JOURNAL OF APPLIED MEASUREMENT, *13*(1),

Formulating Latent Growth Using an Explanatory Item Response Model Approach

Mark Wilson

Xiaohui Zheng
University of California, Berkeley

Leah McGuire
University of Minnesota

In this paper, we present a way to extend the Hierarchical Generalized Linear Model (HGLM; Kamata (2001), Raudenbush (1995)) to include the many forms of measurement models available under the formulation known as the Random Coefficients Multinomial Logit (MRCML) Model (Adams, Wilson and Wang, 1997), and apply that to growth modeling. First, we review two different traditions in modeling growth studies: the first is based in the hierarchical linear modeling (HLM) tradition, and the second, which is the topic of this paper, is rooted in the Rasch measurement tradition—this is the *linear* Latent Growth Item Response Model (LG-IRM). Going beyond the linear case, the LG-IRM approach allows us to considerably extend the range of models available in the HLM tradition to incorporate several of the extensions of IRT models that are used in creating explanatory item response models (EIRM; De Boeck and Wilson, 2004). We next present a number of extensions—including polynomial growth modeling, differential item functioning (DIF) effects, growth functions that can be approximated by polynomial expressions, provision for polytomous responses, person and item covariates (and time varying covariates), and multiple dimensions of growth. We provide two empirical examples to illustrate several of the models, using the ConQuest software (Wu, Adams, Wilson and Haldane, 2008) to carry out the analyses. We also provide several simulations to investigate the success of the estimation procedures.

Requests for reprints should be sent to

Introduction

Individual growth curve models are an important class of longitudinal models used in educational research (e.g., see review in Raudenbush, 2001). Under the No Child Left Behind Act of 2001 (NCLB, 2002), U.S. schools must make "adequate yearly progress" (AYP) in the academic areas of reading and mathematics. While clear in its goals, in practice NCLB allows a variety of ways that this growth may be evaluated, from "status" models to "change" and "value-added" models. The most common practice used so far to demonstrate progress under the NCLB has been to show improvements in test scores from one year to the next for successive cohorts of students in each of several grade levels (using so-called "status" models). In this case, the changes could be attributable to differences in the makeup of a particular cohort rather than to individual progress. Such analyses do not answer questions about how a particular student is learning over time, nor are they rigorous in their attribution of the causes of change, hence their use in evaluation of program effects is also questionable. Cohort-comparison and status models may often be used simply due to lack of longitudinal data (e.g. Thorn and Meyer, 2001). In addition, researchers have presented major concerns about the ability to agree on the many complex methodological (Briggs and Weeks, 2009) and conceptual decisions (Betebenner, 2009) involved in individual growth curve modeling in this area. However, with increasing availability of longitudinal data and available and appropriate methodology, attention is being drawn to models that incorporate individual student growth estimates.

Longitudinal data are also referred to as "repeated measures" data. They arise when measurements are taken repeatedly on the same sample of subjects over multiple time-points or occasions. In a typical example longitudinal data might take the form of a series of academic achievement outcomes collected from a group of students over several years. Various models are available for the analysis of such data where the goal is to track changes in student achievement. In this paper, we first provide a brief review of various longitudinal models developed from different theoretical perspectives. A multidimensional item response model for analyzing repeated measures is then proposed, which is a special case of the Multidimensional Random Coefficients Multinomial Logit (MRCML) Model (Adams, Wilson and Wang, 1997). The proposed Latent Growth Item Response Model (LG-IRM) incorporates dimensions (i.e., random effects) for different aspects of the individual growth curves. Correlations among these dimensions are also estimated to describe the full growth model. The proposed model combines a conventional growth model for change with an item response model for item analysis. Specifically, linear and curvilinear growth models are incorporated into the item response function from a multidimensional perspective. Because the MRCML is a generalized item response model, incorporating the possibility for many classes of measurement models, such as polytomous, multidimensional, latent regression, facets (LLTM) and DIF models, the LG-IRM inherits these possibilities also. All of these models, including the new LG-IRM versions, can be estimated using the ConQuest software (Wu, Adams Wilson and Haldane, 2008). It uses a marginal maximum likelihood procedure to obtain parameter estimates (Adams, Wilson and Wang, 1997).

In this paper, we first review two different traditions in modeling growth studies: the first is based in the hierarchical linear modeling (HLM) tradition, and the second, which is the topic of this paper, is rooted in the Rasch measurement tradition. We show how the two traditions can be combined using the framework of the MRCML approach to produce, first, the *linear* LG-IRM. With the proposed linear LG-IRM model, we are able to examine change over time from the measurement perspective within the IRT framework. The model integrates the HLM formulation for measuring change with various types of IRT models to accommodate measurement complications. It sets a basis for IRT-related analyses in longitudinal settings, such as the DIF analysis and vertical equating. At this point we introduce

two empirical studies from educational and psychological perspectives to illustrate the use of the linear LG-IRM. However, this linear model is just the starting point for the LG-IRM models. The LG-IRM approach allows us to considerably extend the range of models available in the HLM tradition to incorporate several of the extensions of item response theory (IRT) models that are used in creating explanatory item response models (EIRM; De Boeck and Wilson, 2004). Hence, we next present a number of extensions to the traditional HLM-type growth models that can be made using the LG-IRM formulation. These include polynomial growth modeling, differential item functioning (DIF) effects, growth functions that can be approximated by polynomial expressions, provision for polytomous responses, person and item covariates (and time varying covariates), and multiple dimensions of growth. For the first two of these extensions, we provide some illustrative analyses using data from the empirical examples mentioned above. This is followed by a description of simulation studies undertaken to assess the performance of the proposed model and its practicality in the educational context including one where a vertical scale is used to link between test forms from year to year. We also present a simulation that shows that the LG-IRM results can match closely results from a HLM analysis. The paper concludes with a brief discussion that summarizes the salient points in the paper and comments on it.

Background

Hierarchical linear modeling approaches

Hierarchical linear modeling (HLM) has been widely used for analyzing longitudinal data. The simplest model for continuous responses is a linear regression model, where the responses are modeled as a linear combination of covariates and an error term that has an independent normal distribution. In ordinary linear regression models, the coefficients of the covariates are the same for all subjects. In a hierarchical linear model, some coefficients are allowed to vary between subjects. The coefficients are decomposed into a population average effect and a subject-specific random effect. Following the two-stage approach of Raudenbush and Bryk (2002), we see such a model as a two-level HLM. In the level-1 model, individual responses are determined by a set of subject-specific covariates; and in the level-2 model, the coefficients of the covariates become the outcomes and are decomposed into population average effects and subject-specific random effects.

Latent Growth Modeling (LGM) for longitudinal data is developed from a structural equation modeling (SEM) approach. It has been shown that HLM models for measuring change can be mapped onto the SEM framework (Willett and Keiley, 2000; Singer and Willett, 2003; Duncan, Duncan and Strycker, 2006). LGMs specify individual growth trajectories based (typically) on two growth parameters represented by latent variables: the initial status (the intercept) and the growth rate (the slope). The intercept and slope vary across subjects, and can be regressed on time invariant covariates. LGMs can also be formulated as structural equation models with latent variables. The measurement model characterizes person-specific growth over time, and the structural model describes individual differences in the trajectories.

In educational testing, the measurement model typically focuses on a mapping of item responses to a latent variable. Item responses are coded either dichotomously as correct or incorrect, or in terms of quality (i.e., "partial credit"). Generalized Linear Models (GLMs; Nelder and Wedderburn, 1972) extend linear regressions for continuous responses to accommodate categorical responses by using logit (or other) links combined with binomial (or other) distributions. An extension of the GLM for hierarchical data is the Generalized Linear Mixed Model (GLMM; Breslow and Clayton, 1993), or Hierarchical Generalized Linear Model (HGLM; Raudenbush, 1995). Kamata (2001) illustrated the equivalence between the Rasch model and a two-level HGLM for dichotomous responses with a random intercept but no random coefficients. He also extended

the two-level HGLM model by including person-specific variables in the second level of the model. Pastor and Beretvas (2006) presented a three-level HGLM for repeated measures. The model, which they called a multilevel longitudinal Rasch model, was used to investigate changes in the latent trait over time. Fox and Glas (2001; 2003) generalized a conventional random coefficient model by defining covariates as latent variables at any level of a hierarchical structure. A Bayesian approach using the Markov Chain Monte Carlo (MCMC) estimation procedure was used to estimate model parameters. Rabe-Hesketh and her colleagues presented Generalized Linear Latent and Mixed Models (GLLAMMs; Rabe-Hesketh, Skrondal and Pickles, 2004; Skrondal and Rabe-Hesketh, 2004). This framework encompasses a large variety of latent variable models, including multilevel versions of standard IRT models and all of the models mentioned above.

In their review of psychological research using SEM in longitudinal studies McCallum and Austin identified two common approaches that are used separately or in combination; the sequential design and the repeated measures design (2000). In the *sequential* design, a different latent variable is measured at each time point and often the variable measured at one time point is thought to influence a variable measured after some time interval In the *repeated measures* design, the same latent variable is measured over time. This approach includes autorgressive models (see Jöreskog 1979, McArdle and Aber 1990) as well as latent growth curve models, which can take linear or non-linear shapes. The SEM framework is powerful and certainly capable of answering questions about the amount of change over time. However we chose to apply IRT to study change over time because we are also interested in taking advantage of the features of IRT software such as DIF, item bundle-models, and explanatory item response models that can be applied in complex measurement situations.

Item response modeling approaches

Prior to the development of generalized HLMs, measurement researchers developed multidimensional models for latent growth. The models in this area posit that a person's responses at a certain time point are related to the person's location on the latent variable at each time point where the latent ability at each time point is represented by one or more dimensions. Andersen (1985) proposed an extended Rasch model for dichotomous responses in longitudinal form. In his model, a separate factor is estimated for each time point. Figure 1 shows a path diagram of Andersen's model, using an illustrative example with two items and three time points. In the diagram, the circles represent latent variables (θ_{p1}, θ_{p2}, and θ_{p3}) that indicate the abilities of person p at Time 1, Time 2 and Time 3, respectively. The heptagons on the bottom refer to the item parameters δ_1 and δ_2 which are assumed constant across time. Usually there would be many items for each time point—just two items are shown to keep the diagram simple. In the measurement model, item responses (shown in the squares) are related to the time-invariant item difficulties and the time-varying dimensions of person ability. For example, X_{1p1} and X_{2p1} are the responses to item 1 and 2 from person p at Time 1, and X_{1p2} and X_{2p2} are the two item responses from person p at Time 2, and so on. The arrows from the latent variables and the item difficulties to the item responses represent nonlinear (logistic) relationships between the (probabilities of the) item responses and the latent variables. Usually, these straight arrows are labeled with the weights associated with each source of variation—those from the latent variables would be called "factor loadings" in a factor analysis setting, "item discrimination parameters" in a 2PL item response model, or simply, "item scores" in the Rasch setting. In this case (i.e., for the Rash model), all of the weights are equal to 1, and we use a graphical convention of leaving out the labels when the weights are 1. The short arrows are the residuals. In Anderson's model, item responses at *each* time point are modeled with a separate latent variable. In Figure 1, Andersen's model specifies three latent variables (for this case where there are 3 times). The latent variables are assumed to be correlated, as indicated by the curved arrows. Since each time point is modeled with a separate latent variable, Andersen's model

allows for the difference between time points to be non-uniform, but it does not directly parameterize an overall growth estimate. Note that the fact that the "factor loadings" (i.e., "item scores") are all 1 (and the items are also identical) in this example implies that the variables are actually the same in terms of definition, though, of course, individuals may have different values on that variable at different times. Hence, there will be, in general, different distributions on the variable at different times. In this sense, we might say that this is a "multidistributional" model rather than a multidimensional model.

Embretson's (1991) formulation allows for an initial latent trait factor and then a growth dimension between *each* consecutive set of measurement occasions, as illustrated in Figure 2. In this approach, abilities at later time points are decomposed into one dimension for baseline ability and one or more dimensions for the change between successive pairs of times (depending, of course, on how many times are observed). Item difficulties remain constant across time, and latent variables representing initial status and change may be correlated with each other. Of course, for this model, the successive dimensions are distinct, as the loading (i.e., scores) change from time to time.

In both Andersen and Embretson's models, we assume that the item difficulties remain constant over time. This implies that the longitudinal measurement invariance assumption must be met in order to achieve accurate results. Both models specify different variables based on an underlining latent construct representing the student's proficiency in responding to the test. In the Andersen model, one latent variable is modeled for each time point. These dimensions tend to be highly correlated because substantively what is being measured over time is the same. For instance in an application of the Andersen model to PISA data, the correlation between the time-point specific factors was above 0.86 (von Davier, Xu, and Carstensen, 2011). In the Embretson model, a baseline dimension and change dimensions are proposed for the latent construct. The change dimensions in the Embretson model can also be called time-specific dimensions, where each time point has its own additional effect beyond the baseline and the preceding time points. The change dimensions for each time point explains the additional response variance that cannot be explained by the common ability measure (Embreston, 1991; von Davier, Xu, and Carstensen, 2011). Note that, if the set of items that are used at each time point changes, then the interpretational

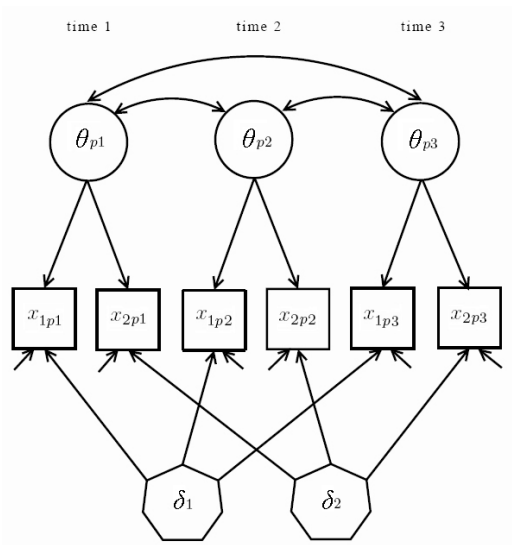

Figure 1. Andersen's IRT-based model for measuring change.

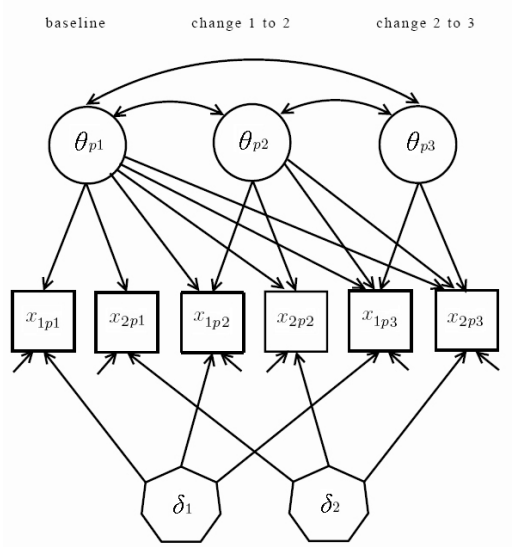

Figure 2. Embretson's IRT-based model for measuring change

context is more complex, with different underlying constructs being measured at different times (for the Andersen model), and different change constructs being measured at different times (for the Embretson model).

The MRCML framework for the LG-IRM

However, with multiple time points, having separately estimated dimensions for each can be too complicated for practical work, hence we seek linear and curvilinear growth models that summarize the pattern of results on a simpler basis. For instance, the growth between each successive pair of time points can be constrained to be equal. In this much more constrained model, both a baseline latent trait and single growth parameter are estimated. This is the formulation of the linear LG-IRM model (Wilson, Zheng, and Walker, 2007). A more complicated formulation allows a curvilinear model, and approximations to other non-polynomial models (Wilson and McGuire, 2010).

The LG-IRM is a multidimensional IRT model constructed within the Multidimensional Random Coefficient Multinomial Logit (MRCML; Adams, Wilson and Wang, 1997) framework. Under the MRCML framework, the probability that person p selects the response in category j to item i is modeled as:

$$P\left(X_{ip}=j\,|\,\theta_p\right)=\frac{\exp\left(\mathbf{b}_i'\theta_p+\mathbf{a}_{ij}'\delta_i\right)}{\sum_{k=1}^{K_i}\exp\left(\mathbf{b}_i'\theta_p+\mathbf{a}_{ik}'\delta_i\right)},$$

where θ_p is a vector of several latent variables being measured on different dimensions, and $\boldsymbol{\delta}_i = (\delta_1,\delta_2,\ldots,\delta_n)'$ is a vector of n item parameters. For dichotomous models such as the Rasch model, δ_i is a vector of I item difficulties, $\boldsymbol{\delta} = (\delta_1,\delta_2,\ldots,\delta_I)$. For polytomous models such as the rating scale model, $\boldsymbol{\delta}_i$ includes an additional set of step parameters, τ_k. For example, in the case of I items each with K categories, $\boldsymbol{\delta}$ may take the form $(\delta_1,\delta_2,\ldots,\delta_I,\tau_1,\tau_2,\ldots,\tau_{K-1})$ for the rating scale model (Andrich, 1978).

The MRCML model includes both "design" and "scoring" matrices. The design vector $\mathbf{a}_{ij}$ of length n defines the linear combination of the n item parameters that corresponds to the observed response $X_{ip}=j$. The collection of all $\mathbf{a}_{ij}$s forms a design matrix $\mathbf{A}$ of the model where $\mathbf{a}_{ij}$ is a row of $\mathbf{A}$. For the Rasch model, $\mathbf{A}$ is matrix of $I \times 2$ rows, $\mathbf{A}=[\mathbf{a}_{11},\mathbf{a}_{12},\ldots,\mathbf{a}_{I1},\mathbf{a}_{I2}]'$. For the rating scale model, $\mathbf{A}$ is matrix of $I \times J$ rows, $\mathbf{A} = [\mathbf{a}_{11},\mathbf{a}_{12},\ldots,\mathbf{a}_{1J},\ldots,\mathbf{a}_{I1},\mathbf{a}_{I2},\ldots,\mathbf{a}_{IJ}]'$. Scores for selecting category j to item i across D dimensions are determined by a column vector of scoring functions, $\mathbf{b}_i=[b_{i1},b_{i2},\ldots,b_{iD}]'$. The collection of all $\mathbf{b}_i$s forms a scoring matrix $\mathbf{B}$ of the model where $\mathbf{b}'_i$ is a row of $\mathbf{B}$, i.e., $\mathbf{B} = (\mathbf{b}'_1,\mathbf{b}'_2,\ldots,\mathbf{b}'_I)$. *ConQuest* software (Wu, Adams, Wilson and Haldane, 2007) has been designed to fit models using any estimable and identifiable pair of item and scoring matrices—in addition, it includes a model-building language that allows the user to specify some models abstractly, and have the program build the required matrices (by far the easier way to use the software).

The linear latent growth model

The proposed linear growth model is a multidimensional IRT model with a particular pattern for the design and scoring matrix following the MRCML framework. Recall that both the Andersen and Embretson models characterize person growth through change dimensions of student proficiency. Similarly in the proposed linear growth model (later this will be generalized to curvilinear models), we specify a linear change dimension in addition to the baseline dimension to explain the additional response variance that cannot be explained by the initial ability measure. Although both person initial status and linear growth dimensions correspond to the same latent construct of student proficiency, we intentionally specify them as two person-specific latent variables as to break the construct into estimable parameters of student proficiency at different points of the time. Specifically, a Time 1 measurement model for dichotomous responses is given as:

$$P\left(x_{ip1}=1\right)=P_{ip1}=\frac{\exp\left(\theta_p-\delta_i\right)}{1+\exp\left(\theta_p-\delta_i\right)}, \quad (2)$$

or equivalently in its logit form logit (P_{ip1}) = $\theta_p - \delta_i$, where P_{ip1} is the probability that person p answers item i (i = 1,2,…,I) correctly at Time 1. The person parameter θ_p is a unidimensional factor that is measured by the I items at Time 1.

For the Time 2 model, following Embretson's formulation, we introduce an additional growth parameter η_p on a different dimension. It represents the change in the person's magnitude of the latent ability from Time 1 to Time 2, logit(P_{ip2}) = $\theta_{p1} + \eta_{p1} - \delta_i$. For the Time 3 model, we initially follow Embretson's approach: logit(P_{ip3}) = $\theta_{p1} + \eta_{p1} + \eta_{p2} - \delta_i$, where η_{p2} is the change in the person's magnitude of the latent ability from Time 2 to Time 3. However, we also add a constraint of $\eta_{p2} = \eta_{p3} = \eta_p$ to obtain a *linear* model of change:

$$\text{logit}\left(P_{ip1}\right) = \theta_p + 0\eta_p - \delta_i,$$
$$\text{logit}\left(P_{ip2}\right) = \theta_p + 1\eta_p - \delta_i, \text{ and}$$
$$\text{logit}\left(P_{ip3}\right) = \theta_p + 2\eta_p - \delta_i,$$

where θ_p and η_p are the person baseline parameter and the person growth parameter each on its own dimension, as shown in Figure 3 (and the coefficient "0" is included in the first equation to make the pattern as explicit as possible) . In the path diagram, the arrows from θ_{p2} to x_{1p2} and x_{2p2} are given for Time 2 data, and the arrows from θ_{p2} to x_{1p3} and x_{2p3} are given for Time 3 data. They are labeled 1 and 2 respectively to represent the coefficient for the corresponding (constant) slope parameter. Note that in the explication in this section we make the assumption that the item set is the same for all the time measurements, but this need not generally be the case (as is illustrated below in one of the examples).

The score matrix is used to map the items onto the dimensions. Following the linear growth formulation, it shows increasing coefficients of the growth dimension in each wave. With three time points, the logit for item responses to item i is determined by the following scoring matrix:

$$\mathbf{B}_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}. \qquad (4)$$

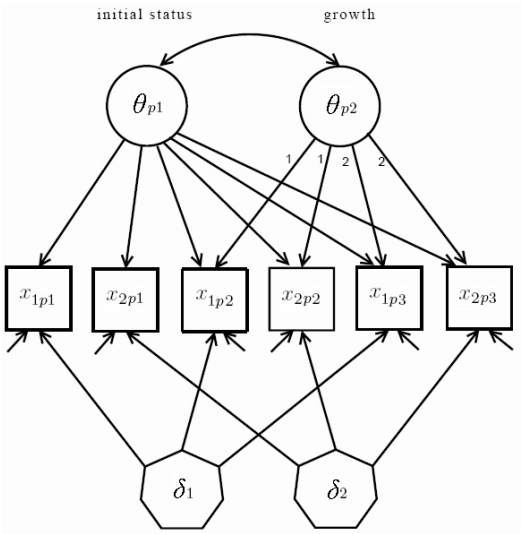

Figure 3. Linear Latent growth IRT model for measuring change.

With two items repeated at three time points, the scoring matrix is simply a collection of $\mathbf{B}_1$ and $\mathbf{B}_2$. The matrix has one row per item from Time 1 to Time 3:

$$\mathbf{B} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \end{bmatrix}. \qquad (5)$$

The design matrix defines linear combinations of the item parameters. In the case where test forms consist only of common items, all items are constrained to have the same difficulty over test administrations through the design matrix. So, the model we propose here also includes a longitudinal measurement invariance assumption. The design matrix **A,** with two dichotomous items repeated at three time points is illustrated below:

$$\mathbf{A} = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \\ 0 \\ -1 \\ 0 \\ 1 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}. \qquad (6)$$

The example design matrix has one row for each item-category combination from Time 1 to Time 3. It shows 12 rows, with each item occupying a total of four rows. The first four rows correspond to the two items given in the first wave of administration. These rows are then repeated for each administration.

While the example design matrix is specific to two items at three time points for simplicity, it is readily generalized for more items and times. Figure 4 shows an expanded example of five items at three time points. By the usual default constraint, the last item is used to constrain the item parameters to sum to zero, for model identification;[1] hence in the design matrix in Figure 4 there are only four columns corresponding to the difficulty parameters for the first four items.

$$A = \begin{array}{c} \begin{matrix} \delta_1 & \delta_2 & \delta_3 & \delta_4 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

Figure 4. Design matrix for five items repeated at three test administrations

Empirical Examples

Cognitive growth in social studies achievement data

An example of using the LG-IRM for measuring student academic achievement and growth is based on data from the National Education Longitudinal study of 1988 (NELS:88). NELS is a major longitudinal study in the U.S. sponsored by the NCES. It follows a nationally representative sample of students from 8th grade in 1988 with four follow-ups. The base-year sample consisted of approximately 25,000 students from 1,052 public and private schools across the U.S. (Curtin et al., 2002). The same cohort was re-surveyed in 1990, 1992, 1994, and 2000. New respondents were freshened into the cohort during the first and second follow-ups to create a grade-representative sample.

NELS:88 administered cognitive tests in four subject areas: reading comprehension, mathematics, science, and social studies to students in grades 8-12 (Curtin et al., 2002). The social stud-

1 Note that the required constraint may be expressed in other ways, but we have found this to be the easiest to implement.

ies assessment tested students on three categories of knowledge including American history, citizenship, and geography. For each test administration, a total of 37 dichotomous items were administered on a single test form. A correct response to an item was scored as a 1 and an incorrect response was scored as a 0. With the dichotomously scored items, raw scores of students could be determined by adding up the number of correct items with higher scores indicating higher levels of achievement in social studies. Although raw scores are commonly used as the outcome variable in traditional growth models, LG-IRM uses IRT to estimate latent abilities and latent growth.

Students of the 8th Grade Cohort who participated in the base-year, first and second follow-up of the social studies assessment were included in this example. The assessment consists of 17 common items across the first three test administrations. These items provided a common score scale for student achievement estimates. For illustrative purposes, we chose to analyze only the common items for this example. Students who skipped one of the three waves of administration, or who did not attempt any of the common items in one administration were dropped from the analysis. This resulted in a final sample of 11,552 students who attempted at least one common item in each of the three waves. For each student, a linear growth parameter was estimated using their baseline achievement and their performance measurements at grade 10 and 12. Using ConQuest[2] (which was also used for the other analyses reported below), weighted likelihood estimation (WLE) estimates of θ_1, the baseline ability parameter, and θ_2, the linear growth parameter were obtained under LG-IRM. A small random sample of 50 linear growth trajectories are shown in Figure 5, based on the WLE estimates of the baseline and growth parameters.

Table 1 provides a summary of item and person parameter estimates. The difficulty measures for the items range from –2.08 to 1.42 (with the mean constrained to zero for identification, as noted above). The order of items from the easiest to most difficult can be determined by the value of the item estimates. Item 12 was the easiest and item 11 was the most difficult. Standard errors for all item estimates were no greater than 0.01. With the mean item difficulty set to zero, the population parameters were also estimated. The mean student baseline estimate was 0.35 logits,

2 Note that the *MPlus* software (Muthén and Muthén, 1998-2007), SAS nlmixed, and the package sem (Fox, 2006) for R could also have been used for these two analyses in this section.

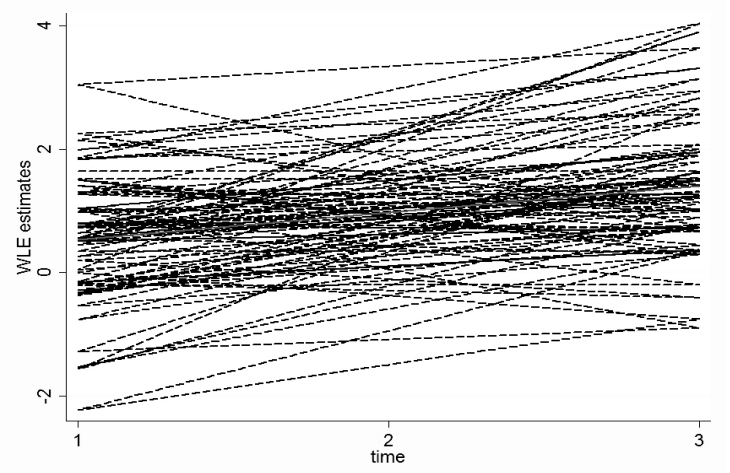

Figure 5. Individual Cognitive Linear Growth Trajectories

Note: This plot shows a random sample of 50 linear growth trajectories plotted using the WLE estimates for the baseline and linear growth parameters from the cognitive growth example.

indicating student initial performance in social studies was slightly higher than test difficulty overall. The standard deviation of the baseline estimate was 0.94. The mean student growth estimate was 0.58 logits with a standard deviation of 0.35. This evidence suggests that students had substantial improvement on social studies learning over each year—that is, approximately 0.37 in standard deviation units (of the baseline estimates) per annum, or 0.74 over the two years. A weak to moderate correlation of 0.33 was observed between the baseline and growth estimates.

Table 1

Parameter Estimates for the Social Studies Assessments

Parameter	Estimate	Error
	Item parameter	
Item 1	–1.168	0.010
Item 2	0.154	0.009
Item 3	–0.106	0.009
Item 4	0.706	0.009
Item 5	–0.285	0.009
Item 6	0.651	0.009
Item 7	0.460	0.009
Item 8	0.746	0.009
Item 9	0.425	0.009
Item 10	–0.566	0.009
Item 11	1.420	0.009
Item 12	–2.082	0.010
Item 13	0.909	0.009
Item 14	–1.359	0.010
Item 15	–0.190	0.009
Item 16	0.108	0.009
Item 17	0.179	0.009
	Person distribution	
	Mean	Variance
Initial location	0.348	0.880
Person slope	0.580	0.123
Correlation	0.325	

Self-esteem example

Another example is the measurement of self-esteem based on data collected over 6 administrations of items[3] from the Rosenberg Self-Esteem scale as part of the Longitudinal Survey of American Youth (LSAY; Miller et al., 1992). The original Rosenberg Self-Esteem scale (Rosenberg, 1965) includes 5 positively worded items and 5 negatively worded items and has been shown to work well for the range of ages of children in high school aged children (Rosenberg 1965; McCarthy and Hoge 1982). It consists of a common set of items used to measure the same construct across a range of ages.

The LSAY data includes 6 of the original Rosenberg items. Respondents are given five response options: "Strongly Agree," "Agree," "Neutral," "Disagree," and "Strongly Disagree." We dichotomized these responses into "Positive" and "Negative," with the "Neutral" category included in "Negative" side for the positively worded items and vice-versa for the negatively worded items.[4] Negatively coded items were reverse coded.

Students from the 7th Grade Cohort who participated in all waves of the Self-Esteem Battery administration were included in this example. This resulted in a sample of 1,572 students. The mean raw score on the self-esteem battery in 7th grade was 4.104 and the mean raw scores for the same students in grades 8-12 were 4.236, 4.231, 4.294, 4.245, and 4.083, showing a slightly decreasing trend. Since the students entered the study at 7th grade, their grade 7 results were set as the baseline. A linear growth parameter was estimated using their baseline self-esteem and their self-esteem measurements in grades 8-12. WLE estimates of θ_1, the initial status Self-Esteem parameter, and θ_2, the linear growth parameter were obtained for each student.

The item parameter estimates for the linear growth example are shown in Table 2. The item aligned with highest self-esteem is "More Self-Respect," an abbreviation for the item asking adolescents whether they wish to have more self-respect has an item parameter estimate of 1.97. Since this was an originally negatively coded item, the location on the logit scale of this item (1.97) is associated with students' disagreement with this statement. The item parameter estimates for the linear model in Table 2 show that students with lower levels of self-esteem would be expected to endorse items "Positive" and "Able."

3 Note that the specific items are shown in footnotes to Table 3.

4 We decided on this strategy based on a factor analysis of the data.

These items have difficulty levels of –0.772 and –0.773. These results are consistent with notions about self-esteem. This example gives more evidence that although the LG-IRM places serious constraints on the item parameter values across years, the results can still show validity with respect to the construct of measure.

Instruments similar to the 6-item self-esteem scale are often used for clinical purposes and are therefore targeted at critically low levels of self-esteem. The LSAY sample however is meant to be nationally representative which means that it would not necessarily include students with below-average self-esteem. Therefore it is not surprising that the average level of self-esteem in this sample is higher than the locations of most of the items. The estimated person parameters are shown in Table 3. The mean baseline self-esteem parameter is 1.203. At this level of self-esteem, students would be expected to respond positively to all the items except for "More Self-Respect." However, it is interesting that the sign of the linear

Table 2

Item Parameter Estimates for the Linear and Quadratic Models

	Linear		Quadratic	
Item Label	Estimate	Error	Estimate	Error
Positive[a]	–0.772	0.021	–0.803	0.021
Worth[b]	–0.236	0.020	–0.246	0.020
Able[c]	–0.733	0.021	–0.763	0.021
Satisfied[d]	–0.030	0.019	–0.031	0.020
More Self–Respect[e]	1.966	0.019	2.044	0.019
Failure[f]	–0.194[g]		–0.202 [g]	

Note: [a] I feel that I am a person of worth, at least on an equal plane with others.
[b] All in all, I am inclined to feel that I am a failure.
[c] I am able to do things as well as most other people.
[d] I take a positive attitude toward myself.
[e] On the whole, I am satisfied with myself.
[f] I wish I could have more self respect.
[g] This value was calculated as a part of the constraint—i.e., to result in an item distribution with mean zero—it was not estimated.

Table 3

Self-Esteem Estimates

	Linear Model		Quadratic Model	
Parameter	Estimate	*se*	Estimate	*se*
$\hat{\mu}_{\theta 1}$	1.203	0.036	1.319	0.041
$\hat{\mu}_{\theta 2}$	–0.009	0.007	–0.110	0.025
$\hat{\mu}_{\theta 3}$	—		0.020	0.005
$\hat{\sigma}_{\theta 1}$	1.416		1.613	
$\hat{\sigma}_{\theta 2}$	0.266		0.979	
$\hat{\sigma}_{\theta 3}$	—		0.190	
$\hat{\sigma}_{\theta 1\theta 2}$	–0.441		–0.507	
$\hat{\sigma}_{\theta 2\theta 3}$	—		–0.958	
$\hat{\sigma}_{\theta 1\theta 3}$	—		0.409	

Note: The number of items on the baseline dimension was $N_{1i} = 36$. The number of items on the growth dimension was $N_{2i} = 30$. For the quadratic growth model, the number of items on the growth acceleration dimension was $N_{3i} = 30$ .The number of students, $N_p = 1{,}572$. ConQuest does not provide standard errors for the elements of the covariance matrix.

growth parameter suggests that the average level of self-esteem decreases over time: It is negative, with a value of –0.009, indicating that the average trend is a slight decrease in self-esteem through the end of high school. The overall correlation between the baseline self-esteem and linear growth parameter is negative, with a value of –0.441. This suggests that when the baseline self-esteem parameter is lower, the positive growth will tend to be larger.

Individual student changes can also be plotted. An example of such a plot is shown in Figure 6 for a small (N = 50) random selection of students using the WLEs for the baseline self-esteem and linear growth parameters from the LSAY example. The plot shows that some students start at lower levels of self-esteem and have steeper slopes upward and vice versa. For practical purposes, by plotting the individual growth trajectories, connections could be drawn between changes in self-esteem and life events or different therapeutic interventions.

Extensions of the Linear LG-IRM

Curvilinear model fitting

The LG-IRM framework presented here also lends itself well to polynomial growth curve fitting. In the linear example, the second dimension represents the rate of change in the latent ability. It can also be thought of as the average rate of change across the time points. It is often the case that the rate of learning will accelerate or decelerate. To quantify these sorts of changes, a growth acceleration term could be added to the model. The growth acceleration term would represent an increase or decrease in the rate of growth. A positive growth acceleration term would signal a speeding up of growth whereas a negative growth acceleration terms would signal a slowing down of growth.

This model is illustrated in Figure 7. In comparison to the illustration of the linear model (Figure 3) a growth *acceleration* dimension, $\theta_{p3,}$ has been added. Note that the value of weights of each path related to θ_{p3} is now squared. Specifi-

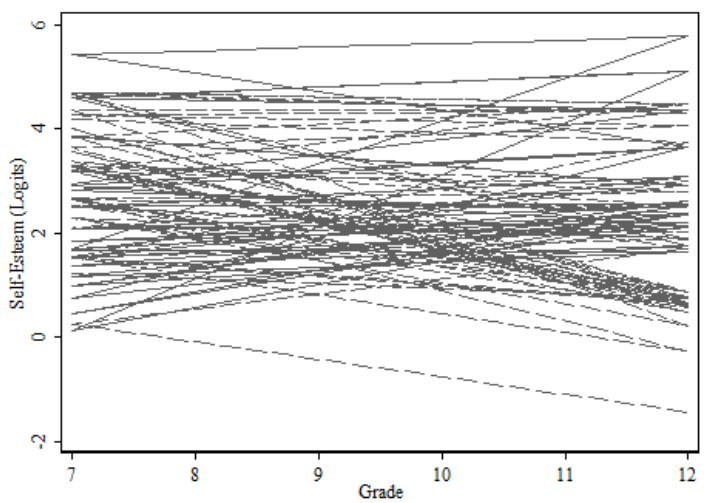

Figure 6. Individual Self-Esteem Linear Growth Trajectories
Note: This plot shows a random sample of 50 linear growth trajectories plotted using the WLEs for the baseline self-esteem and linear growth parameters from the LSAY example. Different line patterns are shown for positive and negative slopes.

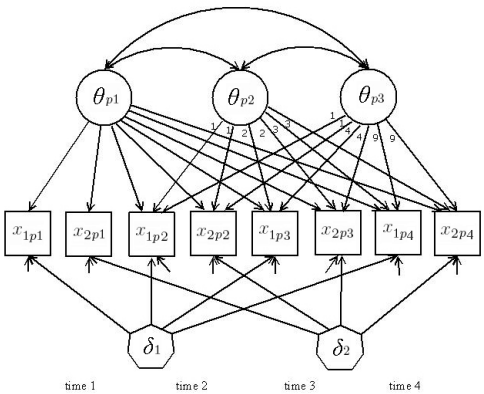

Figure 7. Latent growth IRT model for measuring change with quadratic term

cally, the model shown in Figure 7 illustrates the following set of equations:

$$\begin{aligned}
\text{logit}\left(P_{ip1}\right) &= \theta_{p1} + 0\theta_{p2} + 0^2\theta_p - \delta_i, \\
\text{logit}\left(P_{ip2}\right) &= \theta_{p1} + 1\theta_{p2} + 1^2\theta_{p3} - \delta_i, \\
\text{logit}\left(P_{ip3}\right) &= \theta_{p1} + 2\theta_{p2} + 2^2\theta_{p3} - \delta_i, \\
\text{logit}\left(P_{ip4}\right) &= \theta_{p1} + 3\theta_{p2} + 3^2\theta_{p3} - \delta_i. \qquad (7)
\end{aligned}$$

In practical settings if the data contains only three time points, then it may make sense to summarize the learning using the linear growth parameterization. However when more time points are available and it is suspected that rates of change could fluctuate, then it may be worthwhile to try to quantify these fluctuations. The Self-Esteem data, which consists of 6 time points, is used to illustrate this extension.

The quadratic formulation represented in Equation 6 was used to model changes in self-esteem over the 6 time points. The data used for this example was discussed in the empirical examples section above. The population estimates from the quadratic model are shown in Table 3. The estimate for the mean growth is –0.110. The growth acceleration component has an estimated mean value of 0.020. The estimates from the quadratic growth model suggest a similar relationship to that shown by the linear model, with an estimated correlation between the baseline and growth of –0.507. The correlation between the initial status and quadratic growth is 0.409, suggesting a moderate positive relationship. For this context, an example interpretation could be that for higher levels of initial self-esteem, changes in growth occur less rapidly and for lower levels of initial status, changes in growth occur more rapidly. The correlation between growth and growth acceleration is –0.958, showing a strong negative relationship. In this context, this correlation could mean that for steeper growth, rates of change will tend to occur much less rapidly.

To illustrate some growth trajectory scenarios, a small ($N = 50$) random selection of individual growth trajectories have been plotted in Figure 8. The estimate of the population average for the growth acceleration was nearly zero, therefore many of the students have growth trajectories that are approximately linear. However some students have either large positive or negative growth acceleration term. For those with large positive growth acceleration terms, the growth trajectory tends to curve upwards (concave up) over this time interval whereas for those whose growth acceleration term is negative, the growth trajectory will curve downward (concave down) over this interval.

A comparison of the item estimates from the quadratic model and the linear model shows the same ordering of items in terms of difficulty (Table 2), however the scale is slightly different because of the re-parameterization of the model. In the quadratic formulation, all items after Time 1 are scored on the growth and growth acceleration dimension; the coefficient for the growth acceleration term is squared as shown in Figure 7. In addition, the values of the estimated parameters do not vary so much as to suggest a different interpretation of the dimensions between the two models.

Just as a growth acceleration term was added to the model, cubic and other polynomial terms could be added as well. The higher-order polynomial models can be compared to the lower-order models using familiar tests since the models are considered nested. For the self-esteem example, the final deviance for the linear and quadratic

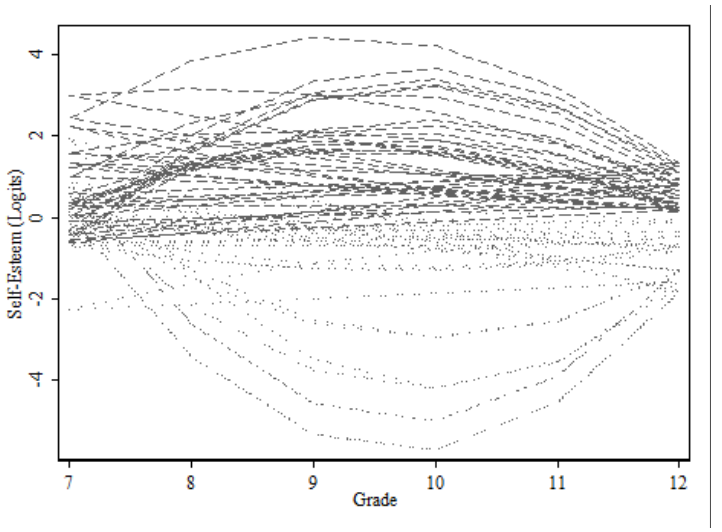

Figure 8. Individual Self-Esteem Quadratic Growth Trajectories
Note: This plot shows a random sample of 50 linear growth trajectories plotted using the WLEs for the baseline self-esteem, linear growth parameters, and quadratic growth parameters from the LSAY example. Different line patterns are shown for trajectories according to the combination of signs for the growth and growth acceleration term.

self-esteem models were used to compare model fit: The likelihood ratio test gave a chi-square statistic that is statistically significant at the $\alpha = 0.01$ significance level ($\chi^2_{df=4} = 476.73$). The results suggest that the quadratic model fits significantly better than the linear model, even though the quadratic model requires the estimation of an additional mean, variance, and two covariances. It may be reasonable to add a cubic growth term in the example of the self-esteem data. However with each additional polynomial term, an additional dimension will also be added to the model. This will introduce costs to the analysis, both in terms of interpretation, and in requirements of sample size.

Differential item functioning

Since the LG-IRM illustrated here is situated within an IRT framework and modeled using IRT software, it is fairly straightforward to incorporate other terms commonly estimated in IRT analyses. A familiar concept is differential item function (DIF). DIF illustrates the degree to which an item would be either more or less difficult to two students of equal ability but differing group characteristics. Often the group characteristics are demographic variables such as gender or race. The results in Table 4 show an example where gender DIF was investigated using the self-esteem instrument that was administered to students in each year. Item parameters, a gender term, and item-by-gender terms were estimated by adding DIF to the LG-IRM using the design matrix. The "Female" term represents a mean difference between genders on the self-esteem scale: The estimate shows that the mean for female respondents was lower than for males, and the standard error shows that this is a statistically significant result at the standard $\alpha = 0.05$ significance level. The "Female*Item" terms represent DIF effects for each item, and several are statistically significant. However, these values are not large in typical interpretations of effect size (Paek, 2002). The only item that may be problematic is "Failure" as suggested by the estimate for the "Female*failure" DIF term.

In the DIF example above, the DIF effect was constrained to be equal over time. Time-varying DIF is another possible extension (e.g., Walker and Wilson, 2008). It could be used in situations

where the item-by-characteristic terms would be expected to change over time. One application of time-varying DIF could be to account for phenomenon like curriculum effects on learning. A curricular indicator variable could be added to distinguish students who have had a certain type of curriculum or curriculum sequence versus those who have not. In the years surrounding the time when the curriculum was taught, it could be reasoned that more DIF might exist for items directly related to, or better covered by that curriculum. Longitudinal studies may also concern whether time-varying covariates are related to DIF. Time-varying DIF indicators could be added to quantify impacts of relevant student attributes that may change over time. Finally, time-wise-DIF may be employed to model longitudinal data in which the item parameters do shift over time, or when the longitudinal measurement invariance assumption may have been violated (Walker and Wilson, 2008). This possibility introduces interesting challenges to interpretation—is the change over time now a change of dimension, as well as a change within a dimension?

Other extensions

Beyond polynomial growth. The curvilinear model fitting and DIF analysis introduced in the previous section offer two examples of model extension. Several additional possibilities are outlined below. All of these possibilities are facilitated through use of design and scoring matrices. For instance, the curvilinear growth example illustrates quadratic growth and suggests how

Table 4

Estimates of Parameters in the DIF Model

Parameter Label	Estimate	*se*
Positive [a]	–0.781	0.021
Able [b]	–0.731	0.021
Satisfied [c]	–0.028	0.019
More Self-Respect [d]	1.968	0.019
Failure [e]	–0.195	0.020
Worth [f]	–0.233 [g]	
Female [h]	–0.036	0.011
Female*positive [i]	–0.200	0.021
Female*able [j]	0.053	0.021
Female*satisfied [k]	–0.033	0.019
Female*more self-respect [l]	–0.061	0.020
Female*failure [m]	0.208	0.019
Female*worth [n]	0.032 [o]	

Note ; [a] I take a positive attitude toward myself.
[b] I am able to do things as well as most other people.
[c] On the whole, I am satisfied with myself.
[d] I wish I could have more self respect.
[e] All in all, I am inclined to feel that I am a failure.
[f] I feel that I am a person of worth, at least on an equal plane with others.
[g] This value was calculated as a part of the constraint—i.e., to result in an item distribution with mean zero—it was not estimated.
[h] mean difference between female and male students.
[i] interaction between gender and "I take a positive attitude toward myself" item.
[j] interaction between gender and "I am able to do things as well as most other people" item.
[k] interaction between gender and "On the whole, I am satisfied with myself" item.
[l] interaction between gender and "I wish I could have more self respect" item.
[m] interaction between gender and "All in all, I am inclined to feel that I am a failure" item.
[n] interaction between gender and "I feel that I am a person of worth, at least on an equal plane with others" item.
[o] This value was calculated as a part of the constraint—i.e., to result in an item DIF distribution with mean zero—it was not estimated.

additional polynomial terms might be added. This sort of extension could also be used to approximate other functions that can be approximated by polynomial expressions, so that, for example, exponential growth over time could be modeled.

Polytomous responses. The LG-IRM can also be generalized to consider polytomously scored items (e.g., see McGuire, 2010). Models for polytmous data such as the rating scale model (RSM; Andrich, 1978), and the partial credit codel (PCM; Masters, 1982), and the graded response model[5] (GRM; Samejima, 1969), can also be specified for longitudinal data (for a full illustration see McGuire, 2010).

Covariates. In the current LG-IRM model, we focus on describing growth in a particular domain using dimensions such as initial status, linear growth, and growth acceleration in that domain. In fact, differences in student development patterns could be associated with other characteristics such as student demographics as well as student and item properties. The characteristics can be either observed or unobserved variables. Latent regression under the MRCML framework enables inclusion of covariates together with other observed person and item covariates as predictors. Also just as time-varying DIF indicators could be added to the model, time-varying characteristics could be added as covariates. Examples that may be relevant in the 7-12 grade population could be variables such as presence of menstrual cycle or puberty, rates of use of alcohol or drugs, changes in family structure, and relocation. Following the model framework, LG-IRM model could also be generalized to introduce time-varying covariates for time-dependent effects within subpopulations. These suggested extensions illustrate how the LG-IRM can be seen as an element of the broader Explanatory Item Response Modeling framework (De Boeck and Wilson, 2004).

Growth in Multiple Domains. A key assumption of the models presented in this paper is that growth is occurring in only one domain. However student progress may often be represented as coordinated growth in multiple domains. In fact many assessments are constructed from a multidimensional theory of developmental growth. These assessments contain items related to one or more domains. The LG-IRM model has the potential advantage of allowing items with multidimensional structure both between and within items (Adams, Wilson and Wang, 1997). This extension would require sufficient data and estimation capabilities because it would require additional dimensions (θ_s) for the added domains. With that practical limitation in mind, it may even be possible to model dimensional shifts over time (as suggested in Walker and Wilson, 2008).

Simulations

Performance assessment of the LG-IRM

This section contains a simulation study to examine the performance of the linear LG-IRM model. Data were simulated for three waves of assessment with 1000 students. The students participated in all three test administrations. They responded to a total of 100 dichotomous items repeated across the three time points. To obtain reliable results, we conducted the simulation with 20 replications.

Student responses across time were determined by three parameters: the item difficulty δ, the person baseline ability θ_1, and the person growth θ_2. They were given values that mimicked the specifications of a real dataset that might be seen in a large-scale longitudinal assessment in an educational or psychological setting. Parameter specifications of the simulation are given in Table 5. To include a wide range of item difficulties, the item parameters were generated from a normal distribution with mean 0 and variance 4. The two person parameters were sampled from a bivariate normal distribution. The mean of the growth parameter was chosen to be larger than the mean of the baseline parameter so that the simulated students will make generally progress (i.e., get better scores) throughout the assessment. In addition, the variance of the growth parameter was chosen to be smaller than that of the baseline parameter so that the simulated students will, in

5 Note that the GRM cannot be estimated by the ConQuest software—MPlus (Muthén and Muthén, 1998-2007) will do so, however.

general, have similar growth patterns. In a large simulation study, we explored alternative parameter specifications under various simulation conditions (Zheng and Wilson, 2009; Zheng, 2009).

Model parameters were estimated using ConQuest (Wu et al., 2008). Across the simulation replications, we fixed the item parameters over all waves of administration (i.e., the design matrix was used to constrain each item to have constant difficulty) and allowed the person parameters to vary based on the bivariate normal distribution. The within-item dimensional structure was specified using the scoring matrix.

Table 5

Parameter Specifications for the Simulation

Parameter	Generating Value
N_i	100
N_p	1000
μ_δ	0.00
$\mu_{\theta 1}$	0.00
$\mu_{\theta 2}$	0.50
σ_δ^2	4.00
$\sigma_{\theta 1}^2$	4.00
$\sigma_{\theta 2}^2$	1.00
$\sigma_{\theta 1 \theta 2}$	0.00

Three indices were computed to assess simulation results in terms of parameter recovery. The first form is the Pearson correlation. The Pearson correlation gives an indication of the overall association between the estimated and generating parameters. The Mean square error (MSE) is the average squared deviation between the estimated and true values of each estimated parameter, averaged over the number of parameters estimated. Since the MSE is calculated as units-squared, the square root or the MSE or the second index, the RMSE is used for interpretation. The RMSE provides a non-directional difference in the same metric as the estimates and can help to quantify overall deviation in terms of the original metric:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{\delta}_i - \delta_i\right)^2}{N}}. \qquad (8)$$

The RMSE does not indicate the direction of differences, or bias. For this reason, the third index, the Average Signed Bias (ASB) was also included. *ASB* indicates general overestimation when positive or underestimation when negative. Like the RMSE, the metric of ASB is the same as the metric of the parameter estimate:

$$ASB = \frac{\sum_{i=1}^{N}\left(\hat{\beta}_i - \beta_i\right)}{N}. \qquad (9)$$

A summary of the simulation results is given in Tables 6 and 7. For the item difficulty parameters, the Pearson correlation value (across 20 replications) is always greater than 0.999. The RMSE values average approximately 0.06 logits, relatively quite small on the scale given that the item difficulties range from around –4 to 4. The average ASB is very close to 0. For person baseline and growth parameters, the RMSEs are 0.057 and 0.043 logits for the means of the parameters, and 0.163 and 0.051 logits for the variance of the parameters. The ASB values show only small bias in the parameter recovery with values of 0.002 and –0.011 logits for the means, and –0.079 and –0.002 logits for the variances. Although the parameter recovery indices are somewhat larger for the person parameters, they are still reasonable for application of the model in typical educational settings. These simulation results provide evidence that for the LG-IRM the person parameters can be estimated with reasonable accuracy for interpretation and descriptive analyses.

LG-IRM in comparison to HLM

We are also interested in comparing parameter estimates from LG-IRM and HLM. A growth analysis of a simulated data was run in both LG-IRM and HLM. The data were simulated for 10,000 persons, each responded to 20 dichotomous items. Scores of the persons were determined by three parameters: the item difficulty, the person initial ability, and the person growth rate. These parameters were given values that mimicked the specifications of the NELS:88 data described above. Specifically, the item difficulty was generated from a normal distribution

Table 6

Item Parameter Recovery Indices

Replication	Correlation	ASB	MSE	RMSE
1	1.000	0.000	0.003	0.059
2	0.999	0.000	0.004	0.062
3	1.000	0.000	0.003	0.050
4	1.000	0.000	0.004	0.060
5	1.000	0.000	0.003	0.056
6	1.000	0.000	0.003	0.050
7	0.999	0.000	0.004	0.063
8	1.000	0.000	0.003	0.057
9	1.000	0.000	0.003	0.055
10	1.000	0.000	0.003	0.058
11	0.999	0.000	0.004	0.062
12	1.000	0.000	0.003	0.051
13	0.999	0.000	0.004	0.061
14	1.000	0.000	0.003	0.058
15	1.000	0.000	0.003	0.056
16	1.000	0.000	0.003	0.058
17	1.000	0.000	0.003	0.052
18	1.000	0.000	0.003	0.054
19	0.999	0.000	0.004	0.062
20	1.000	0.000	0.003	0.056
Average	**1.000**	**0.000**	**0.003**	**0.057**

with mean 0 and variance 1. The two person parameters were sampled from a bivariate normal distribution with means of 0.0 and 0.5, variances of 1.0 and 0.1, and a correlation of 0.3.

A two-step approach was used for the HLM analysis. The first step involved IRT calibration, which put the parameter estimates on the logit scale based on concurrent estimation of the Rasch model. Then, WLE estimates were used as the response variable for the traditional HLM modeling to estimate the means of student initial achievement and change in student achievement, as well as the variances and correlation of the two person parameters. We found that the first step was needed, for otherwise the HLM results were not close to the generating values.

Table 8 lists summary statistics of person estimates from the HLM and the one-step LG-IRM analyses. It shows that both the HLM results and the LG-IRM results successfully retrieve the

Table 7

Person Parameter Recovery Indices

Replication	$\mu_{\theta 1}$	$\sigma_{\theta 1}^2$	Correlation$_{\theta 1}$	$\mu_{\theta 2}$	$\sigma_{\theta 2}^2$	Correlation$_{\theta 2}$
1	0.071	3.856	0.990	0.438	0.909	0.956
2	0.066	4.051	0.983	0.476	0.974	0.957
3	0.072	3.737	0.990	0.499	0.994	0.968
4	–0.032	3.984	0.984	0.533	1.035	0.968
5	–0.003	4.022	0.983	0.493	1.007	0.966
6	0.051	3.793	0.977	0.432	0.895	0.961
7	0.050	3.872	0.989	0.475	0.994	0.963
8	0.013	4.062	0.984	0.412	0.975	0.958
9	–0.118	4.020	0.985	0.442	1.072	0.964
10	0.056	4.024	0.987	0.506	1.045	0.965
11	–0.104	3.848	0.987	0.509	0.985	0.965
12	–0.021	3.922	0.983	0.484	0.979	0.957
13	0.067	4.049	0.975	0.551	1.081	0.965
14	–0.045	3.761	0.981	0.530	1.035	0.964
15	–0.024	3.597	0.989	0.451	0.945	0.970
16	0.016	4.092	0.987	0.437	1.086	0.963
17	0.045	4.051	0.977	0.534	0.983	0.961
18	–0.053	4.098	0.979	0.508	1.018	0.959
19	–0.061	3.875	0.990	0.518	0.999	0.964
20	–0.014	3.707	0.987	0.557	0.942	0.962
ASB	**0.002**	**–0.079**	**–0.016**	**–0.011**	**–0.002**	**–0.037**
MSE	**0.003**	**0.027**	**0.000**	**0.002**	**0.003**	**0.001**
RMSE	**0.057**	**0.163**	**0.016**	**0.043**	**0.051**	**0.037**

generating values. However, LG-IRM analysis provides slightly better estimates as the means and variances of the person estimates are closer to their generating values, except for the variance of the person slopes.

Table 8

Summary statistics of person estimates under HLM and LG-IRM

Statistic	LG-IRM Approach	HLM Approach	Generating Value*
Initial location			
Mean	0.005	0.009	–0.004
Variance	1.013	1.007	1.012
Person slope			
Mean	0.500	0.492	0.498
Variance	0.096	0.092	0.090
Correlation			
Location to slope	0.298	0.280	0.314

Note: *These generating values were the actual values achieved for the generated sample, not those theoretically planned.

An example with vertical scaling

The inspiration for this simulation came from the desire to create better growth models for educational assessment in the era of growth modeling within NCLB. In this section, the model is tested in a situation more similar to a state-wide assessment system, including grade-specific standards and vertical scales.

A vertical scale combines links between forms from year to year. These links together create a scale over a span of years. In a classic educational example, the forms would be increasing in difficulty as we expect the students to increase in ability. This arrangement was followed for the example simulation. Vertical scaling can be achieved through linking based on common items across years, common persons across years, or both. This simulation was designed to match longitudinal data which includes at least some common persons across years. For simplicity, in this simulation we assume that the same set of persons participated in all six waves of assessment. We also used a calibration sample to produce a common scale.

A set of 120 item difficulties was generated from a normal distribution to serve as the bank from which forms of varying difficulty could be assembled. The item parameters were generated from a normal distribution with mean 0 and variance 4. Items were sampled from the bank of 120 items to produce 6 linked forms (Table 9).

ConQuest was used for estimation. Due to local minima in the convergence criterion, the model was rerun three times, starting from the ending parameters from previous runs. Upon reaching convergence the item difficulty estimates and weighted likelihood estimates (WLE) person parameters were obtained. The parameter recovery indices are shown in Table 10. As before, the Pearson correlation is higher for the items than for the persons. In addition, the items show a small but negative ASB value (–0.0002 logits) suggesting slight underestimation. The RMSE for the items shows that overall the item parameter

Table 9

Characteristics of Vertically Scaled Test Forms in Simulation Example

Form	Average Difficulty	Number of items
Wave 0	0.0451	26
Wave 1	0.1196	27
Wave 2	0.1269	27
Wave 3	0.1447	27
Wave 4	0.2208	27
Wave 5	0.2275	26

Note: A bank of 120 items was used to assemble each form. Eight common items were used across all forms.

Table 10

Parameter Recovery Estimates

Type	Pearson Correlation	ASB	MSE	RMSE
Items	0.9990	–0.0002	0.0033	0.0456
(Persons calibration)	0.9895	0.0001	0.0044	0.0520
Persons	0.9542	0.0326	0.0131	0.0915

estimates are recovered within 0.0456 logits. The ASB for the person estimates is negligible. In addition, the RMSE suggests that person parameter abilities are recovered within 0.0915 logits. The results show that reliable estimates can be obtained from the LG-IRM when the data comes from vertically scaled designs.

Discussion

This paper presents a latent growth model for longitudinal data. Based on the initial approaches of Andersen's and Embretson's model for measuring change, it builds a growth model onto the item response framework resulting in a Hierarchical Generalized Linear Model (HGLM). Specifically, a linear growth model is incorporated into the item response function from a multidimensional perspective. Latent traits over time are specified to have a multivariate normal distribution and conceptualized as several dimensions in a multidimensional IRT model, depending on the curvilinear nature of the model.

Many longitudinal assessment studies such as NELS typically follow a multi-step procedure (Ingels et al., 1994) to measure gains in student achievement over time. The first step involves IRT calibration of assessment data. In this step, item parameters and student ability scores are estimated for each wave of data using IRT scaling models. In the second step, base-year item parameters are transformed into the first follow-up scale based on characteristic curve transformation methods. Moreover, base-year ability scores are transformed into the first follow-up scale based on the transformed item parameters. Then, first follow-up gain scores are reported as the difference between the first follow-up ability estimates and rescaled base-year ability estimates. The procedure repeats in the third step to rescale the base-year and the first follow-up ability scores in the second follow-up. Gains in student achievement are then reported based on the new rescaled scores. The multi-step procedure consists of IRT scaling and longitudinal equating based on a scale transformation. The method assumes that a single scale transformation can be applied for all items. However, the assumption could be questionable if, for example, items are more related to the timing of the curriculum rather than the logical structure of the knowledge.

In contrast, the LG-IRM approach uses a single-step procedure for the analysis of longitudinal assessment data. It estimates a single model that combines measurement models with longitudinal equating. Instead of estimating ability scores separately for each time on different metrics, student scores over years are estimated all at once on the same vertical scale.

The model can also be applied to vertically scaled assessment data where a common metric is constructed across years. The vertical scale allows for certain standards to be measured in each grade-level while and over-arching area of achievement is measured across years. The overarching metric would be dominated by the standards that are common to all grades.

The LG-IRM sets a starting point for various IRT-based longitudinal models. The MRCML framework has the flexibility to include many extensions of standard IRT models. The LG-IRM inherits this flexibility: the proposed model can be extended to more complex growth models to cope with more complicated research designs. The model has the capacity to include observed person and item covariates as well as item-by-characteristic interactions. It is appropriate for polytomously scored items by expanding the design matrix to accommodate step parameters. Multidimensional models and various forms of non-polynomial models are also possible by adding extra dimensions in the structural part of the growth model.

Acknowledgements

We would like to thank the following colleagues for their help in completing this work: Sophia Rabe-Hesketh, Karen Draney, and Derek Briggs.

References

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

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

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Education Measurement: Issues and Practice*, *28*, 42-51.

Briggs, D. C., and Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Education Measurement: Issues and Practice, 28,* 3-14.

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9-25.

Curtin, T. R., Ingels, S. J., Wu, S., and Heuer, R. (2002). *National education longitudinal study of 1988: Base-year to fourth follow-up data .le user's manual* (Tech. Rep. No. NCES 2002-323). Washington, DC: Department of Education, National Center for Education Statistics.

De Boeck, P., and Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Duncan, T. E., Duncan, S. C., and Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

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

Fox, J. P. (2006). Structural equation modeling with the sem package in R. *Structural Equation Modeling*, *13*, 465-486.

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

Fox, J. P., and Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, *68*, 169-191.

Ingels, S. J., Scott, L. A., Rock, D., Pollack, J., and Rasinski, K. (1994). *NELS:88 first follow-up final technical report* (Tech. Rep.). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Jöreskog, K. G. (1979). Statistical models and methods for analysis of longitudinal data. In K. G. Jöreskog and D. Sörbom (Ed.), *Advances in Factor Analysis and Structural Equation Models* (pp. 129-69). Cambridge, MA: Abt.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79-93.

MacCallum, R. C., and Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual review of psychology*, *51*(1), 201-226.

McArdle J. J., and Aber M. S. (1990). Patterns of change within latent variable structural equation modeling. In A. von Eye (Ed), *New statistical methods in developmental research* (pp. 151-224). New York: Academic.

McGuire, L. W. (2010). *Practical formulations of the latent growth item response model*. Retrieved from Dissertations and Theses database. (AAT 3413550)

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

Miller, J., Hoffer, T., Suchner, R., Brown, K., and Nelson, C. (1992). *LSAY codebook: Student, parent, and teacher data for cohort two for longitudinal years one through four (1987-1991)* (Vol. 2). DeKalb, IL: Northern Illinois University.

McCarthy, J. D., and Hoge, D. R. (1982). Analysis of age effects in longitudinal studies of adolescent self-esteem. *Developmental Psychology*, *18*, 372-379.

Muthén, L. K., and Muthén, B. O. (1998-2007). *Mplus user's guide* (5th ed.). Los Angeles: Muthén and Muthén.

Nelder, J., and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, *135*, 370-384.

No Child Left Behind Act of 2001, Pub L. No. 107-110. 115 Stat. 1425 (2002).

Paek, I. (2002). Investigation of differential item functioning: Comparisons among approaches, and extension to a multidimensional context. Unpublished doctoral dissertation, University of California, Berkeley, Berkeley, CA.

Pastor, D. A., and Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement*, *30*, 100-120.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167-190.

Raudenbush, S. W. (1995). *Posterior modal estimation for hierarchical generalized linear models with application to dichotomous and count data*. Unpublished manuscript, Michigan State University, College of Education.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inference from longitudinal data. *Annual Review of Psychology, 52,* 501-525.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical linear models*. Thousand Oaks, CA: Sage.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Samejima, F. (1969). *Estimation of latent trait ability using a response pattern of graded scores*. Bowling Green, OH: Psychometric Monograph 17, Psychometric Society.

Singer, J. D., and Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford Press.

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman and Hall/CRC.

Thorn, C. A., and Meyer, R. H. (2006). *Longitudinal data systems to support data-informed decision making: A tri-state partnership between Michigan, Minnesota, and Wisconsin* (WCER Working Paper No. 2006-1). Madison, WI: University of Wisconsin–Madison, Wisconsin Center for Education Research.

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

Walker, L., and Wilson, M. (June, 2008). *Exploring cross-cultural and time-wise item shift: An extension of the Latent Growth Item Response Model*. Invited paper presented at the 6th annual international conference on Survey Methods in Multinational, Multiregional, and Multicultural Contexts (3MC), Berlin, Germany.

Willett, J. B., and Keiley, M. K. (2000). Using covariance structure analysis to model change over time. In H. Tinsley and S. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 665-694). San Diego, CA: Academic Press

Wilson, M., Zheng, X., and Walker, L. (2007, August). *Latent growth item response models.* Paper presented at the IPN Conference on Longitudinal Data Analysis in Educational Studies, Kiel, Germany.

Wilson, M., and McGuire, L. (2010, August). *The latent growth item response model*. Paper presented at the International Conference on Outcome Measurement (ICOM), Bethesda, MD.

Wu, M., Adams, R., Wilson, M., and Haldane, S. (2007). ACER ConQuest version 2.0: Generalised item response modelling software [Computer software and manual]. Camberwell, Victoria, Australia: ACER Press.

Zheng, X. (2009). *Multilevel item response modeling: Applications to large-scale assessment of academic achievement*. Retrieved from Dissertations and Theses database. (AAT 3411238)

Zheng, X., and Wilson, M. (2009, April). *An IRT-based linear growth model for longitudinal data*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.