Measuring the Power of Learning.™

Research Report
ETS RR–16-21

Applications of Multidimensional Item Response Theory Models With Covariates to Longitudinal Test Data

Jianbin Fu

May 2016

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Applications of Multidimensional Item Response Theory Models With Covariates to Longitudinal Test Data

Jianbin Fu

Educational Testing Service, Princeton, NJ

The multidimensional item response theory (MIRT) models with covariates proposed by Haberman and implemented in the *mirt* program provide a flexible way to analyze data based on item response theory. In this report, we discuss applications of the MIRT models with covariates to longitudinal test data to measure skill differences at the individual and group levels. In particular, we describe the differential item functioning procedure to identify common items with item drift across test occasions, and model selection and evaluation based on model comparison, fit statistics, and skill estimates. A real dataset on algebra tests is used to demonstrate the applications.

Keywords Multidimensional item response theory; covariates; longitudinal test data; differential item functioning; model fit

doi:10.1002/ets2.12108

Longitudinal test data contain item scores obtained from a common group of examinees across multiple test occasions. To measure students' growth, tests at different occasions should be placed on the same scale. In practice, the common approach is to use a unidimensional item response theory (IRT) model to calibrate each test separately and then to equate them to a common scale using common items as anchors. Multidimensional IRT (MIRT) models can be used to calibrate longitudinal test data concurrently, which has the potential to increase estimation accuracy. In addition, the mean vector and the covariance matrix of examinees' abilities at the population level across test occasions can be estimated. Applying MIRT models to longitudinal test data has been discussed in, for example, Embretson (1991, 1997); Meiser (1996); Meiser, Stern, and Langeheine (1998); te Marvelde, Glas, Van Landeghem, and Van Damme (2006); and von Davier, Xu, and Carstensen (2011). Recently, Haberman (2013) proposed MIRT models with flexible covariate structures on latent skills and implemented in the computer program, *mirt*. Incorporating covariates into MIRT calibrations has the potential to further increase estimation accuracy because auxiliary information is used (Mislevy, 1991). IRT models with covariates have been studied before in, for example, Adams, Wilson, and Wu (1997); De Boeck and Wilson (2004); and Zwinderman (1997). However, the *mirt* program provides a general treatment of IRT models with covariates and a feasible estimation procedure. In this report, issues related to applying MIRT models with covariates to longitudinal test data are discussed, and a real dataset is used for demonstration.

In the following sections, the MIRT models with covariates proposed by Haberman (2013) are first introduced, followed by discussion of several issues regarding the applications of the MIRT models with covariates to longitudinal test data. A real dataset on algebra tests is then used to demonstrate the applications. In the final section, a summary and discussion are provided.

Multidimensional Item Response Theory Models With Covariates

A MIRT model with covariates includes two basic functions: an item response function and a skills distribution function conditioned on covariates. In Haberman's (2013) computer program, *mirt*, the item response function can be very flexible, for example, (a) one-, two-, and three-parameter logistic models and their multidimensional extensions; (b) a multidimensional random coefficients multinomial logit model (Adams, Wilson, & Wang, 1997) and its submodels, such as the family of linear logistic test models (Fischer, 1973, 1997; Fischer & Ponocny, 1994, 1995); (c) a conjunctive Rasch model (Maris, 1995); (d) a noncompensatory multiple classification latent class model (Maris, 1999) and its submodel, the noisy

Corresponding author: J. Fu, E-mail: jfu@ets.org

inputs, deterministic "and" gate model (Junker & Sijtsma, 2001); and (e) a disjunctive multiple classification latent class model (Maris, 1999). In this report, the commonly used multidimensional generalized partial credit model (MGPCM; e.g., Fu, 2009; von Davier, 2008) is introduced as an example. The item response function of MGPCM for longitudinal test data is written as

$$P\left(x_{jit}=s_{itm}|\theta_{jt}\right)=\frac{\exp\left(\sum_{k\in \mathrm{K}_{it}} a_{ikt}\theta_{jkt}s_{itm}-b_{its_{itm}}\right)}{\sum_{f=0}^{M_{it}-1}\exp\left(\sum_{k\in \mathrm{K}_{it}} a_{ikt}\theta_{jkt}s_{itf}-b_{its_{itf}}\right)}, \tag{1}$$

where $\sum_{k\in \mathrm{K}_{it}} a_{ikt}\theta_{jkt}s_{it0}-b_{its_{it0}}\equiv 0$; x_{jit} is examinee j's score on item i on Test Occasion t; $\boldsymbol{\theta}_{jt}$ is the latent skill vector of examinee j at Test Occasion t with elements θ_{jkt} that denotes the value (continuous or discrete) of examinee j's skill k ($k=1, 2, \ldots, K$) at Test Occasion t; M_{it} is the number of score categories of item i at Test Occasion t; s_{itm} is the score of score category m ($m=1, 2, \ldots, M_{it}$) for item i at Test Occasion t, which can be any real value but is typically the integer $m-1$; $\mathbf{k}_{it}$ is the set of latent skills related to item i at Test Occasion t; a_{ikt} is the discrimination parameter of item i at Test Occasion t for latent skill k; and $b_{its_{itm}}$ is the parameter related to the difficulty of score category m for item i at Test Occasion t. By restricting $a_{ikt}=1$, the MGPCM is reduced to the multidimensional partial credit model.

The distribution function of latent skills conditional on covariates (e.g., demographic variables, previous test scores) is denoted as $P(\boldsymbol{\theta}_j|\mathbf{z}_j)$. Note that $\boldsymbol{\theta}_j$ is the latent skill vector of examinee j including all skill values across test occasions; that is, $\boldsymbol{\theta}_j=\{\boldsymbol{\theta}_{jt}\}$. For convenience, the elements of $\boldsymbol{\theta}_j$ are denoted as θ_{jd}, where d ($d=1, 2, \ldots, D$) refers to a skill at a certain test occasion. Furthermore, $\mathbf{z}_j$ is the covariate vector of examinee j across test occasions with elements z_{jh}, that is, examinee j's covariate h ($h=1, 2, \ldots, H$). Note that the first element in $\mathbf{z}_j$ is the intercept, that is, $z_{j1}=1$. For continuous latent skills, $P(\boldsymbol{\theta}_j|\mathbf{z}_j)$ is assumed to follow a multivariate normal distribution (Haberman, 2013) with mean

$$\mu\left(\theta_j|\mathbf{z}_j\right)=\left[-2\Lambda\left(\eta,\mathbf{z}_j\right)\right]^{-1}\beta\mathbf{z}_j \tag{2}$$

and covariance matrix

$$\sum\left(\theta_j|\mathbf{z}_j\right)=\left[-2\Lambda\left(\eta,\mathbf{z}_j\right)\right]^{-1}, \tag{3}$$

where $\boldsymbol{\beta}$ is a $D\times H$ coefficient matrix for the covariate vector $\mathbf{z}_j$ and elements β_{dh} are the coefficient for the hth covariate z_{jh} on latent skill d; $\boldsymbol{\eta}$ is an array with elements $\eta_{dd'h}$, $1\le d\le d'\le D$, $1\le h\le H$; and $\boldsymbol{\Lambda}(\boldsymbol{\eta},\mathbf{z}_j)$ is a $D\times D$ positive definite matrix with elements

$$\Lambda_{dd'}\left(\eta,\mathbf{z}_j\right)=\begin{cases}(1/2)\sum_{h=1}^{H}\eta_{dd'h}z_{jh}, & d<d',\\ \sum_{h=1}^{H}\eta_{dd'h}z_{jh}, & d=d',\\ (1/2)\sum_{h=1}^{H}\eta_{d'dh}z_{jh}, & d>d',\end{cases} \tag{4}$$

for $1\le d\le D$ and $1\le d'\le D$. Note that the matrix $\boldsymbol{\Lambda}(\boldsymbol{\eta},\mathbf{z}_j)$ is symmetric. The model parameters to be estimated in the skills distribution function are $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$. Note that because both the means and covariance matrix of the latent skills can depend on the covariates, the skills distribution defined here is very general. Thus, for example, the means and covariance matrix of latent skills for examinees with different covariates are different; however, for all examinees, the distributions of latent skills are assumed to be multivariate normal. The mean vector and covariance matrix can be simplified by imposing constraints on $\eta_{dd'h}$. If $\eta_{dd'h}$ is restricted to 0 for $d\ne d'$, then the latent skills for each examinee are independent of each other, conditional on the covariates. If $\eta_{dd'h}$ is restricted to 0 for $1<h\le H$, then the covariance matrix of the latent skills is the same for all examinees across test occasions. If both kinds of restrictions are imposed, then the covariance matrix of the latent skills is a diagonal matrix and is the same for all examinees across test occasions. In this case, the skills distribution function is actually the regular linear regression model.

For discrete latent skills, $\boldsymbol{\theta}_j$ can be any vector $\boldsymbol{\omega}$ in the finite space Ω. Then $P(\boldsymbol{\theta}_j = \boldsymbol{\omega}|\mathbf{z}_j)$ follows a quadratic log-linear model (Haberman, 2013; Haberman, von Davier, & Lee, 2008) and is written as

$$P\left(\theta_j = \omega|\mathbf{z}_j\right) = \frac{W(\omega)\exp\left[\omega'\beta\mathbf{z}_j + \omega'\Lambda\left(\eta, \mathbf{z}_j\right)\omega\right]}{\sum_{\theta_j \in \Omega} W\left(\theta_j\right)\exp\left[\theta_j'\beta\mathbf{z}_j + \theta_j'\Lambda\left(\eta, \mathbf{z}_j\right)\theta_j\right]}, \tag{5}$$

where $W(\boldsymbol{\omega})$ is the known positive weight for $\boldsymbol{\omega}$; the prime indicates the transpose of a matrix, and the other notation is defined in Equations 2–4.

Note that by treating group indicator(s) as covariate(s), the multiple group models (Bock & Zimowski, 1997; Fu, 2009; von Davier, 2008) are special cases of the MIRT models with covariates.

The *mirt* program (Haberman, 2013) can be used to estimate MIRT models with covariates. This program employs log-linear modeling and implements the maximum marginal likelihood method with the stabilized Newton–Raphson algorithm. The stabilized Newton–Raphson algorithm is different from the expectation-maximization algorithm (Fu, 2009), which is the most common algorithm used for maximum marginal likelihood estimation of IRT models. In the *mirt* program, model parameters can be fixed to constants or constrained to be equal to other parameters. For latent skills, maximum a posteriori and expected a posteriori (EAP; Bock & Aitkin, 1981) estimates are provided.

Considerations in Applying Multidimensional Item Response Theory Models With Covariates to Longitudinal Test Data

When calibrating longitudinal test data, each skill at each test occasion is treated as a factor in the MIRT model, and each factor can have a different covariate vector. In the following, we discuss three common issues in calibrating longitudinal test data.

The first issue is regarding separate versus concurrent calibrations. For separate calibrations, test data at each occasion are calibrated in separate runs and then equated to a common scale by an equating method, while a concurrent run calibrates all data across test occasions together to achieve a common scale. Separate calibrations may be necessary in some cases where, for example, the number of items and/or examinees is very large in each test occasion and/or test results need to be reported immediately after each administration. However, unlike concurrent calibration, separate calibrations cannot easily provide the correlation estimates among skills across occasions at the population level. In either case, appropriate constraints on model parameters should be imposed to identify the models.

The second issue is about model selection. Model selection includes the selection of both item response function and skills distribution function. The goal is to find an appropriate item response function and a concise skills distribution function that includes only the important covariates. The model of choice is the most parsimonious model with best model fit. The *mirt* program provides several indexes for comparing models and evaluating model fit that are described in the next section.

The third issue is about item parameter drift across test occasions. In longitudinal test data, anchor items are common items used to link test scale across test occasions. An anchor item's item parameters are kept the same across test occasions to establish a common scale. However, a common item may be functioning differently across test occasions; that is, the item may have differential item functioning (DIF) across test occasions. A common item with DIF should be removed from the anchor set and its item parameters allowed to be different across occasions. Therefore, when calibrating longitudinal test data, we need to identify the common items with DIF across test occasions and remove them from the anchor set.

A Real Data Example

In this section, the process of detecting DIF across test occasions for common items and selecting an appropriate MIRT model with covariates to calibrate longitudinal test data based on concurrent calibration is demonstrated using a real dataset on algebra tests from the *CBAL*™ learning and assessment tool research initiative supported by Educational Testing Service (ETS). Specifically, an IRT DIF procedure was first conducted on all common items across test occasions based on a preselected MIRT model. The common items identified with DIF were removed from the anchor set in the subsequent

Table 1 Design of 2012 CBAL Mathematics Study

	Algebra A (Administration 1)			Algebra B (Administration 2)		
Test sequence	NAEP item blocks	Session 1	Session 2	NAEP item blocks	Session 3	Session 4
1	B1, B2	Moving Sidewalks 1R	All Wet	B3, B4	Moving Sidewalks 2R	Heights and growth
2	B3, B1	Moving Sidewalks 1R	Smartphones	B4, B2	Moving Sidewalks 2R	Heights and growth
3	B2, B4	Moving Sidewalks 1R	All Wet	B1, B3	Moving Sidewalks 2R	Keeping in touch
4	B4, B3	Moving Sidewalks 1R	Smartphones	B2, B1	Moving Sidewalks 2R	Keeping in touch

Note. NAEP = National Assessment of Educational Progress.

Table 2 Distribution of Examinees Across Two Administrations

	Administration 2		
Administration 1	Yes	No	Total
Yes	2,696	159	2,855
No	59	–	59
Total	2,755	159	2,914

model selection process. In model selection, the model fit statistics from several MIRT models with varying levels of complexity were compared, and the most appropriate model was recommended.

CBAL is intended to create a model for an innovative K–12 assessment system that measures students' achievement (*of learning*), provides timely feedback for educational intervention (*for learning*), and is a worthwhile educational experience in and of itself (*as learning*; Bennett, 2010). CBAL tests are developed for mathematics, reading, and science based on underlying cognitive competency models that incorporate curriculum standards with the results of learning sciences research. CBAL tests are administered online and include innovative "technology-enhanced" items that are typically organized under a common scenario and gauge higher order critical thinking abilities.

Data

The CBAL data used in this study were obtained from Grade 8 algebra tests (van Rijn, Wise, Yoo, & Chung, 2013). These tests comprised six tasks; the number of items in each task ranged from 16 to 27, based on a common scenario. In addition, four blocks of 10 discrete items were used to establish links between tests and between administrations. These discrete algebra items were taken from the National Assessment of Educational Progress (NAEP) program. All tasks were administered by computer, and some items involved simulations. The item types included constructed response, numeric entry, selected response, and graph drawing. The item scores were dichotomous (0, 1) or polytomous (integers from 0 to 4).

The six tasks and the four blocks of 10 NAEP items were arranged into four test sequences across two test administrations (see Table 1). Each administration comprised two test sessions, and each test session lasted 45 minutes. The two administrations took place in Winter 2011 and Spring 2012, respectively, with an interval of approximately 3 months between the two administrations.

Schools were recruited throughout the United States and received monetary rewards for their participation. Students were randomly assigned to one of the four test sequences. The tests were stored on an ETS server, and students completed the tests online at local schools via Internet connection to the server. In this study, several demographic variables were collected from students, and only gender was used in this demonstration, as this variable was well represented. Two students with missing gender were excluded from the analyses. Omitted item responses (i.e., missing responses in the middle of a test session) were coded as 0. Not reached responses (i.e., missing responses toward the end of a test session) as well as any missing responses due to system administration problems were treated as missing. The resulting dataset used in the analyses included 1,467 males and 1,447 females with a total of 2,914 students, among whom 2,696 students participated in both test administrations (Winter 2011 and Spring 2012); see Table 2 for the distribution of examinees across the two administrations.

Method

CBAL items are organized under a common scenario (task), and therefore task-specific factors besides the general factor, algebra skill, may have potential effects on item responses. A preliminary analysis[1] showed that the task Moving Sidewalks 2R had the most reliable task effect; thus, only the task factor specific to the Moving Sidewalks 2R task was considered in this demonstration. Note that for scaling purposes, the general factor represents the targeted skill of an assessment. A task-specific factor is usually treated as a nuisance dimension that is unrelated to the general factor or to other task-specific factors. The steps made in analyzing this dataset are described in the following paragraphs.

First, DIF analyses were conducted on the 40 common items based on the two-factor MGPCM without covariates. The first factor was algebra skill at the first administration, and the second factor was algebra skill at the second administration. In the baseline model, the item discrimination (a) and difficulty (b) parameters of every common item were constrained to be the same across test occasions. In the augmented model for each common item, both the item discrimination and difficulty parameters of the target item were free to estimate at both test occasions. Then, the two nested models were compared by the likelihood ratio test for each common item, and a test with a significant p value, $p < .05$, indicated DIF on the target item. In practice, effect size measures, for example, the difference of average item scores between two test occasions, may be combined with the significant test to identify the drifted common items with practical importance. In this report, the effect size criterion is not employed.

Second, the following four MIRT models with different complexities were used to calibrate the test data concurrently. The item parameters of common items were constrained to be the same across test occasions, except for the common items that were identified to have DIF at the first step. The results were then evaluated in terms of model comparison and fit statistics and latent skill estimates:

Model 1 (M1). M1 is a three-factor MGPCM with a covariate, gender. The first one is the algebra skill relating to all items at the first test occasion, the second one is the algebra skill relating to all items at the second test occasion, and the third one is the task factor specific to Moving Sidewalks 2R and relates only to items corresponding to the Moving Sidewalks 2R task at the second test occasion. The first two factors are correlated but are independent to the third factor. All factors have the gender covariate plus intercept.

Model 2 (M2). M2 is a submodel of M1 without the gender covariate.

Model 3 (M3). M3 is a submodel of M1, excluding the task-specific factor; that is, M3 is a two-factor MGPCM with the gender covariate plus intercept.

Model 4 (M4). M4 is a submodel of M3 without the gender covariate.

Restrictions should be imposed to identify these models. For M4, the traditional way to identify the model is to fix the mean and variance of the algebra skill at Test Occasion 1 at the group level to, for example, 0 and 1, respectively. The mean and variance of the algebra skill at Test Occasion 2 as well as the covariance of the two factors at the group level are free to estimate, as its scale can be determined by the common item parameters of anchor items across the two test occasions. In *mirt*, the idea is similar; however, the constraints are made on the element(s) related to the algebra skill at Test Occasion 1 in the inverse of the variance–covariance matrix of factors, $\mathbf{\Lambda}(\boldsymbol{\eta}, \mathbf{z}_j)$ (see Equations 3 and 4), and the coefficient matrix for the covariate vector, $\boldsymbol{\beta}$ (see Equation 2). Specifically, the intercept for the algebra skill at Test Occasion 1, β_{11}, is set to 0, and the diagonal element related to the algebra skill at Test Occasion 1, η_{111}, is set to $-1/2$. However, note that these constraints do not lead to the fix of the mean and variance of the algebra skill at Test Occasion 1 at the group level to 0 and 1, respectively. For M3, the same constraints as M4 apply; however, the two constraints in M3 are now related to the algebra skill at Test Occasion 1 for the female group. For M1 and M2, additional constraints are needed: (a) set $\eta_{131} = \eta_{231} = \eta_{311} = \eta_{321} = 0$ in M2 and, additionally, $\eta_{132} = \eta_{232} = \eta_{312} = \eta_{322} = 0$ in M1 so that the task-specific factor is independent of the two algebra factors; (b) set $\beta_{13} = 0$ and $\eta_{331} = -1/2$ to fix the mean and variance, respectively, of the task-specific factor at Test Occasion 1 for the whole group in M2 and for the female group in M1 to 0 and 1.

Three statistics related to the expected log penalty per item response were used for model comparisons (Haberman et al., 2008, pp. 7–8):

1. Estimated expected log penalty per item response. It is equal to the log likelihood of a model divided by the total number of item responses in the dataset. This statistic does not take into account the number of model parameters being estimated.
2. Estimated expected predicted log penalty per item response (Gilula & Haberman, 2001). This statistic treats model complexity as cost and punishes the models with more model parameters.
3. Akaike information criterion (AIC; Akaike, 1974) per item response. The AIC is equal to twice the number of model parameters plus twice the negative log likelihood. AIC per item response is equal to AIC divided by twice the total number of item responses in the dataset. This statistic is a simplified approximation to the second statistic listed here, which assumes that the model is correct.

For all the three statistics, smaller is better. If a complex model has much smaller values on these three statistics than a simpler model, then the complex model would be recommended. If the differences on the three statistics are minor between the two models, then it indicates that the two models have similar model fit, and in this case, the simpler model with fewer model parameters would be preferred.

In addition, four model fit statistics based on generalized residuals (Haberman, 2009; Haberman & Sinharay, 2013) were compared across the four models. The generalized residual is defined as

$$q = \frac{\left(O - \hat{E}\right)}{\sigma} \qquad O = \frac{\sum_{j=1}^{N} d\left(\mathbf{x}_j\right)}{N}, \tag{6}$$

where $\mathbf{x}_j$ is examinee j's item score vector; N is the number of examinees in a sample; $d(\mathbf{x}_j)$ is a real function of $\mathbf{x}_j$; $\hat{E}$ is the estimated expected value of O under a proposed model; and σ is the standard deviation of $O - \hat{E}$ under the proposed model. If the model fits the data and the sample size is sufficiently large, then theoretically the generalized residual q follows an approximate standard normal distribution. Therefore, q can be compared to a standard normal distribution to determine whether a test is significant. If the absolute value of q is larger than 1.960 or 2.576, then the generalized residual test is significant at the .05 or .01 level. In a test dataset, if more than 5% (or 1%) of the generalized residuals are significant at the .05 (or .01) level, the model is considered not to fit the data well. However, in practical use, we also need to consider effect size related to $O - \hat{E}$ to determine practical meanings of a significant test; more discussion about this issue is given in the Results section The generalized residual test has advantages over the contingency table–based Pearson chi-square tests and likelihood ratio chi-square tests commonly used for IRT models. First, the generalized residual test has a solid theoretical basis, whereas all of the chi-square tests for IRT models contain conditions that lead to violations of their presumed chi-square distribution (Glas & Falcon, 2003; Haberman, Sinharay, & Chon, 2013). Second, the generalized residual is very flexible, as $d(\mathbf{x}_j)$ can be any real function of an item score vector. Therefore users can check the particular important properties of a model by choosing the appropriate $d(\mathbf{x}_j)$ functions. In this application, the following four types of generalized residuals were tested:

1. Set $d(\mathbf{x}_j) = 1$ if $x_{ji} = s_{im}$, $d(\mathbf{x}_j) = 0$ if $x_{ji} \neq s_{im}$, where x_{ji} is examinee j's score on item i and s_{im} is the score of item i's score category m; then O is the proportion of examinees in item i's score category m. In this application, the generalized residuals with $s_{im} = 1$ were examined.
2. Set $d\left(\mathbf{x}_j\right) = x_{ji} * x_{ji'}$, where $i \neq i'$; then O is the proportion of the cross product of item scores of the item pair, item i and item i'. This test statistic is usually used to check whether a model deviates from the assumption of local independence.
3. Set $d(\mathbf{x}_j) = 1$ if $\sum_{i \in \mathbf{K}_v} x_{ji} = r$, $d(\mathbf{x}_j) = 0$ otherwise, where $\mathbf{K}_v$ is the set of items in the CBAL task v ($v = 1, 2, \ldots, 6$); then O is the proportion of examinees in total raw score r of task v (i.e., the probability of total raw score r).
4. Set $d(\mathbf{x}_j) = 1$ if $\sum_{i \in \mathbf{K}_v} x_{ji} \leq r$, $d(\mathbf{x}_j) = 0$ otherwise; then O is the cumulative proportion of examinees in total raw score r of task v (i.e., the cumulative probability of total raw score r).

The *mirt* program (Haberman, 2013) was used for all model estimations. The adaptive Gauss–Hermite quadratures with 5 points for each latent skill were used in calculating the marginal likelihoods.

Table 3 Common Items With Differential Item Functioning

	Log likelihood	No. structural parameters	Difference of degree of freedom	*p* Value
Base model	−199,051	432		
Item 2	−199,048	434	2	.045
Item 5	−199,045	434	2	.003
Item 8	−199,040	434	2	.000
Item 10	−199,046	435	3	.033
Item 13	−199,046	434	2	.008
Item 17	−199,047	434	2	.040
Item 25	−199,047	434	2	.022
Item 36	−199,046	434	2	.015

Table 4 Model Comparisons

	Model			
Statistics	1	2	3	4
Log likelihood	−198,465	−198,454	−199,020	−199,017
No. structural parameters	475	468	454	449
Penalty	.617	.617	.618	.618
Akaike	.618	.618	.620	.620
Gilula–Haberman	.618	.618	.620	.620
Algebra reliability, Test Occasion 1	.956	.956	.956	.956
Algebra reliability, Test Occasion 2	.943	.943	.951	.951
Task-specific skill reliability	.502	.511		

Results

Table 3 lists the likelihood ratio test results for the common items with test p values smaller than .05. In Table 3, the number of structural parameters includes all model parameters, except for students' skill parameters. Eight common items appeared to have DIF. These eight common items were removed from the anchor set, and their item parameters were allowed to be different across the two test occasions in the subsequent analyses.

Table 4 lists the three model comparison statistics described previously for each of the four models: estimated expected log penalty per item response ("Penalty"), AIC per item response ("Akaike"), and estimated expected predicted log penalty per item response ("Gilula–Haberman"). One can see that the differences for all three statistics across the four models are minor (no larger than .002), which indicates that the gains in model fit from the three complicated models to the simple Model 4 were negligible. Table 4 also shows the reliabilities of the EAP estimates of the algebra skill at Test Occasions 1 and 2 and/or task-specific skill related to the task Moving Sidewalks 2R for each model. The reliability is calculated as the ratio of the variance of the sample EAP estimates over the variance of the population EAP estimates (e.g., Adams, 2006; Haberman & Sinharay, 2010). All the models have the same reliability estimate of .956 on algebra skill at the first administration. For algebra skill at the second administration, M3 and M4 have the reliability estimate of .951, which is .08 higher than for M1 and M2. The task-specific skill reliability estimate in M2 is .511, which is .09 higher than for M1. Although the differences in reliability estimates are small, they do suggest an appreciable difference in the proportion of population variance explained by the sample EAP estimates.

The four types of generalized residuals as defined previously were used to compare model fit in the four models. Note that a generalized residual test was reported only if the ratio of the standard error of the generalized residual, σ, over the standard error of O in Equation 6 was larger than .1 and the standard error of O was not equal to 0, because if σ is too small due to computational errors that occur within iterative algorithms, a generalized residual test is not meaningful. For the generalized residual tests for single items as described earlier, only 64 out of 205 items are reported. The percentages of the significant generalized residuals in the total reported tests at the .05 level are shown in Table 5 for the four types of generalized residuals in the four models. As can be seen in Table 8, for each type of generalized residual, the percentages of significant generalized residuals are similar across the four models, and the complicated models do not appear to have significantly better fit than the simple M4. However, all the percentages of significant generalized residuals were

Table 5 Percentages of Significant Generalized Residuals

Generalized residual	No. tests	Model 1	2	3	4
Percentage of examinees with score 1 in each item	64	12.5	9.4	12.5	12.5
Average cross-product of item scores of each item pair	17,275[a]	16.5	15.8	17.1	16.2
Percentage of examinees with each total raw score in each CBAL task	222	18.0	18.0	18.9	18.9
Cumulative percentage of examinees with each total raw score in each CBAL task	217	47.9	46.1	48.8	48.4

[a]Excluded tests where 17 examinees or fewer answered both items; for the included tests, at least 551 students responded to both items.

Table 6 Pearson Correlations of Expected a Posteriori Estimates for General Skill Among Four Models

Administration	Model 1	Model 2	Model 3
First			
Model 2	1.000		
Model 3	1.000	1.000	
Model 4	1.000	1.000	1.000
Second			
Model 2	1.000		
Model 3	.992	.993	
Model 4	.993	.993	1.000
Growth (second minus first)			
Model 2	.996		
Model 3	.945	.948	
Model 4	.945	.950	.999

much larger than 5%. For each CBAL task, the percentages of significant generalized residuals for total raw scores and for cumulative total raw scores, while varied across the six tasks, were also much larger than 5%. Note that this result is not unusual. Sinharay, Haberman, and Jia (2011) applied the generalized residual tests to several operational test data sets and found that, for all the datasets, the IRT models used to calibrate the test data did not fit the data well. However, they found that not all the misfit was of practical significance. The same question applies to the CBAL data: Whether the statistically significant generalized residuals are also practically significant is a topic that needs to be further explored based on the intended uses of the test scores. For example, a rule based on effect size, for instance, the residual (i.e., $|O-\hat{E}|$), or the portion of residual over the fitted statistic (i.e., $|(O-\hat{E})/\hat{E}|$), can be combined with the significant test and used to select residuals with practical importance. In Table 5, the maximum residual (i.e., $|O-\hat{E}|$) for single items is only about .03, which may not be considered as large a discrepancy in practice.

To compare the EAP estimates of the algebra skill for each of the four models, Pearson correlations were calculated among the four models for the algebra skill factor at each administration as well as the growth between the two administrations. As shown in Table 6, all correlations are equal to or larger than .992, except that the correlations on the growth measures between one model with the task-specific factor (M1 or M2) and one model without the task-specific factor (M3 or M4) are approximately .95. This indicates that the gender covariate does not have much impact on estimation of the algebra skill, but whether the task-specific factor is taken into account slightly affects the estimation of the algebra skill. This finding is further verified by the analyses of the agreement of the fail/pass classifications among the four methods. The failure rates were set at 10%, 20%, and 25% for the less able students. Tables 7 and 8 show the number of students with different classifications among the four methods, given each of the three failure rates, for the first and second administrations, respectively. One can see that the numbers of students with disagreement between M1 and M2 and between M3 and M4 are much smaller than those between one model with the task-specific factor (M1 or M2) and one model without the task-specific factor (M3 or M4). In addition, the numbers of students with disagreement between one model with the task-specific factor (M1 or M2) and one model without the task-specific factor (M3 or M4) increase from the first test administration to the second administration and with failure rates.

Table 9 lists the factor mean, standard deviation, and correlation estimates at the population level for the whole group as well as for the male and female groups, if applicable, from the four models. One can see that the correlation estimates on algebra skill between the two administrations from the four models are very close at approximately .93.

Table 7 First Administration: Disagreement of Fail/Pass Classification Among Four Models

Failure rate	Model 1	Model 2	Model 3
10% of bottom students			
Model 2	4		
Model 3	8	8	
Model 4	8	8	2
20% of bottom students			
Model 2	4		
Model 3	12	12	
Model 4	14	14	4
25% of bottom students			
Model 2	6		
Model 3	16	14	
Model 4	16	14	2

Note. The value in each cell represents the number of students with different classifications between two methods.

Table 8 Second Administration: Disagreement of Fail/Pass Classification Among Four Models

Failure rate	Model 1	Model 2	Model 3
10% of bottom students			
Model 2	6		
Model 3	22	24	
Model 4	20	22	2
20% of bottom students			
Model 2	6		
Model 3	38	36	
Model 4	38	36	2
25% of bottom students			
Model 2	12		
Model 3	56	52	
Model 4	58	54	6

Note. The value in each cell represents the number of students with different classifications between two methods.

Although the factor mean and standard deviation estimates cannot be compared directly across the models because their scales are different, the observed patterns across the two administrations and the gender groups are the same. Specifically, (a) in all models, students appeared to have slight improvement as well as a little more dispersion on algebra skill at the second administration compared to the first administration; (b) in M1 and M3, females appeared to have a little better algebra skill than males did; and (c) in M1, males showed a slightly higher mean score on the task-specific factor than females did. Note that the way to fix scales in the *mirt* program (see the preceding discussion regarding imposing constraints on $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ for purpose of model identification) is different from the traditional way (e.g., fix the mean and standard deviation of a latent skill in a group to 0 and 1, respectively). That is the reason we see the much larger standard deviation estimates of algebra skills at both administrations in Table 9 than we would normally get from other IRT programs and no group mean of the algebra skill at either administration that is fixed to 0.

In sum, the model comparison and fit statistics show that the model fit of the simple M4 is similar to that observed for the other three more complicated models, and therefore M4 is recommended as the scaling model for the CBAL data from the parsimonious perspective.

Discussion

The MIRT models with covariates proposed by Haberman (2013) and implemented in the *mirt* program provide a more flexible way to analyze data based on IRT. In this report, we discuss the applications of the MIRT models with covariates to longitudinal test data to measure skill differences at the individual and group levels. In particular, we describe the DIF procedure to identify common items with item drift across test occasions, and we describe model selection and evaluation

Table 9 Factor Population Mean, Standard Deviation, and Correlation Estimates

Statistic	Algebra skill at first admin.		Algebra skill at second admin.	Task-specific factor
Model 1				
Male				
M	.045		.075	.038
S.D.	2.561		2.850	.980
Corr.		.935		NA
Female				
M	.231		.279	.000
S.D.	2.537		2.810	1.000
Corr.		.919		NA
All				
M	.193		.238	.018
S.D.	2.556		2.840	.991
Corr.		.928		NA
Model 2				
All				
M	.244		.292	.000
S.D.	2.704		2.999	1.000
Corr.		.929		NA
Model 3				
Male				
M	−.055		−.018	NA
S.D.	2.546		2.757	NA
Corr.		.931		NA
Female				
M	.132		.154	NA
S.D.	2.526		2.716	NA
Corr.		.918		NA
All				
M	.089		.121	NA
S.D.	2.547		2.747	NA
Corr.		.926		NA
Model 4				
All				
M	.247		.287	NA
S.D.	2.642		2.850	NA
Corr.		.926		NA

based on model comparison and fit statistics and skill estimates. A real dataset from CBAL algebra tests has been used to demonstrate the applications.

Many factors should be taken into account when choosing a scaling method for longitudinal test data. Practical considerations may play a more important role than statistical criteria in some contexts. Sometimes the finding of statistical significance between two models does not necessarily represent a practical significance; the practical meaning of model differences should always be investigated. Complicated models often pose challenges for practical use, such as subtle model identification issues, heavy computational burden, unstable parameter estimates, and difficulties in interpreting model parameters for laypersons.

A useful feature of the *mirt* program that is not described in this report is its capacity to handle stratified sampling. Stratified sampling is commonly used in large-scale assessments for survey, for example, in the NAEP and the Program for International Student Assessment. When analyzing data from stratified sampling based on IRT models, the number of strata and primary sampling units should be taken into account. See Haberman (2013) for more details and Qian (2015) for an application.

Acknowledgments

The author thanks Carolyn Wentzel, Shelby Haberman, Peter van Rijn, Ayleen Gontz, and Yue Jia for their helpful suggestions and edits on early versions of this report.

Note

1 In the preliminary analysis, a bifactor MGPCM with eight factors was fitted to the CBAL data. First was the general factor representing algebra skills and including all items; the other seven factors were specific to the six CBAL algebra tasks and the NAEP items (all NAEP items were treated as one task), respectively. A task-specific factor included only the items in that task. The eight factors were assumed to be independent. The general factor and the task-specific factor for the NAEP items had one covariate plus intercept, test occasion, whereas the other task-specific factors did not have the covariate of test occasion, as those tasks were only administered once. The simplex quadrature (Haberman, 2013) was used to calculate the marginal likelihood for a faster calibration. The results show that the EAP estimates of the task-specific factor for Moving Sidewalks 2R had the highest reliability, at 0.501, among the seven task-specific factors.

References

Adams, R. J. (2006, April). *Reliability and item response modeling: Myths, observations, and applications*. Paper presented at the 13th International Objective Measurement Workshop, Berkeley, CA. Retrieved from http://www.slideserve.com/kerryn/reliability-and-item-response-modelling-myths-observations-and-applications

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational & Behavioral Statistics*, *22*(1), 47–76.

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

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, *8*, 70–91.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

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, NY: Springer.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: 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. J. Engelhard, & K. L. Draney (Eds.), *Objective measurement* (Vol. 4, pp. 223–236). Greenwich, CT: Ablex.

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

Fischer, G. H. (1997). Unidimensional linear logistic Rasch models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 225–244). New York, NY: Springer.

Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, *59*, 177–192.

Fischer, G. H., & Ponocny, I. (1995). Extended rating scale and partial credit models for assessing change. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 181–202). New York, NY: Springer.

Fu, J. (2009, April). *Marginal likelihood estimation with EM algorithm for general IRT models and its implementation in R*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Gilula, Z., & Haberman, S. J. (2001). Analysis of categorical response profiles by informative summaries. *Sociological Methodology*, *31*, 129–187.

Glas, C. A. W., & Falcon, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 87–106.

Haberman, S. J. (2009). *Use of generalized residuals to examine goodness of fit of item response models* (Research Report No. RR-09-15). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02172.x

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton–Raphson algorithm* (Research Report No. RR-13-32). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02339.x

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209–227.

Haberman, S. J., & Sinharay, S. (2013). Generalized residuals for general models for contingency tables with application to item response theory. *Journal of the American Statistical Association*, *108*, 1435–1444. http://dx.doi.org/10.1080/01621459.2013.835660

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, *78*, 417–440.

Haberman, S., von Davier, M., & Lee Y.-H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous distributions* (Research Report No. RR-08-45). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02131.x

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–273.

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, *60*, 523–547.

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

Meiser, T. (1996). Loglinear Rasch models for the analysis of stability and change. *Psychometrika*, *61*, 629–645.

Meiser, T., Stern, E., & Langeheine, R. (1998). Latent change in discrete data: Unidimensional, multidimensional, and mixture distribution Rasch models for the analysis of repeated observations. *Methods of Psychological Research Online*, *3*(2), 75–93. Retrieved from http://www.dgps.de/fachgruppen/methoden/mpr-online/issue5/art6/meiser.pdf

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

Qian, J. (2015, April). *Multidimensional latent linear model for measuring growth for longitudinal samples*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Sinharay, S., Haberman, S. J., & Jia, H. (2011). *Fit of item response theory models: A survey of data from several operational tests* (Research Report No. RR-11-29). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02265.x

te Marvelde, J. M., Glas, C. A. W., Van Landeghem, G., & Van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement*, *66*, 5–34.

van Rijn, P. W., Wise, M., Yoo, H., & Chung, S. (2013). *Statistical report: Summary statistics, local dependence, and differential item functioning in the CBAL mathematics 2012 study*. Unpublished manuscript.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*, 318–336.

Zwinderman, A. A. (1997). Response models with manifest predictors. In W. J. van der Lindern & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 245–257). New York, NY: Springer.

Suggested citation:

Fu, J. (2016). *Applications of multidimensional IRT models with covariates to longitudinal test data* (Research Report No. RR-16-21). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12108

Action Editor: Shelby Haberman

Reviewers: Yue Jia and Peter van Rijn