ABSTRACT

Title of Document: An Integrated Item Response Model for Evaluating Individual Students' Growth in Educational Achievement

Jennifer Koran, Ph.D., 2009

Directed By: Professor Gregory R. Hancock, Department of Measurement, Statistics and Evaluation

Measuring continuous change or growth in individual students' academic abilities over time currently uses several statistical models or transformations to move from data representing a student's correct or incorrect responses on individual test items to inferences about the form and quantity of changes in the student's underlying ability. This study proposed and investigated a single integrated model of underlying growth within an Item Response Theory framework as a potential alternative to this approach. A Monte Carlo investigation explored parameter recovery for marginal maximum likelihood estimates via the Expectation-Maximization algorithm under variations of several conditions, including the form of the underlying growth trajectory, the amount of inter-individual variation in the rate(s) of growth, the sample size, the number of items at each time point, and the selection of items administered across time points. A real data illustration with mathematics assessment data from the Early Childhood Longitudinal Study showed the practical use of this integrated model

for measuring gains in academic achievement. Overall, this exploration of an integrated model approach contributed to a better understanding of the appropriate use of growth models to draw valid inferences about students' academic growth over time.

AN INTEGRATED ITEM RESPONSE MODEL FOR EVALUATING INDIVIDUAL STUDENTS' GROWTH IN EDUCATIONAL ACHIEVEMENT

By

Jennifer Koran

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Gregory R. Hancock, Chair
Professor Patricia A. Alexander
Professor Jeffrey R. Harring
Professor Robert W.Lissitz
Professor Robert J. Mislevy

Acknowledgements

This work was supported by a 2008-2009 Harold Gulliksen Psychometric Research Fellowship Award from the Educational Testing Service, Princeton, NJ.

Early Childhood Longitudinal Study – Kindergarten Cohort item response data were provided by the National Center for Education Statistics with permission from Pro Ed, Inc., Simon & Schuster, David LaRochelle, National Geographic World Magazine, CTB McGraw Hill, Riverside Publishing, Pearson, and HarperCollins Publishers. The ideas herein are those of the author and do not reflect official positions or policies of any of the above-named organizations.

Credit is given to Riverside Publishing for permission to use the Early Childhood Longitudinal Study – Kindergarten Cohort item response data as stipulated in the following required form:

Credit is given to HarperCollins Publishers for permission to use the Early Childhood Longitudinal Study – Kindergarten Cohort item response data as stipulated in the following required form:

"TEXT COPYRIGHT © 1997 BY SEYMOUR SIMON. Used by permission of HarperCollins Publishers."

Table of Contents

List of Tables

List of Figures

Chapter 1: Background and Justification

As the fundamental goal of the educational enterprise, learning is often defined in terms of a gain in knowledge, understanding, or skill. Thus, the study of change in the quantity and quality of individual students' knowledge and abilities over time is at the heart of research in education. The measured change along a continuum of progress in a domain is often called growth.

In recent years there has been an increasing opportunity and great need to appropriately model individual growth over time in educational research, not only in nationally representative longitudinal studies but also increasingly to track and promote educational program improvements in states and large school districts. Most of these applications are focused on change as measured by individual assessments of academic achievement, primarily measures of cognitive knowledge, skills, or abilities. Further, educational accountability legislation has been enacted in the last ten years that demands the tracking of organizational growth over time to judge the effectiveness of education systems. Thus, the growth of individuals can have important implications for the status or growth of an organization as well.

However, historically the research methodology for studying growth and change has been plagued with paradoxes and inadequacies (for an overview see Harris, 1963; Rogosa, 1995), and methodological challenges in modeling growth have continued to the present day. The need for increased rigor in assessing individual growth over time in education exists independent of legislative initiatives, but it has been viewed with greater urgency in recent years due to greater demand for growth modeling to meet the requirements of educational accountability policy.

Methodological Background

Many methods have been proposed to model growth and change over time at the levels of populations and individuals. The growth modeling method and the alternatives considered in this dissertation are for repeated measurements made on the same individuals over time, some times called a panel design to distinguish it from other longitudinal designs. Repeated surveys where different samples of respondents from the same population are surveyed at regular intervals and methods for analyzing this type of data are outside the scope of this work. Rather, all methods described here assume that multiple measurements are made on individuals over time.

The major characteristics of these methods reflect the underlying theory about the nature of the construct that is changing. In some methods the focus is on a latent construct that is hypothesized to be categorical. This leads to approaches, such as latent transition analysis, in which the focus is on the probability of an individual transitioning from one latent category to another. Other methods focus growth in a latent construct that is hypothesized to be continuous. It is these latter methods and the consideration of how they handle both continuous and categorical observed indicators that will be covered in detail in this section on methodological background.

Most reviews and didactic approaches to modeling continuous growth begin with the simple case involving two time points. When two time points are considered in isolation, a gain score (the difference in scores between the two measurements) may be computed when comparable assessments are used to produce scores at an interval level of measurement at each individual time point. Despite its conceptual and computational simplicity, the use of gain scores has traditionally been noted as

highly problematic in the psychometric literature (Harris, 1963; Rogosa, 1995). First, many authors have noted that under certain circumstances the higher the correlation between the measures at the two time points, the lower the reliability vs. *accuracy* of the gain scores (Bereiter, 1963). Measures at two time points may often be highly correlated if the underlying process is slow, if the two measurements are taken close together, or if individuals tend to grow at the same rate. Second, the true amount of underlying change does not necessarily have a linear relationship with gain scores. Gains in the extremes of the score distribution may represent a greater amount of true change than gains near the middle of the score scale, particularly when number correct scores are used to compute the gain scores. Bereiter (1963) showed that a change in the observed score could mean different things depending on the initial score level. Lord (1963) likewise showed that it is necessary to look at predictors of gain scores after partialling out the effect of the initial measurement. Finally, gain scores have been shown to have a negative correlation with initial scores, even when there was no underlying relation. Lord (1963) showed that this was due to the regression effect. Psychometric problems with measuring change were so great that Cronbach and Furby (1970) even recommended abandoning change measurement altogether. Other researchers subsequently revised this viewpoint, concluding instead that "two time points provide an inadequate basis for studying change" (Bryk & Raudenbush, 1987, p.147) and that "investigations of growth that rely upon a two-wave design are particularly weak approaches to the study of individual change" (Willett, 1988, p. 371).

From an early date, however, approaches have also been pursued for handling interval-level measurements at three or more time points simultaneously. Many of the earliest methods were focused on modeling changes over time in groups rather than in individuals (Bollen & Curran, 2006). The focus was on the mathematical form of the trajectory that best summarized growth for the entire group (Bollen & Curran, 2006). Over time methods were developed for studying interindividual differences in individual trajectories over time, and the focus of these methods shifted to the study of deviations from the group trajectory (Bollen & Curran, 2006). The 1930's gave rise to some of the first growth modeling work fitting unique trajectories to individuals and then using a separate analysis to summarize these individual trajectories for a group of individuals using analysis of variance-type approaches (e.g., Wishart, 1938).

Developments for modeling individual growth trajectories progressively moved toward more integrated models within latent variable frameworks. Bollen and Curran (2006) in their review of the historical developments of growth modeling using structural equation models noted that developments by Baker (1954) and Tucker (1958) first used exploratory factor analysis to model trajectories of growth. It was the seminal work by Meredith and Tisak (1990), however, that proposed growth modeling using confirmatory factor models and gave rise to current methods for Latent Growth Curve Modeling using structural equation models.

Current Approaches to Modeling Growth in Continuous Constructs over Time

Numerous overviews of the Latent Growth Curve Modeling approach for modeling growth have been published in recent years (Bollen & Curran, 2006;

Duncan, Duncan, & Strycker, 2006; Hancock & Lawrence, 2006). In this approach latent (unobserved) variables are defined to postulate the form of a latent trajectory for a given individual's observed measurements over time. A commonly chosen latent trajectory is a linear model for growth in which a latent variable α (intercept factor) represents individuals' true status at the initial measured time point and a latent variable β (slope factor) represents individuals' true rate of change per unit time. Loadings for the measured variables at each time point on the slope factor define the amount of change (amount of β) that is added to an individual's initial status to arrive at that individual's subsequent level. In the linear model these loadings are fixed to the amount of time that has elapsed since the initial (or reference) point. Additional forms for the trajectory can be accommodated either by changing the values of the loadings on β or by incorporating additional latent growth factors into the model to describe the rate of change. Loadings for the latent intercept α are typically fixed to unity, and the latent growth factors are typically allowed to correlate. Individuals typically must be measured at defined time points (certain patterns of missing data can be accommodated, however) since the fixed values of the loadings do not vary across individuals. As in other structural equation models, the structural relations among latent and observed variables are used to define model-implied mean and covariance matrices among the observed measures at different time points.

The means and covariances of the observed measurements at each time point are calculated and parameters are estimated using an optimization routine that iteratively works toward finding values of the parameter estimates that cause the model-implied correlation matrix to most closely mimic the values in the observed

correlation matrix. The estimated mean values of the latent growth factors provide insight into the mean trajectory for the group, whereas the estimated variances of the latent growth factors reflect the interindividual variability in both initial status and rate of change. In latent growth curve modeling it is also possible to accommodate multiple measures (indicators) of a latent construct at each time point. A curve-of-factors model, or second-order latent growth curve model (Hancock, Kuo, & Lawrence, 2001; Sayer & Cumsille, 2001), can be postulated to explain growth in several indicators of a latent construct over time. In this model a factor is formed for each time point to represent the latent construct measured without error. Growth in the factors at each time point is then modeled using second-order growth factors to define the trajectory.

Concurrent developments have also occurred for modeling growth using Linear Mixed Effects Models. These models and their variants are also called by a number of other names, including: multilevel linear models, random coefficients models, random effects models, and hierarchical linear models (Fitzmaurice, Laird, & Ware, 2004; Raudenbush & Bryk, 2002). In this approach a regression-style model is postulated to define each observed outcome as a function of a growth trajectory for each individual (Level 1 model). The terms in the model (e.g., linear or quadratic) are functions of the time that has elapsed since the initial or baseline measurement. The terms entered into the model to define the form of the trajectory are the same over individuals. The coefficients for these terms, however, are random variables that may take on different values for different individuals. In the most basic construction the coefficients are then further defined in terms of a mean value for the population

and a deviation from that mean for the individual (Level 2 model). A researcher-defined covariance structure is imposed on the residuals in the Level 1 model and on the deviations in the Level 2 model. This approach has the advantage of allowing for individuals to be measured at very different time points throughout the study, but in practice often imposes a quite restrictive covariance structure on the observations. Multiple measurements at each time point can be incorporated by using a design matrix to define which terms in the Level 1 model are to be applied in explaining a particular observation. It is important to note that although nonlinear trajectories (e.g., quadratic) are accommodated in the Linear Mixed Effects Modeling framework, the model is still called "linear" because it is intrinsically linear in its terms.

Developments for modeling individual growth trajectories using latent variable methods, such as Latent Growth Curve Modeling and Linear Mixed Effects Modeling, have overcome some historical methodological difficulties but have also raised new questions and avenues for studying change over time (Collins & Horn, 1991; Collins & Sayer, 2001; Gottman, 1995). Willett (1988) argued, "Many of the problems that appear to beset the measurement of change are, in fact, artifacts of an inappropriate perspective" (p. 353). Indeed, Willet demonstrated that the difference score can be highly reliable when interindividual variability in underlying growth is large. Both Latent Growth Curve Modeling and Linear Mixed Effects Modeling provide a perspective that allows the underlying growth to be differentiated from measurement error and the interindividual variability in this growth to be explicitly modeled.

Among other advantages, growth models in both the Latent Growth Curve Model and Linear Mixed Effects Model frameworks provide an integrated approach for studying multiple aspects of change, including its structure, group trend, interindividual differences, and predictors of individual change (Bollen & Curran, 2006; Raudenbush & Bryk, 2002). Further, a number of researchers are currently investigating links to bridge the differences between the two so that the advantages of both frameworks can be leveraged for modeling growth.

Growth Models for Categorical Data

Both the Latent Growth Curve Model and Linear Mixed Effects Model frameworks for modeling growth or change over time were originally developed to model data at an interval level of measurement. More recently both methods have been extended to better accommodate ordinal level measurements.

In Latent Growth Curve Modeling, the growth model is typically fit to the means and covariances or correlations of observed interval variables. To extend this framework to ordinal measurements, it is necessary to use methods for obtaining means and correlations from ordinal data (e.g., Olsson, 1979). This is accomplished by assuming that each manifest ordinal indicator corresponds to a latent continuous response variable. In this latent response variable formulation, each underlying continuous variable is considered to have a univariate normal distribution. A monotonic transformation matches the density of the observed categorical distribution to the density of the continuous distribution:

$$y = c \text{ if } \tau_{c-1} < y^* \leq \tau_c \ .$$

That is, an ordinal variable *y* is equal to some category *c* if the underlying continuous variable *y** is greater than a lower threshold τ_{c-1} and less than or equal to the next threshold τ_c.

The underlying variable *y** is assumed to have a range from negative infinity to positive infinity, but its scale is arbitrary. Several options are available for fixing the metric(s) of the underlying response variables (Jöreskog, 1990; B. O. Muthén, 1984, 1996; B. O. Muthén & Asparouhov, 2002). In growth modeling applications with variables having three or more categories, this is most often accomplished by setting the first threshold equal to zero and the second threshold equal to one. By fixing the values of two thresholds equal to two different constants, both the location and the scale of the metric are defined. When binary variables are used, there is only a single threshold, and fixing a single threshold will not fully define the underlying scale. In growth modeling applications with binary data, the scale can be defined instead by fixing the variance of the error terms to a constant (theta method) or using a scaling factor on the variance of the error terms (delta method, B. O. Muthén, 1984, 1996; B. O. Muthén & Asparouhov, 2002).

Once the scale has been defined, the values of any remaining free thresholds must be estimated, as well as the mean and variance of the underlying variable(s). First, standardized values of the thresholds are typically estimated from the data by considering a threshold τ_c to be the point on the underlying standard normal distribution (mean = 0 and variance = 1) that has the same cumulative probability for a response of *c* or below. The value of the standardized threshold estimate, denoted $\hat{z}_c$, is found by using the inverse cumulative standard normal distribution:

$$\hat{z}_c = \Phi^{-1}(p_1 + p_2 + ... + p_C),$$

where p_c is the observed proportion of responses in category c. Then the polychoric correlation between any two underlying variables y_1* and y_2* is estimated by maximizing the log-likelihood of the multinomial distribution for the standardized threshold estimates from the first step:

$$\sum_{i=1}^{C_1}\sum_{j=1}^{C_2} n_{ij} \ln \int_{z_i}^{z_{i+1}} \int_{z_j}^{z_{j+1}} \Phi_2(y_1^*, y_2^*) dy_1^* \, dy_2^*,$$

where C_1 and C_2 are the number of categories in the two respective items and dy* is the variable of integration representing the difference in the integration over both latent continua. The function Φ_2 is the bivariate normal distribution:

$$\Phi_2(\mathbf{Y}^*) = \frac{1}{2}\pi|\mathbf{P}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{Y}^* - \boldsymbol{\mu})'\mathbf{P}^{-1}(\mathbf{Y}^* - \boldsymbol{\mu})\right],$$

where **Y*** is the vector [y_1* y_2*]', **P** is the correlation matrix of the unobserved continuous variables, and $\boldsymbol{\mu}$ is [0 0]'.

The standardized thresholds, mean, and variance for each variable are then converted to the defined scale. For example, for variables with three or more ordered categories, fixed values for the first two thresholds τ_1 and τ_2 are used as the basis for establishing the value of the mean and variance of the underlying variables on this scale using the following transformations:

$$\mu = \tau_1 - \hat{z}_1\left(\frac{\hat{z}_1 - \hat{z}_2}{\tau_1 - \tau_2}\right) \text{ and } \sigma^2 = \left(\frac{\tau_1 - \tau_2}{\hat{z}_1 - \hat{z}_2}\right)^2$$

These values for the mean and variance can then be used to transform the remaining standardized threshold estimates to the scale determined by the fixed thresholds.

Finally, typically one of several varieties of weighted least squares (WLS) estimation (Jöreskog, 1990; B. O. Muthén, 1993; B. O. Muthén, du Toit, & Spisic, 1997) is used to fit the Latent Growth Curve Model to the means and polychoric correlations. In WLS estimation the model parameters are estimated by minimizing a fit function that incorporates a weight matrix **W** as in Browne's (1984) asymptotic distribution free estimation technique,

$$F_{\text{WLS}} = \sum (\mathbf{r} - \hat{\boldsymbol{\rho}})' \mathbf{W}^{-1} (\mathbf{r} - \hat{\boldsymbol{\rho}}),$$

where the vector **r** contains the nonduplicated elements of the polychoric correlation matrix,[1] the vector $\hat{\boldsymbol{\rho}}$ contains the nonduplicated elements of the model-implied correlations among the underlying *y** variables, and the weight matrix **W** is the asymptotic covariance matrix, which needs to be inverted during each iteration of the estimation process. Variations on this procedure include diagonally weighted least squares (DWLS) estimation (Jöreskog, 1990) and WLSM (Mean) and WLSMV (Mean, Variance) estimation (B. O. Muthén, 1993). Maximum likelihood estimation methods for categorical data also exist (see Jöreskog & Moustaki, 2001; S. Lee, Poon, & Bentler, 1990, 1992; Mehta, Neale, & Flay, 2004), but have limitations due to the computational intensity of the estimation algorithms.

The Linear Mixed Effects Model framework has likewise been extended to Generalized Linear Mixed Effects Models (Breslow & Clayton, 1993; Schall, 1991). Here the notable addition to the model that allows for categorical data is the introduction of a logit link function in the Level 1 model that transforms the metric of the categorical outcome to a continuous logit metric, whose trajectory can be modeled

[1] Some methods also include the threshold values in the **r** vector (Muthén, 1984).

using the same terms as in the traditional Linear Mixed Effects Model. A binary outcome variable is assumed to have a Bernoulli distribution with probability π for the occurrence of a value of one and probability 1-π for the occurrence of a value of zero. The log odds of π,

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right),$$

is used in place of the dependent variable in the Level 1 model. The Level 2 model remains the same as in the Linear Mixed Effects Model.

Although the concept of this model is relatively straightforward, ongoing technical developments have been necessary to extend methods for estimating the model parameters. Lee and Nelder (1996) proposed a hierarchical likelihood method for estimating the parameters of this model. Other recent developments in estimation methods have added to the practicality of using these models (Coull, Houseman, & Betensky, 2006). Some authors have noted the capacity of this model for modeling categorical variables over time (Raudenbush & Bryk, 2002). Other authors have noted the correspondence between this model and the Rasch item response model (Kamata, 2001; Raudenbush, Johnson, & Sampson, 2003). However, few authors have used this model as a means for combining data from multiple categorical response variables (e.g., responses to multiple items) at each time point (McArdle, Grimm, Hamagami, Bowles, & Meredith, in press).

Timing of Measurements in Longitudinal Study Design

In the most common scenario measures of academic achievement are collected on a yearly basis. In the era of individual assessment-supported

accountability, it is becoming more common to have interim assessments, that is, to assess students at two to three additional occasions leading up to the end of the year assessment. In some national longitudinal studies conduced by the National Center for Education Statistics (NCES), measurements at subsequent time points may be as far as two years apart. In a rapidly-growing population this may indeed strain the measurement of growth. One potential solution that was implemented in one NCES study, the Early Childhood Longitudinal Study – Kingergarten Cohort 1998-1999 (Tourangeau, Nord, Lê, Pollack, & Atkins-Burnett, 2006), was the use of a "bridge" sample in the skipped year. It is anticipated that this type of longitudinal design leads to stronger vertical scaling.

Problem Statement

Both Latent Growth Curve Modeling and growth models within the Linear Mixed Effects Modeling framework are techniques originally intended to model growth with interval level observed variables that have been extended to accommodate ordinal level observed variables. Most growth modeling applications in these frameworks give little consideration to the measurement model for the ordinal item responses. Some applications of these methods can be considered to use a limited measurement model where there are multiple indicators at each time point, as in cases where multiple test items (either polytomous or binary) are administered at each time. However, this is not generally how these methods are used currently. Typically, a separate measurement model is used. On the other hand one particularly well-developed and flexible measurement model framework is the Item Response

Theory framework. This model framework includes measurement models that are more sophisticated than those used in Latent Growth Curve Models or Linear Mixed Effects Models. What is more, this modeling framework was originally developed for categorical responses to test items, and so it incorporates ordinal data quite naturally. Although this framework also makes use of transformations as used in Generalized Linear Mixed Effects Models, such as the logit transformation, this transformation is used to produce a non-linear model, and a well-developed lexicon exists for interpreting the model parameters and communicating the meaning of these nonlinear model parameters to others in ways that are somewhat intuitive.

The great irony is that although categorical data have been viewed as a complication in other frameworks used to model growth, one of the best-developed frameworks for categorical data has practically been ignored when it comes to growth modeling. Item response theory has not traditionally been used to model growth. Further, the few developments that would allow for the modeling of growth in this framework have occurred exclusively within the Rasch family of IRT models (Embretson, 1991; te Marvelde, Glas, Van Landeghem, & Van Damme, 2006; Wang, Wilson, & Adams, 1998). This family of models, however, is limited in its appropriateness for applications in education. This is because these models may not appropriately describe data where items differ in their discrimination among examinees or where the probability of answering an item correctly may be affected by the possibility of the examinee guessing. Thus, despite the greater measurement model sophistication in the Item Response Theory framework, this level of development has not been leveraged in modeling growth. Latent Growth Curve

Modeling and growth models within the Linear Mixed Effects Model framework remain further developed in this regard.

Current practice in modeling growth in education often begins by estimating measures of overall performance for a student at several specific time points using an IRT model and then treating these estimates as known values in a separate Latent Growth Curve Model or Linear Mixed Effects Model for the growth or change in each student's knowledge or ability over time. The consequences of taking an estimated value and using it as a known value in a growth model are not well understood. Indeed, some research studies suggest that growth models can be sensitive to the scale that is used to measure the growth (Goldschmidt, Choi, & Martinez, 2004; Seltzer, Frank, & Bryk, 1994), but few studies have systematically studied this phenomenon.

Current practice with separate models often involves several manipulations or transformations of the score scale: first an IRT model to estimate scores, then a vertical scaling transformation to put the scores from increasingly difficult tests on a developmental scale, and finally a model of the change or growth in the developmental scale scores for individual students over time.

A single integrated IRT growth model that can handle all these aspects simultaneously holds the potential to keep the determination of the growth scale close to the data. Indeed, McArdle, Grimm, Hamagami, Bowles, and Meredith (in press) found that longitudinal models within the IRT framework exhibited more robust performance than other methods for analyzing latent growth curves using different age-appropriate measures of the same construct over time. How would such a model

be developed and estimated outside of the Rasch family of models, and under what circumstances will the parameter estimates be accurate in their recovery of both item characteristics and growth characteristics underlying the item response data?

An Integrated Model

This study proposes and investigates parameter recovery for an integrated IRT model that is appropriate for modeling individual growth in educational achievement data. There are two types of item parameters and two types of person parameters in the proposed model. The two item parameters are typically found in item response models in education. The first is the item location parameter, represented by the letter *b*, which models the difficulty of the item. The second is the item discrimination parameter, represented by the letter *a*, which models the extent to which the item clearly discriminates between examinees who have the requisite skill(s) or ability to answer the item correctly and those who do not. The Rasch model is the simplest IRT model because it excludes the item discrimination parameter and instead assumes that all items have equal discrimination. The proposed model builds upon prior IRT longitudinal model research with Rasch models by expanding the two-parameter logistic model to model growth in latent ability over time.

Typically in item response models there is a single person parameter, represented by the Greek letter θ, which models the level of skill or ability of the examinee. There may be multiple person parameters if multiple skills or abilities are being assessed in the same test, as in Multidimensional IRT. In the proposed model θ represents the examinee's level of skill or ability at the initial time of assessment. An additional person parameter, represented by the Greek letter δ, models the change in

this ability as an increment (or decrement) over time. It should be clarified that the person parameters θ and δ are both random variables in the model. That is, their values vary across individuals so that not all individuals need start with the same level of ability θ nor increase at the same rate δ. However, the proposed model of linear growth may be too simplistic in some applications. Indeed, growth in areas such as elementary reading may be characterized by a change in the rate of growth over time at some critical transition point(s). This can be modeled using a piecewise linear growth trajectory. Other more gradual changes in the rate may be modeled by instead extending the model to include a quadratic term.

Because the observed item response data have only two discrete categories (1 or 0, right or wrong), it is necessary to transform the data onto a continuous scale. This is achieved using the logit transformation. This transformation divides the probability of observing a "1" by the probability of observing a "0" and then takes the natural logarithm of the result. The proposed model may be viewed as a member of the family of nonlinear mixed models (De Boeck & Wilson, 2004). Within the nonlinear mixed model framework a version of the model expressing linear growth would be denoted as:

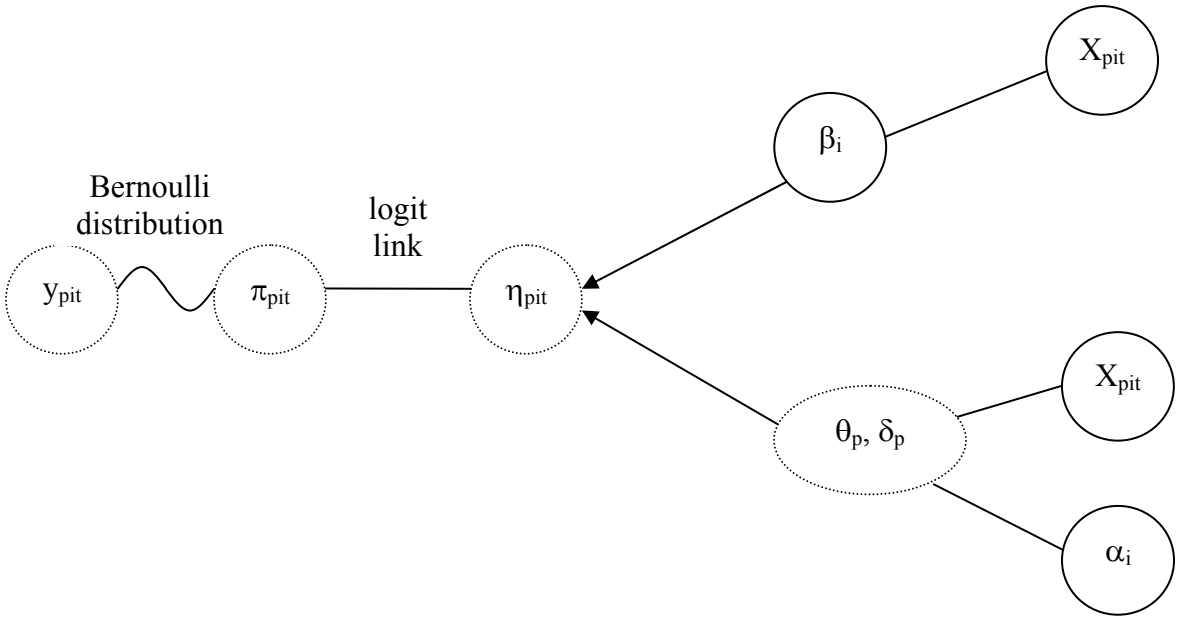
$$\text{logit}\left(y_{pit}\right) = \eta_{pit} = \sum_{i=1}^{I} a_i \theta_p X_{pit} + t\sum_{i=1}^{I} a_i \delta_p X_{pit} - \sum_{i=1}^{I} \beta_i X_{pit}$$

where y_{pit} is the observed binary response of zero or one from person $p = 1, 2, \ldots, N$ on indicator $i = 1, 2, \ldots, I_{pt}$ at a duration of time t since the first measurement, a_i is the item discrimination parameter for item i, θ_p represents the examinee's level of skill or ability at the initial time of assessment, X_{pit} is an item indicator that takes the value one if person p was administered item i at time t and zero otherwise, δ_p models the

change in examinee ability as an increment per unit of time, and $\beta_i = a_i b_i$, where b_i is the item location parameter. The notation η_{pit} represents the expected value of a latent continuous variable underlying the observed response y_{pit}. This expected value is predicted by the systematic component of the model, which is a function of the item indicator variables X_{pit} as well as the latent item and person parameters.

A symbolic representation of this model is shown in Figure 1. In this representation random variables are represented by circles drawn with broken lines. Fixed variables are represented by circles drawn with solid lines. The curved connecting line between y_{pit} and π_{pit} represents the Bernouli distribution. π_{pit} represents the probability that the observed response y_{pit} is one. The straight connecting line represents the logit transformation of π_{pit} to the linear predictor η_{pit}. The arrows represent additive portions composing the linear predictor. Two such additive components are depicted: one containing the item location within the parameter β and the other containing the structural parameters for the growth trajectory, depicted here for the linear model as including the parameters θ_p and δ_p.

Figure 1: Symbolic Model for the Two-Parameter Logistic Item Response Growth Model



Note: The convention used in this diagram closely follows that used in De Boeck and Wilson (2004).

The person parameters θ and δ are normally distributed random variables with mean vector $[0, \mu_\delta]$' and variance/covariance matrix $\mathbf{T} = \begin{bmatrix} 1 & \\ \tau_{10} & \tau_{11} \end{bmatrix}$. The mean and variance of θ are fixed to 0 and 1, respectively, to identify the model. In addition, the variance of the residuals is assumed to be normally distributed with mean zero and variance one. It should be noted that the linear version of the model has been

presented here and that a nonlinear version will also be investigated as part of the proposed Monte Carlo study and real data illustration discussed below.

Environmental Context and Significance

In January of 2002 the No Child Left Behind Act of 2001 (NCLB) was enacted by Congress and became a dramatic force in the development of school accountability systems across the nation. As education agencies have sought to meet the provisions of this legislation and as Congress has considered the reauthorization of this bill in 2007, there has been a great surge of interest in identifying appropriate methods for measuring student progress in academic achievement, particularly as it applies to the adequate yearly progress provisions of the law. A number of alternative approaches have been suggested for measuring student progress and carrying out adequate yearly progress calculations, and limitations of cross-sectional methods for measuring adequate yearly progress have been discussed in technical reports and education journals (Arce-Ferrer, Frisbie, & Kolen, 2002). Some authors argue that the adequate yearly progress provision be modified to focus exclusively on growth (Peterson, 2007) or at least to include growth with other considerations (Thum, 2003). The fundamental argument for modeling growth in school accountability is that the student's own performance is the most appropriate baseline against which to judge whether effective learning is occurring. When growth over time for many individual students is combined in a single model, a pattern of growth for a larger entity, such as a school or district, can be readily observed. Recognizing the desire to incorporate growth into adequate yearly progress calculations, the U.S.

Department of Education has solicited and accepted proposals from several state education agencies to conduct pilot studies using growth models (2007, July 3).

Historically the research methodology for studying individual growth and change has been plagued with paradoxes and apparent dilemmas (Harris, 1963; Rogosa, 1995). Attempts to measure change at the aggregate level, such as a school or district, are particularly sensitive to unintended influences when the methodology does not have an appropriate basis in modeling changes in the individuals that make up the aggregate. Growth models are likely to continue to be sought after as an important tool for monitoring educational systems by tracking changes in individual students' academic achievement. At the same time, however, more work is needed on methodologies for modeling growth, and these methodological issues are inherently tied to the psychometric scale(s) used to measure change over time. These issues are particularly critical in situations where there is a great deal of variability across students and a great deal of growth occurring between measurements, as occurs when modeling student growth across grades in an inclusive testing program.

Despite ongoing developments in growth modeling methodology, however, modeling growth in individual students' academic abilities over time currently typically uses several statistical models or transformations to move from data representing a student's correct or incorrect responses on individual test items to inferences about changes in the student's underlying ability. First, a measurement model, typically from the IRT family of models, is used to move from a student's responses to many individual items to a single test score that provides an estimate of the student's proficiency at the time he or she answered the items. Then a vertical

scaling transformation may be needed to put scores from increasingly difficult tests on a developmental scale. Finally, a trajectory for the individual scaled estimates of proficiency may be estimated in a growth model, typically from the Latent Growth Curve Model or Linear Mixed Effects Model frameworks.

Each of these steps has received a great deal of focused research attention isolated from the other steps. Indeed, many dissertations in the psychometrics field and its related disciplines focus on a specific problem or issue involved in item response models, or vertical scaling, or longitudinal models of growth or change. However, more research is needed to address the limitations that yet exist because of the relative isolation of these topics from one another. All three are important for effective and reliable interpretation of results arising from studies of growth. There are two important reasons why this research is needed.

First, the choice of metric used for growth modeling can make an important difference in the policy implication of the results of a growth study. This is especially critical given the movement toward using growth models to meet the requirements for educational accountability (e.g., NCLB). The misallocation of educational resources can have tremendous consequences for communities. Decisions regarding the allocation of educational resources are sufficiently contentious without compounding the situation with methodological ambiguity. Seltzer, Frank, and Bryk (1994) provided a concrete example of the problem in their fitting of a two piece growth model to grade equivalent scores and IRT scores from Iowa Test of Basic Skills data for children in grades 1-6 in the Chicago Public Schools. The model with the grade equivalent scores suggested that resources should be focused on grades 4-6; the

model with the IRT-based scores suggested that resources should be focused on grades 1-3. Goldschmidt, Choi, and Martinez (2004) likewise compared growth models using normal curve equivalents and IRT scale scores for the purposes of evaluating school performance and program effectiveness. Goldschmidt et al. found that although several statistical inferences were unaffected by the difference in metric, conclusions about the magnitude of growth were greatly affected. Goldschmidt et al. did not investigate inferences regarding the shape of the growth pattern, however. Both the Goldschmidt et al. and the Seltzer et al. studies used existing data and psychometric scales to demonstrate their points. Both studies are subject to the limitation of not knowing with certainty what the true underlying mechanism was in each case. However, a simulation study by Leite (2007) investigating growth models in the Latent Growth Curve Modeling framework controlled the underlying growth mechanism and found that a latent growth curve approach fit to the composite means of groups of five point Likert items resulted in positively biased estimates of the mean of the latent intercept. Decisions about latent scaling affect the growth model results and the interpretation of those results. Thus, the implications of these decisions need to be thoroughly researched and understood so that conscious decisions may be made in the context of a growth study.

If the choice of the score scale can make such a difference in the interpretation of growth model results, how much more of an influence might there be in other aspects of the multi-step process of moving from item responses to growth model parameters? In addition to the observed differences due to the choice of score scale, other factors may affect the resulting parameter estimates from a growth model. For

example, the consequences of taking an estimated scale score value and using it as a known value in a growth model are not well understood, and few studies have systematically studied this phenomenon.

Second, whereas a great deal of development has taken place in modeling frameworks such as Latent Growth Curve Modeling and Linear Mixed Effects Modeling to model growth over time, the IRT framework has not traditionally been used to model growth. Some limited research has shown, however, that it is possible to use an IRT growth model to model growth directly from the item response data (e.g., Embretson, 1991). Published research studies of longitudinal item response models seem to be rare with little systematic development over the past 20 years. Further, developments that would allow for the modeling of growth in this framework have occurred almost exclusively within the Rasch family of IRT models (Embretson, 1991; te Marvelde, et al., 2006; Wang, et al., 1998). The Rasch family of models, however, is limited in its appropriateness for applications in education, as achievement test items often differ in the extent to which they discriminate among examinees at different levels of achievement. Expansions of the IRT growth model concept to the two-parameter and three-parameter models would expand the potential applicability of these models in education.

A single integrated IRT growth model that can handle item response modeling, vertical scale transformation, and growth modeling simultaneously is worth further investigation. Such a model holds the potential to keep the determination of the growth scale close to the data, and an understanding of the consequences of this option may yield further insight into the issues surrounding

vertical scaling and the choice of growth metric. The value of the proposed study is its ability to contribute to the growth modeling literature from the standpoint of an underdeveloped framework for growth. Further it approaches the problem of growth modeling with categorical data from the opposite standpoint than the bulk of the research on this topic, which occurs primarily in the Latent Growth Curve Model and Linear Mixed Effects Model frameworks.

Chapter 2: Literature Review

Item Response Models for Longitudinal Data

Early models for repeated measures in the IRT framework

Fischer's (1976) linear logistic latent trait model with relaxed assumptions and Andersen's (1985) multidimensional Rasch model are often acknowledged as presenting some of the first IRT models to accommodate repeated measures data. Fischer's approach models a different theta for each item as well as each person and incorporates one or more treatment effects, which could include the effect of time to create a growth model. However, Embretson (1991) noted that since the treatment effects are assumed to be the same for all individuals measured at the same time intervals, this model is not appropriate for measuring individual differences in change over time. This is limitation shows that the level of development of this model is analogous to early growth model developments in other frameworks that focused on the mean trajectories for groups of individuals.

Andersen (1985) included a different θ for each person at each time point but maintained the same item difficulty value over repeated administrations of the same items. The model accommodates correlation among the thetas at the different time points. Anderson's multidimensional model parameterizes ability at each time point as a different dimension, rather than parameterizing change at each subsequent time point. Thus, change scores are computed outside of the model, leaving them vulnerable to some of the same limitations as traditional difference scores. Roberts

and Ma (2006) noted that when the difference scores are computed using this model there is a negative relationship between the reliability of the difference score and the correlation between the two measures used to construct it. Thus, authors, such as Embretson (1991), have noted that this model too is inappropriate for measuring individual differences in change over time. Andersen's model is also restricted to situations in which all examinees are measured at the same time points.

The multidimensional IRT approach for longitudinal data

Proposed methods for handling longitudinal data within an IRT framework have generally drawn upon the literature for Multidimensional IRT models. In fact, Andersen's (1985) model is one such example. More recent attempts to take this approach have drawn from the multidimensional random coefficients multinomial logit model framework as outlined by Adams, Wilson, and Wang (1997). Adams et al. (1997) focused on what they call a "between-item" multidimensional model, where a test can be divided into groups of items with each group represented by a unidimensional IRT model. However, performances on these dimensions measured by different latent variables were correlated. This was contrasted with a "within-item" multidimensional test, where individual items measured multiple dimensions. In applying the former type of multidimensional model to longitudinal data, Wang, Wilson, and Adams (1998) considered measurements at different time points to be different dimensions in the multidimensional model. te Marvelde, Glas, Van Landeghem, and Van Damme (2006) likewise showed the application of the multidimensional generalized partial credit model for repeated measures. Unlike Anderson's model, the Wang et al. and te Marvelde et al. approaches did not require

that item parameters be constant over time. However, the restriction remained that all examinees be measured at the same time points.

The latent change model approach for longitudinal data

Embretson's (1991) multidimensional Rasch model for learning and change presented perhaps the first IRT growth model consistent with major advances in growth modeling in other frameworks (i.e., Latent Growth Curve Modeling and Generalized Linear Mixed Modeling). As with the Fischer (1976), Andersen (1985), and Wang, Wilson, and Adams (1998) models, Embretson's model was developed within the Rasch family of IRT models. Rather than model separate abilities at each time point, however, Embretson proposed modeling a latent ability at the initial time point and the change score, or modifiability, at each subsequent time point. This overcame a limitation of the multidimensional IRT approach for longitudinal data, namely that the change scores are computed within the model, and thus avoided an inverse relationship between the reliability of the change score and the correlation between the two measures. Embretson demonstrated that model parameters were accurately recovered when estimating this model using maximum likelihood estimation and showed that the standard error of the change estimates did not depend on the correlation between theta estimates at adjacent time points. A design matrix was used to match the correct modifiability to the specified time point for the observed item response. Roberts and Ma (2006) likewise followed Embretson's (1991) approach to extend the generalized partial credit model for multiple measurements over time. Unlike the te Marvelde et al. (2006) model, which was also a multivariate extension of the generalized partial credit model, this model was

parameterized in terms of difference scores. Person parameters were parameterized as an initial theta and change scores between adjacent time points.

Roberts and Ma (2006) noted the advantages of using IRT for measuring change over the gain score method. In IRT the reliability of the gain score is of less interest than the precision, as represented by the standard error, of the estimate of change in the IRT model. This overcame the reliability paradox of the gain score. Roberts and Ma (2006) showed that the test characteristic curve modeled the nonlinear relationship between changes in expected observed score and changes in latent trait (true) scores. This overcame the limitation that true change was not necessarily linearly related to gain scores.

When contrasted against the most recent growth modeling developments in other frameworks, however, the IRT models presented by Embretson (1991) and Roberts and Ma (2006) have two notable limitations. First, these models are restricted to circumstances in which all examinees are measured at the same time points. They cannot be used in situations in which examinees are tested at different points in time, as in some formative or embedded assessment programs. Second, they do not impose a functional form on the growth. Thus, they do not allow for model-based testing of a particular functional form for growth, as might be desired for studies in education and human development. Both models employ a piecewise-defined trajectory, in which each measured time point defines its own piece in the model. It is ironic that IRT authors, such as Embretson (1991) and Roberts and Ma (2006), have extolled the virtues of using IRT to measure change, and yet so little systematic development has taken place for growth models in this framework!

Explanatory item response model approach

In 2004 DeBoeck and Wilson edited a book on explanatory item response models. They reframed Rasch and two-parameter logistic IRT models as members of the family of nonlinear mixed models and showed how covariates could be incorporated to explain item parameters, person parameters, or both. Wilson, Zheng, and Walker (2007) built upon this conception of IRT within the nonlinear mixed model framework by proposing an extension of the Rasch item response model to incorporate growth parameters in a manner similar to multilevel model approaches for growth. Wilson et al. called this a Latent Growth Item Response Model (LG-IRM). They showed that parameters were successfully recovered in this model using standard Rasch model software (ConQuest) and illustrated the application of the model with NELS data. By positing relationships among latent person parameters, IRT growth models are one manifestation of explanatory item response models, although DeBoeck and Wilson did not explicitly include growth models in their book. As with the multilevel approach for modeling growth, these models allowed examinees to be measured at different time points and allow an a priori hypothesized model for growth to be imposed on the data.

Table 1 summarizes the three different approaches for longitudinal item response models by listing the most salient papers representing each approach.

Table 1: Summary of Longitudinal IRT Models

Correlated abilities over time	Latent difference scores	Functional form for growth
Andersen's (1985) multidimensional Rasch model	Embretson's (1991) multidimensional Rasch model for learning and change	Fischer's (1976) linear logistic latent trait model with relaxed assumptions
Wang, Wilson, & Adams (1998)	Roberts & Ma (2006)	Wilson, Zheng, & Walker (2007)
te Marvelde, Glas, Van Landeghem, & Van Damme (2006)		

Connections with Multidimensionality and Vertical Scaling

Conceptually, the proposed IRT model (and its predecessors discussed here) is a unidimensional model. However, the formulation of this model draws upon prior work in Multidimensional IRT (MIRT). True IRT growth models (those that impose a functional form for growth on the data) actually create a distinct category of MIRT models. Here the vector of thetas can be reduced to a vector containing a (potentially) smaller number of parameters that define a functional relationship among the elements of the original vector.

Estimation in Longitudinal IRT Models

By far the most common estimation method for longitudinal item response models previously reviewed is marginal maximum likelihood estimation. This estimation method was adopted in the longitudinal item response models of Wang, Wilson, & Adams (1998), te Marvelde, et al., and Wilson, Zheng, and Walker (2007).

One notable exception is Embretson (1991), who used a two step procedure estimating item parameters first using conditional maximum likelihood and then estimating the person parameters using the estimates of the item parameters as known values. The next several sections describe marginal maximum likelihood estimation in greater detail.

The Marginal Likelihood Function

In the linear trajectory version of the proposed model, the likelihood of response for an individual person *p* is the product of the model-implied probability of each observed item response given the unobserved values of the person parameters θ and δ. Thus, the likelihood function for an individual person *p* is constructed as:

$$L(\mathbf{y_p} \mid \theta_p, \delta_p) = \prod_{i=1}^{I} \left(\Xi\left(\eta_{pit}\right)\right)^{y_{pit}} \left(1 - \Xi\left(\eta_{pit}\right)\right)^{1-y_{pit}}$$

(Hambleton, Swaminathan, & Rogers, 1991), where Ξ is the inverse logit transformation. Recall that η_{pit} represents the expected value of a latent continuous variable that is predicted by the systematic component of the model. For the linear trajectory version of the model, this systematic component is defined as:

$$\eta_{pit} = \sum_{i=1}^{I} a_i \theta_p X_{pit} + t \sum_{i=1}^{I} a_i \delta_p X_{pit} - \sum_{i=1}^{I} \beta_i X_{pit} \ .$$

To obtain the marginal maximum likelihood the likelihoods of response for all of the individuals in the sample are multiplied. The joint likelihood

$$\prod_{p=1}^{N} \prod_{i=1}^{I} \left(\Xi\left(\eta_{pit}\right)\right)^{y_{pit}} \left(1 - \Xi\left(\eta_{pit}\right)\right)^{1-y_{pit}}$$

is marginalized by integrating over the values of person parameters:

$$\prod_{p=1}^{N} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} L(\mathrm{y_p} \mid \theta_p, \delta_p) \phi(\theta_p, \delta_p \mid 0, \Sigma) d\theta d\delta .$$

where $\phi\left(\theta_p, \delta_p \mid \mathbf{0}, \Sigma\right)$ is the bivariate normal probability density function for the random effects (Tuerlinckx et al., 2004).

The estimation algorithm then seeks parameter estimates that will maximize the value of this marginal likelihood expression. There are several estimation algorithms that may be used (for an overview see Tuerlinckx et al., 2004). Maximization of the likelihood can be achieved indirectly using the Expectation-Maximization (EM) algorithm with Gauss-Hermite quadrature. This full-information approach, which iteratively alternates between computing and maximizing the expected value of the marginal log likelihood, will be explained in detail in the following section.

Indirect Maximization of the Marginal Likelihood

In the EM algorithm estimation approach the values of the fixed and random effects are considered to be missing data and the item responses are considered to be observed data. Given some potential values for the missing parameter values, a likelihood for the complete (missing and observed) data can be constructed as shown in the previous section. The expected value of the complete data loglikelihood is:

$$\sum_{p=1}^{N} \int \int \left(\left\{ \log\left(\left(\Xi(\eta_{pit}) \right)^{y_{pit}} \left(1 - \Xi(\eta_{pit}) \right)^{1-y_{pit}} + \log\left(\phi\left(\theta_p, \delta_p \mid 0, \mu_\delta, 1, \tau_{01}, \tau_{11} \right) \right) \right) \right\} h\left(\theta_p, \delta_p \mid \mathbf{y}, \boldsymbol{\beta}, \mu_\delta, \tau_{01}, \tau_{11} \right) d\theta d\delta \right)$$

(Tuerlinckx et al., 2004; Tuerlinckx, Rijmen, Verbeke, & De Boeck, 2006) where $\phi\left(\theta_p, \delta_p \mid 0, \mu_\delta, 1, \tau_{01}, \tau_{11}\right)$ is the bivariate normal probability density function for the random effects and $h\left(\theta_p, \delta_p \mid \mathbf{y}, \boldsymbol{\beta}, \mu_\delta, \tau_{01}, \tau_{11}\right)$ is the marginal posterior density of the

random effects (growth) parameters given the item responses (**y**), as well as the current proposed values of the item parameters (**β**; fixed effects) and the parameters defining the mean and covariance structure of the random effects (Adams, et al., 1997). This is what Tuerlinckx et al. (2004) refer to as the conditional density of the random effects given the observed data.

The expected value of the complete data log likelihood can be divided into two major parts, which can be tackled separately in the maximization step. The first part is the fixed effect (item parameter) component

$$\sum_{p=1}^{N} \int\int \left(\log\left(\left(\Xi\left(\eta_{pit} \right) \right)^{y_{pit}} \left(1 - \Xi\left(\eta_{pit} \right) \right)^{1-y_{pit}} h\left(\theta_p, \delta_p \mid \mathbf{y}, \boldsymbol{\beta}, \mu_\delta, \tau_{01}, \tau_{11} \right) d\theta d\delta \right) \right),$$

which can be further divided by item and maximized separately for each item due to the conditional independence assumption of the item response model. This quantity reflects the expected proportion of observed values of 1 on each item at each level of the latent propensity θ (Wirth & Edwards, 2007). The second part is the random effect (growth) component

$$\sum_{p=1}^{N} \int\int \left(\log\left(\phi\left(\theta_p, \delta_p \mid 0, \mu_\delta, 1, \tau_{01}, \tau_{11} \right) h\left(\theta_p, \delta_p \mid \mathbf{y}, \boldsymbol{\beta}, \mu_\delta, \tau_{01}, \tau_{11} \right) d\theta d\delta \right) \right),$$

(Adams, et al., 1997). This quantity reflects the expected proportion of people at each level of the latent propensity θ (Wirth & Edwards, 2007).

In the maximization step the individual components of the expected likelihood for the fixed- and random-effects parameters are maximized to obtain estimates of the item parameters (also called structural parameters; Wirth & Edwards, 2007) and the random effects (growth) parameters. The optimization, or maximization, is carried out using a quasi-Newton approach, which begins by using initial values of the

parameter estimates to generate first-order and estimates of second-order derivatives of the expected log likelihood. Then, a system of equations based on setting these first-order and estimates of second-order derivatives of the expected log likelihood equal to zero is solved to obtain closer estimates of the parameter values. These new estimates are used to repeat the step of generating the first-order and estimates of second-order derivatives of the expected log likelihood function. The process is repeated until the new estimates are not very different from the previous estimates.

A numerical approximation is used to approximate the integrand before integrating or approximate the integral itself in the expected log likelihood (Tuerlinckx et al., 2004; Wirth & Edwards, 2007). Several numerical approximation approaches have been suggested. Investigating estimation from the perspective of the nonlinear mixed effects model framework, Pinheiro and Bates (1995) examined four numerical approximation approaches for dealing with the integral and found that three methods: linear mixed-effects approximation, Laplacian approximation, and Gaussian quadrature centered at the conditional modes of the random effects, to be both accurate and computationally efficient. The Gauss-Hermite quadrature method will be described here (Tuerlinckx et al., 2004; Wirth & Edwards, 2007). This method specifies a number of quadrature points, and a rectangular area is estimated at each point. The areas of the rectangles are summed over all the quadrature points to approximate the area under the distribution. For multidimensional integrals, such as in this model, which has two dimensions for the linear trajectory, the number of quadrature points must be specified for both dimensions, and quadrature is used to estimate the multidimensional integral. The estimated values of the item parameters

are then used to repeat the expectation step and the whole process begins again and continues until there is very little change in the values of the parameters (Wirth & Edwards, 2007).

Estimation of Person Parameters

The EM algorithm does not result in estimated values for the person parameters defining individual growth patterns. However, empirical Bayes estimates, also sometimes called predictions, for these values may be computed using the results of the Marginal Maximum Likelihood Estimation. The empirical Bayes estimates of θ_p and δ_p (in the case of a linear trajectory model) are the values which maximize the log likelihood

$$L(\mathrm{y}_\mathrm{p} \mid \hat{\boldsymbol{\beta}}, \hat{\mu}_\delta, \hat{\tau}_{01}, \hat{\tau}_{11}, \theta_p, \delta_p)\phi\left(\theta_p, \delta_p \mid 0, \Sigma\right),$$

where the previously-obtained estimates of the fixed effects for the item parameters and of the parameters defining the random effects distribution from the marginal maximum likelihood estimation are substituted as known values.

Estimation Alternatives

Fully Bayesian estimation approaches, including those implemented with Markov Chain Monte Carlo (MCMC) methods, have also been gaining popularity in the literature (Roberts & Ma, 2006; Tuerlinckx et al., 2004; Wirth & Edwards, 2007). Roberts and Ma (2006) used a fully Bayesian estimation procedure to estimate parameters for their model. Wirth and Edwards (2007) particularly note the potential for MCMC because it avoids instances of multiple integration. Multiple integration is a particular issue in longitudinal IRT models because the dimension of the integration

is increased for each person parameter (random effect) added to the model. In this model Bayesian estimation via MCMC is computationally intensive, but may be considered as a viable alternative where traditional estimation approaches break down for more complex models.

The generalized estimating equations (GEE) approach, as found in available programs such as Proc Genmod in SAS, may sometimes be considered as an alternative to marginal maximum likelihood estimation. This estimation approach iteratively solves a system of equations based on the partial derivatives of the log likelihood with respect to the linear predictor η_{pit} to obtain estimates of the fixed effects. These partial derivatives are weighted by the deviations of the observations from their predicted value, and these predicted values are initially generated by assuming all observations are independent (even if they were responses by the same examinee at the same point in time). For more details and a gentle introduction to GEE, see Hanley, Negassa, Edwardes, and Forrester (2003). This approach has the advantage of allowing for correlated data as found in panel studies without specifying an explicit model for the covariance structure of the random effects. Rather, a "working" correlation matrix of the manifest observations is created and used in the estimation. However, there are two limitations to this approach for the item response model in this study.

First, the GEE approach cannot be used to fit random effects models. The application of GEE leads to a situation in which only the mean structure for the growth model is estimated, and inferences are only made for the mean growth in the population. For this reason this approach is sometimes called a population-averaged

approach (Dmitrienko, Molenberghs, Chuang-Stein, & Offen, 2005). Thus, some of the parameters of interest in the proposed model, such as the variances and covariances of the random person effects, will not be estimated. Second, in order for the model to be identified, constraints are needed to identify the latent scale. Typically in item response theory this is handled by fixing the mean and variance of the person parameters, in other words, the random effect. In the absence of this random effect, an alternative means of imposing this constraint would need to be applied, such as constraining the average of the discrimination parameters to one. Due to these limitations the model that would be specified and estimated in Proc Genmod would not be comparable to the item response growth model proposed in this study.

Marginal maximum likelihood estimation with 2I item parameters, where I is the number of items, is quite computationally intensive because each of these item parameters adds one more fixed effect to be estimated. Could the two different item parameters (item location and item discrimination) instead be incorporated into the model as random effects? Unfortunately, the currently available program Proc Nlmixed in SAS does not support estimation with random effects at multiple levels. This prohibits the specification of a model in which both person and item parameters are treated as random effects in a nonlinear mixed effects model.

Hypotheses

The goal of this research project was not only to propose a new model but also to explore the effects of examinee population and item characteristics on the estimation of model parameters. These examinee population and item characteristics

determine the context for a longitudinal study design. Two broad research questions were asked. First, which characteristics in a longitudinal study design affect the variability in the estimates that are produced for the item parameters and the means, variances, and covariances of the random effect distributions? Second, under what conditions does the proposed model and estimation algorithm produce unbiased estimates for the item parameters and the means, variances, and covariances of the random effect distributions?

In addition a real data illustration was introduced to show that the utility of the model generalizes beyond a tightly controlled laboratory experiment.

Chapter 3: Monte Carlo Investigations of Parameter Recovery

Rationale

The goal of this research project was not only to propose a new model but also to explore the effects of examinee population and item characteristics on the estimation of model parameters. These examinee population and item characteristics determine the context for a longitudinal study design. Two broad research questions were asked. First, which characteristics in a longitudinal study design affect the variability in the estimates that are produced for the item parameters and the means, variances, and covariances of the random effect distributions? Second, under what conditions does the proposed model and estimation algorithm produce unbiased estimates for the item parameters and the means, variances, and covariances of the random effect distributions?

What might these characteristics be and what answers to these questions may be anticipated? Based on previous research with IRT growth models within the Rasch family (Embretson, 1991), it was expected that under reasonable conditions the item parameters in the model will be accurately recovered. However, it was also anticipated that item and growth model parameters may not be as robust under more extreme combinations of conditions. The series of three related Monte Carlo studies to follow investigated the effects of a total of five factors that could influence parameter recovery: the form of the growth trajectory, the variance of the rate of growth, sample size, test length, and the item selection design. It was anticipated that

variability in the item and growth model parameters would be reduced as sample size was increased. Item parameter estimates for items with extremely high or extremely low item location parameters would be recovered with less bias when item selection overlaps across time points. The form of the trajectory itself should not directly influence parameter recovery, but it was anticipated that the difference in the number of random effects that comes with differing trajectories may have an effect due to the increase in information needed to estimate additional parameter values.

Manipulated Factors

To explore the quality of parameter recovery in the new model, longitudinal data was simulated for examinee responses to sets of items at five time points. By knowing the true values of all person and item parameters used to generate the data, a clear baseline was established for judging the parameter estimates produced by the model that are used to make inferences about individual growth.

It will be helpful here to clarify the meanings of some terms that will be used frequently in describing the analysis and results. *Factors* are the experimental variables being directly manipulated in the generation of data for the study: form of growth trajectory, sample size, test length, etc. A *level* refers to a particular manifestation of a specific factor, such as a linear or nonlinear form of the growth trajectory. A *condition* refers to a unique combination of levels of the factors, such as the linear trajectory with rate variance 0.20, sample size 500, eight item test length, and a full100 item selection design.

A total of five factors were manipulated in the study. However, a full factorial design incorporating all combinations of the levels of all five factors was not feasible.

Some levels of some of the factors were particularly intensive to compute, and combining two or more such levels in the same condition led to a particularly intensive situation that could not be accommodated in this exploratory study. Thus, the results will be reported in terms of three studies that isolated and investigated different combinations of the five factors. Each of these five factors and their levels are summarized in Table 2 and then each factor is explained in more detail below.

Table 2: The Five Factors and Their Levels in the Monte Carlo Studies

Form of trajectory	Variance of δ	Sample size	Number of items	Item selection
Linear (θ, δ)	τ_{11}=0.20	n=500	I=8	Full range & 100% constrained
Piecewise linear (θ, δ_1, δ_2)	τ_{11}=0.50	n=1000	I=16	Full range & 50% constrained
		n=2000		Targeted & 50% constrained
				Adapted & 50% constrained

Form of the Growth Trajectory

The first three factors to be discussed relate directly to persons in the study. The first factor is the form of the growth trajectory: a linear model with two random

effects defining the growth trajectory (as described when the integrated model was introduced in Chapter 1) or a piecewise linear model with three random effects defining the growth trajectory. The additional random effect in the piecewise linear model comes from the addition of a second rate parameter δ_2. In the piecewise linear model the first rate parameter δ_1 represents the linear growth rate from time points one through three and the second rate parameter δ_2 represents the linear growth rate from time points three through five.

Although somewhat limited, the choice of the linear and piecewise linear forms for the growth trajectory reflects the current state of estimation development in the nonlinear mixed model framework, which will currently accommodate only intrinsically linear trajectories while allowing for random effects in the model parameters. Nonlinear mixed models with intrinsically nonlinear trajectories do not have closed-form solutions for the parameters, and this currently presents an estimation challenge. A polynomial trajectory, such as a quadratic, could also have been chosen as a trajectory form for this study. However, the piecewise linear trajectory was chosen over the quadratic form because this parameterization is easier to interpret.

The generating distribution for all person parameters (random effects) was a normal distribution. An underlying normal distribution is a common assumption in parametric IRT models. The mean and variance of the θ parameter distribution were fixed to 0 and 1, respectively, as described when the model was introduced in Chapter 1. These are common choices for fixing parameters to identify IRT models. The mean of the generating δ parameter distribution was 0.50 in the linear growth

condition. This was also the value for δ_1 in the piecewise linear model. The mean of the generating δ_2 distribution was 0.25 in the piecewise linear growth condition. In considering a context of annual achievement testing, a mean annual growth rate of 0.5 of a standard deviation reflects modest growth. The lower value for the generating mean for the second piece in the piecewise linear condition reflects a situation in which growth in a domain slows after an initial period of rapid growth. Further, these values straddle the value of 0.385, which was used by Embretson (1991) in simulating a pretest/posttest design for a similar growth model in the Rasch model family.

Variance of the Growth Rate

The variance of the δ distribution(s) was controlled as a second factor. The first level of the variance for all δ parameter distributions was 0.20 (low variance condition). The second level was 0.50 (high variance condition). Although there is little guidance regarding the values of growth parameter variances in the IRT literature, Muthén and Muthén (2002) note that it is common in applied growth models in the Structural Equation Modeling framework to observe a five to one ratio between the variance of the intercept and the variance of the slope. In this study the variance of the intercept was fixed to one, so the generating variance of 0.20 for the underlying slope reflects a reasonable amount of variance according to the five-to-one ratio. The generating variance of 0.50 for the underlying slope reflects an unusually high level of variability.

In practice fitting growth models with real data, each of three scenarios can occur. The covariance between the initial status and the rate of growth can be

positive, negative, or zero. For linear or piecewise linear trajectories which of the three relations is observed in the data depends on the time frame in which measurements are taken for the purpose of fitting the growth model. Unless there is zero variability in the rate of growth in the population of interest, extrapolated linear growth trajectories will cross at some point. If study measurements are taken before the time period in which the trajectories cross, then a negative covariance will be observed. If study measurements are taken in the time period in which trajectories cross, then a (nearly) zero covariance will be observed. If study measurements are taken after the time period in which the trajectories cross, then a positive covariance will be observed. In this study a generating value of zero was chosen out of convenience to help control the effects of censoring on the modeling of growth.

The amount of shared variance (r^2) between θ and δ was held constant at zero percent. That is, the intercept and slope did not covary. Thus, the ability level at which individuals start at the first time point had no systematic relationship with the rate at which they grow according to the mechanism used to generate the data. This also applied to δ_1 in the piecewise model. In the piecewise model the amount of shared variance between the δ_1 and δ_2 was held constant at 0 percent. In the piecewise model θ and δ_2 were also statistically independent according to the generating mechanism.

Sample Size

The third factor was the sample size to be generated. Three levels were used: 500, 1000, and 2000. Roberts and Ma (2006) encountered less than desirable parameter recovery for 500 cases despite using an MCMC estimation approach in a

fully Bayesian model with their multivariate extension of the Generalized Partial Credit Model. The value of 1000 cases matches the value used by Embretson (1991) in her Monte Carlo study with a similar growth model in the Rasch framework. Embretson found adequate parameter recovery in her study but did not explore parameter estimation with smaller samples and did not produce sample size recommendations for her growth model. It is possible that a smaller sample size would also produce adequate parameter estimates for the Rasch growth model. The model under investigation in the present study was more sophisticated, as it is an outgrowth of the two-parameter logistic model. It was possible that 1000 cases would be adequate for this model, but it was also necessary to test the smaller sample size of 500 in order to draw a conclusion about the adequacy of smaller sample sizes for parameter recovery. The sample size of 2000 was included to verify parameter recovery with larger sample sizes in the event that 1000 cases was not sufficient for adequate parameter recovery.

Test Length

The last two factors focused on the items in the study. The first of these two factors was the number of item responses observed at each time point, set at either eight or 16. The value of 16 items reflects a fairly short subject matter test. The value of eight items is substantially less and may be closer to representing a scenario of embedded assessment where only a few items are embedded in the student's regular activities at any given assessment point.

In each condition three values for the item discrimination, 0.85, 1.10, and 1.35, were used in the ratio 3:4:1 items, respectively. These values were loosely based

on the academic achievement application that was intended for the proposed model. Although a direct comparison could not be made due to differences in the calibration, the chosen values are modest when compared to previously estimated item discrimination values for items in the Early Childhood Longitudinal Study – Kindergarten Cohort, which was used for the real data illustration of this model. These values also appear to be reasonable because they are situated around the approximate value of 1.0, which is considered a customary value for the item discrimination parameter in item response models.

The item difficulty parameters were distributed across a range that was defined by the values of the mean(s) and standard deviation(s) of the person parameter distribution(s), which will be described in more detail shortly. The item location parameters were evenly distributed within the defined range. This reflected an achievement test that was constructed to provide a reasonably uniform amount of information across a desired range. The definition of this desired range was manipulated in the final factor under investigation.

Item Selection Design

The final factor in the design systematically defined how items are selected to build the test at each time point. Three aspects of item selection were manipulated to create the distinct levels of this factor. The first was the proportion of items whose parameters were constrained to be equal across time points (i.e., the proportion of common items across tests at different time points). The second had to do with whether the difficulty of the items on the test was adjusted over time to account for the general anticipated growth in ability within the *population* of examinees. The

third was whether the difficulty of the items on the test was adjusted over time to account for the ability level of each *individual* examinee at each time of testing. In this scenario, the number correct score on a routing test consisting of 50 percent of the test items was used to determine the difficulty level of the remaining 50 percent of the test items for the individual examinee. These three variables were manipulated together since shifting difficulty values of the items affects the proportion of overlapping items that can be accommodated across tests at different time points. Combinations were selected that imitated real-life designs with vertical scales and facilitated meaningful comparisons between manipulated conditions.

The difficulty of the items comprising the test at each time point was controlled by defining a range for the item location parameters. The distribution of item difficulties with respect to this range is discussed in more detail below. Although variability in the population is changing over time as described earlier, the standard deviation of the initial score distribution was used to set the target ranges for the item difficulties for simplicity and consistency in the design. Conditions were further distinguished by specifying which items would have their item parameters constrained across tests at different time points in the model-fitting process to simulate a common item design. The four levels of the item selection factor are briefly outlined in Table 3.

Table 3: Brief Descriptions of Four Item Selection Schemes

Level	Common item constraints	Range of generating values of item difficulty parameters	Adapted to the individual
Full100	All items constrained across all time points	From 2.5 units below initial population mean to 2.5 units above final population mean	No
Full50	50 percent of items constrained across pairs of tests at adjacent time points	From 2.5 units below initial population mean to 2.5 units above final population mean	No
Target50	50 percent of items constrained across pairs of tests at adjacent time points	From 2.5 units below to 2.5 points above the population mean at the given time point	No
Adapt50	50 percent of items constrained across pairs of tests at adjacent time points	From 2.5 units below to 2.5 points above the population mean at the given time point for the routing test. High, medium, and low difficulty tests covered the range from 2.5 units below to 2.5 points above the population mean at the given time point	Yes

In the "full100" level the same set of items were administered at each time point. The generating values of the item location parameters in this condition were within the range from 2.0 points (2.0 standard deviations of initial scores) below the

mean θ level to 2.0 points above the mean θ level plus four times the mean δ level (linear condition) or 2.0 points above the mean θ level plus two times the mean of the δ_1 level plus two times the mean of the δ_2 level (piecewise condition). The estimated values of the item parameters were then constrained to be equal across all time points.

In the "full50" level different tests were administered at each time point with a set of 50 percent common items constrained across adjacent test forms. Item difficulty parameters for each test continued to cover the full range of difficulty for the entire study. Thus, the generating values of the item location parameters in this condition were the same values as in the previous level. The estimated values of the item parameters for the common items would be constrained to be equal by the model. Common items were selected in such a way as to avoid having any item constrained across three adjacent time points.

In the "target50" level different tests were administered at each time point, but the values of the item location parameters were tailored for the general ability level of the population at each time point with 50 percent of the items common to tests at adjacent time points. Thus, the range of item locations at the first time point was 2.0 points below the mean θ level to 2.0 points above the mean θ level. The range of item locations at the second time point was 2.0 points below to 2.0 points above the mean of θ plus the mean of δ. The range for the third time point was 2.0 points above and below the mean of θ plus two times the mean of δ, and so on as appropriate depending on whether the model is linear or piecewise linear.

The adapt50 level was designed to mimic the adaptive procedure that was used in the Early Childhood Longitudinal Study – Kindergarten Cohort. For the design used

in the simulation, 50 percent of the items were administered to all examinees as a routing test at each time point. The routing test was designed much as in the target50 level, except with only half as many items, including the 50 percent overlap (25 percent of the full test length) across test forms at adjacent time points. The remaining 50 percent of the items were selected from either a high, medium, or low difficulty test form individually for each examinee based on the examinee's number correct score on the routing test.

The specific values marking the boundaries of the ranges and the distribution of item location parameters within each range under each combination of conditions and at each time point will be given in more detail next.

Tables 4 and 5 display the generating item parameter values for the linear and piecewise models, respectively, in the conditions with eight items. The column labeled "Disc" contains the item discrimination values. The column labeled "Loc" contains the item location values. These are the generating values for the level in which the same items are administered at each time point (full100) as well as for the level in which different items are administered but the test is not targeted at each time point (full50). In both of these levels the generated item parameters for items at each time point are expected to nearly cover the full range of latent abilities anticipated to be encountered over the duration of the panel study. In the full50 design the unshaded item parameter values in Table 4 were used for the common items across the tests at time one and time two. The shaded item parameter values were used for the common items across the tests at times two and three. The unshaded item parameter values were used for the common items across the tests at times three and

four, and so on. However, the model did not constrain the parameter estimates for the set of common items for the test forms at times one and two to be equal to the parameter estimates for the set of common items for the test forms at times three and four. To control the feasibility of the design of the Monte Carlo studies, the full50 level was not tested in combination with the nonlinear trajectory.

Table 4: Generating Item Parameter Values for Full Range Item Selection Designs with Eight Items and Linear Model

Item	Disc	Loc
Upper bound		*-2.000*
1	1.10	-1.625
2	0.85	-0.875
3	1.35	-0.125
4	1.10	0.625
5	1.10	1.375
6	0.85	2.125
7	1.10	2.875
8	0.85	3.625
Lower bound		*4.000*

Table 5: Generating Item Parameter Values for Eight Item Condition with Piecewise Model

Item	Disc	Loc
Upper bound		*-2.000*
1	1.10	-1.656
2	0.85	-0.969
3	1.35	-0.281
4	1.10	0.406
5	1.10	1.094
6	0.85	1.781
7	1.10	2.469
8	0.85	3.156
Lower bound		*3.500*

For the level in which test difficulty is targeted at each time point (target50), Table 6 shows five columns – one containing the generating item parameter values for each time point. In this level the tests at each time point contain an overlapping set of anchor items common to two adjacent tests. The item parameters for items that were common with items in the test preceding the given time point are shaded.

Table 6: Generating Item Parameter Values for Targeted Item Selection Design with Eight Items and Linear Model

	Time 1		Time 2		Time 3		Time 4		Time 5	
Item	Disc	Loc	Disc	Loc	Disc	Loc	Disc	Loc	Disc	Loc
Upper bound		*-2.00*		*-1.50*		*-1.00*		*-0.50*		*0.00*
1	0.85	-1.75	0.85	-1.25	1.10	-0.75	1.10	-0.25	0.85	0.25
2	0.85	-1.25	1.10	-0.75	1.10	-0.25	0.85	0.25	1.35	0.75
3	1.10	-0.75	1.10	-0.25	0.85	0.25	1.35	0.75	1.10	1.25
4	1.10	-0.25	0.85	0.25	1.35	0.75	1.10	1.25	1.10	1.75
5	0.85	0.25	1.35	0.75	1.10	1.25	1.10	1.75	0.85	2.25
6	1.35	0.75	1.10	1.25	1.10	1.75	0.85	2.25	0.85	2.75
7	1.10	1.25	1.10	1.75	0.85	2.25	0.85	2.75	1.10	3.25
8	1.10	1.75	0.85	2.25	0.85	2.75	1.10	3.25	1.10	3.75
Lower bound		*2.00*		*2.50*		*3.00*		*3.50*		*4.00*

Table 7 displays the generating item parameter values for the linear model with a test length of eight items for the level in which item selection was adapted to the individual examinee at each time point (adapt50). However, in this level there are four sets of items at each time point. Routing items were administered to all examinees at each time point. Low, medium, and high items were administered to examinees who answered 0-1, 2, and 3-4 items correctly, respectively, on the routing test form. As in Table 6 the item parameters for items that were common with items in the test preceding the given time point are shaded.

Table 7: Generating Item Parameter Values for Adapted Item Selection Design with Eight Items and Linear Model

	Time 1		Time 2		Time 3		Time 4		Time 5	
Item	Disc	Loc	Disc	Loc	Disc	Loc	Disc	Loc	Disc	Loc
Upper bound		*-2.000*		*-1.500*		*-1.000*		*-0.500*		*0.000*
Routing items										
1	1.10	-0.840	0.85	-0.280	0.85	0.280	1.10	0.840	1.10	1.400
2	0.85	-0.280	0.85	0.280	1.10	0.840	1.10	1.400	0.85	1.960
3	0.85	0.280	1.10	0.840	1.10	1.400	0.85	1.960	0.85	2.520
4	1.10	0.840	1.10	1.400	0.85	1.960	0.85	2.520	1.10	3.080
Low items										
1	1.10	-1.717	0.85	-1.217	1.35	-0.717	1.10	-0.217	1.10	0.283
2	0.85	-1.217	1.35	-0.717	1.10	-0.217	1.10	0.283	0.85	0.783
3	1.35	-0.717	1.10	-0.217	1.10	0.2826	0.85	0.783	1.35	1.283
4	1.10	-0.217	1.10	0.282	0.85	0.7826	1.35	1.283	1.10	1.783
Medium items										
1	1.10	-0.750	0.85	-0.250	1.35	0.250	1.10	0.750	1.10	1.250
2	0.85	-0.250	1.35	0.250	1.10	0.750	1.10	1.250	0.85	1.750
3	1.35	0.250	1.10	0.750	1.10	1.250	0.85	1.750	1.35	2.250
4	1.10	0.750	1.10	1.250	0.85	1.750	1.35	2.250	1.10	2.750
High items										
1	1.10	0.217	0.85	0.717	1.35	1.217	1.10	1.717	1.10	2.217
2	0.85	0.717	1.35	1.217	1.10	1.717	1.10	2.217	0.85	2.717
3	1.35	1.217	1.10	1.717	1.10	2.217	0.85	2.717	1.35	3.217
4	1.10	1.717	1.10	2.217	0.85	2.717	1.35	3.217	1.10	3.717
Lower bound		*2.000*		*2.500*		*3.000*		*3.500*		*4.000*

Table 8 displays the generating item parameter values for the linear model with a test length of 16 items. For feasibility the test length of 16 items was only estimated for the level in which the same items are administered at each time point (full100). Thus, all items are common across all time points.

Table 8: Generating Item Parameter Values for 16 Item Condition with Linear Model

Item	Disc	Loc
Upper bound		*-2.000*
1	1.10	-1.812
2	0.85	-1.438
3	1.35	-1.062
4	1.10	-0.688
5	1.10	-0.312
6	0.85	0.062
7	1.10	0.438
8	0.85	0.812
9	1.10	1.188
10	0.85	1.562
11	1.10	1.938
12	0.85	2.312
13	1.10	2.688
14	0.85	3.062
15	1.35	3.438
16	1.10	3.812
Lower bound		*4.000*

The four levels of this item selection design factor were carefully chosen to investigate three specific comparisons. The distinction between the first two levels,

full100 and full50, was the proportion of common items across time points, either 100 percent or 50 percent. The distinction between the full50 and target50 levels of this factor was whether the difficulty of the items on the test was adjusted over time to account for the general anticipated growth in the population of examinees. The distinction between the target50 and adapt50 levels was whether the test was targeted for the population only or whether an additional attempt was made to target the exam for each individual examinee at each time point.

Design

This was essentially an exploratory study with a new model. Thus, some levels of some factors were specifically chosen to potentially strain the model to see where it would break down. As the combination of extreme levels of multiple conditions could possibly lead to low convergence rates in these cells, all levels of the five factors were not fully crossed in the research design. Instead, conditions were designed that would be examined in three separate but related analyses such that the levels of some factors would be held constant while others were varied.

Study 1

In Study 1 all conditions included eight items with a linear growth trajectory and a moderate (0.20) rate variance. Three levels of sample size (500, 1000, and 2000 examinees) were crossed with four levels of item selection design (full100, full50, target50, and adapt50) for a total of 12 cells in this study.

Study 2

In Study 2 all conditions included a linear growth trajectory with a sample size of 500 examinees, a moderate (0.20) rate variance, and an item selection design in which all examinees were administered all items at all time points (full100). Varied in this study was the test length, either eight or 16 items, for a total of two cells.

Study 3

In Study 3 all conditions included eight items where all items were administered at all five time points (full100 item selection design). Two levels of form of trajectory (linear or nonlinear), two levels of sample size (500 or 1000 examinees), and two levels of variance of the growth rate (0.20 or 0.50) were crossed for a total of eight cells in this study.

Methods of Analysis

The simulated data sets were generated using SAS commands. One hundred replications were carried out for each condition. The proposed model was estimated using SAS Proc NLMIXED (SAS Institute, 1999). The estimation algorithm followed a marginal maximum likelihood approach, a common estimation approach for item response theory models (Tuerlinckx et al., 2004). The Expectation-Maximization (EM) algorithm was used to carry out the marginal maximum likelihood estimation of the model parameters.

To address the two research questions described earlier, two dependent variables were computed from the recovered parameter estimates and their generating

values: bias and error of the estimate. Bias was computed as the difference between the parameter estimate and the generating value of the parameter. Error of the estimate was computed as the absolute value of the difference between the parameter estimate and the mean of the estimated values across replications.

For each of the two dependent variables, three general linear models were fit, one for each of the three types of model parameters: item location, item discrimination, and random effect (growth) parameters. The effects of each of the manipulated factors in each study and their interactions were included in the models. In addition the models for the item parameters included as a continuous covariate the generating value of the item location since a potential trend by item location was suspected based on plots of the data. For random effects (growth) parameters the general linear model included the effect of the specific random effect parameter (nested within the type of trajectory in Study 3). Type III sums of squares were used in all analyses to account for potential differences in cell size in the case of cells where some replications did not converge. The cell means that were compared in the general linear models are provided in tables in the Appendix.

In addition, the standardized effect size index partial eta-squared (η^2) was used to clarify the practical meaning of statistically significant effects. Partial eta-squared is the proportion of effect plus error variance that can be attributed to the effect for which the effect size is calculated. Partial eta-squared was chosen due to different error sums of squares for the different effects in the model due to the clustering by item. In addition, partial eta-squared has the advantage that its value for any given effect is independent of the other effects in the model.

The reporting of the results for each of the studies below begins with a report of the convergence rates for each of the conditions in the study. As the Monte Carlo investigation was conducted, convergence behavior for the conditions was closely observed. Any condition that failed to converge in the first three replications had its remaining 97 replications suspended. This decision was made in light of the computational intensity of many of the conditions and the time commitment of running each replication to the maximum number of iterations without obtaining valid parameter estimates for the investigation of parameter recovery.

Convergence behavior was also evaluated at the individual replication level. For the individual replications that failed to converge, this data was excluded from subsequent analysis. Generally, replications that failed to converge had multiple item parameter estimates at boundary values. Thus, it would be inappropriate to include these boundary values in subsequent analysis because they would inappropriately reduce the variability associated with the estimates.

Recall that the simulation is divided into three studies. In Study 1 the sample size and item selection design were analyzed. In Study 2 the test length was studied. In Study 3 the variance of the growth rate, the form of the trajectory, and the sample size were investigated. Results are presented here separately for each study.

Study 1

Results of Study 1

In Study 1 all conditions included eight items with a linear growth trajectory and a moderate (0.20) rate variance. The number of completed replications in each cell and their convergence behavior are reported in Table 9. All conditions involving

the full100 item selection design converged. The conditions involving the full50 and adapt50 item selection designs were suspended after the first three replications in each condition failed to converge. The convergence rate for the target50 condition was affected by the sample size. A little more than half of the replications with 500 examinees converged, but the convergence rate improved with 1000 examinees and continued to improve with 2000 examinees. Results from analysis of the converged replications are provided in this section.

Table 9: Number of Replications and Convergence Behavior by Condition for Study 1

Factors		Number of replications		
Sample size	Item selection	Converged	Not converged	Total
500	adapt50	0	3	3
1000	adapt50	0	3	3
2000	adapt50	0	3	3
500	full50	0	3	3
1000	full50	0	3	3
2000	full50	0	3	3
500	full100	100	0	100
500	target50	52	48	100
1000	full100	100	0	100
1000	target50	81	19	100
2000	full100	100	0	100
2000	target50	92	8	100

Item Location Parameter Recovery for Study 1

The results of the analysis of bias in the item location parameters are reported in Table 10. The generating value for the item location parameter was used as an

item level covariate. This continuous covariate was statistically significant ($F(1, 7268) = 2422.64$, $p < 0.0001$) with the trend reflecting a gradual transition from negative bias for easy items to positive bias for difficult items. The partial eta-squared value of 0.25 suggests that there is enough variability attributable to this effect for it to be practically meaningful. The effect of the item selection design was statistically significant ($F(1, 1898.6) = 1146.00$, $p < 0.0001$), reflecting a the difference between a substantial positive average bias in the target50 item selection design and a small positive average bias in the full100 item selection design. The partial eta-squared value of 0.38 suggests that a great deal of variability in the bias of the item location parameter estimates is attributable to this effect. None of the other effects were significant.

Table 10: Bias of Item Location Parameter Estimates for Study 1

Source	df	F	η^2	p
Sample size	2	0.8500	0.0009	0.4281
Error	1965.3	(1.5198)		
Item selection	1	1146.0000	0.3764	<.0001
Error	1898.6	(1.4971)		
Sample size x Item selection	2	2.4900	0.0026	0.0835
Error	1934.7	(1.5094)		
Location	1	2422.6400	0.2500	<.0001
Error	7268	(6.2359)		

Note: Values enclosed in parentheses represent mean square errors.

The results of the analysis of error of estimate in the item location parameters are reported in Table 11. There was a fourth degree polynomial effect of item location with all four terms statistically significant (in order of decreasing degree $F(1, 7265) = 383.55$, $p < 0.0001$; $F(1, 7265) = 176.03$, $p < 0.0001$; $F(1, 7265) = 574.38$, $p < 0.0001$; $F(1, 7265) = 251.16$, $p < 0.0001$) with the trend reflecting greater error of estimate for items with extreme locations, particularly for difficult items. The linear and the quadratic terms are the most meaningful, however, as suggested by partial eta-squared values of 0.40 and 0.07, respectively. Clearly, the linear influence of generating item location is dominant over the contribution of the remaining polynomial terms in the trend. The interaction of item selection design x sample size was statistically significant ($F(2, 605.86) = 22.06$, $p < 0.0001$). While the target50 item selection design had greater average error of estimate than the full100 item selection design and this error of estimate decreased with increasing sample size, the interaction appears to be due to a more dramatic difference in the error of the estimate between the target50 and full100 designs as the sample size increases. That is, the error of the estimate decreases faster with increasing sample size in the full100 design than in the target50 design. The value of 0.07 for partial eta-squared suggests that this interaction has practical meaning, but is far less influential than the main effects of sample size and item selection (eta-squared values 0.22 and 0.49, respectively) in accounting for differences in the amount of error in the estimates. The main effects of the item selection design ($F(1, 96) = 144.41$, $p < 0.0001$) and the sample size ($F(2, 96) = 32.07$, $p < 0.0001$) were also statistically significant.

Table 11: Error of Estimate of Item Location Parameter Estimates for Study 1

Source	df	F	η^2	p
Sample size	2	84.7800	0.2182	<.0001
Error	607.54	(0.3803)		
Item selection	1	605.7500	0.4893	<.0001
Error	632.27	(0.3760)		
Sample size x Item selection	2	22.0600	0.0679	<.0001
Error	605.86	(0.3806)		
Location	1	251.1600	0.3957	<.0001
Location x Location	1	574.3800	0.0733	<.0001
Location x Location x Location	1	176.0300	0.0237	<.0001
Location x Location x Location x Location	1	383.5500	0.0501	<.0001
Error	7265	(0.2402)		

Note: Values enclosed in parentheses represent mean square errors.

Item Discrimination Parameter Recovery for Study 1

The results of the analysis of bias in the item discrimination parameters are reported in Table 12. The covariate item location was statistically significant ($F(1, 7268) = 341.74$, $p < 0.0001$) with a slight trend reflecting a gradual transition from negative bias of lesser magnitude for easy items to negative bias of greater magnitude for difficult items. The partial eta-squared value of 0.04 reinforces the fact that this trend is not nearly as dramatic as the generating item location trend for the bias in the item location parameter estimates described previously. A statistically significant

interaction between the item selection design and the sample size was found ($F(2, 687.74) = 12.1500$, $p < 0.0001$). However, the partial eta-squared value of 0.03 suggests that this is not a very substantial effect. The average magnitude of the bias decreased with increasing sample size in the full100 item selection design but increased with increasing sample size for the target50 item selection design. This result should be interpreted with extreme caution due to the differences in convergence rates not only between the two item selection designs but also across sample sizes within the target50 item selection design. The main effect of the item selection design ($F(1, 683.83) = 8033.5200$, $p < 0.0001$) was also statistically significant, and with a very large eta-squared value of 0.92 should be considered a much more important influence on the bias in the item discrimination parameter estimates.

Table 12: Bias of Item Discrimination Parameter Estimates for Study 1

Source	df	F	η^2	p
Sample size	2	2.9800	0.0086	0.0514
Error	691.05	(0.0633)		
Item selection	1	8033.5200	0.9216	<.0001
Error	683.83	(0.0633)		
Sample size x Item selection	2	12.1500	0.0341	<.0001
Error	687.74	(0.0633)		
Location	1	341.7400	0.0449	<.0001
Error	7268	(0.0699)		

Note: Values enclosed in parentheses represent mean square errors.

The results of the analysis of error of estimate in the item discrimination parameters are reported in Table 13. There was a fourth degree polynomial effect of item location with three of the four terms statistically significant (in order of decreasing degree $F(1, 7265) = 301.84$, $p < 0.0001$; $F(1, 7265) = 28.26$, $p < 0.0001$; $F(1, 7265) = 56.70$, $p < 0.0001$; $F(1, 7265) = 44.87$, $p < 0.0001$) with the trend reflecting greater error of estimate for items with extreme locations, particularly for easy items. Again, the linear influence in this trend predominates with an eta-squared value of 0.87. A statistically significant effect of sample size was found ($F(2, 559.46) = 110.44$, $p < 0.0001$). This appears to be due to lower average error of estimate with increasing sample size, and the partial eta-squared value of 0.28 suggests that a great deal of variability in the bias of the item discrimination parameter estimates can be attributed to this effect.

Table 13: Error of Estimate of Item Discrimination Parameter Estimates for Study 1

Source	df	F	η^2	p
Sample size	2	110.4400	0.2831	<.0001
Error	559.46	(0.0195)		
Item selection	1	7.1200	0.0123	0.0078
Error	570.8	(0.0191)		
Sample size x Item selection	2	0.7600	0.0027	0.4676
Error	558.68	(0.0196)		
Location	1	301.8400	0.8706	<.0001
Location x Location	1	28.2600	0.0039	<.0001
Location x Location x Location	1	56.7000	0.0077	<.0001
Location x Location x Location x Location	1	44.8700	0.0061	<.0001
Error	7265	(0.2758)		

Note: Values enclosed in parentheses represent mean square errors.

Random Effect (Growth) Parameter Recovery for Study 1

The results of the analysis of bias in the random effect (growth) parameters are reported in Table 14. A statistically significant interaction was found between the sample size and the item selection design ($F(2, 1554) = 22.54$, $p < 0.0001$). However, this interaction accounted for only a small amount of variance as suggested by the partial eta-squared value of 0.03. The main effects of sample size ($F(2, 1554) = 16.91$, $p < 0.0001$) and item selection design ($F(1, 1554) = 6050.34$, $p < 0.0001$) were also statistically significant. The partial eta-squared value of 0.80 suggests that the

item selection design accounts for a very large proportion of the variability in the bias in the estimates of the random effects (growth) parameters.

The average bias also differed according to the specific random effect (growth) model parameter ($F(2, 1554) = 1195.35$, $p < 0.0001$), accounting for a large proportion of the variability in the bias as suggested by the partial eta-squared value of 0.61. All of the significant effects reported in the previous paragraph also interacted with the effect of the individual parameter. The interaction among sample size, item selection design, and parameter was statistically significant ($F(4, 1554) = 6.70$, $p < 0.0001$). However, the partial eta-squared value of 0.02 suggests that this effect is not very meaningful. For the full100 item selection design the average bias was generally quite small but was especially small for the covariance between the initial ability level and the rate of growth. For this design average bias decreased as sample size increased, especially between 500 and 1000 examinees. For the target50 item selection design the average bias increased as sample size increased for the variance of the growth rate. However, for the covariance and the mean of the growth rate, the bias was largest in the 1000 examinee condition and smaller in the conditions with larger and smaller sample sizes. The results for conditions with a target50 item selection design should be interpreted with extreme caution since the convergence rate was a function of sample size in the target50 item selection design.

Table 14: Bias of Random Effect (Growth) Parameter Estimates for Study 1

Source	df	F	η^2	p
Sample size	2	16.9100	0.0213	<.0001
Item selection	1	6050.3400	0.7956	<.0001
Sample size x Item selection	2	22.5400	0.0282	<.0001
Parameter	2	1195.3500	0.6061	<.0001
Sample size x Parameter	4	6.2500	0.0158	<.0001
Item selection x Parameter	2	1208.9700	0.6088	<.0001
Sample size x Item selection x Parameter	4	6.7000	0.0170	<.0001
Error	1554	(0.0571)		
Corrected total	1571			

Note: Values enclosed in parentheses represent mean square errors.

The results of the analysis of error of estimate in the random effect (growth) parameters are reported in Table 15. A statistically significant interaction was found between the sample size and the item selection design ($F(2, 1554) = 40.17$, $p < 0.0001$). This interaction accounted for a modest amount of the variability in the average error of the estimate as suggested by the partial eta-squared value of 0.05. In the target50 item selection design average error of estimate decreased as sample size increased. In the full100 item selection design the average error of estimate decreased as sample size increased from 500 examinees to 1000 examinees but did not increase substantially from 1000 examinees to 2000 examinees. The main effects of sample size ($F(2, 1554) = 13.40$, $p < 0.0001$) and item selection design ($F(1, 1554) = 112.45$, $p < 0.0001$) were also statistically significant and accounted for larger

amounts of variability as suggested by their partial eta-squared values of 0.08 and 0.29, respectively.

The average error of estimate also differed according to the specific random effect (growth) model parameter ($F(2, 1554) = 229.34$, $p < 0.0001$), which also accounted for a substantial amount of variability as suggested by the eta-squared value of 0.23. The statistically significant effects reported in the previous paragraph also interacted with the effect of the individual parameter. The sample size x parameter interaction was statistically significant ($F(4, 1554) = 15.31$, $p < 0.0001$) but accounted for a small amount of variability in the error of the estimate as suggested by the eta-squared value of 0.04. The item selection design x parameter interaction was also statistically significant ($F(4, 1554) = 2.66$, $p < 0.0001$), and the partial eta-squared value of 0.22 suggests that this interaction accounts for substantial variability in the error of estimate of the random effects (growth) model parameters. Although the error of estimate was generally larger for the target50 item selection design than for the full100 item selection design, the average error of estimate for the variance of the growth rate was unusually large. The sample size x item selection design x parameter interaction was statistically significant ($F(4, 1554) = 13.74$, $p < 0.0001$), but the partial eta-squared value of 0.03 suggests that this interaction is not practically meaningful.

Table 15: Error of Estimate of Random Effect (Growth) Parameter Estimates for Study 1

Source	df	F	η^2	P
Sample size	2	72.1700	0.0850	<.0001
Item selection	1	622.3600	0.2860	<.0001
Sample size x Item selection	2	40.1700	0.0492	<.0001
Parameter	2	229.3400	0.2279	<.0001
Sample size x Parameter	4	15.3100	0.0379	<.0001
Item selection x Parameter	2	220.5500	0.2211	<.0001
Sample size x Item selection x Parameter	4	13.7400	0.0342	<.0001
Error	1554	0.0232		
Corrected total	1571			

Note: Values enclosed in parentheses represent mean square errors.

Commentary on Study 1

Study 1 attempted to examine the interaction of various item selection designs with differing numbers of examinees on parameter recovery in the item response growth model. From the results it appears that model estimation is extremely sensitive to the item selection design. The estimation readily converged in the case of a very simple item selection design in which the same items are administered at all time points. Convergence rates decreased as the arrangement of selected items and the sample size resulted in inadequate information for estimation of model parameters, particularly item parameters. An arrangement in which items at each time point were targeted to the ability distribution of the population and 50 percent of

the items were common across adjacent time points exhibited fairly high convergence rates when the sample size was 1000 examinees or more. The poor convergence rate for this design with 500 examinees suggests that convergence is not strictly a function of the item selection design but is rather a matter of the amount of information available to estimate model parameters. There were two item selection designs in which item information was so poor that the estimation failed to converge with as many as 2000 examinees. A design in which 50 percent of the items were common across time points but the difficulty level of the items was not targeted to the ability level of the population at each time point failed to converge. This appears to be due to insufficient information to estimate parameters for items that were too easy or too difficult for the population at the specific time point(s) in which they were administered. An adaptive design attempted not only to target the difficulty level to the population but also to the individual student. However, because some items were administered to only a portion of the examinees at any given time point, this model too failed to converge with as many as 2000 examinees overall.

Due to the issues with convergence behavior, the parameter recovery results of Study 1 can only be interpreted with great caution. Because convergence rates were very different for the two item selection designs, perhaps it would be most enlightening to discuss each item selection design separately. Results differed noticeably for the two item selection designs, particularly with regard to the effects on bias in the parameter estimates.

In the simpler of the two item selection designs, all items were administered at all time points, and convergence rates were quite high. Thus, the results pertaining to

this item selection design can be interpreted with confidence. The average bias of 0.06 for the item location parameter estimates was small enough to perhaps be acceptable for some applications. The average bias decreased with increasing sample size, particularly as sample size increased from 500 examinees to 1000 examinees. Error of estimate for the item location parameters decreased notably with increasing sample size. For item discrimination parameters the average magnitude of bias decreased with increasing sample size, and there was lower average error of estimate with increasing sample size. For the random effects (growth) parameters the average bias was generally quite small but was especially small for the covariance between the initial ability level and the rate of growth. For this design average bias decreased as sample size increased, especially between 500 and 1000 examinees. The error of estimate likewise was quite small, especially for the covariance parameter, and decreased steadily with increasing sample size.

In the more complex targeted item selection design items increased in difficulty over time with 50 percent of the items overlapping across adjacent time points. The convergence rate was poor for the condition with 500 examinees but increased with increasing sample size. There was substantial positive bias in the item location parameters on the magnitude of 1.10 on average, which is unacceptable in any application. Error of estimate decreased somewhat with increasing sample size, but not as much as it decreased for the simpler item selection design. Unlike in the simpler item selection design, the average magnitude of the bias for the item discrimination parameters *increased* with increasing sample size for the targeted item selection design. This apparent increase in the bias with increasing sample size may

be attributed to a selection effect from the substantially lower convergence rate with lesser sample size. Indeed, had convergence been achieved in the replications that failed to converge, the bias for these particular replications may have been large enough for the average bias to decrease with increasing sample size. As in the simpler item selection design, there was lower average error of estimate with increasing sample size for the item discrimination parameters in the targeted item selection design.

For the targeted item selection design the average bias for the random effects (growth) parameters increased as sample size increased for the variance of the growth rate. However, for the covariance and the mean of the growth rate, the bias was largest in the 1000 examinee condition and smaller in the conditions with larger and smaller sample sizes. The effect of sample size in the covariance and mean of the growth rate for the targeted item selection design is likely a complex combination of the true pattern and the influence of lack of convergence. It is quite possible that bias would have decreased with sample size had more of the replications in the condition with 500 examinees had converged, as these replications may have exhibited the largest bias. The average error of estimate steadily decreased as sample size increased.

Study 2

Results of Study 2

In Study 2 all conditions included a linear growth trajectory with a sample size of 500 examinees, a moderate (0.20) rate variance, and an item selection design in which all examinees were administered all items at all time points. The number of

completed replications in each cell and their convergence behavior are reported in Table 16. All replications of the condition involving eight items converged. Nearly one-quarter of the replications involving 16 items did not converge. Results from analysis of the converged replications are provided in this section.

Table 16: Number of Replications and Convergence Behavior by Condition for Study 2

Factor	Number of replications		
Test length	Converged	Not converged	Total
8	100	0	100
16	76	24	100

Item Location Parameter Recovery for Study 2

The results of the analysis of bias in the item location parameters are reported in Table 17. The generating value for the item location parameter was used as an item level covariate. This continuous covariate was statistically significant ($F(1, 1847) = 1733.44$, $p < 0.0001$) with the trend reflecting a gradual transition from negative bias for easy items to positive bias for difficult items. This trend accounted for a substantial amount of the variability in the bias in the item location parameter estimates as supported by a partial eta-squared value of 0.48. The effect of test length on average bias of the item location parameters was statistically significant ($F(1, 176.47) = 16.78$, $p < 0.0001$), but accounted for a relatively smaller proportion of bias. The partial eta-squared value for this effect was 0.08.

Table 17: Bias of Item Location Parameter Estimates for Study 2

Source	df	F	η^2	p
Test length	1	16.7800	0.0868	<.0001
Error	176.47	(0.2114)		
Location	1	1733.4400	0.4841	<.0001
Error	1847	(0.0600)		

Note: Values enclosed in parentheses represent mean square errors.

The results of the analysis of error of estimate in the item location parameters are reported in Table 18. There was a third degree polynomial effect of item location with all three terms statistically significant (in order of decreasing degree $F(1, 1845) = 30.17$, $p < 0.0001$; $F(1, 1845) = 224.09$, $p < 0.0001$; $F(1, 1845) = 8.26$, $p = 0.0041$) with the trend reflecting greater error of estimate for items with extreme locations, particularly for difficult items. The quadratic term was important in driving this trend. The corresponding partial eta-squared value for the quadratic term was 0.10, the only substantial value for the three terms in the trend. The effect of test length on average error of estimate was statistically significant ($F(1, 175.09) = 5.27$, $p = 0.0228$), accounting for a rather small proportion of the variability as suggested by the partial eta-squared value of 0.03. The eight item level reflected somewhat larger error of estimate on average than the 16 item level.

Table 18: Error of Estimate of Item Location Parameter Estimates for Study 2

Source	df	F	η^2	p
Test length	1	5.2700	0.0292	0.0228
Error	175.09	(0.1480)		
Location	1	8.2600	0.0045	0.0041
Location x Location	1	224.0900	0.1083	<.0001
Location x Location x Location	1	30.1700	0.0161	<.0001
Error	1845	(0.1513)		

Note: Values enclosed in parentheses represent mean square errors.

Item Discrimination Parameter Recovery for Study 2

The results of the analysis of bias in the item discrimination parameters are reported in Table 19. The covariate item location was statistically significant ($F(1, 1847) = 12.13$, $p = 0.0005$), but did not account for any meaningful amount of variability in the bias as suggested by the partial eta-squared value of 0.01. The effect of test length on average bias of the item discrimination parameters was significant ($F(1, 174.53) = 24.60$, $p < 0.0001$). This accounted for a more substantial proportion of the variability in the bias as suggested by the partial eta-squared value of 0.12. This effect reflected a negative bias of larger magnitude for the 16 item level than for the eight item level.

Table 19: Bias of Item Discrimination Parameter Estimates for Study 2

Source	df	F	η^2	p
Test length	1	24.6000	0.1236	<.0001
Error	174.53	(0.0948)		
Location	1	12.1300	0.0065	0.0005
Error	1847	(0.0058)		

Note: Values enclosed in parentheses represent mean square errors.

The results of the analysis of error of estimate in the item discrimination parameters are reported in Table 20. There was a statistically significant quadratic polynomial effect of item location (in order of decreasing degree $F(1, 1846) = 34.68$, $p < 0.0001$; $F(1, 1846) = 24.16$, $p < 0.0001$). However, the partial eta-squared values of 0.02 for the quadratic term and 0.01 for the linear term suggest that this is not a meaningful trend. The effect of test length on average error of estimate was statistically significant ($F(1, 175.15) = 14.71$, $p = 0.0002$), reflecting slightly higher error of estimate for the item discrimination parameters in the eight item level than in the 16 item level. This accounted for a modest proportion of variability as suggested by the partial eta-squared value of 0.08.

Table 20: Error of Estimate of Item Discrimination Parameter Estimates for Study 2

Source	df	F	η^2	p
Test length	1	14.7100	0.0775	0.0002
Error	175.15	(0.0247)		
Location	1	24.1600	0.0129	<.0001
Location x Location	1	34.6800	0.0184	<.0001
Error	1846	(0.0033)		

Note: Values enclosed in parentheses represent mean square errors.

Random Effect (Growth) Parameter Recovery for Study 2

The results of the analysis of bias in the random effect (growth) parameters are reported in Table 21. The test length x parameter interaction was statistically significant ($F(1, 105) = 6.13$, $p = 0.0057$). However, this interaction accounted for a small proportion of variability in the bias as suggested by the partial eta-squared value of 0.02. The main effects of test length ($F(1, 105) = 6.13$, $p < 0.0001$) and parameter ($F(2, 105) = 7.27$, $p < 0.0001$) were also statistically significant. Both accounted for small to modest proportions of variability with partial eta-squared values of 0.04 and 0.06, respectively. Overall, the average bias for the mean and variance of the growth rate parameters was larger than for the covariance parameter, and this effect was more pronounced in the 16 item level than in the eight item level.

Table 21: Bias of Random Effect (Growth) Parameter Estimates for Study 2

Source	df	F	η^2	p
Test length	1	21.0300	0.0385	<.0001
Parameter	2	17.1000	0.0612	<.0001
Test length x Parameter	2	5.2100	0.0195	0.0057
Error	525	(0.0036)		
Corrected total	530			

Note: Values enclosed in parentheses represent mean square errors.

The results of the analysis of error of estimate in the random effect (growth) parameters are reported in Table 22. The effect of test length was statistically significant ($F(1, 525) = 8.62$, $p = 0.0035$), but was not practically meaningful as suggested by the small partial eta-squared value of 0.02. The effect of parameter on error of estimate was also statistically significant ($F(1, 525) = 11.90$, $p < 0.0001$). This was due to a smaller magnitude of average error of estimate for the covariance parameter than for the variance or mean of the growth rate parameters and accounted for a small proportion of variability in the error of estimate as suggested by the partial eta-squared value of 0.04. The test length x parameter interaction was not statistically significant.

Table 22: Error of Estimate of Random Effect (Growth) Parameter Estimates for Study 2

Source	df	F	η^2	p
Test length	1	8.6200	0.0162	0.0035
Parameter	2	11.9000	0.0434	<.0001
Test length x Parameter	2	0.1100	0.0004	0.8931
Error	525	(0.0017)		
Corrected total	530			

Note: Values enclosed in parentheses represent mean square errors.

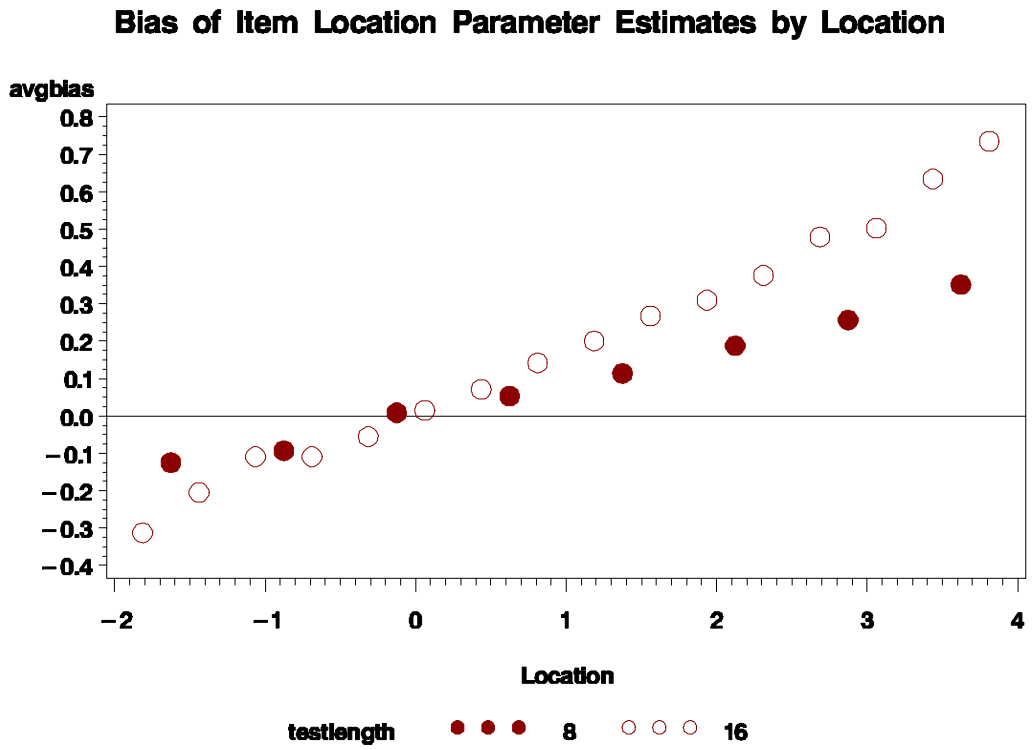
Commentary on Study 2

Study 2 examined the effect of differing numbers of items on parameter recovery in the growth model. All replications in the eight item condition converged properly. However, a modest number on the magnitude of 24 percent of the replications for the 16 item condition failed to converge. Generally, there was bias of greater magnitude for the item parameter estimates for the condition with 16 items than for the condition with 8 items. However, there was greater average error of estimate for the condition with eight items than for the condition with 16 items. This latter pattern may not have held had all replications converged in the 16 item condition as the replications that failed to converge may have otherwise produced estimates far from the mean value. Whereas the covariance parameter estimates had similar average bias in both conditions, the parameter estimates for the mean and variance of the growth rate exhibited substantially larger bias in the 16 item

condition. The average error of estimate displayed differences among the random effect (growth) parameters and also differed according to the number of items.

Indeed, the greater bias for the item parameters in the 16 item condition is puzzling. Figure 2 shows the bias in the item location parameter estimates by the generating item location. Although the range of the generating item location parameters for the 16 item condition is slightly broader than for the 8 item condition, the graph clearly shows that excluding these items would not account for the difference in average bias. The distinction between the two conditions is particularly pronounced for the more difficult items. The greater bias in the estimates of the mean and variance of the growth rate for the 16 item condition reflects this same phenomenon. The greater positive bias in the more difficult item location parameter estimates essentially leads to a stretching of the underlying vertical scale. Thus, the estimates of the mean and variance of the growth rate are likewise inflated.

Figure 2: Bias of the Item Location Parameter Estimates by Location



Study 3

Results of Study 3

In Study 3 all conditions included eight items and an item selection design in which all examinees were administered all items at all time points. Three factors were manipulated simultaneously: the form of the growth trajectory, the sample size, and the variance of the growth rate. The number of completed replications in each cell and their convergence behavior are reported in Table 23. Almost all replications converged properly. An occasional non-converging replication can be found in three of the cells, but there is no discernible pattern in the lack of convergence.

Table 23: Number of Replications and Convergence Behavior by Condition for Study 3

Factors			Number of replications		
Trajectory	Rate variance	Sample size	Converged	Not converged	Total
Linear	0.20	500	100	0	100
Linear	0.20	1000	100	0	100
Linear	0.50	500	98	2	100
Linear	0.50	1000	100	0	100
Non-linear	0.20	500	99	1	100
Non-linear	0.20	1000	99	1	100
Non-linear	0.50	500	100	0	100
Non-linear	0.50	1000	100	0	100

Item Location Parameter Recovery for Study 3

The results of the analysis of bias in the item location parameters are reported in Table 24. The generating value for the item location parameter was used as an item level covariate. This continuous covariate was statistically significant ($F(1, 5587) = 1450.49$, $p < 0.0001$) with the trend reflecting a gradual transition from negative bias for easy items to positive bias for difficult items. This trend accounts for substantial variability in the average bias as suggested by the partial eta-squared value of 0.21. A statistically significant interaction effect between growth trajectory and sample size was found ($F(1, 788.78) = 3.89$, $p = 0.0489$). However, the partial eta-squared value of less than 0.01 reveals that this interaction is not practically

meaningful. A statistically significant interaction between trajectory and the variance of the growth rate was found ($F(1, 788.78) = 7.73$, $p = 0.0056$), but this too was not practically meaningful as the partial eta-squared was less than 0.01. There were statistically significant main effects of trajectory ($F(1, 792.49) = 15.38$, $p < 0.0001$), sample size ($F(1, 788.78) = 6.58$, $p = 0.0105$), and rate variance ($F(1, 788.78) = 12.15$, $p = 0.0005$). Again, however, none of these effects had any practical significance since their corresponding partial eta-squared values were all less than 0.02.

Table 24: Bias of Item Location Parameter Estimates for Study 3

Source	df	F	η^2	p
Trajectory	1	15.3800	0.0190	<.0001
Error	792.49	(0.2583)		
Sample size	1	6.5800	0.0083	0.0105
Error	788.78	(0.2591)		
Trajectory x Sample size	1	3.8900	0.0049	0.0489
Trajectory x Sample size x Rate variance	1	1.3000	0.0016	0.2555
Error	788.78	(0.2591)		
Rate variance	1	12.1500	0.0152	0.0005
Error	788.78	(0.2591)		
Trajectory x Rate variance	1	7.7300	0.0097	0.0056
Error	788.78	(0.2591)		
Sample size x Rate variance	1	2.3300	0.0029	0.1273
Error	788.78	(0.2591)		
Location	1	1450.4900	0.2061	<.0001
Error	5587	(0.1060)		

Note: Values enclosed in parentheses represent mean square errors.

The results of the analysis of error of estimate in the item location parameters are reported in Table 25. There was a statistically significant cubic polynomial effect of item location (in order of decreasing degree $F(1, 5585) = 15.39$, $p < 0.0001$; $F(1, 5585) = 326.12$, $p < 0.0001$; $F(1, 5585) = 2.49$, $p = 0.1147$) with the trend reflecting greater error of estimate for items with extreme locations, particularly for difficult

items. The quadratic term had a partial eta-squared value of 0.06, suggesting a small but practically meaningful effect on the error of the estimate. The three-way interaction of trajectory x sample size x variance of the growth rate was statistically significant ($F(1, 788.2) = 4.48$, $p = 0.0345$). However, the partial eta-squared value of 0.02 suggests that this interaction is not practically meaningful. Statistically significant two-way interactions were also found between the growth trajectory and the sample size ($F(1, 788.2) = 9.75$, $p = 0.0019$), between the growth trajectory and the growth rate variance ($F(1, 788.2) = 12.31$, $p = 0.0005$), and between the sample size and the growth rate variance ($F(1, 788.2) = 7.99$, $p = 0.0048$). However, the partial eta-squared values for all of the two-way interactions were less than 0.02, and thus they are not practically meaningful. The main effects of trajectory ($F(1, 792.85) = 17.67$, $p < 0.0001$) and growth rate variance ($F(1, 788.20) = 24.89$, $p < 0.0001$) were statistically significant but were of little practical value with partial eta-squared values of 0.02 and 0.03, respectively. The main effect of sample size ($F(1, 788.20) = 51.51$, $p < 0.0001$) was statistically significant, and its partial eta-squared value of 0.06 suggests that it may have some practical value. As sample size increased, the average error of estimate in the item location parameters decreased.

Table 25: Error of Estimate of Item Location Parameter Estimates for Study 3

Source	df	F	η^2	p
Trajectory	1	17.6700	0.0218	<.0001
Error	792.85	(0.3115)		
Sample size	1	51.5100	0.0613	<.0001
Error	788.2	(0.3192)		
Trajectory x Sample size	1	9.7500	0.0122	0.0019
Trajectory x Sample size x Rate variance	1	4.4800	0.0057	0.0345
Error	788.2	(0.3192)		
Rate variance	1	24.8900	0.0306	<.0001
Error	788.2	(0.3192)		
Trajectory x Rate variance	1	12.3100	0.0154	0.0005
Error	788.2	(0.3192)		
Sample size x Rate variance	1	7.9900	0.0100	0.0048
Error	788.2	(0.3192)		
Location	1	2.4900	0.0004	0.1147
Location x Location	1	326.1200	0.0552	<.0001
Location x Location x Location	1	15.3900	0.0027	<.0001
Error	5585	(0.0339)		

Note: Values enclosed in parentheses represent mean square errors.

Item Discrimination Parameter Recovery for Study 3

The results of the analysis of bias in the item discrimination parameters are reported in Table 26. The linear effect of the covariate item location was statistically

significant ($F(1, 5587) = 18.33$, $p < 0.0001$). However, this accounted for very little of the variability in bias as the partial eta-squared value was less than 0.01. A statistically significant interaction between trajectory and the variance of the growth rate was found ($F(1, 788.08) = 7.24$, $p = 0.0073$). However, the partial eta-squared value of 0.01 suggests that the interaction is not practically significant. The main effects of trajectory ($F(1, 788.46) = 9.73$, $p = 0.0019$), sample size ($F(1, 788.08) = 5.89$, $p = 0.0154$), and growth rate variance ($F(1, 788.08) = 9.35$, $p = 0.0023$) were statistically significant. However, each of these effects had a corresponding partial eta-squared value that was less than 0.02, and thus these three factors do not seem to have any practically meaningful influence on the bias in the item discrimination parameter estimates.

Table 26: Bias of Item Discrimination Parameter Estimates for Study 3

Source	df	F	η^2	p
Trajectory	1	9.7300	0.0122	0.0019
Error	788.46	(0.0985)		
Sample size	1	5.8900	0.0074	0.0154
Error	788.08	(0.0990)		
Trajectory x Sample size	1	0.3600	0.0005	0.5505
Trajectory x Sample size x Rate variance	1	0.1500	0.0002	0.697
Error	788.08	(0.0990)		
Rate variance	1	9.3500	0.0117	0.0023
Error	788.08	(0.0990)		
Trajectory x Rate variance	1	7.2400	0.0091	0.0073
Error	788.08	(0.0990)		
Sample size x Rate variance	1	0.8900	0.0011	0.3451
Error	788.08	(0.0990)		
Location	1	18.3300	0.0033	<.0001
Error	5587	(0.0042)		

Note: Values enclosed in parentheses represent mean square errors.

The results of the analysis of error of estimate in the item discrimination parameters are reported in Table 27. The effect of item location was statistically significant ($F(1, 5587) = 10.79$, $p = 0.0010$). However, the nature of this trend was not discernible from graphs of the data, and the corresponding partial eta-squared value less than 0.01 suggests that the trend is not practically meaningful. Statistically

significant interactions between the growth trajectory and the sample size ($F(1, 788.18) = 7.52$, $p = 0.0062$), between the growth trajectory and the growth rate variance ($F(1, 788.18) = 6.05$, $p < 0.0001$), and between the sample size and the growth rate variance ($F(1, 788.18) = 5.21$, $p = 0.0228$). However, the corresponding effect size partial eta-squared was less than 0.01 for each of these effects, suggesting that they have no practical influence on the error of the estimate. The main effects of trajectory ($F(1, 789.02) = 19.63$, $p < 0.0001$) and growth rate variance ($F(1, 788.18) = 29.76$, $p < 0.0001$) were likewise statistically significant but had only very small partial eta-squared values on the magnitude of 0.02 to 0.03. The main effect of sample size was statistically significant ($F(1, 788.18) = 88.70$, $p < 0.0001$), but also had a modest partial eta-squared value of 0.10. Thus, as sample size increased, the average error of the estimate decreased.

Table 27: Error of Estimate of Item Discrimination Parameter Estimates for Study 3

Source	df	F	η^2	p
Trajectory	1	19.6300	0.0243	<.0001
Error	789.02	(0.0300)		
Sample size	1	88.7000	0.1012	<.0001
Error	788.18	(0.0301)		
Trajectory x Sample size	1	7.5200	0.0095	0.0062
Trajectory x Sample size x Rate variance	1	1.9900	0.0025	0.1592
Error	788.18	(0.0301)		
Rate variance	1	29.7600	0.0364	<.0001
Error	788.18	(0.0301)		
Trajectory x Rate variance	1	6.0500	0.0076	0.0142
Error	788.18	(0.0301)		
Sample size x Rate variance	1	5.2100	0.0066	0.0228
Error	788.18	(0.0301)		
Location	1	10.7900	0.0019	0.001
Error	5587	(0.0028)		

Note: Values enclosed in parentheses represent mean square errors.

Random Effect (Growth) Parameter Recovery

The results of the analysis of bias in the random effect (growth) parameters are reported in Table 28. The three way interaction among the three factors, trajectory, sample size, and rate variance, was statistically significant ($F(1, 3943) =$

8.58, $p = 0.0034$) but not practically meaningful since the corresponding partial eta-squared value was less than 0.01. All of the two-way interactions between trajectory and sample size ($F(1, 3943) = 13.56$, $p = 0.0002$), between sample size and parameter ($F(8, 3943) = 2.28$, $p = 0.0199$), between trajectory and rate variance ($F(1, 3943) = 45.19$, $p < 0.0001$), between rate variance and parameter ($F(8, 3943) = 9.42$, $p < 0.0001$), and between sample size and rate variance ($F(1, 3943) = 14.28$, $p = 0.0002$) were likewise statistically significant but not practically meaningful with partial eta-squared values less than 0.01. The main effect of sample size was statistically significant ($F(1, 3943) = 25.66$, $p < 0.0001$) but, again, not practically meaningful due to a partial eta-squared value less than 0.01. The main effects of trajectory ($F(1, 3943) = 98.26$, $p < 0.0001$) and variance of the growth rate ($F(1, 3943) = 92.75$, $p < 0.0001$) were statistically significant. Both effects had eta-squared values of 0.02 suggesting that the practical influence of these effects on bias is quite small.

Table 28: Bias of Random Effect (Growth) Parameter Estimates for Study 3

Source	df	F	η^2	p
Trajectory	1	98.2600	0.0243	<.0001
Parameter(trajectory)	8	11.2800	0.0224	<.0001
Sample size	1	25.6600	0.0065	<.0001
Trajectory x Sample size	1	13.5600	0.0034	0.0002
Sample size x Parameter(trajectory)	8	2.2800	0.0046	0.0199
Rate variance	1	92.7500	0.0230	<.0001
Trajectory x Rate variance	1	45.1900	0.0113	<.0001
Rate variance x Parameter(trajectory)	8	9.4200	0.0188	<.0001
Sample size x Rate variance	1	14.2800	0.0036	0.0002
Trajectory x Sample size x Rate variance	1	8.5800	0.0022	0.0034
Sample size x Rate variance x Parameter(trajectory)	8	1.8400	0.0037	0.0657
Error	3943	(0.0222)		
Corrected total	3982			

Note: Values enclosed in parentheses represent mean square errors.

The results of the analysis of error of estimate in the random effect (growth) parameters are reported in Table 29. The three-way interaction of trajectory x sample size x variance of the growth rate was statistically significant ($F(1, 3943) = 25.05$, $p < 0.0001$) but not practically meaningful with partial eta-squared equal to 0.01. Likewise, the three-way interaction of trajectory x sample size x parameter was statistically significant ($F(8, 3943) = 5.33$, $p < 0.0001$) but not practically meaningful

with partial eta-squared equal to 0.01. The two-way interactions between trajectory and sample size ($F(1, 3943) = 33.57$, $p < 0.0001$), between sample size and parameter ($F(8, 3943) = 6.19$, $p < 0.0001$), between trajectory and rate variance ($F(1, 3943) = 63.14$, $p < 0.0001$), and between sample size and rate variance ($F(1, 3943) = 42.90$, $p < 0.0001$) were likewise statistically significant but not practically meaningful with partial eta-squared values 0.02 or less. However, the rate variance x parameter interaction was statistically significant ($F(8, 3943) = 20.81$, $p < 0.0001$), and the partial eta-squared value of 0.05 suggested that this interaction may have a small but meaningful effect. Overall, the error of estimate for the random effects (growth) parameters was higher for the extreme 0.50 generating rate variance than for the more moderate 0.20 rate variance. Also, the error of estimate appears to be higher on average for the variance of the growth rate parameter(s) in the 0.50 level. However, the interaction can be seen in the error of estimate for the mean and variance of the growth rate parameters with the linear growth trajectory with the 0.50 rate variance level, which are particularly large on average. The main effect of trajectory was statistically significant ($F(1, 3943) = 71.52$, $p < 0.0001$) but having very little practical meaning with a partial eta-squared of 0.02. The main effect of parameter was statistically significant ($F(8, 3943) = 25.71$, $p < 0.0001$) and also showed a small practical meaning with a partial eta-squared value of 0.05. Generally, this was due to estimates for variances of the growth rates having larger average error of the estimate than for the other random effects (growth) parameters. The main effect of sample size was statistically significant ($F(1, 3943) = 104.23$, $p < 0.0001$) with a small partial eta-squared value of 0.03.. This was due to the sample size of 500 examinees having

greater average error of estimate for the random effects (growth) parameter estimates than 1000 examinees. The main effect of variance of growth rate was statistically significant ($F(1, 3943) = 199.02$, $p < 0.0001$) with a small partial eta-squared value of 0.05. This effect is due to the high rate variance (0.50) having greater average error of estimate than the low rate variance (0.20) condition.

Table 29: Error of Estimate of Random Effect (Growth) Parameter Estimates for Study 3

Source	df	F	η^2	p
Trajectory	1	71.5200	0.0178	<.0001
Parameter(trajectory)	8	25.7100	0.0496	<.0001
Sample size	1	104.2300	0.0258	<.0001
Trajectory x Sample size	1	33.5700	0.0084	<.0001
Sample size x Parameter(trajectory)	8	6.1900	0.0124	<.0001
Rate variance	1	199.0200	0.0480	<.0001
Trajectory x Rate variance	1	63.1400	0.0158	<.0001
Rate variance x Parameter(trajectory)	8	20.8100	0.0405	<.0001
Sample size x Rate variance	1	42.9000	0.0108	<.0001
Trajectory x Sample size x Rate variance	1	25.0500	0.0063	<.0001
Sample size x Rate variance x Parameter(trajectory)	8	5.3300	0.0107	<.0001
Error	3943	(0.0152)		
Corrected total	3982			

Note: Values enclosed in parentheses represent mean square errors.

Commentary on Study 3

Study 3 examined three factors simultaneously: the form of the trajectory of growth, sample size, and the variance of the growth rate in the examinee population. Although there were many effects under consideration in this study and many statistically significant results were found, few of these effects were large enough to suggest any practical significance. As could be anticipated, the sample size influenced the error of the estimate in the item location and item discrimination parameters. As sample size increased, the average error of estimate in the item parameters decreased. However, none of the three factors had any meaningful influence on the bias of the parameter estimates and that factors that would not be expected to influence the error of the estimate, such as the form of the trajectory and the population variance in the growth rate, indeed have no meaningful effect.

The generating value of the variance of the growth rate played a small role in influencing the error of the estimate for the random effects (growth) parameters. Overall, the error of estimate for the random effects (growth) parameters was higher for the extreme 0.50 generating rate variance than for the more moderate 0.20 rate variance. Also, the error of estimate appeared to be higher on average for the variance of the growth rate parameter(s) in the 0.50 level. A small but meaningful interaction was seen in the error of estimate for the mean and variance of the growth rate parameters with the linear growth trajectory with the 0.50 rate variance level, which was particularly large on average.

Upon examination of the mean rates of bias and error of estimate, it is apparent that parameter recovery in the item response growth model is generally

acceptable under reasonable conditions, such as with adequate sample size. The effect of variability in the growth trajectory may in fact be a reflection of the extreme value of 0.50 setting up a situation in which the growth of many individuals is not adequately measured by the items selected due to censoring effects.

Chapter 4: Illustration Using ECLS-K Data

The Monte Carlo simulation approach has particular usefulness in its ability to demonstrate the characteristics of the model in light of the true patterns in the data. However, the sole presentation of simulated results can raise suspicion about whether the advocated technique will work in the world of messy "real life" data. To offer some perspective on the use of the item response growth model in practice, this chapter provides an illustration of the model with item response data from the Early Childhood Longitudinal Study – Kindergarten Cohort (Tourangeau, et al., 2006)[2] [3].

ECLS-K is one of several nationally-representative longitudinal studies conducted by the National Center for Education Statistics (NCES). The primary purpose of this study was to provide researchers with data for investigating children's transition from their early childhood environments into school. The kindergarten cohort contributed to this purpose by following a nationally representative sample of American school children from the start of their formal schooling in kindergarten.[4] Both cognitive and non-cognitive measures were collected. Children were tested directly; their parents were interviewed; and their teachers completed questionnaires. This data is made available to researchers to conduct research in education at a scale that would be difficult for any individual researcher to coordinate. Thus, ECLS-K is one of the most readily-available sources of longitudinal data for illustrating the item

[2] Used with permission. Reid, Hresko, Hammill, ProEd, Inc. International, 2008, TERA-3.

[3] Used with permission. Ginsburg, Baroody, ProEd, Inc. International, 2008, tema-3.

[4] A separate cohort was followed from birth through early childhood.

response growth model. Of the longitudinal data sets compiled by NCES, ECLS-K is one of the more recent as well as one of the more complete studies offering measurements at several time points to allow for modeling alternative trajectories for growth. The existence of measurements at more than two time points is especially important to demonstrate the utility of the item response growth model distinct from other longitudinal item response models that parameterize growth differently. With only two time points, the item response growth model proposed here cannot be distinguished from other longitudinal item response models.

The illustration models growth in math achievement in a sample from a cohort of elementary children from the beginning of kindergarten through the end of fifth grade. Although the ECLS-K cognitive battery included math, reading, and general knowledge assessments, the math assessment was chosen for this illustration due to the recent interest and push for more emphasis on STEM (science, technology, engineering, and mathematics) education across the United States' educational systems.

With this real life data set two research questions were investigated using the item response growth model. The first two questions relate directly to factors that were investigated in the Monte Carlo simulation study. First, is an item response growth model based on Rasch measurement or based on two-parameter logistic measurement a better fit to the data? Second, is an item response growth model with a linear or nonlinear trajectory of growth a better fit to the data?

Methods

Participants

The data for this illustration consisted of a random subsample of 2000 examines from ECLS-K. The decision to fit the model to a subsample of examinees was driven by concerns about the computational intensity of the model arising from the simulation study and the extreme amount of time it would likely take to estimate the model with the full sample of approximately 22,000 children. A random sample was chosen to reflect the composition, including patterns of missing data, as found in the full data set.

Data Collection Design

The children in the study's primary cohort were tested in what was for most of them the fall of kindergarten, spring of kindergarten, spring of first grade, spring of third grade, and spring of fifth grade[5]. In addition, item responses were collected from a supplementary "bridge" sample of second grade children. Fall first grade testing included only a subsample of the original primary sample. However, these differences in the sampling design were easily absorbed by the model since the model accommodates students who take different items at different time points. ECLS-K also uses a complex sampling design. However, information about this design, such as sampling and replication weights, was not included in the item response data file and thus was not incorporated in the results described here. Thus, the conclusions

[5] After being recruited as kindergarteners, the children continued to be tested at the same intervals even if they skipped a grade or were held back a grade. Thus, not all children were in the anticipated grade for their cohort at subsequent assessments.

about growth in this sample should not be generalized to the national population of school children.

Instruments

The cognitive assessments in ECLS-K were specially assembled for the study because shelf tests did not meet the desired content standards for the study (Rock & Pollack, 2002). Mathematics items in the cognitive assessment are from an individually-administered adaptive test (Rock & Pollack, 2002). Thus, items given to the student at each time point were selected to match the student's approximate level of achievement. Performance on a short routing test was used to determine if the student would complete items from a low, middle, or high achievement form of the math assessment (Rock & Pollack, 2002). Adjacent level forms of the test contained a block of items common to both tests (anchor items). Due to the poor results with a similar design in the simulation study, for this illustration only the 14 common items were used. These items were only used through third grade so there are no fifth grade responses to the items in this particular set. All items were dichotomously scored, short answer items from the mathematics assessments in grades K-5.

Procedures

The proportion of subsample examinees who were part of the bridge sample was calculated. In addition the proportions of examinees responding to each item at each time point were calculated to demonstrate the item selection design for the common item set. Four different versions of the model were fit to the item response data, which included the bridge sample: a Rasch linear trajectory model, a Rasch

nonlinear trajectory model, a two-parameter logistic linear trajectory model, and a two-parameter logistic nonlinear trajectory model.

In all cases the model was fit using the same estimation procedure used in the simulation study and described previously in Chapter 2. For the item response growth model comparison, the -2 log likelihood, AIC, and BIC model fit statistics were compared to decide on the best fitting model. Empirical Bayes estimates of the parameters for individual growth trajectories were computed for the best fitting model. These growth trajectories are interpreted few examples of individual cases.

Results

Descriptive Statistics

Of the 2000 examinees in the selected subsample, 85 examinees, or 4.25 percent of them, were members of the second-grade bridge sample. The use of only the common items from the mathematics assessment resulted in a design that was similar to the targeted design used in the simulation study. Table 30 shows the proportion of the sample that took each of the 14 items at each of the time points. The distributions of the proportions at each time point suggest that fewer students were given difficult items in the lower grades and easy items in the higher grades. Proportions are lower overall for the Fall 1 and Spring 2 time points because only a fraction of the full sample was tested at these time points.

Table 30: Proportion of Examinees Responding to Each of the Common Items at Each Time Point

Item	Fall K	Spring K	Fall 1	Spring 1	Spring 2	Spring 3
1	0.84	0.89	0.24	0.76	0.02	0.19
2	0.84	0.88	0.24	0.76	0.02	0.18
3	0.84	0.89	0.24	0.76	0.02	0.19
4	0.84	0.89	0.24	0.76	0.02	0.19
5	0.21	0.52	0.18	0.71	0.04	0.65
6	0.20	0.52	0.18	0.71	0.04	0.65
7	0.21	0.52	0.18	0.71	0.04	0.65
8	0.21	0.52	0.18	0.71	0.04	0.65
9	0.21	0.52	0.18	0.71	0.02	0.19
10	0.07	0.24	0.11	0.60	0.02	0.18
11	0.07	0.24	0.11	0.60	0.02	0.19
12	0.07	0.24	0.11	0.60	0.02	0.18
13	0.07	0.24	0.11	0.60	0.04	0.65
14	0.07	0.24	0.11	0.60	0.02	0.19

Model Comparisons

Model fit statistics from fitting four different item response growth models to the random subsample are shown in Table 31. The two-parameter logistic versions of the item response model fit this data better than the Rasch version as suggested by lower values of the AIC and BIC for the two-parameter logistic models. The

piecewise linear trajectory fit the data better than the linear trajectory as suggested by lower values of the AIC and BIC for the piecewise-linear models.

Table 31: Model Fit Statistics for Four Item Response Growth Models fit to ECLS-K Data

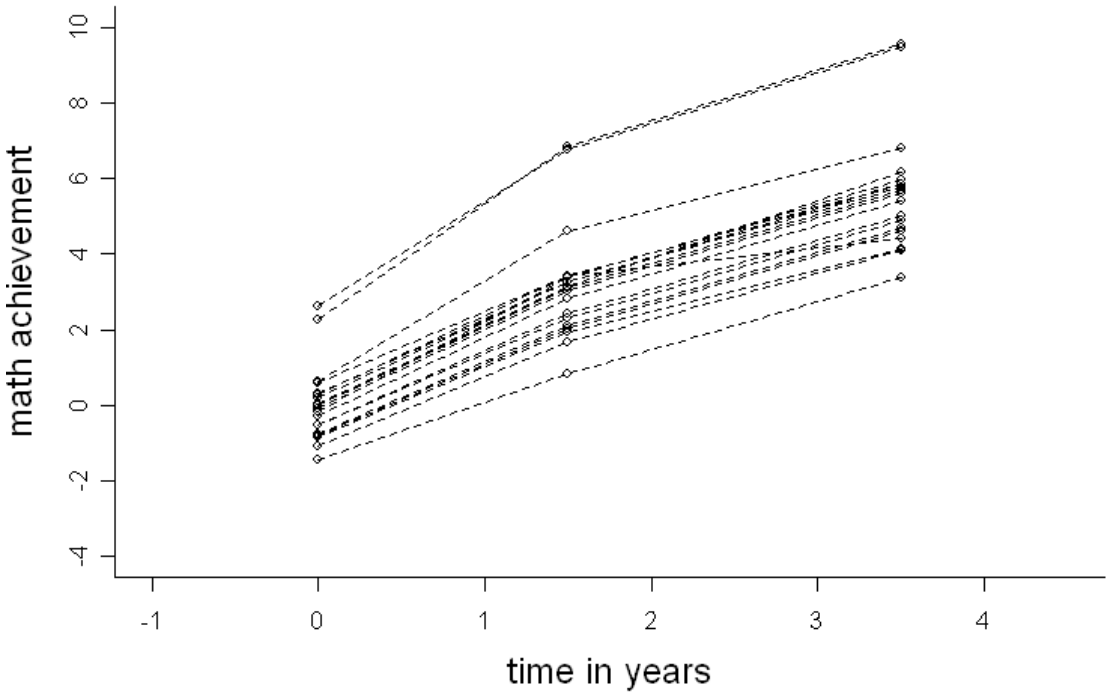
Measurement model	Growth trajectory	Parameters	-2 log likelihood	AIC	BIC
Rasch	Linear	17	63051	63085	63107
Rasch	Piecewise linear	21	62835	62877	62904
Two-parameter logistic	Linear	31	56842	56904	56944
Two-parameter logistic	Piecewise linear	35	56313	56383	56429

Individual Growth Trajectories

Figure 3 shows graphed individual trajectories from a subsample of 20 of the 2000 examinees in the real data illustration for the two-parameter logistic item response model with a piecewise linear growth trajectory form. Notice that in general the lines were not as steep for the period from 1.5 to 3.5 years (late growth period) as they were for the period from zero to 1.5 years (early growth period). This reflects the population mean rates of change that were estimated to be 2.09 for the early growth period and 1.31 for the late growth period. Notice that there is modest variability in the examinees' initial levels of math achievement θ, as suggested in the

plot by the vertical spread in the individual points at zero years and represented by a fixed variance value of one in the model. However, there is less variability in their rates of growth, as suggested in the plot by the nearly parallel lines for most individuals during both growth periods and reflected in the estimated growth rate variances of 0.18 and less than 0.01, respectively. The covariances of the growth rates with the initial status θ in math achievement were 0.26 and 0.15, respectively, reflected in Figure 3 as a slight fanning of the individual trajectories from left to right. Examinees that started out with higher initial status in math achievement tended to have somewhat higher growth rates than average; examinees that started out with lower initial status in math achievement tended to have somewhat lower growth rates than average. The covariance between the growth rates was 0.01, suggesting that examinees who grew at a faster than average rate in the first growth period tended to also grow at a faster than average rate in the second growth period.

Figure 3: Trajectories for a Random Sample of 20 Examinees



Conclusions can also be drawn for individual examinees depicted in Figure 3. Student 1, whose trajectory can be seen at the top of the figure, had an estimated initial math achievement of 2.62 standard deviations above the mean initial math achievement of the subsample of 2000 examinees. Student 1 also grew at a tremendous rate estimated to be 2.78 units per year during the first growth period and 1.34 units per year during the second growth period. Student 2, whose trajectory is also shown in the figure at the top of the large cluster of trajectories, started with an estimated initial math achievement level that was substantially lower at 0.61. However, Student 2 grew at similar estimated rates of 2.67 units per year and 1.10 units per year during the first and second growth periods, respectively, as can be noted by a trajectory that is nearly parallel to Student 1's trajectory in Figure 3. Thus, despite differences in math achievement status at any single point in time, these two students are exhibiting similar growth in math achievement.

Student 3 has a trajectory that is embedded in the large cluster of trajectories in Figure 3. Student 3 had an initial estimated math achievement level of 0.28, about average. This student also had an estimated early period growth rate of 2.09 units per year, which perfectly matched the average for the full sample of 2000 examinees. However, a clear change is noted in Student 3's trajectory during the second growth period, where the growth rate is a mere 0.51 units per year, less than half of the average for the full sample of 200 examinees. Indeed, this can be seen in Figure 3 where Student 3's trajectory crosses those of the other examinees during the second growth period. This student has clearly experienced a dramatic change in growth rate that is causing him or her to fall behind the other students, and thus Student 3 would

likely be identified as struggling and in need of educational intervention despite having a math achievement status at 3.5 years that is not readily distinguished from the math achievement of the other students at this point in time. Student 3's trajectory may be contrasted with that of Student 4, whose trajectory is at the bottom of Figure 3. Student 4 started with an initial math achievement of 1.47 standard deviations *below* average and at 3.5 years is at a lower level of math achievement than Student 3. However, it is clear from Student 4's estimated growth rates of 1.52 and 1.28, respectively, that this student is growing steadily and is probably already benefiting from successful educational intervention.

Discussion

Overall, the real data illustration suggests that the model estimation performs similarly when fit to real data as when fit to simulated data. It also shows the value of the two-parameter logistic item response model over the Rasch version, as the two-parameter logistic version was a better fit to the real-life empirical data in this illustration. It can also be observed from this illustration that a linear trajectory for growth may not be the most appropriate trajectory for math achievement growth in the early elementary years. In this example the item selection design, which was similar to the targeted design used in the simulation study, was realistic for annual testing, where students may be expected to forget the specifics of test items between assessments. For assessments that are spaced less than several months apart, however, a more sophisticated item selection design would be needed.

Chapter 5: Concluding Remarks

Discussion

In recent years there has been an increasing opportunity and great need to appropriately model individual growth trajectories in educational research. This study proposed and explored the technical adequacy of an integrated IRT growth model that combined the typically separate data transformations involved in applying item response model(s), vertical scaling (where appropriate) of scores from tests at different levels, and growth model components. The general motive for undertaking the present study was to lay a foundation for future research on IRT growth models by means of a broad investigation of the performance of the two-parameter logistic version of the model in a variety of situations. Indeed, this research has shown that parameter recovery in this model is sensitive to some important considerations in the process of modeling growth. This section is organized into subsections addressing lessons learned about the estimation of the model, its convergence behavior, the inflation of the underlying vertical scale, the effects of the five manipulated factors on parameter recovery, and the application of the model to real data.

Model Estimation

Although a thorough investigation of estimation methods was not proposed as part of this study, an unsolicited result of carrying out this study has been some insight into the demands of estimating this model. In the original research proposal the plan was to implement marginal maximum likelihood estimation to

simultaneously obtain estimates for all model parameters in a single run of SAS Proc NLMIXED. Although it was known that the model would be computationally intensive to estimate using this approach, the full extent of this intensive demand was not known prior to commencing the research. After starting to carry out the study, it quickly became clear that the computational intensity of the estimation approach rendered it practically infeasible for use with IRT growth models. As a result of this finding, the original SAS code for the study was modified to implement marginal maximum likelihood estimation via the Expectation-Maximization (EM) algorithm. This approach, described in detail in Chapter 2, estimated model parameters using an iterative approach that incorporated strategic isolation of portions of the model for the purposes of updating estimated values of the parameters to reduce unnecessary computational demand and speed the estimation process.

Case study trials suggested that on average the new estimation approach produced parameter estimates that were about as close to the generating values as the originally proposed approach. However, it was noted through observation of the patterns of the difference of the estimate from the true value across items that the two estimation approaches displayed different patterns. For example, estimation in a single run of SAS Proc NLMIXED appeared to produce better estimates of discrimination parameters for items in the middle of the ability distribution. However, the EM algorithm appeared to produce better estimates of discrimination parameters for items located at the extremes of the ability distribution.

In addition to a more thorough investigation comparing different estimation approaches, there are many factors within the EM algorithm estimation approach that

remain to be formally investigated. Since this was not a proposed goal of the present research study, these issues received only superficial investigation to quickly assess adequately performing and reasonable approaches for the present study. These issues include further study of stopping rule alternatives for the EM cycles, stopping rules for updating the parameter estimates for parts of the model for internal cycles within the EM cycle, different value updating approaches for the fixed effects and random effects parts of the model, different numbers and approaches for establishing the number of internal iterations in the internal cycles to update the parameters, and the potential use of a Bayesian framework to incorporate prior distributions for the structural parameters. Further, it is recognized that SAS is not a very effective platform for algorithms involving looping structures, and future research in this area should proceed with this estimation approach on a different platform.

Convergence Behavior

Some of the best lessons to be learned from this study have been in the areas where the model has not performed according to expectation. No where is this more true than in the observed convergence behavior for different conditions in the study. Convergence in this model is most strongly influenced by the number and arrangement of items and the number of examinees providing item responses. In short some item selection designs failed to converge at all. One item selection design had a convergence rate that depended on the sample size. In addition there was a lowered convergence rate as the number of items increased. The implications of each of these situations will be discussed in turn.

First, the location of the items selected must be appropriately targeted to the population ability distribution at any given point in time. In this study one item selection design was attempted in which some items were generated with locations that were not a good match to the ability distribution in the population at the time(s) the item was administered. For example, a very difficult item was administered at the first time point when virtually none of the examinees would generate a correct response to such an item, and then this item was not included on any of the test forms at the remaining time points. The opposite situation also occurred where a new very easy item was introduced on the test form at the final time point when virtually none of the examinees would generate an incorrect response. This led to a situation where the model failed to converge because there was little empirical basis for pinpointing precise item locations. In addition item discrimination values for such items tended toward infinity due to an extreme imbalance between numbers of correct and incorrect responses on a particular item. Items of equal generating difficulty that were administered at a time when the generating difficulty was a good match for the population ability distribution generally maintained values within the established boundary values, even while the item parameter estimates for poorly matched items hit the boundary values.

Second, it is not sufficient to simply target the item location. Scale inflation and lack of convergence also tends to occur where there are more items to be estimated. An adaptive design was attempted in which some items served as a routing test to determine whether an examinee would be best match to items at the low, medium, or high locations on the latent ability continuum. At the point at which

the estimation algorithm was stopped for this adaptive item design, some discrimination parameters and many item location parameters were at the established boundary values. Easy items, whose generating values were always greater than -2, had item location estimates at the boundary value of -4; difficult items, whose generating values were always less than four, had item location estimates a the boundary value of ten. The most instances for item locations at boundary values were observed in the "low" and "high" ability test forms, which contained the easiest and most difficult items, respectively, which naturally would be most likely to reach the boundaries in the case of scale inflation. Discrimination values were generally underestimated in the adaptive design, further reflecting the tendency toward scale inflation.

Third, the number of examinees relative to the number of item parameters to be estimated in the model may also play a role in convergence. The lowered convergence rate with 16 items compared to eight items may be attributable to a need for a larger number of examinee responses to estimate a larger number of item parameters. The estimation with 16 items was only attempted in this study with 500 examinees, which was found to be insufficient for satisfactory parameter recovery in other components of this study. This suggests a need for balance between the parameters to be estimated and the number of examinees providing item responses.

Bias by Item Location and Stretching of the Vertical Scale

The magnitude and direction of bias in item location parameter estimates was directly influenced by the generating value of the item location. Thus, bias was more extreme for items with extreme locations on the underlying scale, and the direction of

bias was related to the end of the scale where the item was located. Concurrent with this phenomenon was the tendency for item discrimination parameters to be underestimated and for random effects (growth) parameters estimates to be overestimated, particularly the mean and variance of the growth rate. These results point to a general tendency for inflation in the underlying vertical scale that should be further studied to better understand the impact of the unified growth modeling approach on the characteristics of the underlying vertical scale.

Effect of Item Selection Design on Parameter Recovery

In order for the IRT growth model estimation to converge and the parameter estimates be accurate, it is necessary that items be well-suited to the difficulty levels of the examinees at each of the time points. This is more critical in item response models that include a discrimination parameter than it is in the Rasch family of item response models. This can be explained as what may be termed the *Discrimination-Censoring Paradox*.

The Discrimination-Censoring Paradox:

To properly estimate item discrimination, we need sufficient numbers of responses from persons located above and below the item location. In order to properly estimate the person's location, we need sufficient items (with adequate discrimination) above and below the person's location *at each time point* to avoid the effects of censoring on the estimation of the growth parameters. The effect of censoring is generally to decrease the spread of person locations, leading to greater difficulty estimating item discrimination parameters.

In this study the spread of the generating item locations was chosen to be quite broad in an attempt to avoid the well-known effect of censored item response data in contributing to the underestimation of the growth rate and the covariance between the growth rate and initial status. Due to a lack of full information about the true growth, in censoring there is a shrinking or contracting of the underlying vertical scale. However, an opposite effect, an overestimation of the growth rate and covariance between the growth rate and initial status, a "stretching," or inflation of the underlying vertical scale was instead observed in this study, as described in the previous section. Thus, an effect opposite to the effect of censoring is observed.

Effect of Number of Items on Parameter Recovery

The effect of the number of items on parameter recovery will require further investigation. No difference in item parameter bias was anticipated, although it was thought that the error of estimate of the random effects (growth) parameters might decrease as the number of item responses available to estimate each individual's growth trajectory increased. In fact, meaningful differences in the bias were observed, suggesting again that test construction has an important role to play in the performance of the item response growth model, particularly when item discrimination parameters are included.

Effect of Variance of the Growth Rate on Parameter Recovery

The extreme variability in the growth rate in the population has the potential to trigger the effect of the Discrimination-Censoring Paradox. As the variance of the growth rate was increased from a typical rate of 0.20 to an extreme rate of 0.50, the

average error of estimate in the random effects (growth) parameters increased. Fortunately, however, the magnitude of this effect as observed in this study suggested only a small effect for a level of population growth variability that was intended to be quite extreme. The error of the estimate in the random effects (growth) parameters may be influenced when variability in the growth rate produces a situation in which the growth of many individuals is not adequately measured by the items selected. Thus, it may be necessary to include more items covering a broader range of item locations if there is dramatic heterogeneity in the rate of growth in the examinee population.

Effect of Growth Trajectory on Parameter Recovery

Although there were statistically significant differences between the linear growth trajectory and the nonlinear (piecewise) growth trajectory, the magnitude of these effects was too small to have practical meaning. It is possible that the statistically significant differences found can be attributed to the effect of the variance of the growth rate discussed in the preceding section, as this should have been more prevalent for the linear trajectory in this study than in the nonlinear trajectory due to the way in which the generating parameters for the items were controlled. A more sophisticated design that accounts for the variability in the growth rate in arranging the spread of generating item locations would be needed to test whether this is the case.

There are also some possible explanations that stem from the relationship between the form of the growth trajectory and the number of time points used to measure the growth. It is possible that the bias in the parameter estimates is related to

the ratio between the number of time points and the number of random effects defining the growth trajectory. For example in the linear trajectory level in the present study, there are five time points and two random effect (growth) parameters. In the nonlinear trajectory level, there are five time points and three random effect (growth) parameters. Additional levels of trajectory and different numbers of time points could be used to further investigate potential associations with differences in bias. It is also possible that the bias in the item location parameters is related to the number of time points per segment of the piecewise trajectory. Perhaps as fewer points are used to define each linear segment in the trajectory, there is less room for bias in the item location estimates. These theories may be checked by evaluating the performance of the model with additional forms of growth trajectory that are not piecewise defined, such as a quadratic trajectory.

Effect of Sample Size on Parameter Recovery

Both Studies 1 and 3 investigated the effect of differing sample sizes on parameter recovery in the item response growth model. Both studies found a substantial drop in the average magnitude of bias as sample size increased from 500 examinees to 1000 examinees. However, Study 1 also examined a condition with 2000 examinees and did not find a significant further decrease in the average magnitude of bias. This suggests that the sample size of 500 is too small in that it produces estimates with substantially more bias. This also suggests that the bias in the item location parameter estimates may be influenced by the amount of information available to estimate each item's location. As items become more extreme in their difficulty, there is less information in the data to estimate that item

parameter. This could potentially lead to items having estimated locations that are even more extreme than the generating values.

Application of the Model in Practice

The application of the model with real data supported the validity of the claims made on the basis of the simulated data. Overall, the real data illustration suggested that the computational intensity of the model estimation is similar when fitting real data as when fitting simulated data. Fitting the model with real data reflected a proportional level of computational intensity to the estimation with simulated data.

With 14 items and 2000 examinees only four models were fit to the data. Thus, this illustration is quite limited. However, the two conclusions that it does support are important for demonstrating the value of exploring the new model proposed in this study. First, the empirical illustration showed the value of the two-parameter logistic item response model over the Rasch version, as the two-parameter logistic version was a better fit to the real-life empirical data in this illustration. It can also be observed from this illustration that a linear trajectory for growth may not be the most appropriate trajectory for math achievement growth in the early elementary years.

Scope and Limitations

To understand the results of this study in context, consider current practice in modeling growth in individual students' academic abilities over time. This process typically uses several statistical models or transformations to move from data

representing a student's correct or incorrect responses on individual test items to inferences about changes in the student's underlying ability. First, a measurement model, typically from the IRT family of models, is used to move from a student's responses to many individual items to a single test score that provides an estimate of the student's proficiency at the time he or she answered the items. Then a vertical scaling transformation may be needed to put scores from increasingly difficult tests on a developmental scale. Finally, a trajectory for the individual scaled estimates of proficiency may be estimated in a growth model, typically from the Latent Growth Curve Model or Linear Mixed Effects Model frameworks.

Each of these steps has received a great deal of focused research attention isolated from the other steps. For example much measurement research focuses on the estimation and fit of item response models. Researchers concerned with vertical scaling are working to address substantive concerns regarding the vertical alignment of content in a way that makes a vertical scale meaningful. Likewise, growth modeling researchers study issues such as the effect of misspecification of the growth trajectory. Issues such as these that are focused on isolated steps in the process are outside the scope of the present study. This study attempts to draw connections among these steps by investigating a unified model that conducts all three steps simultaneously.

The proposed integrated model extends beyond previous item response models for longitudinal data by specifying a functional form for the growth of individual examinees. In addition this model includes item discrimination

parameters, thus extending the item response growth model beyond similar models in the Rasch family that are currently being investigated by other researchers.

Although this study expands item response growth models outside of the Rasch family of models, there do remain some limitations on the model used in this study. The model under consideration is limited to binary item responses. However, extension of this approach to models with polytomous data should be fairly straightforward. Although it includes the item discrimination parameter, the model under consideration is also limited to two item parameters. It is theoretically possible to extend this model to accommodate three item parameters. However, the estimation of such a model becomes much more difficult.

In addition to possible extensions of the measurement portion of the integrated model, it is also possible to extend the growth component. The use of a piecewise trajectory may be used to test the effectiveness of an intervention by strategic placement of the point at which one growth rate ends and the next growth rate begins at the point at which the intervention began following baseline observations. The model may also be expanded to include observed variables, or covariates, that help predict or explain the parameters of the growth trajectory. Extended work with the model could also further develop the methodology for making individual projections of growth, including potentially linking these projections to normative patterns of growth (Betebenner, 2008). However, the use of this approach to make such inferences would require more research to identify the inferences that are robust to monotonic transformations of the vertical scale. The model could also be extended to model the effects of schools on the means and standard deviations of the initial

measurement and trajectory parameters by adding a third level. However, because this is primarily a methodological investigation of a new approach for analyzing longitudinal growth, a simpler model was investigated in this study.

There is one limitation of the longitudinal study design that can be overcome in practice. The characteristics of the items that are administered repeatedly may change over time due to memory effects, practice, and influences that are separate from changes in the underlying latent construct of interest. There are two proposed strategies for overcoming this problem. One is to avoid administering the same item to the same examinee at multiple time points. This can be accomplished by incorporating a more sophisticated vertical linking design involving multiple forms of the test at a single measurement occasion and randomly equivalent groups. An alternative strategy is to incorporate within the model some additional item parameters that account for changes in the characteristics of the items over time (Wang, et al., 1998). Both strategies, however, were originally intended for situations in which examinees are tested at fixed time points. These two approaches may require further development or modification to accommodate situations in which examinees are tested at varying time points.

Areas for Future Research

In the immediate future the computational intensity of the model will be further addressed by transferring the estimation algorithm to a more efficient platform. SAS was used for convenience in this study, but this is not necessarily an efficient environment for carrying out the EM algorithm.

Since the item selection design has such a dramatic influence on the convergence behavior in this model, future research should more closely investigate the more precise effects of modifying different aspects of the item selection design. In this study a wide variety of characteristics of the item selection design were varied, including whether the design is adaptive to the latent abilities of individuals, targeted to the population ability at each time point, and the proportion of common items across tests administered at adjacent time points. However, other aspects that were held constant in this study, such as the spread of the locations of the items on the latent continuum, should also be considered for experimental manipulation. In combination these aspects of the item selection design provide an important component in controlling the amount of empirical information available to estimate the item parameters.

The other critical component of information is related to the examinees who provide responses to the items that are selected. This includes the sample size, the spread in the parameters of latent growth in the population, as well as the consideration of potential missing data (planned or unplanned) in which there is variability among individual examinees in the number of measurement occasions. More focused examination of the effect of manipulating these characteristics is also needed. Clearly, this interacts with the item characteristics in that proper alignment between the item characteristics and examinee characteristics produces a situation in which optimal information is available to estimate the parameters of the item response growth model.

As this study has demonstrated, parameter recovery in the item response growth model is very sensitive to the alignment between the items selected and the parameters of latent growth in the examinee population. However, it is not known whether this model is any more sensitive to this alignment than competing methods in other frameworks, such as latent growth curve modeling in the SEM framework or mixed effects models. Presently, other researchers are investigating a closely-related Rasch version of this model. However, a comparative study focusing on the specification of item discrimination parameters in the model is needed to gain a better understanding of how much more information is needed to estimate the IRT growth model item discrimination parameters. It would be interesting to observe the differences in bias between the Rasch growth model and the two-parameter logistic growth model to see what effect including the discrimination parameters has on the bias in the estimation of the location parameters. An additional point of investigation relevant to the role of the discrimination parameter may be role of the magnitude of the discrimination of the items with extreme item locations in influencing the bias in the item location parameter estimates.

Broader Relevance and Conclusion

The integrated model investigated in this study addresses the entire growth modeling process that currently involves several transformations to move from student responses to individual test items to implications about growth in a student body. Thus it ties together several distinct areas of active research in the psychometric literature that deal with issues in each transformation in isolation from

the others. The integrated approach provides a leverage point for consideration of the quality of the interaction among multiple components in the process. The further investigation of some of the results of this study may be key to understanding this interaction. For example, the effects of the Discrimination-Censoring Paradox suggest a particular relationship between the characteristics of measurement at the item level and the resulting quantification of growth, as mediated by the latent vertical scale. Thus, continued research of the item response growth model concept will have implications for ongoing research in item response modeling, vertical scaling, and growth modeling in other frameworks.

The present study answered two small but important questions regarding the quality of the estimates produced using an integrated model approach. The resolution of these questions lays a necessary foundation for future studies to follow. By providing a foundation for research addressing the entire growth modeling process, this study presents yet another step toward resolving nagging questions about the source of methodological ambiguity in the use of growth modeling for educational accountability.

Subsequent research would compare the integrated approach, which incorporates the measurement model and the growth model in a single model, with the current approach used in practice, which separates the measurement model estimation from the growth model estimation. One important point of comparison, for example, will be to compare the two different approaches under less than ideal conditions to evaluate their relative robustness. By avoiding multiple transformations it is anticipated that the integrated approach will provide more accurate estimates of

the variability in students' academic growth over time. If this suggestion is supported by the research results, then practical outcomes of applying an integrated item response growth model would include more accurate reports of individual students' academic growth as well as more accurate results of research studies examining the effects of additional variables on learning. Increased understanding of how to draw valid inferences about students' academic growth and change over time lends greater credibility to the study of initiatives intended to affect learning and has direct implications for discussions of educational accountability.

Appendix: Means and Standard Deviations for Bias and Error of Estimate

This appendix provides tables of means and standard deviations for the bias and error of estimate for the item response growth model parameters in the Monte Carlo investigation of parameter recovery. The mean values provided in the tables support reported conclusions about the direction of statistically significant effects and the nature of statistically significant interactions in the general linear model analyses.

Study 1

Table A1: Means and Standard Deviations (SD) for Bias of Item Location Parameter Estimates for Study 1

Sample size	Item selection	Mean	SD
500		0.65	2.60
1000		0.82	2.88
2000		0.85	2.94
	full100	0.07	0.25
	target50	1.11	3.35
500	full100	0.09	0.35
500	target50	1.01	3.27
1000	full100	0.06	0.21
1000	target50	1.12	3.37
2000	full100	0.05	0.14
2000	target50	1.14	3.39

Table A2: Means and Standard Deviations (SD) for Error of Estimate of Item Location Parameter Estimates for Study 1

Sample size	Item selection	Mean	SD
500		0.49	0.72
1000		0.34	0.53
2000		0.25	0.42
	full100	0.13	0.17
	target50	0.44	0.64
500	full100	0.20	0.24
500	target50	0.68	0.85
1000	full100	0.12	0.13
1000	target50	0.43	0.60
2000	full100	0.08	0.08
2000	target50	0.31	0.48

Table A3: Means and Standard Deviations (SD) for Bias of Item Discrimination Parameter Estimates for Study 1

Sample size	Item selection	Mean	SD
500		-0.39	0.37
1000		-0.47	0.37
2000		-0.48	0.37
	full100	-0.06	0.10
	target50	-0.63	0.32
500	full100	-0.07	0.13
500	target50	-0.59	0.34
1000	full100	-0.06	0.09
1000	target50	-0.63	0.31
2000	full100	-0.05	0.06
2000	target50	-0.64	0.31

Table A4: Means and Standard Deviations (SD) for Error of Estimate of Item Discrimination Parameter Estimates for Study 1

Sample size	Item selection	Mean	SD
500		0.11	0.12
1000		0.07	0.08
2000		0.05	0.07
	full100	0.07	0.07
	target50	0.07	0.10
500	full100	0.10	0.08
500	target50	0.11	0.14
1000	full100	0.07	0.05
1000	target50	0.07	0.09
2000	full100	0.04	0.04
2000	target50	0.05	0.08

Table A5: Means and Standard Deviations (SD) for Bias of Random Effects (Growth) Parameter Estimates for Study 1

Sample size	Item selection	Parameter	Mean	SD
500			0.32	0.59
1000			0.49	0.72
2000			0.51	0.70
	full100		0.03	0.05
	target50		1.01	0.72
500	full100		0.04	0.07
500	target50		0.88	0.73
1000	full100		0.03	0.04
1000	target50		1.05	0.75
2000	full100		0.02	0.02
2000	target50		1.05	0.69
		TauD	0.82	1.01
		TauTD	0.24	0.26
		mudelta	0.29	0.31
500		TauD	0.57	0.90
500		TauTD	0.17	0.24
500		mudelta	0.23	0.29
1000		TauD	0.90	1.06
1000		TauTD	0.26	0.27
1000		mudelta	0.30	0.31

2000		TauD	0.95	1.01
2000		TauTD	0.27	0.27
2000		mudelta	0.32	0.32
	full100	TauD	0.03	0.05
	full100	TauTD	0.02	0.03
	full100	mudelta	0.03	0.05
	target50	TauD	1.89	0.61
	target50	TauTD	0.52	0.14
	target50	mudelta	0.62	0.14
500	full100	TauD	0.04	0.08
500	full100	TauTD	0.03	0.04
500	full100	mudelta	0.05	0.08
500	target50	TauD	1.60	0.86
500	target50	TauTD	0.45	0.20
500	target50	mudelta	0.57	0.22
1000	full100	TauD	0.02	0.04
1000	full100	TauTD	0.02	0.03
1000	full100	mudelta	0.03	0.05
1000	target50	TauD	1.98	0.59
1000	target50	TauTD	0.54	0.13
1000	target50	mudelta	0.63	0.12
2000	full100	TauD	0.02	0.02
2000	full100	TauTD	0.02	0.02

2000	full100	mudelta	0.03	0.03
2000	target50	TauD	1.96	0.38
2000	target50	TauTD	0.54	0.09
2000	target50	mudelta	0.64	0.10

Table A6: Means and Standard Deviations (SD) for Error of Estimate of Random Effects (Growth) Parameter Estimates for Study 1

Sample size	Item selection	Parameter	Mean	SD
500			0.15	0.27
1000			0.11	0.21
2000			0.07	0.14
	full100		0.03	0.04
	target50		0.21	0.29
500	full100		0.05	0.05
500	target50		0.34	0.39
1000	full100		0.03	0.03
1000	target50		0.21	0.29
2000	full100		0.02	0.02
2000	target50		0.13	0.19
		TauD	0.20	0.33
		TauTD	0.06	0.07
		mudelta	0.06	0.08
500		TauD	0.27	0.42
500		TauTD	0.08	0.09
500		mudelta	0.09	0.11
1000		TauD	0.21	0.33
1000		TauTD	0.06	0.07
1000		mudelta	0.06	0.06

2000		TauD	0.14	0.22
2000		TauTD	0.04	0.05
2000		mudelta	0.04	0.06
	full100	TauD	0.03	0.04
	full100	TauTD	0.02	0.02
	full100	mudelta	0.04	0.04
	target50	TauD	0.43	0.41
	target50	TauTD	0.10	0.09
	target50	mudelta	0.10	0.10
500	full100	TauD	0.05	0.06
500	full100	TauTD	0.03	0.03
500	full100	mudelta	0.05	0.05
500	target50	TauD	0.69	0.50
500	target50	TauTD	0.17	0.11
500	target50	mudelta	0.16	0.14
1000	full100	TauD	0.03	0.03
1000	full100	TauTD	0.02	0.02
1000	full100	mudelta	0.03	0.03
1000	target50	TauD	0.43	0.40
1000	target50	TauTD	0.10	0.08
1000	target50	mudelta	0.09	0.08
2000	full100	TauD	0.02	0.01
2000	full100	TauTD	0.02	0.01

2000	full100	mudelta	0.02	0.02
2000	target50	TauD	0.27	0.26
2000	target50	TauTD	0.06	0.07
2000	target50	mudelta	0.07	0.07

Study 2

Table A7: Means and Standard Deviations (SD) for Bias of Item Location Parameter Estimates for Study 2

Test length	Mean	SD
8	0.08	0.35
16	0.17	0.36

Table A8: Means and Standard Deviations (SD) for Error of Estimate of Item Location Parameter Estimates for Study 2

Test length	Mean	SD
8	0.20	0.24
16	0.16	0.16

Table A9: Means and Standard Deviations (SD) for Bias of Item Discrimination Parameter Estimates for Study 2

Test length	Mean	SD
8	-0.07	0.13
16	-0.14	0.10

Table A10: Means and Standard Deviations (SD) for Error of Estimate of Item Discrimination Parameter Estimates for Study 2

Test length	Mean	SD
8	0.10	0.08
16	0.08	0.06

Table A11: Means and Standard Deviations (SD) for Bias of Random Effects (Growth) Parameter Estimates for Study 2

Test length	Parameter	Mean	SD
8		0.04	0.07
16		0.06	0.05
	TauD	0.05	0.07
	TauTD	0.03	0.04
	mudelta	0.06	0.07
8	TauD	0.04	0.08
8	TauTD	0.03	0.04
8	mudelta	0.05	0.08
1000	TauD	0.07	0.05
2000	TauTD	0.03	0.03
2000	mudelta	0.08	0.05

Table A12: Means and Standard Deviations (SD) for Error of Estimate of Random Effects (Growth) Parameter Estimates for Study 2

Test length	Parameter	Mean	SD
8		0.05	0.05
16		0.04	0.03
	TauD	0.05	0.05
	TauTD	0.03	0.02
	mudelta	0.05	0.04
8	TauD	0.05	0.06
8	TauTD	0.03	0.03
8	mudelta	0.05	0.05
1000	TauD	0.04	0.03
2000	TauTD	0.02	0.02
2000	mudelta	0.04	0.03

Study 3

Table A13: Means and Standard Deviations (SD) for Bias of Item Location Parameter Estimates for Study 3

Trajectory	Sample size	Rate variance	Mean	SD
Linear			0.11	0.49
Non-linear			0.04	0.24
	500		0.09	0.48
	1000		0.06	0.26
Linear	500		0.14	0.62
Linear	1000		0.08	0.32
Non-linear	500		0.04	0.28
Non-linear	1000		0.03	0.18
		0.20	0.05	0.26
		0.50	0.10	0.49
Linear		0.20	0.07	0.29
Linear		0.50	0.15	0.64
Non-linear		0.20	0.03	0.22
Non-linear		0.50	0.04	0.26
	500	0.20	0.06	0.31
	500	0.50	0.12	0.61
	1000	0.20	0.05	0.19
	1000	0.50	0.07	0.31
Linear	500	0.20	0.08	0.35

Linear	500	0.50	0.20	0.80
Linear	1000	0.20	0.06	0.21
Linear	1000	0.50	0.11	0.40
Non-linear	500	0.20	0.04	0.25
Non-linear	500	0.50	0.05	0.31
Non-linear	1000	0.20	0.03	0.18
Non-linear	1000	0.50	0.04	0.19

Table A14: Means and Standard Deviations (SD) for Error of Estimate of Item Location Parameter Estimates for Study 3

Trajectory	Sample size	Rate variance	Mean	SD
Linear			0.22	0.38
Non-linear			0.14	0.16
	500		0.23	0.36
	1000		0.13	0.19
Linear	500		0.29	0.47
Linear	1000		0.15	0.24
Non-linear	500		0.17	0.18
Non-linear	1000		0.11	0.12
		0.20	0.14	0.17
		0.50	0.21	0.38
Linear		0.20	0.16	0.20
Linear		0.50	0.28	0.50
Non-linear		0.20	0.13	0.14
Non-linear		0.50	0.15	0.17
	500	0.20	0.18	0.20
	500	0.50	0.29	0.47
	1000	0.20	0.11	0.12
	1000	0.50	0.14	0.24
Linear	500	0.20	0.20	0.24
Linear	500	0.50	0.39	0.61

Linear	1000	0.20	0.12	0.13
Linear	1000	0.50	0.17	0.31
Non-linear	500	0.20	0.15	0.16
Non-linear	500	0.50	0.18	0.20
Non-linear	1000	0.20	0.10	0.11
Non-linear	1000	0.50	0.12	0.12

Table A15: Means and Standard Deviations (SD) for Bias of Item Discrimination Parameter Estimates for Study 3

Trajectory	Sample size	Rate variance	Mean	SD
Linear			-0.09	0.14
Non-linear			-0.06	0.11
	500		-0.08	0.15
	1000		-0.06	0.10
Linear	500		-0.10	0.17
Linear	1000		-0.07	0.10
Non-linear	500		-0.07	0.13
Non-linear	1000		-0.05	0.09
		0.20	-0.06	0.11
		0.50	-0.09	0.14
Linear		0.20	-0.06	0.11
Linear		0.50	-0.11	0.16
Non-linear		0.20	-0.06	0.10
Non-linear		0.50	-0.06	0.12
	500	0.20	-0.07	0.13
	500	0.50	-0.10	0.17
	1000	0.20	-0.06	0.09
	1000	0.50	-0.07	0.11
Linear	500	0.20	-0.07	0.13
Linear	500	0.50	-0.13	0.20

Linear	1000	0.20	-0.06	0.09
Linear	1000	0.50	-0.09	0.12
Non-linear	500	0.20	-0.07	0.12
Non-linear	500	0.50	-0.07	0.14
Non-linear	1000	0.20	-0.06	0.08
Non-linear	1000	0.50	-0.05	0.10

Table A16: Means and Standard Deviations (SD) for Error of Estimate of Item Discrimination Parameter Estimates for Study 3

Trajectory	Sample size	Rate variance	Mean	SD
Linear			0.10	0.09
Non-linear			0.08	0.07
	500		0.11	0.10
	1000		0.07	0.06
Linear	500		0.13	0.11
Linear	1000		0.08	0.07
Non-linear	500		0.10	0.08
Non-linear	1000		0.07	0.06
		0.20	0.08	0.07
		0.50	0.11	0.09
Linear		0.20	0.09	0.07
Linear		0.50	0.12	0.11
Non-linear		0.20	0.08	0.06
Non-linear		0.50	0.09	0.07
	500	0.20	0.10	0.08
	500	0.50	0.13	0.11
	1000	0.20	0.07	0.05
	1000	0.50	0.08	0.07
Linear	500	0.20	0.10	0.08
Linear	500	0.50	0.16	0.13

Linear	1000	0.20	0.07	0.05
Linear	1000	0.50	0.09	0.08
Non-linear	500	0.20	0.09	0.07
Non-linear	500	0.50	0.11	0.08
Non-linear	1000	0.20	0.07	0.05
Non-linear	1000	0.50	0.07	0.06

Table A17: Means and Standard Deviations (SD) for Bias of Random Effects (Growth) Parameter Estimates for Study 3

Trajectory	Sample size	Rate variance	Parameter	Mean	SD
Linear				0.07	0.26
Non-linear				0.02	0.07
Linear			TauD	0.12	0.42
Linear			TauTD	0.04	0.08
Linear			mudelta	0.06	0.12
Non-linear			TauD1	0.03	0.10
Non-linear			TauD2	0.02	0.11
Non-linear			TauDD	0.02	0.05
Non-linear			TauTD1	0.05	0.07
Non-linear			TauTD2	0.00	0.05
Non-linear			mudelt1	0.03	0.06
Non-linear			mudelt2	0.01	0.05
	500			0.05	0.20
	1000			0.03	0.10
Linear	500			0.10	0.33
Linear	1000			0.05	0.15
Non-linear	500			0.03	0.09
Non-linear	1000			0.02	0.06
Linear	500		TauD	0.17	0.54
Linear	500		TauTD	0.05	0.09

Linear	500		mudelta	0.08	0.15
Linear	1000		TauD	0.07	0.24
Linear	1000		TauTD	0.03	0.06
Linear	1000		mudelta	0.04	0.08
Non-linear	500		TauD1	0.04	0.12
Non-linear	500		TauD2	0.03	0.13
Non-linear	500		TauDD	0.02	0.05
Non-linear	500		TauTD1	0.05	0.08
Non-linear	500		TauTD2	0.00	0.05
Non-linear	500		mudelt1	0.04	0.08
Non-linear	500		mudelt2	0.02	0.05
Non-linear	1000		TauD1	0.02	0.07
Non-linear	1000		TauD2	0.02	0.08
Non-linear	1000		TauDD	0.01	0.04
Non-linear	1000		TauTD1	0.04	0.06
Non-linear	1000		TauTD2	0.00	0.04
Non-linear	1000		mudelt1	0.03	0.05
Non-linear	1000		mudelt2	0.01	0.04
		0.20		0.02	0.05
		0.50		0.06	0.21
Linear		0.20		0.03	0.06
Linear		0.50		0.12	0.36
Non-linear		0.20		0.01	0.05

Non-linear		0.50		0.03	0.09
Linear		0.20	TauD	0.03	0.06
Linear		0.20	TauTD	0.03	0.04
Linear		0.20	mudelta	0.04	0.06
Linear		0.50	TauD	0.21	0.58
Linear		0.50	TauTD	0.05	0.10
Linear		0.50	mudelta	0.08	0.15
Non-linear		0.20	TauD1	0.01	0.04
Non-linear		0.20	TauD2	0.00	0.05
Non-linear		0.20	TauDD	0.01	0.03
Non-linear		0.20	TauTD1	0.04	0.06
Non-linear		0.20	TauTD2	0.00	0.04
Non-linear		0.20	mudelt1	0.03	0.05
Non-linear		0.20	mudelt2	0.01	0.04
Non-linear		0.50	TauD1	0.05	0.13
Non-linear		0.50	TauD2	0.04	0.14
Non-linear		0.50	TauDD	0.02	0.06
Non-linear		0.50	TauTD1	0.05	0.08
Non-linear		0.50	TauTD2	0.00	0.05
Non-linear		0.50	mudelt1	0.03	0.07
Non-linear		0.50	mudelt2	0.02	0.05
	500	0.20		0.02	0.06
	500	0.50		0.07	0.27

	1000	0.20		0.02	0.04
	1000	0.50		0.04	0.13
Linear	500	0.20	TauD	0.04	0.08
Linear	500	0.20	TauTD	0.03	0.04
Linear	500	0.20	mudelta	0.05	0.08
Linear	500	0.50	TauD	0.30	0.75
Linear	500	0.50	TauTD	0.06	0.12
Linear	500	0.50	mudelta	0.11	0.19
Linear	1000	0.20	TauD	0.02	0.04
Linear	1000	0.20	TauTD	0.02	0.03
Linear	1000	0.20	mudelta	0.03	0.05
Linear	1000	0.50	TauD	0.13	0.33
Linear	1000	0.50	TauTD	0.04	0.08
Linear	1000	0.50	mudelta	0.06	0.09
Non-linear	500	0.20	TauD1	0.01	0.05
Non-linear	500	0.20	TauD2	0.00	0.06
Non-linear	500	0.20	TauDD	0.01	0.03
Non-linear	500	0.20	TauTD1	0.05	0.06
Non-linear	500	0.20	TauTD2	-0.01	0.05
Non-linear	500	0.20	mudelt1	0.03	0.07
Non-linear	500	0.20	mudelt2	0.01	0.04
Non-linear	500	0.50	TauD1	0.06	0.15
Non-linear	500	0.50	TauD2	0.06	0.18

Non-linear	500	0.50	TauDD	0.02	0.07
Non-linear	500	0.50	TauTD1	0.06	0.09
Non-linear	500	0.50	TauTD2	0.00	0.06
Non-linear	500	0.50	mudelt1	0.04	0.09
Non-linear	500	0.50	mudelt2	0.02	0.06
Non-linear	1000	0.20	TauD1	0.00	0.03
Non-linear	1000	0.20	TauD2	0.00	0.04
Non-linear	1000	0.20	TauDD	0.01	0.03
Non-linear	1000	0.20	TauTD1	0.04	0.05
Non-linear	1000	0.20	TauTD2	0.00	0.04
Non-linear	1000	0.20	mudelt1	0.03	0.04
Non-linear	1000	0.20	mudelt2	0.01	0.03
Non-linear	1000	0.50	TauD1	0.03	0.09
Non-linear	1000	0.50	TauD2	0.03	0.10
Non-linear	1000	0.50	TauDD	0.02	0.04
Non-linear	1000	0.50	TauTD1	0.04	0.06
Non-linear	1000	0.50	TauTD2	0.00	0.05
Non-linear	1000	0.50	mudelt1	0.03	0.05
Non-linear	1000	0.50	mudelt2	0.01	0.04

Table A18: Means and Standard Deviations (SD) for Error of Estimate of Random Effects (Growth) Parameter Estimates for Study 3

Trajectory	Sample size	Rate variance	Parameter	Mean	SD
Linear				0.08	0.23
Non-linear				0.05	0.05
Linear			TauD	0.15	0.38
Linear			TauTD	0.04	0.06
Linear			mudelta	0.06	0.09
Non-linear			TauD1	0.06	0.07
Non-linear			TauD2	0.07	0.08
Non-linear			TauDD	0.03	0.03
Non-linear			TauTD1	0.05	0.04
Non-linear			TauTD2	0.04	0.03
Non-linear			mudelt1	0.05	0.04
Non-linear			mudelt2	0.04	0.03
	500			0.08	0.17
	1000			0.04	0.08
Linear	500			0.12	0.30
Linear	1000			0.05	0.14
Non-linear	500			0.06	0.06
Non-linear	1000			0.04	0.04
Linear	500		TauD	0.22	0.48
Linear	500		TauTD	0.05	0.07

Linear	500		mudelta	0.09	0.11
Linear	1000		TauD	0.08	0.22
Linear	1000		TauTD	0.03	0.05
Linear	1000		mudelta	0.04	0.06
Non-linear	500		TauD1	0.08	0.08
Non-linear	500		TauD2	0.09	0.09
Non-linear	500		TauDD	0.04	0.04
Non-linear	500		TauTD1	0.06	0.05
Non-linear	500		TauTD2	0.04	0.03
Non-linear	500		mudelt1	0.06	0.05
Non-linear	500		mudelt2	0.04	0.03
Non-linear	1000		TauD1	0.05	0.05
Non-linear	1000		TauD2	0.06	0.05
Non-linear	1000		TauDD	0.03	0.02
Non-linear	1000		TauTD1	0.04	0.04
Non-linear	1000		TauTD2	0.03	0.03
Non-linear	1000		mudelt1	0.04	0.03
Non-linear	1000		mudelt2	0.03	0.02
		0.20		0.04	0.03
		0.50		0.08	0.19
Linear		0.20		0.04	0.04
Linear		0.50		0.13	0.32
Non-linear		0.20		0.04	0.03

Non-linear		0.50		0.06	0.06
Linear		0.20	TauD	0.04	0.05
Linear		0.20	TauTD	0.03	0.03
Linear		0.20	mudelta	0.04	0.04
Linear		0.50	TauD	0.26	0.51
Linear		0.50	TauTD	0.05	0.08
Linear		0.50	mudelta	0.09	0.12
Non-linear		0.20	TauD1	0.03	0.03
Non-linear		0.20	TauD2	0.04	0.03
Non-linear		0.20	TauDD	0.02	0.02
Non-linear		0.20	TauTD1	0.04	0.03
Non-linear		0.20	TauTD2	0.03	0.03
Non-linear		0.20	mudelt1	0.04	0.04
Non-linear		0.20	mudelt2	0.03	0.02
Non-linear		0.50	TauD1	0.09	0.08
Non-linear		0.50	TauD2	0.11	0.10
Non-linear		0.50	TauDD	0.04	0.04
Non-linear		0.50	TauTD1	0.06	0.05
Non-linear		0.50	TauTD2	0.04	0.03
Non-linear		0.50	mudelt1	0.05	0.05
Non-linear		0.50	mudelt2	0.04	0.03
	500	0.20		0.04	0.04
	500	0.50		0.11	0.24

	1000	0.20		0.03	0.02
	1000	0.50		0.06	0.11
Linear	500	0.20	TauD	0.05	0.06
Linear	500	0.20	TauTD	0.03	0.03
Linear	500	0.20	mudelta	0.05	0.05
Linear	500	0.50	TauD	0.38	0.64
Linear	500	0.50	TauTD	0.08	0.09
Linear	500	0.50	mudelta	0.12	0.15
Linear	1000	0.20	TauD	0.03	0.03
Linear	1000	0.20	TauTD	0.02	0.02
Linear	1000	0.20	mudelta	0.03	0.03
Linear	1000	0.50	TauD	0.13	0.30
Linear	1000	0.50	TauTD	0.03	0.07
Linear	1000	0.50	mudelta	0.05	0.08
Non-linear	500	0.20	TauD1	0.04	0.03
Non-linear	500	0.20	TauD2	0.05	0.03
Non-linear	500	0.20	TauDD	0.03	0.02
Non-linear	500	0.20	TauTD1	0.05	0.04
Non-linear	500	0.20	TauTD2	0.04	0.03
Non-linear	500	0.20	mudelt1	0.05	0.04
Non-linear	500	0.20	mudelt2	0.04	0.03
Non-linear	500	0.50	TauD1	0.12	0.10
Non-linear	500	0.50	TauD2	0.13	0.11

Non-linear	500	0.50	TauDD	0.05	0.04
Non-linear	500	0.50	TauTD1	0.07	0.06
Non-linear	500	0.50	TauTD2	0.05	0.04
Non-linear	500	0.50	mudelt1	0.07	0.06
Non-linear	500	0.50	mudelt2	0.05	0.04
Non-linear	1000	0.20	TauD1	0.03	0.02
Non-linear	1000	0.20	TauD2	0.03	0.02
Non-linear	1000	0.20	TauDD	0.02	0.02
Non-linear	1000	0.20	TauTD1	0.04	0.03
Non-linear	1000	0.20	TauTD2	0.03	0.02
Non-linear	1000	0.20	mudelt1	0.03	0.03
Non-linear	1000	0.20	mudelt2	0.03	0.02
Non-linear	1000	0.50	TauD1	0.07	0.06
Non-linear	1000	0.50	TauD2	0.08	0.07
Non-linear	1000	0.50	TauDD	0.03	0.03
Non-linear	1000	0.50	TauTD1	0.05	0.04
Non-linear	1000	0.50	TauTD2	0.04	0.03
Non-linear	1000	0.50	mudelt1	0.04	0.03
Non-linear	1000	0.50	mudelt2	0.03	0.02

Glossary

Achievement: describes a measure of knowledge, skills, or abilities at the time of an assessment, as opposed to measures of academic aptitude or non-cognitive areas, such as attitude.

Change: a difference in the quantity or quality of a measurement in a domain

Common item design: an item selection scheme in which a subset of items on one test is incorporated into another test for the purpose of linking scores on the tests. The common items are also sometimes called anchor items.

Growth: movement or change along a continuum on which progress in a domain can be measured.

Item discrimination parameter: parameterizes the item's capacity to distinguish between examinees of ability below and above the difficulty level of the item.

Item location parameter: parameterizes the relative difficulty of an item; the higher the value, the more difficult the item; also sometimes called the item difficulty.

Latent variable: any variable that is not directly observed but is hypothesized as part of a structural model.

Nonlinear mixed effects model: a multilevel model wherein effects may have both a fixed and a random component. Random components vary over units in the study.

Repeated measures data: data that are measured on the same examinees across time points; also known as panel data

Trajectory: Raudenbush (2001) provides an excellent definition: "Whereas a person's history can capture many domains of change – for example, changes in cognitive skill, emotional self-regulation, mood, and social behavior – *a trajectory describes a person's development in one well-defined domain*" (italics added; p. 502).

References

Adams, R. J., Wilson, M., & Wang, W. (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.

Arce-Ferrer, A., Frisbie, D. A., & Kolen, M. J. (2002). Standard errors of proportions used in reporting changes in school performance with achievement levels. *Educational Assessment, 8*, 59.

Baker, G. A. (1954). Factor analysis of relative growth. *Growth, 18*, 137-143.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3-20). Madison: University of Wisconsin Press.

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155-169). New York: Taylor & Francis.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, N.J: Wiley-Interscience.

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

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*, 147-158.

Collins, L. M., & Horn, J. L. (1991). *Best methods for the analysis of change: Recent advances, unanswered questions, future directions*. Washington, DC: American Psychological Association.

Collins, L. M., & Sayer, A. (2001). *New methods for the analysis of change*. Washington, DC: American Psychological Association.

Coull, B. A., Houseman, E. A., & Betensky, R. A. (2006). A computationally tractable multivariate random effects model for clustered binary data. *Biometrika, 93*, 587.

Cronbach, L. J., & Furby, L. (1970). How we should measure 'change': Or should we? *Psychological Bulletin, 74*, 68-80.

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

Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., & Offen, W. (2005). *Analysis of clinical trials using SAS: A practical guide*. Cary, NC: SAS Publishing.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (Eds.). (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, N.J: Lawrence Erlbaum Associates.

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

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

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. New York: Chichester: Wiley-Interscience.

Goldschmidt, P., Choi, K., & Martinez, F. (2004). *Using hierarchical growth models to monitor school performance over time: Comparing NCE to scale score results* (Report No. 618). Los Angeles, CA: CRESST/University of California, Los Angeles.

Gottman, J. M. (1995). *The analysis of change*. Mahwah, N.J: L. Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hancock, G. R., Kuo, W., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling, 8*, 470-489.

Hancock, G. R., & Lawrence, F. R. (2006). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 171-196). Greenwood, CT: Information Age Publishing, Inc.

Hanley, J. A., Negassa, A., Edwardes, M. D. d., & Forrester, J. E. (2003). Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology, 157*, 364-375.

Harris, C. W. (1963). *Problems in measuring change*. Madison: University of Wisconsin Press.

Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality & Quantity, 24*, 387.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36*, 347-387.

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

Lee, S., Poon, W., & Bentler, P. M. (1990). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics & Probability Letters, 9*, 91-97.

Lee, S., Poon, W., & Bentler, P. M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika, 57*, 89-105.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society.Series B (Methodological), 58*, 619-678.

Leite, W. L. (2007). A comparison of latent growth models for constructs measured by multiple items. *Structural Equation Modeling, 14*, 581-610.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21-38). Madison: University of Wisconsin Press.

McArdle, J. J., Grimm, K., Hamagami, F., Bowles, R., & Meredith, W. (in press). Modeling lifespan growth curves of cognition using longitudinal data with changing measures. *Psychological Methods*.

Mehta, P. D., Neale, M. C., & Flay, B. R. (2004). Squeezing interval change from ordinal panel data: Latent growth curves with ordinal outcomes. *Psychological Methods, 9*, 301-333.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107-122.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115-132.

Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205-234). Newbury Park, CA: Sage.

Muthén, B. O. (1996). Growth modeling with binary responses. In A. von Eye & C. C. Clogg (Eds.), *Categorical variables in developmental research: Methods of analysis* (pp. 37-54). San Diego: Academic Press.

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (Report No. 4).

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Retrieved from http://www.gseis.ucla.edu/faculty/muthen/categorical.htmfrom www.statmodel.com

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599-620.

NCLB. (2002). No Child Left Behind Act of 2001 (Vol. 20, pp. 6301 et seq.).

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443-460.

Peterson, P. E. (2007). A lens that distorts. *Education Next, 7*, 46-51.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics, 4*, 12-35.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: Sage Publications.

Raudenbush, S. W., Johnson, C., & Sampson, R. J. (2003). A multivariate, multilevel Rasch model with application to self-reported criminal behavior. *Sociological Methodology, 33*, 169-211.

Roberts, J. S., & Ma, Q. (2006). IRT models for the assessment of change across repeated measurements. In R. W. Lissitz (Ed.), *Longitudinal and value added modeling of student performance* (pp. 100-127). Maple Grove, MN: JAM Press.

Rock, D. A., & Pollack, J. M. (2002). *Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K) psychometric report for kindergarten through first grade* (Report No. 2002–05). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Rogosa, D. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3-66). Mahwah, N.J.: Lawrence Erlbaum Associates.

SAS Institute. (1999). *SAS/STAT user's guide version 8*. Cary, N.C.: SAS Institute, Inc.

Sayer, A., & Cumsille, P. E. (2001). Second-order latent growth models. In L. M. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 179-200). Washington, DC: American Psychological Association.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika, 78*, 719-727.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis, 16*, 41-49.

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

Thum, Y. M. (2003). *No child left behind: Methodological challenges & recommendations for measuring adequate yearly progress* (Report No. 590). Los Angeles, CA: Center for the Study of Evaluation (CSE), National Center for Research on Evaluation,Standards, and Student Testing (CRESST), Graduate School of Education & Information Studies, University of California, Los Angeles.

Tourangeau, K., Nord, C., Lê, T., Pollack, J. M., & Atkins-Burnett, S. (2006). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) combined user's manual for the ECLS-K fifth-grade data files and electronic*

codebooks (Report No. 2006–032). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika, 23*, 19-23.

Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D. C., Van den Noortgate, W., et al. (2004). Estimation and software. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 343-373). New York: Springer.

Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology, 59*, 225-255.

U. S. Department of Education. (2007, July 3). Secretary Spellings approves additional growth model pilots for 2006-2007 school year. Retrieved from http://www.ed.gov/news/pressreleases/2007/07/07032007.html

Wang, W., Wilson, M., & Adams, R. J. (1998). Measuring individual differences in change with Rasch models. *Journal of Outcome Measurement, 2*, 240-265.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education, 15*, 345-422.

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

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58-79.

Wishart, J. (1938). Growth-rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrika, 30*, 16-28.