and the Grade 10 math curriculum. The 2003 assessment used 77 items; the 2004 assessment used the same items plus 22 more.

The sample used in this study is representative of ninth graders in Germany and includes all types of schools. Our study included all students promoted from Grade 9 to Grade 10. We also tried to find students who had moved to a different school, whenever possible, so we could include their data. The sample of students tested at both times is not representative for 10th grade students. Therefore, all results refer to our sample of students. Table 2 gives the number of schools, classrooms, and students in each assessment. Our analyses were 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 types are the main sources 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. Our analysis took this fact into account by incorporating multiple-group models that reflect the 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, 6 different booklets were used. 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 as well as items unique to different time points. We computed our analyses using the same survey weights for data from both assessments.

Table 2

Number of Schools, Classrooms, and Students in the Study Sample

	2003 assessment			2004 assessment		
Type of school by instructional track	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.

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 used in this study. By including items repeated over time, we were able to apply models that use these items as the anchor set; we therefore could link scales over time points. items unique to different time points. Although there were 77 items in common across the two time points, each student was administered only a small number of these common items over time.

The PISA-I-Plus math data share some features with data 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 "multi-matrix 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. Therefore, neither model as originally specified is completely appropriate for data collected with a pBIB.

However, in large-scale survey assessments, the focus is on group-level differences in skills between subgroups of interest, not in change at the level of the individual. We used aggregates of change based on groups of students so that the effects of potential individual biases would be greatly reduced by cancellation effects when reporting group-level proficiency distributions and measures of change.

As indicated above, an Andersen-type model can be used within a GDM framework (von Davier, 2005). In the Andersen approach, items that appear at different time points (occasions) are assumed to represent different dimensions. Also, items repeatedly administered across time points are assumed to have the same item parameters and therefore act as linking items across measurement occasions. This constraint enables construction of a common scale for comparisons between the two dimensions defined by Time Points 1 and 2. Table 3 shows the structure of the Andersen-type model within a GDM framework.

Table 3

Andersen-Type Model Within a General Diagnostic Model (GDM) Framework

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

We also applied a GDM framework to the Embretson-type model. Within such a framework, the entire set of items from both time points is specified to measure a single main dimension. The items that reappear at Time Point 2 are specified to measure a second dimension that is assumed to be uncorrelated with the main dimension. With regard to parameters that refer to the main dimension, items common to both time points are assumed to have the same item-parameter values across the two time points. Table 4 shows the structure of the Embretson-type model under the linkage design within a GDM framework.

Table 4

Embretson-Type Model Within a 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

As mentioned before, the PISA-I-Plus situation differs from those for which Embretson (1991) developed her model. First, whereas Embretson’s data set did not contain common items over time, the PISA-I-Plus data set contains a large proportion of common items. Second, whereas Embretson’s MRMLC allows correlations between dimensions, the application shown in Table 4 assumes no correlations between the two dimensions. Third, our analysis entails a multidimensional generalization of the 2PL GPCM (Muraki, 1992) in addition to the Rasch versions of the above model variants. The Rasch model served as the basis for Embretson’s model. In our study, the average score on the second dimension accounts for differences in change over time.

To reflect the fact that a relatively large proportion of the items were presented at two time points, we incorporated various constraints on item parameters. In the Andersen (1985) approach, all items that were presented on both occasions were assumed to have the same item parameters. In the Embretson (1991) approach, the repeated items received the same item difficulty, and the same discrimination parameter on the main dimension, as the first occurrence of that item. The second discrimination parameter is unique to the second occasion; it consequently is an unconstrained parameter in our extension of Embretson's model.

We applied three approaches when modeling the proficiency distributions' dependence on type of school. As a baseline, we assumed no differences between schools leads to models with one common proficiency distribution. This baseline model was compared to a multiple-group version of the 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 mixture distribution longitudinal IRT model. The last model in the comparisons is an approach that takes the hierarchical structure into account.

Results

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 population 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 class derived from school types. In the multiple-group 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 GPCM variants (Table 5) and the other using the Rasch model (Table 6). Each set included six types of analysis: (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 show the goodness-of-fit index, in the form of the Akaike information criterion (AIC; Akaike, 1974), and the log likelihood for each model.

Table 5

Two-Parameter Logistic/Generalized Partial Credit Models: Akaike Information Criterion (AIC) and Log Likelihood

Model	Number of parameters	AIC	Log likelihood	AIC per response
Andersen				
Single-group	213	308,525.93	–154,049.97	0.5324
Multiple school type	244	30,6637.17	–153,074.58	0.5284
Hierarchical-mixture	224	306,805.36	–153,178.68	0.5295
Embretson				
Single-group	319	307,281.50	–153,321.75	0.5295
Multiple school-type	324	305,348.21	–152,350.11	0.5262
Hierarchical-mixture	326	305,540.05	–152,444.02	0.5276

Table 6

Rasch-Type Models: Akaike Information Criterion (AIC) and Log Likelihood

Model	Number of parameters	AIC	Log likelihood	AIC per response
Andersen				
Single-group	121	310,884.51	–155,321.26	0.5357
Multiple school type	143	308,863.21	–154,288.61	0.5322
Hierarchical-mixture	123	309,038.75	–154,396.37	0.5330
Embretson				
Single-group	127	310,930.45	–155,338.23	0.5358
Multiple school type	128	308,935.25	–154339.62	0.5328
Hierarchical-mixture	130	309,111.00	–154,425.50	0.5331

Of the models shown in Table 5, the Embretson-type model with multiple-groups has the smallest AIC and the largest log likelihood. Also, it shows a slightly better fit in terms of AIC and log likelihood than the Embretson-type hierarchical-mixture model (with school type as the clustering variable).

A comparison of the data in Tables 5 and 6 shows that the Rasch-type models have worse fit than their 2PL GPCM counterparts. Although Table 5 shows that the Andersen-type 2PL GPCM has worse fit than the Embretson-type 2PL GPCM in terms of AIC and log likelihood, Table 6 shows the opposite for the corresponding Rasch models. By using the 2PL GPCM as the basis for analysis, we improved the model fit and changed the order of preference of the Andersen and Embretson models. The worst-fitting 2PL GPCM (Table 5) outperforms the best-fitting Rasch model (Table 6).

The reason for the superior performance of the 2PL GPCM may be that some, but not all, of the tasks readministered at Time Point 2 may be 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, change). If so, those items would be poorly represented by the Rasch-type approach, in which all items receive the same discrimination parameter (in multidimensional approaches, one per dimension).

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 the second dimension. However, because this dimension measures only change in ability over time, it has inherently lower reliability than the first (main) dimension; therefore, it is not useful for reporting amount of change for individual students. In contrast, measures of group-level distributions can reliably 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 and standard deviation in the growth dimension (Dimension 2), as shown in Table 7. To determine the scale used in the multiple-group IRT model, we constrained the main dimension to a mean of zero and a standard deviation of 1.0 for medium-level schools.

Table 7

Mean and Standard Deviation of Multiple Groups Under the Embretson-Type Two-Parameter Logistic/Generalized Partial Credit Model

Model type	School type	Main dimension				Change at Time 2			
		Mean	s.e.	*SD*	s.e.	Mean	s.e.	*SD*	s.e.
Multiple-groups	Low	–0.482	0.10	0.950	0.04	0.864	0.09	1.009	0.11
	Medium	–0.005	0.07	0.981	0.03	0.926	0.08	1.144	0.12
	Integrative	–0.627	0.18	1.139	0.09	0.555	0.11	0.878	0.11
	High	1.069	0.05	0.945	0.02	1.011	0.05	0.939	0.06
Hierarchical-mixture	Class 1	1.042		0.910		1.012		0.935	
	Class 2	–0.417		0.940		0.764		0.995	

Note. s.e. = standard error.

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 integrative schools show similar improvement at Time Point 2.

In our analysis we used the hierarchical-mixture IRT version of the Embretson model with two mixture components (latent IRT classes). The resulting, somewhat different means of these two components show that students are divided into two groups: (a) low-performing with moderate growth between Time Points 1 and 2 and (b) high-performing with larger growth between Time Points 1 and 2. Notice that the variances of the two mixture components are similar at the same time point.

For the Andersen model, the difference between Dimensions 1 and 2 can be viewed as a measure of growth over time because the common items are constrained to have equal parameters across time. Table 8 shows the average growth measures under an Andersen-type 2PL GPCM with school type as the grouping variable.

Table 8

Mean and Standard Deviation of Multiple Groups Under the Andersen-Type Two-Parameter Logistic/Generalized Partial Credit Model

		Dimension 1				Dimension 2				Change	
Model type	School type	Mean	s.e.	*SD*	s.e.	Mean	s.e.	*SD*	s.e.		s.e.
Multiple-groups	Low	–0.531	0.09	0.958	0.04	0.025	0.13	1.017	0.05	0.556	0.07
	Medium	0.001	0.07	0.968	0.03	0.474	0.07	1.073	0.03	0.473	0.03
	Integrative	–0.713	0.17	1.107	0.11	-0.256	0.20	1.246	0.08	0.457	0.07
	High	1.078	0.05	1.012	0.02	1.552	0.05	0.934	0.02	0.474	0.02
Hierarchical-mixture	Class 1	–0.466		0.917		0.001		1.046		0.467	
	Class 2	1.080		0.971		1.557		0.903		0.477	

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 model, low-track schools show the largest growth, whereas under the Embretson model, high-track schools do. Recall that the Embretson model fits the data better than the Andersen model. Also, the Embretson model specified as a 2PL-based MIRT model estimates a separate growth-discrimination parameter for all items assessed at Time Point 2. In contrast, the discrimination parameters on the first (stability) dimension are constrained to be the same over time, much as the Andersen model is constrained in terms of parameters of items assessed at both time points.

These results lead us to conjecture that the Embretson model’s larger number of parameters results in improved fit. An inspection of the resulting parameters reveals that the way the Embretson model is specified allows some items to receive 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. Figure 1 shows the empirical distribution of slopes for the first (main) and second (change) dimensions.

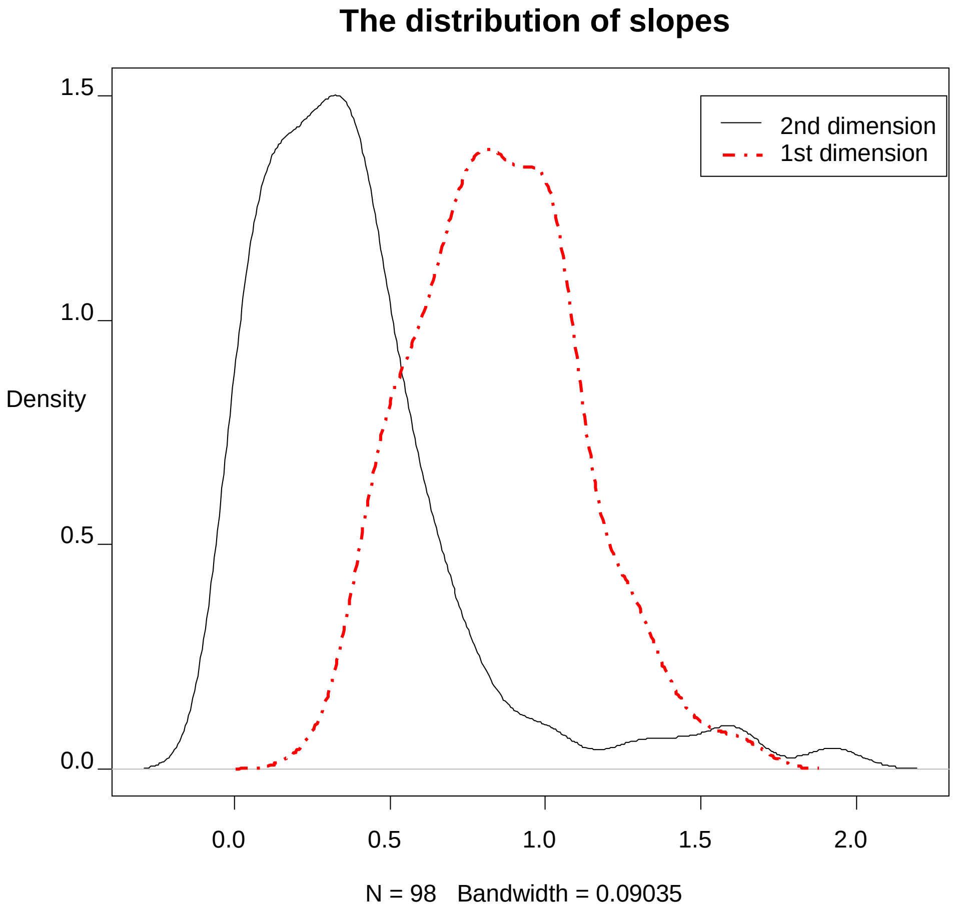

***Figure 1*. Slope parameters for Dimensions 1 and 2 based on estimates of slope parameters obtained with the Embretson-type two-parameter-logistic, generalized-partial-credit, hierarchical-growth mixture model.**

Most of the slope estimated for the second (growth) dimension falls between 0.0 and 0.5, indicating that some items do not load on the dimension unique to Time Point 2. In contrast, most of the slope estimated for the main dimension (across Time Points 1 and 2) falls between 0.5 and 1.5. Given that the variances of the ability estimates for both dimensions are of comparable size, it appears that the change dimension specific to Time Point 2 does not affect all items measured at that time point. In other words, for items with a slope close to zero on Dimension 2, the main ability estimate across the two time points suffices to fit the response behavior observed in this study.

Discussion

In this study, we analyzed a longitudinal data set using two models to measure change within the context of IRT: (a) the Andersen model, with a unique dimension per testing occasion, and (b) the Embretson model, which assumes an overall dimension across 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). The results on model data fit indicated that the Embretson-type 2PL model, extended to a multiple-group 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 Embretson model's fit to the data indicates that an overall dimension that cuts across time points, when used with a specific dimension that quantifies change, appropriately describes the observed student performance. This conclusion is supported by the finding that fitting the Andersen model (which assumes a unique dimension per time point) results in two highly correlated abilities. Specifically, the ability distributions estimated for each type of school show correlations above 0.8. The change-dimension average, estimated with the multiple-group Embretson-type GDM, students in all types of schools grow but students in lower performing school-types grow somewhat more slowly than those in higher performing school types. The type of school with the highest average proficiency also shows the highest average growth.

In the Embretson approach, the change dimension is specific to Time Point 2 and is designed to detect systematic differences in response behavior that cannot be explained by the overall ability variable. Items that carry substantive loading are most sensitive to change over time. Results presented in Figure 1 indicate that, for the PISA-I-Plus data, a number of items show this resistance to growth. These items may cover topics that were not taught during the year or that were taught before the first assessment.

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.

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.

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

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.), *Lecture notes in statistics: Vol. 157. Essays on item response theory* (pp. 43–68). New York: Springer-Verlag.

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

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

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

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*] (S. 29-63). Münster, Germany: Waxmann.

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

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

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

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

von Davier, M. (2007). *Hierarchical general diagnostic models* (ETS Research Rep. No. RR-07-19). Princeton, NJ: ETS.

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., & 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.

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, CT: Ablex.

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