Research Report

Modeling Growth With Adaptive Longitudinal Large-Scale Assessments

ETS RR–18-34

Jiahe Qian

December 2018

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

Modeling Growth With Adaptive Longitudinal Large-Scale Assessments

Jiahe Qian

Educational Testing Service, Princeton, NJ

The 2-parameter logistic multidimensional item response theory (MIRT) model was employed to model growth for the National Education Longitudinal Study of 1988 (NELS:88). The 3 measurement waves of NELS:88 (base year, first follow-up, and second follow-up) represented 3 dimensions. The inquiry aimed to improve modeling performance growth based on examinees' responses to the test items in each wave, with item location parameters set to be invariant across the 3 waves (instead of using item changes) and the latent mean of the first wave set to 0. The yielded scores of 3 waves were thus placed on approximately the same scale; the changes of the scores across waves could be measured. Moreover, the growth models for longitudinal data were improved by using auxiliary information such as gender, race, school location, and parents' education. In the study, the results of the growth pattern were compared with those yielded by the Embretson models.

Keywords Multidimensional item response theory (MIRT); longitudinal data; survey with adaptive testing; National Education Longitudinal Study of 1988 (NELS:88)

doi:10.1002/ets2.12220

Latent logistic regression models are often used to model growth assessing student learning and change over time (Millsap, 2008; Mislevy & Bock, 1982; Rasch, 1960); multidimensional item response theory (MIRT) models are used in particular (Embretson, 1991; Fischer & Seliger, 1997). Normally the MIRT models provide an appropriate way to measure multiple layers of skills with a multidimensional data structure (Haberman & Sinharay, 2010; Haberman, von Davier, & Lee, 2008; Reckase, 2009; te Marvelde, Glas, van Landeghem, & van Damme, 2006; von Davier, 2005).

As a longitudinal survey, the National Education Longitudinal Study of 1988 (NELS:88) was designed to evaluate the growth of skills of the same collection of individuals with periodic assessments, including reading, mathematics, and science (Rock, Pollack, & Quinn, 1995). This program examined the educational, vocational, and personal development of an eighth-grade cohort of students[1] in 1988 at various stages during their educational years, as well as the personal, familial, social, institutional, and cultural factors that may have affected that development.

In this study, the three measurement waves of NELS:88 (base year, first follow-up [F1], and second follow-up [F2]) represented three dimensions in the MIRT growth models. The NELS:88 reading assessment adopted an adaptive design, beginning with a standard test booklet for the first wave and branching into two test booklets of different difficulty levels for each of the second and third waves. Each participant was assigned to a test booklet in a subsequent wave based on his or her performance (high/low) in the prior wave.

Much research has focused on issues related to model fitting for unidimensional or multidimensional item response theory modeling and the goodness-of-fit of models (Haberman, 2009; Haberman, Sinharay, & Chon, 2013; von Davier & Sinharay, 2010). van Rijn (2008) employed a quadrature Kalman filter to improve the estimation of longitudinal IRT models. Alternatively, Almond (2011) proposed applying the particle filter expectation – maximization (PFEM) algorithm to estimating parameters of longitudinal IRT models. Fu (2016) applied MIRT models with flexible covariate structures to longitudinal data to measure skill differences at the individual and group levels. In application, Yao and Boughton (2007) used MIRT models to improve subscale proficiency estimation and classification. von Davier, Xu, and Carstensen (2011) used a general latent variable model to measure growth for the Programme for International Student Assessment (PISA). Mislevy and Zwick (2012) discussed the scaling and linking issues in reporting the results of a periodic assessment system. Some growth models assessed growth by the changes of item parameters (Doran & Jiang, 2006; Kaplan & Sweetman, 2006; Reckase, 2009; von Davier, 2005), for example, the Embretson model (Embretson, 1991; Xu & Qian, 2009).

Corresponding author: J. Qian, E-mail: jqian@ets.org

This study aimed to construct conditional two-parameter logistic (2PL) MIRT models to measure growth based on examinees' responses in testing under the invariant constraints; that is, the location parameters of the same items were set invariant across the measurement waves of longitudinal assessments (Almond, 2011; Qian, 2015; Rijmen, 2010; van Rijn, 2008; von Davier et al., 2011). In this study, the item discrimination parameters of the same items across the measurement waves were not fixed. The marginal maximum likelihood (MML) approach (Bock & Aitkin, 1981; Zhang, 2012) is widely used in most MIRT packages to calibrate item parameters for the response data with multidimensional structure and estimate the correlation coefficients between abilities, including the discrimination parameters of the overlapping items. Without loss of generality, one can also standardize the latent traits of each measurement wave. In this study, the latent means of the first wave were set to zero; then, based on the EM algorithm for marginal likelihood, the latent means and variances of the other waves could also be estimated. Moreover, under this design, together with the invariant constraints for overlapping items, the model places the scale scores of different waves on the same scale; the changes of the scores across waves can be measured and compared; and the correlations between dimensions can also be estimated because the same examinees were measured across waves. Although the latent traits of three waves are not identical and measure different dimensions, as described above, this design yields comparable scores across different waves. Under this design, the MIRT growth models can be used to create a scale for measuring the changes of groups and individuals for longitudinal data, such as the NELS:88 (Millsap, 2008; Rock, 2012).

In this study, the constraints imposed on the proposed MIRT model were fulfilled through its design matrix that defined a covariate structure of all the parameters in the model; the location parameters of the same items were set invariant across three waves, the latent mean and variance of the first wave were standardized, the correlations of across-wave scales were specified, and the parameters for each demographic group were designated. A methodical description of the 2PL MIRT model used in this study can be found in the "Methodology" section of this report.

In this study, the psychometric specification of the model was also enhanced with incorporated demographic variables such as gender, race, school location, and father's and/or mother's education. The results were compared with those yielded from the Embretson models. Although growth models can be promoted by use of auxiliary variables, these models are mainly recommended for the data of low-stakes assessments such as NELS:88 for reporting group scores. When such models are applied to estimating individual scores, fairness issues can arise because, for example, a boy and girl who give identical item responses could get different scores.

The software package, *mirt*, used in this study (Haberman, 2013) to calibrate the MIRT models employs the stabilized Newton–Raphson algorithm (Haberman, 1988). In addition, the software allows users to define their own design matrixes to build new models, including models using the incorporated demographic variables in this study. Other software packages (Cai, 2013; von Davier, 2005; Yao, 2008) for MIRT analysis are based on different algorithms, such as the expectation–maximization algorithm using MML and an approach applying the Markov chain Monte Carlo (MCMC) method.

The report is organized as follows. The next section introduces the main conditional 2PL MIRT model and the MIRT models enhanced by incorporating auxiliary information. The third section reviews the instrument design, data collection, and the dimensionality structure of the reading assessments of NELS:88. The fourth section presents the results of the empirical analysis for different subsets of NELS:88 data, and the final section summarizes the findings.

Methodology

The Two-Parameter Logistic Multidimensional Item Response Theory Model With Use of Auxiliary Information

In this study, the 2PL MIRT model was used to fit the adaptive longitudinal assessments of NELS:88; by the NELS:88 design, all 54 items in the reading item pool were multiple-choice items. The 3PL model was not employed because, in addition to the difficulty in estimation of item parameters (Holland, 1990), a 2PL describes real test data as well as the 3PL model (Haberman, 2006a; Sinharay & Holland, 2007).

The 2PL MIRT model is enhanced by incorporating demographic variables as the predictors. The MIRT model assumes that an r-dimensional latent skill vector $\boldsymbol{\theta}_i$ with elements θ_{ik}, $1 \le k \le r$ ($r = 3$ for NELS:88) is associated with each examinee i at wave (or time) k. Let $\mathbf{Y}_i$ be the response vector for each examinee i ($1 \le i \le n$). The pairs $(\mathbf{Y}_i, \boldsymbol{\theta}_i)$, $1 \le i \le n$, are independently and identically distributed and, given $\boldsymbol{\theta}_i$, the response variables Y_{ij}, $1 \le j \le q$ are conditionally independent.

Let τ_j be the item intercept (difficulty), and let $\boldsymbol{a}_j$ be the vector with all parameters, including the parameters of slopes and the nonconstant elements of the predicting variables at different test waves of item j, $1 \le j \le q$. The 2PL MIRT model for dichotomous item j can be expressed as

$$P_j(1|\boldsymbol{\theta}) = \frac{\exp\left(\tau_j + \boldsymbol{a}_j'\mathbf{A}\boldsymbol{\theta}\right)}{1 + \exp\left(\tau_j + \boldsymbol{a}_j'\mathbf{A}\boldsymbol{\theta}\right)}, \tag{1}$$

where latent vector $\boldsymbol{\theta} \sim MN(\boldsymbol{\mu}(z), \boldsymbol{\Sigma}(z))$, $\mathbf{A}$ is a defined matrix, z is a matrix that contains U predictors including examinee's covariate vector across waves. The model in Equation (1) provides a general framework for 2PL multidimensional IRT models (Reckase, 2009), including those making use of auxiliary information (Mislevy, 1991) and those employing algorithms that can appropriately estimate parameters and statistics of interest for complex survey data such as NELS:88 in this study. Because the latent vector $\boldsymbol{\theta}$ is assumed to be multivariate normal, the K-dimensional latent mean and $K \times K$-dimensional latent covariance matrix are

$$\boldsymbol{\mu}(z) = [-2\boldsymbol{\Lambda}]^{-1}\boldsymbol{\Psi}z$$

and

$$\boldsymbol{\Sigma}(z) = [-2\boldsymbol{\Lambda}(\boldsymbol{\lambda}, z)]^{-1},$$

respectively, where $\boldsymbol{\Psi}$ is a $K \times U$ matrix of regression parameters and $-\boldsymbol{\Lambda}(\boldsymbol{\lambda}, z)$ is a $K \times K$ positive definite matrix. The elements of $\boldsymbol{\Lambda}$ are defined as

$$\Lambda_{kk'}(\boldsymbol{\lambda}, z) = \begin{cases} (1/2)\sum_{u=1}^{U} \lambda_{kk'u} z_u, k < k', \\ \sum_{u=1}^{U} \lambda_{kk'u} z_u, k = k', \\ (1/2)\sum_{u=1}^{U} \lambda_{k'ku} z_u, k > k', \end{cases} \tag{2}$$

where $\boldsymbol{\lambda}$ is an array with elements $\lambda_{kk'u}$, $1 < k \le k' \le K$, $1 \le u \le U$ (Haberman, 2013).

The CMIRT model in Equation (1) can also be expressed as a logit model:

$$\log\left[\frac{P_j(1|\boldsymbol{\theta})}{P_j(0|\boldsymbol{\theta})}\right] = \tau_j + \boldsymbol{a}_j'\mathbf{A}\boldsymbol{\theta}. \tag{3}$$

Because the scale structure for the model is defined by a covariance matrix with the item location parameters invariant across the three waves, it is called a 2PL conditional MIRT (CMIRT) model; its scale differs from the one yielded by concurrent calibration of IRT models (Lord, 1980; Wingersky & Lord, 1984). The scale yielded by concurrent calibration can fail to account for construct shift in scale across dimensions (Patz & Yao, 2007).

The 2PL CMIRT model for item j can be extended to a very general model:

$$P_j\left(y|\boldsymbol{\theta}\right) = \frac{\exp\left(\tau_{yj} + \boldsymbol{a}_{yj}'\mathbf{A}\boldsymbol{\theta}\right)}{\sum_{m=0}^{M_j-1}\exp\left(\tau_{mj} + \boldsymbol{a}_{mj}'\mathbf{A}\boldsymbol{\theta}\right)}, \tag{4}$$

where $\boldsymbol{\theta} \sim MN(\boldsymbol{\mu}(z), \boldsymbol{\Sigma}(z))$, $\mathbf{A}$ is a defined matrix, and $y = 0, 1, \ldots, M_j - 1$ (Haberman, 2013). This model is an extension of the model in Equation (1) and covers different types of items, including dichotomous and polytomous items.

Design Matrix for the Linear Model of Latent Vectors

Let $\boldsymbol{\beta}$ be a vector of dimension B with elements of the parameters of locations, slopes, and the nonconstant elements of the predicting variables and the quadratic terms in the upper triangular of $\boldsymbol{\Sigma}^{-1}$. If a model is designed to include auxiliary information in calibration, as in this study, demographic variables need to be recoded and included in $\boldsymbol{\beta}$ as well. Let $\mathbf{o}$ be a known offset vector of dimension B with the elements listed in the same sequence as $\boldsymbol{\beta}$. Let $\mathbf{T}$ be a $B \times C$ ($C > 0$) design matrix. Let Γ be a nonempty open subset of the space R^{C} and $\boldsymbol{\gamma} \in \Gamma$ be a C-dimensional vector. A linear model is defined as

$$\boldsymbol{\beta} = \mathbf{o} + \mathbf{T}\boldsymbol{\gamma} \tag{5}$$

(Haberman, 2013). Let w_i be the case weight for examinee i ($1 \le i \le n$). Let the log-likelihood component $l_i(\boldsymbol{\beta})$ for examinee i be the logarithm of the probability $P(\boldsymbol{Y}_i|\, \boldsymbol{z}_i, \boldsymbol{\beta})$ under conditions in Equations (1) – 33 of Haberman (2013). Then, the weighted log-likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i l_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \log P\left(\boldsymbol{Y}_i|\, \boldsymbol{z}_i, \boldsymbol{\beta}\right). \tag{6}$$

The stabilized Newton – Raphson algorithm is the procedure used to solve Equation (6) for MML estimation.

To reduce the computational complexity, the transition matrix $\mathbf{T}$ and vector $\mathbf{o}$ are decomposed: $\mathbf{o} = \mathbf{T}^1\mathbf{o}^2$, $\mathbf{T} = \mathbf{T}^1\mathbf{T}^2$, and $\boldsymbol{\beta} = \mathbf{T}^1\boldsymbol{\beta}^1$. Thus the model can be simplified as

$$\boldsymbol{\beta}^1 = \mathbf{o}^2 + \mathbf{T}^2\boldsymbol{\gamma}, \tag{7}$$

where $\mathbf{T}^2$ is a $B_1 \times C$ design matrix. Therefore the main task of modeling MIRT is to define its design matrix by specifying covariates and taking account of auxiliary information. When Equation (7) has been defined, the *mirt* program can automatically implement Equation (5); the stabilized Newton – Raphson algorithm is then used to obtain MML estimation for Equation (6) (Haberman, 2013). A partitioned $\mathbf{T}^2$ provides a modular design technique that emphasizes separating the functionality in modeling. The block matrix $\mathbf{T}^2$ in the appendix, provides the block design details, defining unknown location and slope parameters in $\mathbf{T}^2_{\tau}$ and $\mathbf{T}^2_{ad}$, specifying the means and covariates across waves in $\mathbf{T}^2_{\lambda}$, and designating auxiliary variables to the appropriate parameters in $\mathbf{T}^2_{\lambda,*}$.

The model can be nonidentifiable unless some relationships are well defined between the elements in $\boldsymbol{a}_j$ and the elements in $\boldsymbol{\Sigma}$. The identification conditions are set by making the first element of $\boldsymbol{\mu}$ equal to 0 and the element in the first row and the first column of $\boldsymbol{\Sigma}$ equal to 1. The specific setup for the reduced vector in linear model can be found in the Appendix.

Estimation for the 2PL Conditional MIRT Models

In this study, the scale scores were *expected a posteriori* (EAP) scores (Mislevy & Bock, 1982), which were used as the estimated theta scores in item analysis. In the *mirt* output of estimated EAP scores for examinee ($1 \le i \le n$), the row vector (Bellman, 1970) of the estimated scores for three waves is $\widehat{\boldsymbol{\theta}}_{(i)} = \left(\widehat{\theta}_{i1}, \widehat{\theta}_{i2}, \widehat{\theta}_{i3}\right)$; vector $\boldsymbol{\theta}_{(i)} = (\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2}, \boldsymbol{\theta}_{i3})$ is a corresponding row random vector. The column vector of estimates is $\widehat{\boldsymbol{\theta}}_k = \left(\widehat{\theta}_{1k}, \ldots, \widehat{\theta}_{ik}, \ldots, \widehat{\theta}_{nk}\right)'$ for wave ($1 \le k \le 3$). Based on column vectors, the matrix of estimated scale scores is $\widehat{\boldsymbol{\theta}}^c = \left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2, \widehat{\boldsymbol{\theta}}_3\right)$.

The EAP scores, augmented based on Kelley's formula (Kelley, 1947), are regressed estimates of the true score that shrink proportionally toward the aggregate average; moreover, in *mirt,* the accuracy and efficiency of estimation are improved by use of the subtest score mean through methods of numerical adaptive quadrature procedure (Haberman, 2008; Haberman, 2013). Because NELS:88 is a low-stakes assessment, this study is focused on the performance of group-level achievement. Maximum likelihood estimator (MLE) is an alternative choice. Because it was unavailable in *mirt,* MLE was not chosen in studying.

The conditional mean of the EAP scores, given $\mathbf{Y}_i$ and $\mathbf{Z}_i$, for kth test wave is $\bar{\widehat{\theta}}_k = \frac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_{ki}$. The *mirt* output includes the estimated conditional covariance matrix of conditional means: $\mathrm{V}\left(\widehat{\boldsymbol{\theta}}^{c\prime}|\mathbf{Y}_i, \mathbf{Z}_i\right)_{3\times3}$ with element $\mathrm{Cov}^c_{k,k'}$ $\left(1 \le k, k' \le 3\right)$. Note that these estimates can be also derived from $\widehat{\boldsymbol{\theta}}_k$ ($1 \le k \le 3$) provided by *mirt*. Let $\Delta\widehat{\boldsymbol{\theta}}_{(i)} = \boldsymbol{\theta}_{(i)} - \widehat{\boldsymbol{\theta}}_{(i)}$. The *mirt* output also includes the estimated conditional covariance matrix: $\mathrm{V}\left(\Delta\widehat{\boldsymbol{\theta}}_{(i)}|\mathbf{Y}_i, \mathbf{Z}_i\right)_{3\times3}$ ($1 \le i \le n$) with element $\mathrm{Cov}_{(i),\,k,\,k'}$ of waves k and k' $(1 \le k, k' \le 3)$. The average across examinees is

$$\overline{\mathrm{V}}\left(\Delta\widehat{\boldsymbol{\theta}}_{(\cdot)}|\mathbf{Y}_i, \mathbf{Z}_i\right)_{3\times3} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{V}\left(\Delta\widehat{\boldsymbol{\theta}}_{(i)}|\mathbf{Y}_i, \mathbf{Z}_i\right)$$

with element $\overline{\mathrm{Cov}}_{(\cdot),k,k'}$ $(1 \le k, k' \le 3)$. Because the unconditional covariance matrix for $\boldsymbol{\theta}_{(i)}$ equals the conditional covariance matrix of conditional means and the conditional covariance matrix (Haberman, 2013, p. 44), the estimated unconditional covariance matrix is

$$\widehat{\mathrm{V}}\left(\boldsymbol{\theta}_{(\cdot)}\right) = \mathrm{V}\left(\widehat{\boldsymbol{\theta}}^{c\prime}|\mathbf{Y}_i, \mathbf{Z}_i\right) + \overline{\mathrm{V}}\left(\Delta\widehat{\boldsymbol{\theta}}_{(\cdot)}|\mathbf{Y}_i, \mathbf{Z}_i\right). \tag{8}$$

Therefore, the correlation between latent traits of waves k and k' $(1 \le k, k' \le 3)$ is.

$$\widehat{\rho}_{k,k\prime} = \frac{\mathrm{Cov}^c_{k,k\prime} + \overline{\mathrm{Cov}}_{(\cdot),k,k\prime}}{\sqrt{\mathrm{Cov}^c_{k,k}\ \mathrm{Cov}^c_{k\prime,k\prime}}} = \frac{\mathrm{Cov}^c_{k,k\prime} + \overline{\mathrm{Cov}}_{(\cdot),k,k\prime}}{\sqrt{\mathrm{V}^c_{k,k}\ \mathrm{V}^c_{k\prime,k\prime}}}. \tag{9}$$

Embretson Model and Its Expansion

Embretson (1991) proposed a multidimensional Rasch (1960) model for learning and change (MRMLC) by employing a Wiener simplex structure for skills to provide parameters for individual differences in changes. The Wiener simplex model (Anderson, 1960; Joreskog, 1970) is appropriate for structuring parameters in Rasch models because the properties of equivalency of measurement scales between tests and the additivity of these effects are compatible between the models. In MRMLC, Embretson postulated the involvement of M skills in item responses within K occasions, that is, measurement waves. Specifically, assume in the first wave ($k = 1$) that only an initial skill is involved in the item responses; assume further in later waves ($k > 1$) $k - 1$ additional skills as well as the initial skill factored into examinees' skill scores (Embretson, 1991). Thus the number of skills tested increases in each wave over time. Combining these assumptions with the one-dimensional Rasch model, MRMLC is given as

$$P\left(X_{i(k)j} = 1|\theta_{im}, b_j\right) = \frac{\exp\left(\sum_{m=1}^{k}\theta_{im} - b_j\right)}{1 + \exp\left(\sum_{m=1}^{k}\theta_{im} - b_j\right)}, \tag{10}$$

where θ_{im} and b_j are the skill of person i at test wave m and the difficulty parameter for item j, respectively.

It is worth mentioning that Embretson (1991) developed MRMLC for situations where items are not repeated across waves to avoid the well-known effects of repeated item presentation (practice effects and memory effects, among others) and local dependency among item responses. Under such situations, Equation (5) tells us that for an item j observed at wave k, all tested skills up to wave k are involved. So an item observed at wave k measures k skills, including initial skill (θ_{i1}) as well as $k - 1$ modifiabilities (i.e., $\theta_{i2}, \ldots, \theta_{ik}$). The change between condition $k - 1$ and k is equal to the kth modifiability (θ_{ik}). An Embretson model can be extended and has the following form:

$$P\left(X_{i(k)j} = 1|\theta_{im}, b_j\right) = \frac{\exp\left(\sum_{m=1}^{k} a_j\left(\theta_{im} - b_j\right)\right)}{1 + \exp\left(\sum_{m=1}^{k} a_j\left(\theta_{im} - b_j\right)\right)}, \tag{11}$$

where a_j is the slope parameter (Xu & Qian, 2009). The parameters of the Embretson models in Equations (10) and (11) can be calibrated through many software packages, such as *mirt* and *mdltm* (von Davier, 2005). The results of the six ability groups reported in Xu and Qian (2009) were yielded by *mdltm*.

Data Resources

Survey Design and Data Collection for the National Education Longitudinal Study of 1988

NELS:88 was initiated by the National Center for Education Statistics (NCES). Based on a two-stage stratified, clustered-sample design, the base-year survey for NELS:88 was carried out during the 1988 spring semester (Spencer, Frankel, Ingels, Rasinski, & Tourangeau, 1990). A representative sample of 24,242 students was selected among eighth graders from 1,052 public and private schools. On average, 23 student participants represented each of the participating schools.

The base-year study design consisted of four data components split by the role targeted individuals played in the educational system; these components were the surveying and testing of students along with the surveying of parents, school administrators, and teachers.

The first follow-up study was conducted in 1990, when most of the students were high school sophomores, and included all the components in the base-year study except the parent survey. In addition, a freshened sample was added to the student component to enhance the representativeness of the sample of the nation's sophomores. A total of 21,474 students participated in the first follow-up study.

Table 1 Number of Overlapped Items Across Booklets for the National Education Longitudinal Study of 1988 Reading Assessments

		Wave 2 (F1)		Wave 3 (F2)	
Test wave	Wave 1 (base year)	Low	High	Low	High
Wave 1 (base year)	–	20	8	15	4
Wave 2 (F1)					
Low	–	–	7	15	4
High	–	–	–	7	4
Wave 3 (F2)					
Low	–	–	–	–	8
High	–	–	–	–	–

Note. This table is the upper triangular of a symmetric matrix. Low/high denotes the booklet difficulty level. F1 = first follow-up; F2 = second follow-up.

The second follow-up took place in early 1992, when most sampled students were in the second semester of their senior year. Again, a freshened sample was added in 1992 to better represent 12th graders during the spring term of the 1991 – 1992 school year. As in the previous waves, students were asked to complete a questionnaire and a series of cognitive tests. Overall, this second follow-up study included a total of 20,923 students, which was made up of both in-school students and dropouts.

Adaptive Reading Assessment Design

The construction of the NELS:88 eighth-grade battery was in some sense a delicate balancing act between several competing objectives suggested by the NELS Technical Review Panel and/or NCES project staff during base-year development (Rock et al., 1995). To improve the measuring accuracy of the NELS:88 reading assessment, the program was designed to contain two test booklets at different difficulty levels for the second and third waves, respectively, and each participant was assigned to a booklet based on his or her performance (high/low) on the assessment of the previous wave. Therefore the examinees in the first follow-up assessment were partitioned into two ability groups (high and low); based on their base-year and F1 performances, the examinees in the second follow-up assessment were partitioned into six ability groups (high/high, high/low, high/not present, low/high, low/low, and low/not present).

The item pool of the NELS:88 reading assessment comprised 54 multiple-choice items. Each booklet contained 21 items. By the NELS:88 design, some items were overlapped across the two booklets at the second and third waves; some items were also overlapped through two or three waves. Table 1 presents the number of the items overlapped across booklets from base year to F2. For example, seven and eight items were overlapped across the two booklets in the second and third wave, respectively. For further information on item overlapping, see the test item map in Rock et al. (1995). Such a design, with common items across the booklets within each wave and through different waves, guaranteed development of a scale (Kolen & Brennan, 2004; Stocking & Lord, 1983; Vale, 1986) that could measure changes over time, even though participants at different levels answered a different assortment of test questions across the measurement waves.

Dimensionality Structure of the National Education Longitudinal Study of 1988 Reading Assessments

In this study, the MIRT models were applied to model growth for the NELS:88 reading assessment, and the structure of the multiwave assessments was treated as multidimensional. Thus, verifying the dimensional structure of multiwaves was essential to the study. Qian (2010), based on the dimensionality evaluation to enumerate contributing traits (DETECT) procedure (Zhang & Stout, 1999), analyzed the dimensionality structure of NELS:88 data. The results of the DETECT confirmatory analysis validated that the three waves were approximately three dimensional and that these findings were in line with the conclusions drawn by the DETECT exploratory analysis. Therefore, NELS:88 is one of the longitudinal assessments with a dimension structure appropriate for modeling growth by applying multidimensional logistic models (Embretson, 1991; Fischer & Seliger, 1997; Sijtsma, 2001). Otherwise, the conclusions drawn from the MIRT model may be without basis (Roberts & Ma, 2006) with the dimensional structure not confirmed.

Results

Model Calibration With Complex Survey Data

In modeling MIRT, the common items were set with the same item parameters across test occasions, and the dimensions defined by the test occasions were assumed to be dependent. The *mirt* program also enabled incorporating the complex sampling features such as primary sampling units in model calibration and variance estimation. Four sets of standard error (*SE*) estimates—regular *SE*, Louis *SE*, sandwich *SE*, and complex *SE* for the parameter estimates were computed based on the estimates of the Hessian matrix in the Newton–Raphson algorithm (Haberman, 2006b; Louis, 1982).

In this study, four demographic variables (DVs) correlated with growth were included in the CMIRT model: gender, race, school location, and father's education. They were coded as 11 dummy variables (Draper & Smith, 1981); for details, see Table 2.

For NELS:88, a large proportion of missing data was embedded due to the adaptive design of booklets illustrated in the Adaptive Reading Assessment Design section in this report. In response to this issue, the data employed in the analysis were extracted from the original data set with the criterion that each student in the data set took at least one assessment.

The Criterion Functions of Evaluation

The chi-square test based on the log-likelihood ratio statistic (Wilks, 1938) can be used to compare the models. The chi-square test statistic for comparison is

$$G^2(\mathrm{A},\mathrm{B}) = -2\ \left(l\left(X,\boldsymbol{\theta}_{\mathrm{A}},\boldsymbol{\beta}_{\mathrm{A}}\right) - l\left(X,\boldsymbol{\theta}_{\mathrm{B}},\boldsymbol{\beta}_{\mathrm{B}}\right)\right), \tag{12}$$

where $l(X,\boldsymbol{\theta}_{\mathrm{A}},\boldsymbol{\beta}_{\mathrm{A}})$ and $l(X,\boldsymbol{\theta}_{\mathrm{B}},\boldsymbol{\beta}_{\mathrm{B}})$ are the log-likelihood ratio statistics for two models, respectively. Let p be the difference in the number of parameters between two models. The discrepancy measure $G^2(\mathrm{A},\mathrm{B})$ has an asymptotic chi-square distribution with p degrees of freedom.

For comparison of models by entropy, the estimated expected log penalty per presented item (PE) can also be used for model selection. Let J_{ik} be the number of items presented to an examinee i corresponding to the kth measurement wave. The PE index is defined as

$$\mathrm{PE} = \frac{-l\left(\hat{\boldsymbol{\beta}}\right)}{\sum_{i=1}^{N} w_i \sum_{k=1}^{K} J_{ik}},$$

which can be improved with a bias correction (Gilula & Haberman, 1994). The Gilula–Haberman index (GH) is defined as

$$\mathrm{GH} = \frac{-l\left(\hat{\boldsymbol{\beta}}\right) + \mathrm{tr}\left\{\left[-\nabla^2 l\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\Phi\left(\hat{\boldsymbol{\beta}}\right)\right\}}{\sum_{i=1}^{N} w_i \sum_{k=1}^{K} J_{ik}},$$

where $\mathrm{tr}\{M\}$ denotes the trace of the matrix M and $-\nabla^2 l\left(\hat{\boldsymbol{\beta}}\right)$ is the estimated Hessian matrix of the weighted log-likelihood (Gilula & Haberman, 1995; Haberman, 2013). The computation of the Hessian matrix is based on the estimated

Table 2 Eleven Dummy Variables Recoded From the Demographic Variables (Gender, Race, School Location, and Father's Education)

Dummy variable	Definition
1	D1 = 1, if sex = 2 (female); otherwise 0
2	D2 = 1, if race2 = 2 (Black); otherwise 0
3	D3 = 1, if race2 = 3 (Hispanic); otherwise 0
4	D4 = 1, if race2 = 4 (White); otherwise 0
5	D5 = 1, if region = 2 (urban); otherwise 0
6	D6 = 1, if region = 3 (suburban); otherwise 0
7	D7 = 1, if region = 4 (rural); otherwise 0
8	D8 = 1, if father's education = 2 (< high school); otherwise 0
9	D9 = 1, if father's education = 3 (high school ≤ & < college); otherwise 0
10	D10 = 1, if father's education = 4 (college ≤ or master's); otherwise 0
11	D11 = 1, if father's education = 5 (doctor, professional, etc.); otherwise 0

Table 3 Model Prediction Statistics for Growth Models for the National Education Longitudinal Study of 1988 Reading Data

2PL MIRT models	No. of predictors	DVs included	Model dimen.	Log-likelihood	Penalty	SE of penalty	Akaike	Gilula–Haberman
CMIRT	1		115	−609019.31	0.5483	0.00045	0.5484	0.5484
$CMIRT_{DVA}$	12	G, R, L, F_edu	148	−606050.51	0.5456	0.00158	0.5458	0.5458
$CMIRT_{DVB}$	11	R, L, F_edu	145	−606248.23	0.5458	0.00158	0.5459	0.5459
$CMIRT_{DVC}$	9	G, R, F_edu	139	−606203.60	0.5458	0.00158	0.5459	0.5459
$CMIRT_{DVD}$	9	G, L, F_edu	139	−606730.48	0.5462	0.00158	0.5464	0.5464
$CMIRT_{DVE}$	8	G, R, L	136	−607536.83	0.5470	0.00158	0.5471	0.5471
$CMIRT_{DVF}$	8	R, F_edu	136	−606401.54	0.5459	0.00158	0.5461	0.5461
$CMIRT_{DVG}$	8	F_edu	127	−607080.14	0.5465	0.00158	0.5467	0.5467
$CMIRT_{DVH}$	8	R	124	−607965.68	0.5473	0.00158	0.5475	0.5475

Note. The sample size of the base-year sample was 24,242, and a total of 90 items were used in the NELS:88 reading assessments. 2PL = two-parameter logistic; CMIRT = conditional multidimensional item response theory; DV = demographic variable; Model dimen. = model dimensions; F_edu = father's education; G = gender; L = school location; R = race. "Stage 1 iterations" and "Stage 2 iterations" refer to the iterations used in the first and second stages of the computation of the maximum-likelihood estimates.

covariance matrix with built-in complex sampling effects. The Akaike information criterion (AIC; Akaike, 1973) is also used for model selection; the AIC index is provided with the model dimension added to the numerator of the PE index. However, AIC can misrepresent the disparity between models, in particular for large samples (Gilula & Haberman, 1994).

Nine Calibrated Conditional MIRT Models

Nine CMIRT models were fitted to the NELS:88 data: one 2PL CMIRT model and eight models used DVs, $CMIRT_{DVA}$ to $CMIRT_{DVH}$. Table 3 presents the model prediction indexes for the nine models fitted for the NELS:88 data (sample size $N = 24{,}242$), including log-likelihood, PE, GH, and AIC.

The test chi-square statistic $G^2(\mathrm{A}, \mathrm{B})$ in Equation (12) was used to compare the models. For example, in Table 3, $CMIRT_{DVA}$, with all the DVs included, was compared with $CMIRT_{DVB}$. Because $G^2(\mathrm{A}, \mathrm{B}) = 395.4$ and $p = 1$, the model $CMIRT_{DVA}$ was significantly improved from $CMIRT_{DVB}$. Although the structures of the models $CMIRT_{DVB}$ and $CMIRT_{DVC}$ were close, the $CMIRT_{DVB}$ was significantly improved from $CMIRT_{DVC}$ because $G^2(B, C) = 89.3$ and $p = 2$. By comparing the DVs used in the models, it was determined that father's education and race were important predictors.

The outcomes of the model selection based on different criteria, either PE, GH, or AIC, were consistent with those by the log-likelihood ratio tests. Except for 2PL CMIRT, the eight other models, from $CMIRT_{DVA}$ to $CMIRT_{DVH}$, all made use of auxiliary information and had a better fit than the 2PL CMIRT model. Clearly, the conditional MIRT models were improved by the incorporation of background variables, analogous to the estimation procedures of the National Assessment of Educational Progress (NAEP; Allen, Donoghue, & Schoeps, 2001). Among the models, $CMIRT_{DVA}$ fit the data best. The nine 2PL CMIRT models can be sorted from best to worst as follows: $CMIRT_{DVA}$, $CMIRT_{DVC}$, $CMIRT_{DVB}$, $CMIRT_{DVF}$, $CMIRT_{DVD}$, $CMIRT_{DVG}$, $CMIRT_{DVE}$, $CMIRT_{DVH}$, and CMIRT.

Figure 1 presents the cumulative frequency curves (CFCs) and the conditional CFCs per booklet at each score point on the frequency curves for three waves yielded by the 2PL CMIRT and 2PL $CMIRT_{DVA}$, respectively. Define their deviations as the difference between CFC and the conditional CFC and the residuals as the difference between FC and the conditional FC at each score point. The patterns of the CFCs across the ability continuum, measured by EAP, were comparable with the two models; the curves shifted to the right end from base year to F1, and then to F2, which suggested growth in examinees' ability over time. On the second and third rows of Figure 1, the standardized deviations, equal to the deviation divided by the standard error at each score point, and standardized residuals, equal to the residual divided by the standard error at each score point, on the low end of the continuum were mostly larger than those on the high end, reflecting the guessing effects on the performance of examinees with low scale scores; it can also be due to memory effects on overlapping items across waves. The plots also exhibited the differences between CMIRT and $CMIRT_{DVA}$ models; for example, the curves of the conditional CFCs per booklet yielded by $CMIRT_{DVA}$ were smoother than those yielded by CMIRT. Table 4 presents the estimated correlations between latent traits among three waves for the 2PL CMIRT and 2PL $CMIRT_{DVA}$, respectively.

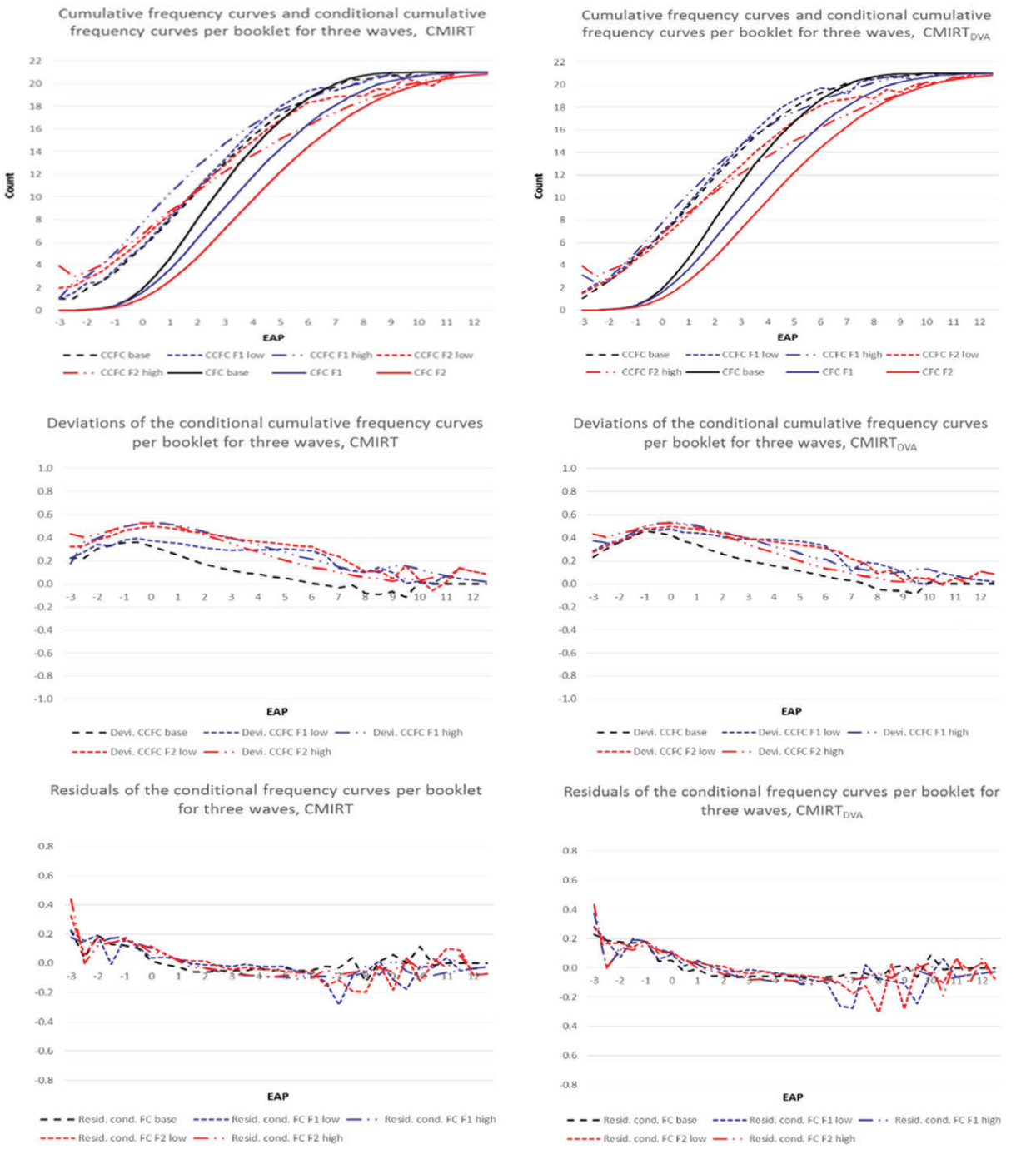

Figure 1 Cumulative frequency curves, conditional cumulative frequency curves per booklet, their deviations, and the residuals at each score on the frequency curves for three test waves of the National Education Longitudinal Study of 1988 reading (*left* CMIRT, *right* CMIRT$_{DVA}$).

The corresponding correlations, of CMIRT and CMIRT$_{DVA,}$ were very close, ranging from .86 to .90. The calculations were based on Equation 9. As expected, the correlation between the traits of base year and F2 was slightly lower than the other two correlations.

Figure 2 presents the Q–Q plots (Filliben, 1975) of the scale scores of base year, F1, and F2 yielded by the two models. The plots in the second row were slightly closer to the lines of the normal distribution in the Q–Q plots than the corresponding plots in the first row. Figure 3 shows the scatterplots of a pair of the score sets (base year vs. F1, base year vs. F2,

Table 4 Estimated Correlations Between Latent Traits Among Three Waves (*upper triangular* CMIRT, *lower triangular* $CMIRT_{DVA}$)

Test wave	Wave 1 (base year)	Wave 2 (F1)	Wave 3 (F2)
Wave 1 (base year)	–	0.902	0.866
Wave 2 (F1)	0.899	–	0.896
Wave 3 (F2)	0.864	0.900	–

Note. F1 = first follow-up; F2 = second follow-up.

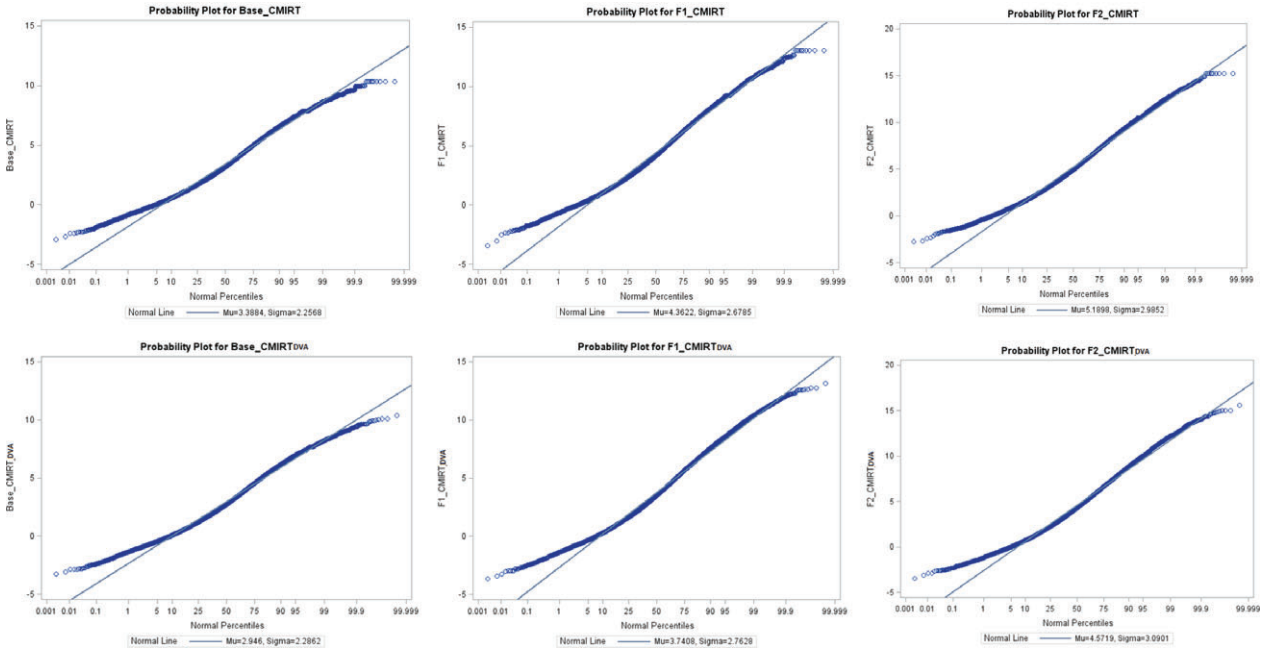

Figure 2 Probability plots of the scale scores of base year, first follow-up, and second follow-up for the National Education Longitudinal Study of 1988 reading (*first row* CMIRT, *second row* $CMIRT_{DVA}$).

and F1 vs. F2) for the two models. Although the patterns of the plots were compatible with the two models, the slopes of the trend lines in the scatterplots in the first row were about 1–2% smaller than those in the second row. A majority of the points in the plots were above the diagonal lines: approximately 91%, 95%, and 90% for the corresponding plots in the first row and 87%, 95%, and 90% for the plots in the second row. Evidently, most of the examinees made progress in reading over time because the scores assessed were higher in a later measurement wave in the comparisons of base year versus F1, base year versus F2, and/or F1 versus F2. Note that the progress stated here was measured by the changes of scores on the same scale for three dimensions; this did not suggest that the latent traits of the three dimensions were identical and measuring the identical ability of reading at different grades.

The skill scores of seven demographic groups for six ability patterns (i.e., high/high, high/low, high/not present, low/high, low/low, and low/not present) provided comprehensive information on evaluating nine CMIRT models calibrated with the NELS:88 data. Table A1 in the appendix presents the detailed results, including the total sample yielded by the 2PL CMIRT and 2PL $CMIRT_{DVA}$ models. For the groups with a high/high pattern, Figure 4 presents the skill scores of seven demographic groups yielded by the 2PL CMIRT and $CMIRT_{DVA}$ model. In the left plot in Figure 4, the means of all seven groups with high/high pattern at base year ranged from 174 to 188; all the groups gain scores between F1 and F2, though growth rates were different. In general, the results on the right plot, yielded by the $CMIRT_{DVA}$ model, had a similar pattern to those yielded by the 2PL MIRT model, although the models with the use of DVs always fit the data better.

The plots in Figure 5 present the skill scores of seven demographic groups with a high/low pattern yielded by the 2PL CMIRT and $CMIRT_{DVA.}$ The means of all seven groups with a high/low pattern were homogeneous at the base year, ranging from 147 to 152 for the 2PL CMIRT; most of the groups experienced a 1–2 point loss in F1, except for the

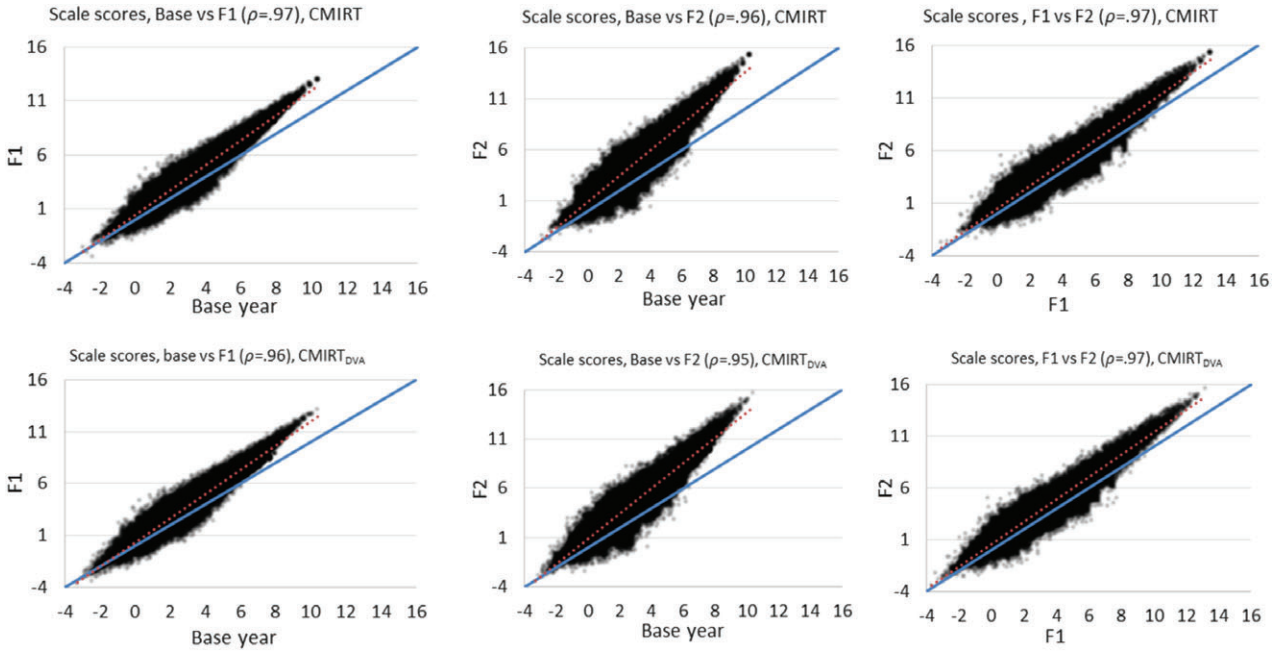

Figure 3 Plots of the comparisons of the EAP scores (base year vs. first follow-up, base year vs. second follow-up, and first vs. second follow-up) for the National Education Longitudinal Study of 1988 reading (*first row* CMIRT, *second row* $\text{CMIRT}_{\text{DVA}}$).

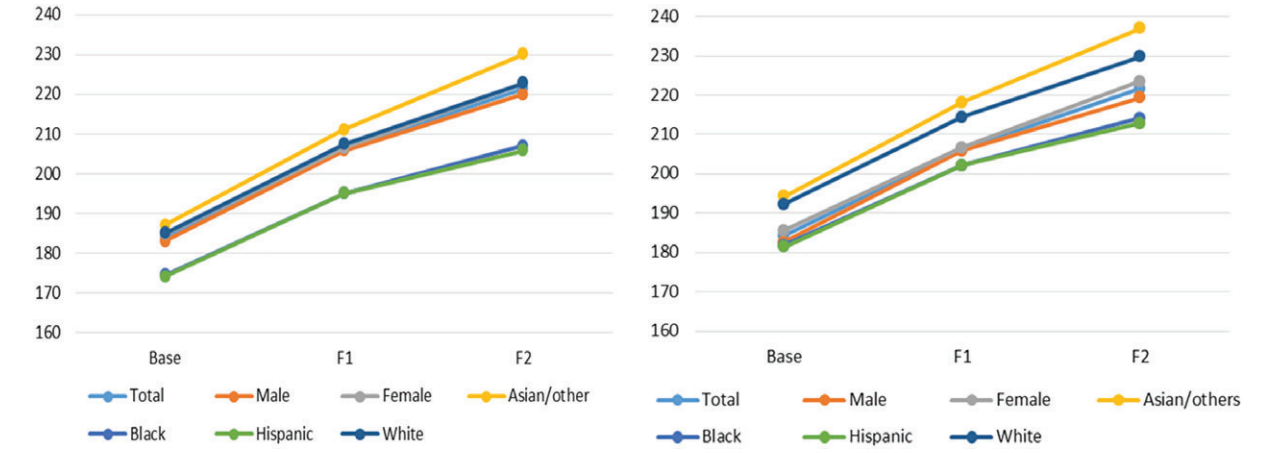

Figure 4 Performance of seven demographic groups with high/high pattern for the National Education Longitudinal Study of 1988 reading (*left* CMIRT, *right* $\text{CMIRT}_{\text{DVA}}$).

female and Hispanic groups. These students, with a high/low pattern, had outperformed other students on the test in the base-year wave. The right plot in Figure 5 shows a similar pattern for the $\text{CMIRT}_{\text{DVA}}$ model. All the demographic groups with a high/low pattern gained progress in the F2 wave test after 2 years in the study.

Model Comparison: Embretson Model Versus CMIRT Model

Table 5 presents the results of the comparison of the average expected a posteriori (EAP) scores of the six ability pattern groups yielded by the 2PL CMIRT and Embretson models. For the sake of comparison, the EAP scores were aligned with the same target mean and standard deviation (150, 35) by a linear transformation of each set of results for CMIRT and Embretson (Xu & Qian, 2009), respectively. When the whole base-year samples (with no ability group yet classified) were transformed, the same transformation was then applied to the EAP scores of F1 and F2 for CMIRT and Embretson, respectively. Although this transformation is neither linking nor equating, it enables us to better perceive trends or patterns

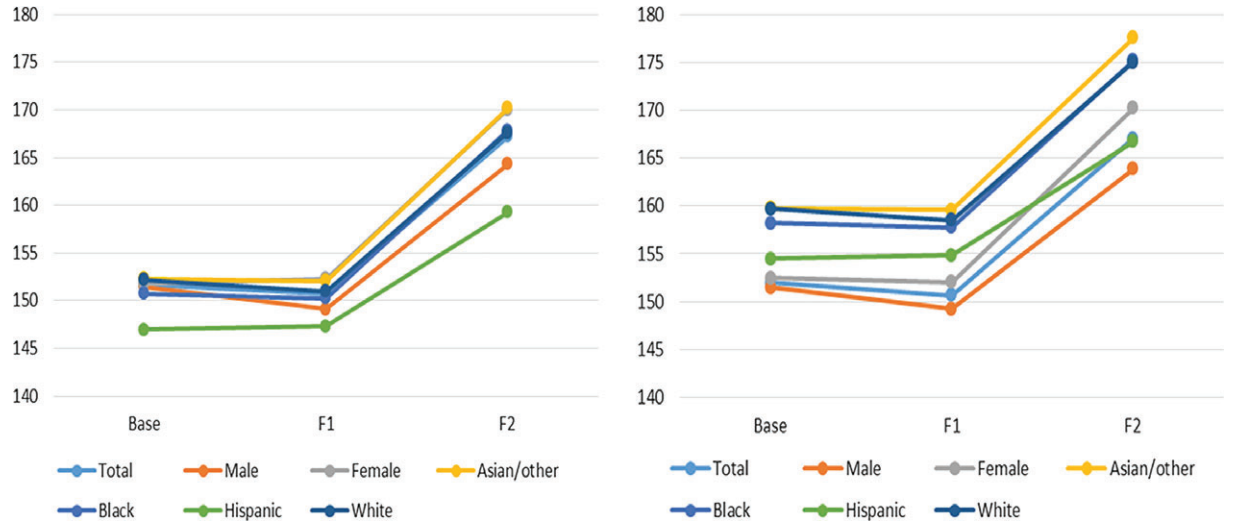

Figure 5 Performance of seven demographic groups with a high/low pattern for the National Education Longitudinal Study of 1988 reading (*left* CMIRT, *right* $\text{CMIRT}_{\text{DVA}}$).

Table 5 Comparison of the Performance of Six Ability Groups Yielded by Multidimensional Item Response Models (*left* CMIRT, *right* Embretson model)

		2PL CMIRT Model			Extended Embretson Model		
	Size	Base	F1	F2	Base	F1	F2
High/high	5,373	183.98	206.25	221.50	228.10	284.32	293.31
High/low	1,076	151.61	150.70	167.19	186.05	152.00	183.68
High/not present	1,713	165.60	181.38	195.87	217.56	245.03	204.36
Low/high	1,103	146.06	174.46	183.01	154.07	211.39	230.92
Low/low	4,009	119.01	127.81	138.87	108.74	128.55	144.68
Low/not present	2,091	119.99	130.52	140.70	109.37	132.01	122.28

Note. All expected a posteriori (EAP) mean scores are transformed by a linear transformation that sets base-year EAP distribution with mean and *SD* of [150, 35]. 2PL = two-parameter logistic; CMIRT = conditional multidimensional item response theory; F1 = first follow-up; F2 = second follow-up.

based on the transformed mean EAPs. In addition, the objective information on the student performance in the United States reported by The Nation's Report Card and other educational surveys (Allen et al., 2001; Neidorf, Binkley, Gattis, & Nohara, 2006) supports the following two facts: (a) Most or almost all groups of students between Grade 8 and Grade 12 make some progress (no stepping backward by group), and (b) the changes for the same group across two waves are not drastic or like outliers. The educational reality forms the basis to be used to judge if an outcome yielded by a model conflicts with reality. For example, in a longitudinal probability sample, it is unlikely to have a group that performs worse in Grade 12 after 4 years of studying.

In Table 5, the results yielded by the CMIRT model were consistent with the educational reality we observed. Among the second ability group (high/low), some may have been classified as high level by luck by the base-year measurement and then received slightly poor skill scores in the F1 measurement. Nonetheless, these students, as a group, should have made some progress in the 12th grade. The Embretson model yielded different results. For example, in Table 5, the mean score of the second ability group (high/low) at the F2 measurement yielded by the extended Embretson model is 183.68, that is, still lower than the base-year mean of 186.05; this can also be found in the left plot in Figure 6. It seems unrealistic that a group of students failed to make any progress on the test after 4 years of study. However, the results yielded by both 2PL CMIRT models, as illustrated by the right plot in Figure 6, did not show such contradictions. Other numbers yielded by the Embretson model were difficult to explain, in particular the extreme changes of mean EAPs between two waves. For example, the students in the high/high group gained approximately 60 points between base year and F1. Considering

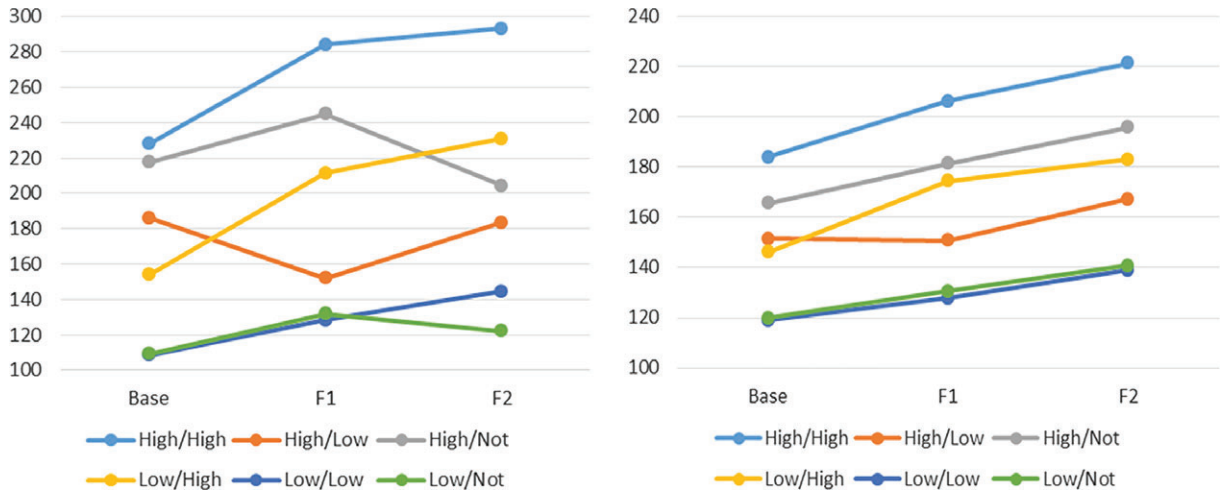

Figure 6 Performance of six ability groups for the National Education Longitudinal Study of 1988 reading (*left* Embretson model, *right* CMIRT).

that the jackknifed standard error for six ability groups ranges from 4.5 to 9.0 (Qian, 2015; Wolter, 1985), such a large jump of 60 points is hard to explain.

Some Remarks on Model Comparisons

The rationale for the structure of the Embretson model in Equation (10) can reveal the causes of the issues in fitting the NELS:88 data. First, as an extension of Rasch models, the Embretson (1991, p. 495) "model specifies a Wiener simplex pattern for the involvement of initial ability and one or more modifiabilities in response potential for successive measurement occasions (waves)." The model yields latent scales for each wave with Rasch models with comparable properties; although the latent mean and variance of the first wave are fixed, the model specifies no covariance structures across the scales of each wave. In comparison, the covariate structures between the scales for the CMIRT model are fully defined. Second, the Embretson model measures growth based on changes of item parameters for the Rasch models for each wave; additionally, the properties of the Rasch models for each wave are compatible, and the growth effects are additive. Instead, the CMIRT model measures growth by examinees' responses to items in each wave and constrains the item parameters to be invariant across the waves of the longitudinal assessment (Almond, 2011; van Rijn, Sinharay, Haberman, & Johnson, 2016).

As a high quality longitudinal survey, the NELS:88 program collected data through repeated assessments using booklets at different levels assembled from the same item pool; the growth model with the constraints seems appropriate for the NELS:88 longitudinal data.

Summary

In this study, a conditional 2PL MIRT model was proposed to fit the three waves of the NELS:88 reading data. In the model, the IRT parameters were constrained to be invariant across the three waves, and the growth in examinee skills was measured by the progress of examinees' performance on the assessments. The design of the adaptive NELS:88 reading assessment provided a basis for modeling growth with the constraints, setting the item location parameters of the same item invariant across measurement waves. The assessment instrument, also useful for developing a scale for growth, contained a standard test booklet for the first wave and two test booklets of different difficulty levels for each of the second and third waves; furthermore, each pair of the five booklets shared common items. For discussion of item parameter invariance, see Lord (1980).

In general, the results yielded by the CMIRT models provide a more reasonable explanation for the performance of different groups than those that had been yielded by the Embretson models. The empirical results yielded by the 2PL CMIRT model are consistent with the educational reality in the United States; most student groups made progress

in 4 years' study in schools, and the progress made across two waves was gradual, without radical changes. However, some of the results yielded by the Embretson models seem implausible. For instance, for the students in the second ability group (high/low), their mean EAP score in 12th grade was still lower than their base-year scores after 4 years' study. In addition, the students in the high/high group gained approximately 60 points between base year and eighth grade.

The results also show that the growth models for longitudinal data can certainly benefit from making use of demographic variables (DVs) such as gender, race, school location, and father's education. Evaluated by the model prediction indexes of log-likelihood, PE, GH, and AIC, all the models with incorporated DVs had an improved fit over the 2PL CMIRT model without using demographic information. Among the nine models fitted, the 2PL $\text{CMIRT}_{\text{DVA}}$ model demonstrated the best fit of the data. In this study, the growth models for NELS:88 were improved by use of auxiliary information because NELS:88 was a low-stakes assessment and the analysis was focused on reporting group scores. However, when the goal of an assessment is to estimate individual scores, major fairness issues can arise for the models with incorporated auxiliary information because two examinees with different demographic backgrounds could get different scores, even if they give identical item responses.

The *mirt* program offers flexibility in defining the design matrix of the covariate structure for the linear model of latent vectors. The 2PL $\text{CMIRT}_{\text{DVA}}$ model provides a specific example of how to specify a complex design matrix $\mathbf{T}^2$, which is expressed in a partitioned matrix, including the blocks of setting the item parameters equal across waves, defining the across-wave correlations, and designating parameters for a specific demographic group. Moreover, the stabilized Newton – Raphson algorithm provides an efficient way to calibrate the MIRT models and create a scale that measures growth at both group and individual levels. The flexibility in defining a design matrix within the efficient stabilized Newton – Raphson algorithm offers an edge over other programs based on the expectation – maximization (EM) algorithm using MML and the approach employing the MCMC method.

Several issues can be explored in future studies. The first one is the conditions under which the CMIRT growth models can yield valid scales in evaluating growth, such as the number of overlapping items and the size of the correlations between the latent traits across waves. Second, the growth model developed in this study, measuring changes in person-related (instead of item-related) parameters, can certainly be extended to other longitudinal assessment data, such as the Education Longitudinal Study of 2002 (Ingels, Pratt, Rogers, Siegel, & Stutts, 2005). Third, the idea of including auxiliary information for inference with latent models, though not MIRT, has been supported by many educational assessments, such as NAEP (Beaton & Zwick, 1990; Mislevy, 1991) and PISA (Turner & Adams, 2007). It is certainly of interest to compare the NAEP results with those yielded by the CMIRT models with auxiliary information.

Acknowledgments

The author thanks Shelby Haberman, Rebecca Zwick, Peter van Rijn, and Hongwen Guo for their suggestions and comments. The author also thanks Kim Fryer and Shuhong Li for editorial help. Any opinions expressed in this paper are those of the author and not necessarily those of Educational Testing Service.

Note

1 The population of NELS:88 covered the eligible eighth graders who were considered capable of participating in the surveys in the United States in 1988. Excluded from the sample were Bureau of Indian Affairs (BIA) schools, special education schools for the handicapped, area vocational schools that did not enroll students directly, and schools for dependents of U.S. personnel overseas.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267 – 281). Budapest, Hungary: Adadeiai Kiado.

Allen, N., Donoghue, J., & Schoeps, T. (2001). *The NAEP 1998 technical report* (NCES Report No. 2001-509). Washington, DC: National Center for Education Statistics.

Almond, R. (2011). *Estimating parameters of periodic assessments* (Research Memorandum No. RM-11-06). Princeton, NJ: Educational Testing Service.

Anderson, T. W. (1960). Some stochastic process models for intelligence test scores. In K. J. Arrow, S. Karlin, & P. Suppes (Eds.), *Mathematical methods in the social sciences* (pp. 205 – 220). Stanford, CA: Stanford University Press.

Beaton, A. E., & Zwick, R. (1990). *The effect of changes in the national assessment: Disentangling the NAEP 1985 – 86 reading anomaly* (Technical Report No. 17-TR-21). Princeton, NJ: Educational Testing Service/National Assessment of Educational Progress.

Bellman, R. (1970). *Introduction to matrix analysis* (2nd ed.). New York, NY: McGraw-Hill.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of the EM algorithm. *Psychometrika*, *46*, 443 – 459. https://doi.org/10.1007/BF02293801

Cai, L. (2013). *flexMIRT: Flexible Multilevel Item Factor Analysis and Test Scoring* (Version 2) [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Doran, H. C., & Jiang, T. (2006). The impact of linking error in longitudinal analysis: An empirical demonstration. In R. Lissitz (Ed.), *Longitudinal and value added models of student performance* (pp. 210 – 229). Maple Grove, MN: JAM.

Draper, N., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York, NY: John Wiley.

Embretson, S. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495 – 515. https://doi.org/10.1007/BF02294487

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics*, *17*, 111 – 117. https://doi.org/10.1080/00401706.1975.10489279

Fischer, G. H., & Seliger, E. (1997). Multidimensional linear logistic models for change. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 323 – 346). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-2691-6_19

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

Gilula, Z., & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association, 89*, 645 – 656. https://doi.org/10.1080/01621459.1994.10476789

Gilula, Z., & Haberman, S. J. (1995). Prediction functions for categorical panel data. *Annals of Statistics, 23*, 1130 – 1142. https://doi.org/10.1214/aos/1176324701

Haberman, S. J. (1988). A stabilized Newton – Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology*, *18*, 193 – 211. https://doi.org/10.2307/271049

Haberman, S. J. (2006a). *An elementary test of the normal 2PL model against the normal 3PL model* (Research Report No. RR-06-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02020.x

Haberman, S. J. (2006b). *Adaptive quadrature for item response models* (Research Report No. RR-06-29). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02035.x

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204 – 229. https://doi.org/10.3102/1076998607302636

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. https://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. https://doi.org/10.1002/j.2333-8504.2013.tb02339.x

Haberman, S. J., & Sinharay, S. (2010). Reporting subscores using multidimensional item response theory. *Psychometrika, 75*, 209 – 227. https://doi.org/10.1007/s11336-010-9158-4

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. https://doi.org/10.1007/s11336-012-9305-1

Haberman, S. J., von Davier, M., & Lee, Y. (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. https://doi.org/10.1002/j.2333-8504.2008.tb02131.x

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika*, *55*, 5 – 18. https://doi.org/10.1007/BF02294739

Ingels, S. J., Pratt, D. J., Rogers, J., Siegel, P. H., & Stutts, E. S. (2005). *Education longitudinal study of 2002: Base-year to first follow-up data file documentation* (Report No. 2006-344). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Joreskog, K. G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology, 23*, 121 – 145. https://doi.org/10.1111/j.2044-8317.1970.tb00439.x

Kaplan, D., & Sweetman, H. M. (2006). Finite mixture modeling approaches to the study of growth in academic achievement. In R. Lissitz (Ed.), *Longitudinal and value added models of student performance* (pp. 130 – 169). Maple Grove, MN: JAM.

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. https://doi.org/10.1007/978-1-4757-4310-4

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillside, NJ: Lawrence Erlbaum.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44*, 226–233.

Millsap, R. E. (2008). Introduction to the special issue on growth models for longitudinal data in educational research. *Educational Research and Evaluation: An International Journal on Theory and Practice, 14*, 283–285. https://doi.org/10.1080/13803610802249308

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196. https://doi.org/10.1007/BF02294457

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Zwick, R. (2012). Scaling, linking, and reporting in a periodic assessment system. *Journal of Educational Measurement, 49*, 148–166. https://doi.org/10.1111/j.1745-3984.2012.00166.x

Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). Comparing mathematics content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Programme for International Student Assessment (PISA) 2003 assessments (NCES 2006-029). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Patz, R., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 253–272). https://doi.org/10.1007/978-0-387-49771-6_14

Qian, J. (2010, April). *Effects of sampling variability on dimensionality analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

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.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. D. (2009). *Multidimensional item response theory: Statistics for social and behavioral sciences*. New York, NY: Springer. https://doi.org/10.1007/978-0-387-89976-3

Rijmen, F. (2010). *Measuring multidimensional latent growth* (Research Report No. RR-10-24). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02231.x

Roberts, J. S., & Ma, Q. (2006). IRT models for the assessment of change across repeated measurements. In R. Lissitz (Ed.), *Longitudinal and value added models of student performance* (pp. 100–129). Maple Grove, MN: JAM.

Rock, D. A. (2012). *Modeling change in large-scale longitudinal studies of educational growth: Four decades of contributions to the assessment of educational growth* (Research Report No. RR-12-04). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02286.x

Rock, D. A., Pollack, J. M., & Quinn, P. (1995). *Psychometric report for the NELS:88 base year through second follow-up* (Report No. 95-382). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Sijtsma, K. (2001). Developments in measurement of persons and items by means of item response models. *Behaviormetrika, 28*, 65–94. https://doi.org/10.2333/bhmk.28.65

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, *44*, 249–275. https://doi.org/10.1111/j.1745-3984.2007.00037.x

Spencer, B. D., Frankel, M. R., Ingels, S. L., Rasinski, K. A., & Tourangeau, R. (1990). *Base year sample design report, National Education Longitudinal Study of 1988*. Washington, DC: National Center for Education Statistics.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210. https://doi.org/10.1177/014662168300700208

te Marvelde, J. M., Glas, C. A. W., van Landeghem, G., & van Damme, J. V. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement, 66*, 5–34. https://doi.org/10.1177/0013164405282490

Turner, R., & Adams, R. J. (2007). The Program for International Student Assessment: An overview. *Journal of Applied Measurement, 8*, 237–248.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychology Measurement, 10*, 333–344. https://doi.org/10.1177/014662168601000402

van Rijn, P. (2008). *Categorical time series in psychological measurement* (Ph.D. thesis). Amsterdam, Holland: University of Amsterdam.

van Rijn, P., Sinharay, S., Haberman, S. J. & Johnson, M. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-scale Assessments in Education*, *4*(10), 1–23. https://doi.org/10.1186/s40536-016-0025-3

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb01993.x

von Davier, M., & Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics, 35*, 174 – 193. https://doi.org/10.3102/1076998609346970

von Davier, M., Xu, X., & Carstensen, C. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76*, 318 – 336. https://doi.org/10.1007/s11336-011-9202-z

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics, 9*, 60 – 62. https://doi.org/10.1214/aoms/1177732360

Wingersky, M., & Lord, F. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*, 347 – 364. https://doi.org/10.1177/014662168400800312

Wolter, K. (1985). *Introduction to variance estimation*. New York, NY: Springer.

Xu, X., & Qian, J. (2009). Longitudinal data analysis for NELS 88 reading assessment data. Unpublished manuscript.

Yao, L. (2008). *BMIRT: Bayesian multivariate item response theory* (Version 1.0). Monterey, CA: CTB/McGraw-Hill.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*, 83 – 105. https://doi.org/10.1177/0146621606291559

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement*, *36*, 5, 375 – 398. https://doi.org/10.1177/0146621612445904

Zhang, J., & Stout, W. (1999). Conditional covariance structure of compensatory multidimensional items. *Psychometrika, 64*, 129 – 152. https://doi.org/10.1007/BF02294532

Appendix

Structure of the Design Matrix for the Two-Parameter Logistic CMIRT Model

As defined in the Design Matrix for the Linear Model of Latent Vectors section in this report, $\boldsymbol{\beta}$ and $\mathbf{o}$ are the parameter vector of the linear model and the offset vector of dimension B, respectively; $\boldsymbol{\gamma}$ is the vector of dimension C, where C is an integer greater than zero. The parameter vector of the linear model in Equation (5) can be expressed as

$$\boldsymbol{\beta} = \mathbf{o} + \mathrm{T}\boldsymbol{\gamma} = \mathrm{T}^1\left(\mathbf{o}^2 + \mathrm{T}^2\boldsymbol{\gamma}\right),$$

where T^1 is a $B \times B_1$ matrix and T^2 is a $B_1 \times C$ matrix. Because $\boldsymbol{\beta} = \mathrm{T}^1\boldsymbol{\beta}^1$,

$$\boldsymbol{\beta}^1 = \mathbf{o}^2 + \mathrm{T}^2\boldsymbol{\gamma}.$$

When T^2 is partitioned as a block matrix, the main task of modeling is to define the functionality of each block in T^2. Whenever T^2 has been defined, the *mirt* program can automatically generate T^1; the stabilized Newton – Raphson algorithm is the procedure used to solve the log-likelihood function

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i l_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \log P\left(Y_i \middle| z_i, \boldsymbol{\beta}\right)$$

for MML estimation. In the multivariate normal case, the procedure is also featured with the adaptive quadrature technique, which is based on the adaptive Gauss – Hermite integration (Haberman, 2006b).

The NELS:88 reading assessment contains 54 multiple-choice items. By the item map (Rock et al., 1995), the number of distinct items selected for the three waves, that is, base year, F1, and F2, is 21, 35, and 34, respectively; the total number of items is 90. In this study, for the 2PL MIRT model of three dimensions, the structure of T^2 is a $B_1 \times C$ block matrix, that is, with $(90 + 270 + 9) \times (54 + 54 + 7)$ dimensions:

$$\mathrm{T}^2 = \begin{pmatrix} \mathrm{T}^2_{\tau} & \mathrm{O}_{12} & \mathrm{O}_{13} \\ \mathrm{O}_{21} & \mathrm{T}^2_{ad} & \mathrm{O}_{23} \\ \mathrm{O}_{31} & \mathrm{O}_{32} & \mathrm{T}^2_{\lambda} \end{pmatrix},$$

where O_{12}, O_{13}, O_{21}, O_{23}, O_{31}, and O_{32} are zero matrixes.

The matrix block T^2_{τ}, 90×54 is defined for unknown *location parameters*; the dimensions 90 and 54 correspond to the 90 items used for the three waves (=21 + 35 + 34) and the 54 items used in the reading assessment. Note that in this study, the location parameters of the overlapping items are set to be invariant across the three waves; otherwise, the number of the parameters could triple. When an item is used, the identification conditions are set to 1; otherwise, they are set to 0.

For example, because the item with ID 1 is used in all three waves, the elements $t_{\tau,\,1,\,1}$, $t_{\tau,\,22,\,1}$, and $t_{\tau,\,57,\,1}$ in $\mathbf{T}_{\tau}^{2}$ are set to 1; because the item with ID 29 is only used once in the high-level test booklet in the second wave, $t_{\tau,\,45,\,29}$ in $\mathbf{T}_{\tau}^{2}$ is set to 1; and so on.

The matrix block $\mathbf{T}_{ad}^{2}$, 270×54, is defined for unknown *slope parameters*. The vectors of the slope parameters for the three waves are different, so 270 in total for 90 items. Vectors of the slope parameters are $(1\ 0\ 0)'$, $\begin{pmatrix}0 & 1 & 0\end{pmatrix}'$, and $\begin{pmatrix}0 & 0 & 1\end{pmatrix}'$ for the base year, F1, and F2 assessments, respectively. For the case of the first item, (91, 55) in $\mathbf{T}_{ad}^{2}$, because the item with ID 1 is used in all three waves, in $\mathbf{T}_{ad}^{2}$, the elements

$$\begin{pmatrix}t_{ad,91,55} & t_{ad,92,55} & t_{ad,93,55}\end{pmatrix}' = \begin{pmatrix}1 & 0 & 0\end{pmatrix}',$$

$$\begin{pmatrix}t_{ad,154,55} & t_{ad,155,55} & t_{ad,156,55}\end{pmatrix}' = \begin{pmatrix}0 & 1 & 0\end{pmatrix}',$$

$$\begin{pmatrix}t_{ad,259,55} & t_{ad,260,55} & t_{ad,261,55}\end{pmatrix}' = \begin{pmatrix}0 & 0 & 1\end{pmatrix}';$$

because the item with ID 29 is only used once in the high-level test booklet in the second wave, in $\mathbf{T}_{ad}^{2}$, the element

$$\begin{pmatrix}t_{ad,223,84} & t_{ad,224,84} & t_{ad,225,84}\end{pmatrix}' = \begin{pmatrix}1 & 0 & 0\end{pmatrix}',$$

and so on. The vector $\mathbf{o}^{2}$ in the model has 369 elements with zero elements, except $o_{364}=o_{366}=o_{369}=-0.5$. The block matrix $\mathbf{T}_{\lambda}^{2}$ is defined for unknown *predictors and quadratic terms*,

$$\mathbf{T}_{\lambda}^{2}=\begin{pmatrix}\mathbf{T}_{p}^{2} & \mathbf{O}_{\boldsymbol{\lambda}12}\\ \mathbf{O}_{\boldsymbol{\lambda}21} & \mathbf{Q}_{p}^{2}\end{pmatrix},$$

with dimensions $(3+6)\times(2+5)$. In $\mathbf{T}_{\lambda}^{2}$, $\boldsymbol{O}_{\boldsymbol{\lambda}12}$ and $\boldsymbol{O}_{\boldsymbol{\lambda}21}$ are zero matrixes:

$$\mathbf{T}_{p}^{2}=\begin{pmatrix}0 & 0\\ 1 & 0\\ 0 & 1\end{pmatrix}$$

$$\mathbf{Q}_{p}^{2}=\begin{pmatrix}0 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 1\end{pmatrix}.$$

In the setup of $\mathbf{Q}_{p}^{2}$ for the 2PL CMIRT model, the estimated $\Lambda_{11}(\boldsymbol{\lambda},z)$, in (7), is set to 1. Although the initial values of $\Lambda_{22}(\boldsymbol{\lambda},z)$ and $\Lambda_{33}(\boldsymbol{\lambda},z)$ in $\mathbf{o}^{2}$ are also set to 1, conditioned on the F1 estimates, the values of $\Lambda_{22}(\boldsymbol{\lambda},z)$ and $\Lambda_{33}(\boldsymbol{\lambda},z)$ are updated in each iteration, and their yielded estimates are not the same as their initial values.

Structure of the Design Matrix for the Two-Parameter Logistic CMIRT Model With Demographic Information

In this study, to improve the accuracy of estimation, the 2PL CMIRT model has made use of auxiliary information.

Extended Block Matrix for CMIRT$_{\mathrm{DVA}}$

The structure of $\mathbf{T}_{DVA}^{2}$, for the model with demographic variables incorporated (2PL CMIRT$_{\mathrm{DVA}}$), is a 468×148 block matrix, that is, with $(90+270+108)\times(54+54+40)$ dimensions. Thus 148 parameters need to be estimated for the model; the parameter vector of the linear model $\boldsymbol{\beta}_{DVA}^{1}$ and the offset vector $\mathbf{o}_{DVA}^{2}$ both have 468 elements. The vector $\mathbf{o}_{DVA}^{2}$ has 465 zero elements, except $o_{397}=o_{399}=o_{402}=-0.5$.

Table A1 Mean Expected A Posteriori Scores and Standard Deviations of Six Ability Groups for Demographic Groups Yielded by the Two-Parameter Logistic CMIRT and Two-Parameter Logistic $CMIRT_{DVA}$ Models

	2PL CMIRT						2PL $CMIRT_{DVA}$					
	Base		F1		F2		Base		F1		F2	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
	Total						Total					
High/high	183.98	23.08	206.25	28.07	221.50	36.13	184.27	23.10	206.35	28.17	221.65	36.36
High/low	151.61	12.17	150.70	17.10	167.19	27.09	151.99	12.20	150.64	17.10	167.06	27.12
High/not present	165.60	31.43	181.38	39.42	195.87	42.98	165.99	31.45	181.27	40.11	196.01	43.71
Low/high	146.06	18.73	174.46	19.69	183.01	28.66	146.46	18.92	174.57	19.85	183.17	28.96
Low/low	119.01	15.35	127.81	19.96	138.87	26.31	119.48	15.36	127.97	19.80	138.91	26.15
Low/not present	119.99	18.83	130.52	26.81	140.70	26.35	120.52	18.89	130.43	26.88	140.30	26.88
	Male						Male					
High/high	182.97	22.77	205.89	27.65	219.96	36.32	182.63	22.57	205.96	27.60	219.34	36.30
High/low	151.43	12.65	149.10	17.67	164.33	29.49	151.49	12.61	149.22	17.62	163.87	29.36
High/not present	163.25	31.89	179.05	40.03	193.53	43.63	162.93	31.58	178.79	40.37	192.45	43.77
Low/high	143.99	17.43	172.60	18.41	180.32	28.12	143.93	17.35	172.74	18.45	179.91	28.16
Low/low	117.37	15.45	125.60	20.12	136.33	26.75	117.57	15.37	125.93	19.97	136.10	26.49
Low/not present	118.04	19.13	127.87	27.50	137.98	27.00	118.12	19.12	127.74	27.51	136.46	27.29
	Female						Female					
High/high	184.82	23.31	206.56	28.43	222.78	35.93	185.63	23.45	206.69	28.64	223.58	36.31
High/low	151.79	11.69	152.27	16.39	170.02	24.19	152.47	11.76	152.05	16.46	170.20	24.33
High/not present	167.73	30.86	183.50	38.74	198.00	42.30	168.78	31.09	183.51	39.76	199.25	43.43
Low/high	148.16	19.75	176.34	20.74	185.72	28.97	149.01	20.08	176.42	21.02	186.45	29.42
Low/low	120.96	15.01	130.42	19.44	141.85	25.47	121.73	15.05	130.37	19.33	142.21	25.34
Low/not present	122.15	18.27	133.45	25.72	143.71	25.29	123.18	18.27	133.42	25.85	144.56	25.77
	Asian/others						Asian/others					
High/high	187.08	23.87	211.14	28.67	230.13	35.68	194.29	23.68	218.16	28.44	237.00	35.39
High/low	152.28	12.54	152.09	16.63	170.20	25.11	159.77	12.44	159.58	16.50	177.55	24.92
High/not present	169.66	32.31	187.71	39.82	202.83	42.88	177.01	32.05	194.92	39.50	209.92	42.54
Low/high	145.84	18.86	175.18	18.08	185.40	27.64	153.38	18.71	182.48	17.94	192.62	27.42
Low/low	117.43	16.50	126.00	21.88	136.79	28.63	125.19	16.37	133.69	21.71	144.40	28.41
Low/not present	118.57	19.44	130.28	27.48	139.97	27.00	126.32	19.29	137.94	27.26	147.55	26.79
	Black						Black					
High/high	174.56	20.92	195.04	25.31	207.07	34.09	181.87	20.75	202.19	25.11	214.12	33.82
High/low	150.72	12.09	150.26	17.14	167.86	23.93	158.22	12.00	157.77	17.01	175.22	23.74
High/not present	147.07	30.99	159.14	37.28	171.20	42.27	154.60	30.75	166.57	36.99	178.53	41.93
Low/high	143.01	17.57	171.45	17.42	179.84	27.77	150.57	17.44	178.78	17.28	187.11	27.55
Low/low	117.79	15.34	125.96	19.67	136.80	24.91	125.55	15.22	133.66	19.51	144.41	24.72
Low/not present	117.41	18.19	126.73	24.78	137.03	24.44	125.17	18.04	134.42	24.58	144.64	24.25
	Hispanic						Hispanic					
High/high	174.11	21.39	195.04	25.43	205.78	34.08	181.43	21.22	202.19	25.23	212.84	33.81
High/low	146.96	11.07	147.29	14.50	159.27	24.66	154.49	10.98	154.82	14.38	166.70	24.46
High/not present	148.75	32.22	161.92	39.08	173.35	44.49	156.26	31.97	169.34	38.78	180.67	44.14
Low/high	139.83	14.09	165.91	15.33	168.95	24.26	147.41	13.98	173.28	15.21	176.31	24.06
Low/low	114.82	15.61	122.73	19.64	130.99	24.57	122.60	15.48	130.45	19.48	138.65	24.38
Low/not present	114.93	18.15	124.19	24.44	133.88	24.60	122.71	18.01	131.90	24.25	141.52	24.41
	White						White					
High/high	185.06	22.95	207.40	28.01	222.84	35.90	192.29	22.77	214.45	27.79	229.77	35.62
High/low	152.19	12.17	151.00	17.38	167.64	27.89	159.68	12.07	158.49	17.25	175.01	27.67
High/not present	169.86	29.58	186.23	37.84	201.37	40.72	177.21	29.34	193.45	37.55	208.46	40.40
Low/high	147.28	19.17	175.80	20.40	184.74	28.96	154.80	19.02	183.10	20.24	191.97	28.73
Low/low	120.64	14.89	129.87	19.58	141.74	26.35	128.37	14.77	137.53	19.42	149.31	26.15
Low/not present	122.47	18.74	133.58	27.54	143.92	26.83	130.20	18.59	141.21	27.32	151.47	26.62

Note. All expected a posteriori (EAP) mean scores are transformed by a linear transformation, which sets base-year EAP distribution with mean and *SD* of [150, 35]. 2PL = two-parameter logistic; CMIRT = conditional multidimensional item response theory; DV = demographic variable; F1 = first follow-up; F2 = second follow-up.

The matrix blocks $\mathbf{T}^2_{\tau}$ and $\mathbf{T}^2_{ad}$ are the same; $\mathbf{T}^2_{\lambda,\text{DVA}}$ is an extended block matrix for $\text{CMIRT}_{\text{DVA}}$ with *auxiliary information* added; therefore, for three test waves, each of the 11 dummy DVs in Table 1 is coded as a 3×3 identity matrix $I_1, \ldots, I_{10}$, and I_{11},

$$\mathbf{T}^2_{\lambda,\text{DVA}} = \begin{pmatrix} \mathbf{T}^2_{p,\text{DVA}} & \mathbf{O}_{\lambda^*12} \\ \mathbf{O}_{\lambda^*21} & \mathbf{Q}^2_{p,\text{DVA}} \end{pmatrix}$$

with dimensions $(36+72)\times(35+5)$. The matrix

$$\mathbf{T}^2_{p,\text{DVA}} = \begin{pmatrix} \mathbf{T}^2_{p} & \mathbf{O}_{\lambda,1,2} & \cdots & \mathbf{O}_{\lambda,1,11} \\ \mathbf{O}_{\lambda,2,1} & I_1 & \cdots & \mathbf{O}_{\lambda,2,11} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O}_{\lambda,12,1} & \mathbf{O}_{\lambda,12,2} & \cdots & I_{11} \end{pmatrix},$$

where $\mathbf{T}^2_p$ is defined the same as earlier; 11 identity matrixes, $I_1, \ldots, I_{10}$, and I_{11}, are set on the diagonal of $\mathbf{T}^2_{p,\text{DVA}}$, and 132 other matrixes, in symbols of $O_{\lambda,\star,\star}$ in $\mathbf{T}^2_{p,\text{DVA}}$, are zero matrixes with appropriate dimensions. The matrix $\mathbf{Q}^2_{p,\text{DVA}}$ is defined as

$$\mathbf{Q}^2_{p,\text{DVA}} = \begin{pmatrix} \mathbf{Q}^2_{p} \\ \mathbf{O}_{p^*1} \\ \vdots \\ \mathbf{O}_{p^*11} \end{pmatrix},$$

where $\mathbf{Q}^2_p$, 6×5 dimensions, is defined the same as earlier; 11 matrixes, $\mathbf{O}_{p^*1}, \ldots, \mathbf{O}_{p^*11}$, are zero matrixes with 6×5 dimensions.

Extended Block Matrix for $\text{CMIRT}_{\text{DVB}}$

The structure of $\mathbf{T}^2_{DVB}$, for the model with demographic variables incorporated (2PL $\text{CMIRT}_{\text{DVB}}$), is a 459×145 block matrix, that is, with $(90+270+99)\times(54+54+37)$ dimensions. Thus 145 parameters need to be estimated for the model; the parameter vector of the linear model $\boldsymbol{\beta}^1_{DVB}$ and the offset vector $\mathbf{o}^2_{DVB}$ both have 459 elements. The vector $\mathbf{o}^2_{DVB}$ has 456 zero elements, except $o_{394} = o_{396} = o_{399} = -0.5$.

The matrix blocks $\mathbf{T}^2_{\tau}$ and $\mathbf{T}^2_{ad}$ are the same; $\mathbf{T}^2_{\lambda,\text{DVB}}$ is an extended block matrix with *auxiliary information* added; therefore, for three test waves, each of the 10 dummy DVs in Table 1 is coded as a 3×3 identity matrix $I_1, \ldots, I_9$, and I_{10},

$$\mathbf{T}^2_{\lambda,\text{DVB}} = \begin{pmatrix} \mathbf{T}^2_{p,\text{DVB}} & \mathbf{O}_{\lambda^*12} \\ \mathbf{O}_{\lambda^*21} & \mathbf{Q}^2_{p,\text{DVB}} \end{pmatrix}$$

with dimensions $(33+66)\times(32+5)$. The matrix

$$\mathbf{T}^2_{p,\text{DVB}} = \begin{pmatrix} \mathbf{T}^2_{p} & \mathbf{O}_{\lambda,1,2} & \cdots & \mathbf{O}_{\lambda,1,10} \\ \mathbf{O}_{\lambda,2,1} & I_1 & \cdots & \mathbf{O}_{\lambda,2,10} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O}_{\lambda,11,1} & \mathbf{O}_{\lambda,11,2} & \cdots & I_{10} \end{pmatrix},$$

where $\mathbf{T}^2_p$ is defined the same as earlier; 10 identity matrixes, $I_1, \ldots, I_9$, and I_{10}, are set on the diagonal of $\mathbf{T}^2_{p,\text{DVB}}$, and 110 other matrixes, in symbols of $O_{\lambda,\star,\star}$ in $\mathbf{T}^2_{p,\text{DVB}}$, are zero matrixes with appropriate dimensions. The matrix $\mathbf{Q}^2_{p,\text{DVB}}$ is defined as

$$\mathbf{Q}^2_{p,\text{DVB}} = \begin{pmatrix} \mathbf{Q}^2_{p} \\ \mathbf{O}_{p^*1} \\ \vdots \\ \mathbf{O}_{p^*10} \end{pmatrix},$$

where $\mathbf{Q}_p^2$, 6×5 dimensions, is defined the same as earlier; 10 matrixes, $\mathbf{O}_{p^*1}, \ldots, \mathbf{O}_{p^*10}$, are zero matrixes with 6×5 dimensions.

For $\text{CMIRT}_{\text{DVC}}$ to $\text{CMIRT}_{\text{DVH}}$, based on the demographic included variables, their extended block matrixes with auxiliary information ($\mathbf{T}^2_{\lambda,\text{DVC}}$ to $\mathbf{T}^2_{\lambda,\text{DVH}}$) can be also constructed.

Suggested citation:

Qian, J. (2018). *Modeling growth with adaptive longitudinal large-scale assessments* (Research Report No. RR-18-34). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12220

Action Editor: Barbara Zwick

Reviewers: Hongwen Guo and Peter van Rijn