

# Modeling Growth in Student Achievement: Psychometric Considerations, Communicating Growth, and Standards-based Applications

**Harold C. Doran**

American Institutes for Research

Computer and Statistical Sciences Center

[hdoran@air.org](mailto:hdoran@air.org)

## Introduction and Objectives

- Discuss the precision of gain scores obtained from vertically linked scales;
- Discuss applications of Literate Statistical Computing (LSC) and how this concept can be applied to create variable information growth reports;
- Discuss a method for judging the adequacy of student growth trajectories within a standards-based context; and
- Scalable software for estimating models with crossed random effects.

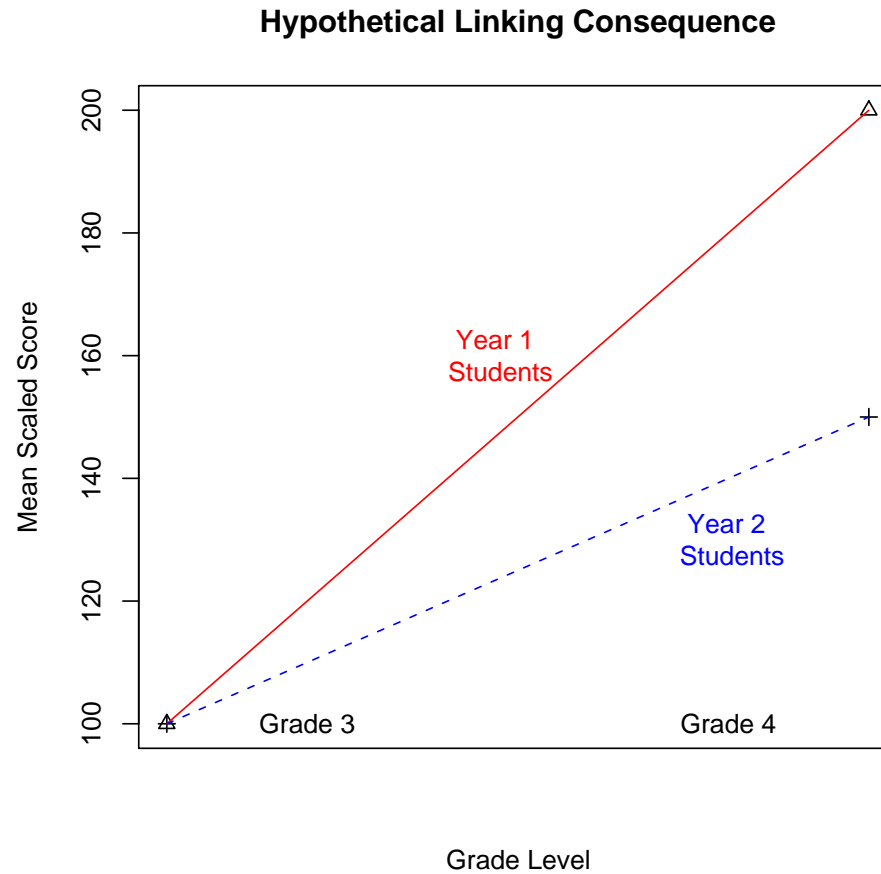
## Vertical Linking: An Overview of the Issues

- A vertically linked scale is commonly identified as a prerequisite for measuring change
- This process places the scores across test forms on a common developmental scale
- However, the processes used to construct vertical scales introduce a new variance component due to the sampling of common items with item parameters obtained from a sample of students
- This variance component is commonly ignored in longitudinal models treating the vertical trait as if it were measured with exact precision
- Consequently, the standard errors of growth parameters are commonly underestimated

## How Is Growth Operationalized?

- The within-grade trait,  $\theta$ , is often well defined in a test blueprint and is based on the content standards and performance objectives
- Measuring growth is somewhat ambiguous - no test blueprint and often no clear conceptualization of what it actually is
- Test developers sample test items from a population of (hypothetical) common items thought to operationalize this vertical trait
- Assume a teacher emphasizes geometry more than algebra. Now assume in Year 1 the vertical trait is more sensitive to geometry but in Year 2 the vertical trait is more sensitive to algebra.
- The average score in Year 1 (the geometry year) would be higher than the average score in Year 2 (the algebra year)

# Hypothetical Consequences For Same Grade 4 Teacher



## Why Not Ignore Vertical Linking?

- One might ask, “Why not use other scores, such as raw or a standardized metric like NCEs?”
- Standardized metrics only establish numerical equivalences, they do not create equivalences in the measurement of a latent trait.
- For example, you can standardize height and weight to be on the same scale, but you’re not measuring the same thing
- Even if other scores are used this issue still exists. Measuring change assumes growth has been properly defined and operationalized.

## Vertical Linking

- There are many techniques that can be used to create a vertical scale (Kolen & Brennan, 2004)
- For the current discussion we will limit our attention to the common item non-equivalent groups design
- For other discussions of equating variance see Michaelides & Haertel (2004), Cohen, Johnson, & Angeles (2001), von Davier & Kong (2005), Sheehan & Mislevy (1988).

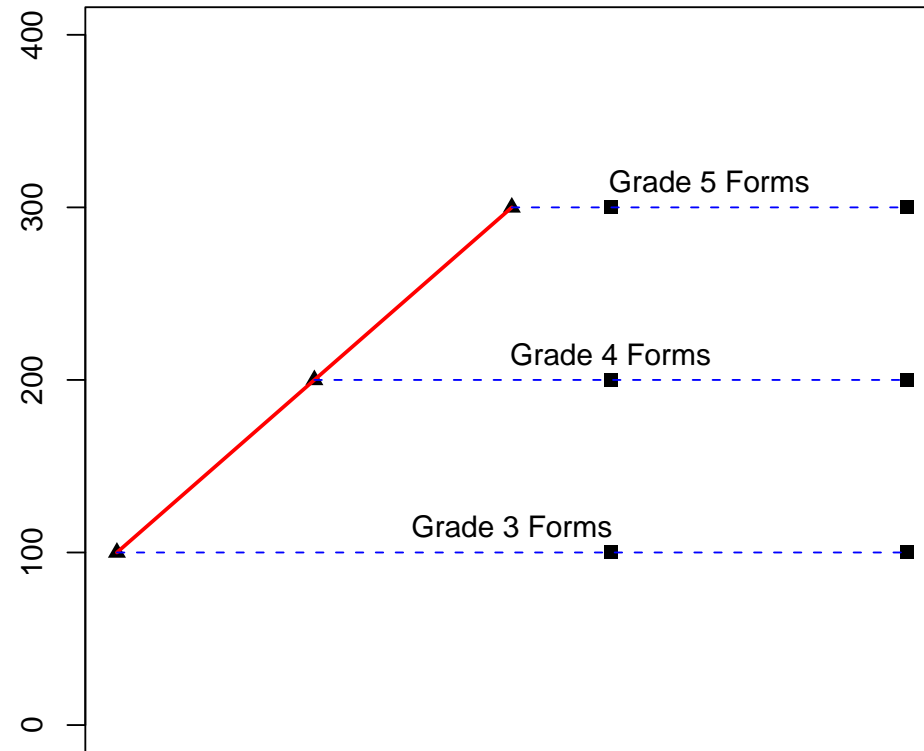
## The Common Item Non-Equivalent Groups Design

- Embed common items across adjacent test forms (e.g., grade 3 and 4 share similar “anchor” items)
- Estimate item parameters for the common items from an IRT model (either separately or jointly)

$$P(x = 1|\theta, b, a) = \frac{1}{1 + e^{-Da(\theta-b)}} \quad (1)$$

- Under separate calibration one obtains two sets of parameter estimates, one set for the base grade and another for the form in the next higher grade

# Equating and Linking Approach



## Linking Notation

- Let  $\mu(b_g)$  represent the mean of  $\mathbf{b}_g = (b_{g1}, \dots, b_{gn})'$ , the difficulty parameters for the common items on the base test form
- Let  $\mu(b_{g+1})$  represent the mean of  $\mathbf{b}_{g+1} = (b_{g+1,1}, \dots, b_{g+1,n})'$ , the difficulty parameters for the common items on the test form in the next higher grade
- Let  $\mu(a_g)$  represent the mean of  $\mathbf{a}_g = (a_{g1}, \dots, a_{gn})'$ , the discrimination parameters for the common items on the base test form
- Let  $\mu(a_{g+1})$  represent the mean of  $\mathbf{a}_{g+1} = (a_{g+1,1}, \dots, a_{g+1,n})'$ , the discrimination parameters for the common items on the test form in the next higher grade

## General Linking Transformation

- The following is the transformation to place the scores on a common scale

$$Y_i^{vl} = \hat{A}\theta_{g+1,i} + \hat{B} \quad (2)$$

where  $\theta_{g+1,i}$  is derived from the within-grade scaling of the test items and the constants for the linking transformation are obtained via the means of the common item parameters:

$$\begin{aligned} \hat{A} &= \frac{\mu(a_g)}{\mu(a_{g+1})} \\ \hat{B} &= \mu(b_g) - \hat{A}\mu(b_{g+1}) \end{aligned} \quad (3)$$

## Rasch Transformation

- In the Rasch model  $a_k = a \forall k$ .
- Therefore, standard Rasch practice is to simply add a constant to the estimate from the within-grade scaling

$$Y_i^{vl} = \theta_{g+1,i} + \hat{B} \quad (4)$$

$$\hat{B} = \mu(b_g) - \mu(b_{g+1}) \quad (5)$$

- Hence the term *mean* equating (and not *mean/mean*)
- Note that the variance of  $Y_i^{vl}$  is a function of its distance from the mean
- In this design all values are equidistant from the mean and the variance is constant across all values of  $\theta$
- This would not be true if  $a_k \neq a$ .  $\hat{A}$  would then be a contributing factor in the linking transformation

## General Issues with Equating Variance

- The vertically linked scaled score is a composite of two estimates: a within-grade trait ( $\theta$ ) and growth measured along this dimension
- Clearly,  $Y_i^{vl}$  depends the estimate of  $\hat{B}$
- However,  $\hat{B}$  is only derived from a sample of common items used to operationalize the vertical trait
- In addition, the parameter estimates for  $\hat{B}$  are derived from a sample of students
- Consequently, if the items that operationalize the growth dimension were more sensitive to different curricular aspects across test forms, then teacher and school effects may fluctuate over time

## An Illustration of the Implications

- Assume the linking process was replicated  $R$  times where each replication used a different set of common items and the item parameters were obtained from a different sample of students.
- Replicating the linking process would result in different realizations of  $\mathbf{B} = (\hat{B}_1, \dots, \hat{B}_R)'$ .
- Relating this back to the linking transformation suggests:

$$\begin{aligned}
 Y_i^{vl} &= \theta_{g+1,i} + \hat{B}_1 \\
 Y_i^{vl} &= \theta_{g+1,i} + \hat{B}_2 \\
 &\vdots \\
 Y_i^{vl} &= \theta_{g+1,i} + \hat{B}_R
 \end{aligned} \tag{6}$$

- The random deviations of  $\hat{B}$  from  $B$  would result in a different linking transformation and different estimates of  $Y_i^{vl}$ .

## Summary of Equating Variance

- In operational testing programs, only a single instance of  $\hat{B}$  is observed.
- $\hat{B}$  is only an estimate of the unknown, true parameter  $B$  and is subject to sampling variance from two sources: the sampling of common items and the sampling of students.
- Therefore, the observed  $\hat{B}$  can be expressed as a linear combination of a true, unknown  $B$  plus random error:

$$\hat{B} = B + \eta_{(l)i} \quad (7)$$

where the notation  $(l)i$  denotes that student  $i$  took form  $l$

## Consequences of Linking Variance in Longitudinal Modeling

- The vertically linked scaled score,  $Y_i^{vl}$ , can be expressed as a linear combination of:

$$\begin{aligned} Y_i^{vl} &= \theta_{g+1,i} + \hat{B} + \epsilon_{g+1,i} \\ &= \theta_{g+1,i} + B + \eta_{(l)i} + \epsilon_{g+1,i} \end{aligned} \quad (8)$$

- However, the score from the base grade is unaffected by the linking transformation and is simply:

$$Y_i = \theta_{gi} + \epsilon_{gi}$$

- Therefore, the difference score is:

$$\begin{aligned} G_i &= Y_i^{vl} - Y_i \\ &= (\theta_{g+1,i} + \hat{B} + \epsilon_{g+1,i}) - (\theta_{gi} + \epsilon_{gi}) \\ &= (\theta_{g+1,i} - \theta_{gi}) + (\epsilon_{g+1,i} - \epsilon_{gi}) + \hat{B} \end{aligned} \quad (9)$$

## Exploring the Consequences of Equating Variance in Longitudinal Models

- Because the difference score contains  $\hat{B}$ , the variance of  $G_i$  must consider the variance component,  $var(\hat{B})$
- In practice, this is difficult to explore because the linking process is performed once and the variance of the linking constants are often not reported
- One way to explore the consequences of linking variance on growth parameters is through the use of a Monte Carlo simulation
- In a simulation we can generate true values of  $\theta$  and  $B$  and explore what would happen if small departures from  $B$  were introduced over many hypothetical replications as discussed in Slide 14.

## Monte Carlo Simulation

1. Generate known “true scores” from a multivariate distribution at various sample sizes (100, 250, 500, 750, 1000)
2. “Contaminate” these true scores by causing the linking constant to vary and create a new set of scores
3. Generate point estimates from a linear growth model for the true scores and the contaminated scores
4. Repeat 250 times at each sample size for both sets of scores
5. Estimate sampling distribution of parameters

## Generate True Scores

- Sample from a multivariate distribution with known parameters to generate true scores

$$\begin{bmatrix} Y_0 \\ Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} \sim \begin{bmatrix} 100 \\ 150 \\ 200 \\ 250 \end{bmatrix} \begin{bmatrix} 400 & 320 & 256 & 204 \\ 320 & 400 & 320 & 256 \\ 256 & 320 & 400 & 320 \\ 204 & 256 & 320 & 400 \end{bmatrix} \quad (10)$$

- Variance was fixed to be constant over time,  $\sigma^2 = 400$
- Correlation was structured to follow an AR(1) pattern,  $\rho = .8, .64, .51$

## Introduce Equating Variance

- The next challenge is to introduce random linking error
- This is accomplished by first sampling from a linking error distribution:

$$\begin{bmatrix} \eta_{(l)0} \\ \eta_{(l)1} \\ \eta_{(l)2} \\ \eta_{(l)3} \end{bmatrix} \sim \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \sigma_{vle.1}^2 & 0 & 0 \\ 0 & 0 & \sigma_{vle.2}^2 & 0 \\ 0 & 0 & 0 & \sigma_{vle.3}^2 \end{bmatrix} \quad (11)$$

- The value realized from this draw is the deviation of the observed  $\hat{B}$  at time  $t$  from its true value,  $B$ , or  $\eta_{(l)t} = \hat{B}_t - B_t$ .
- We then add  $\eta_{(l)t}$  to the true score for each student  

$$Y_{ti}^* = Y_{ti} + \eta_{(l)t}$$

## Choosing the Size of $\sigma_{vle.t}^2$

- One challenge was to choose reasonable values for  $\sigma_{vle.t}^2$
- Other simulation results, as well as another published study by Tsai, Hanson, Kolen, & Forsyth (2001) showed that linking error is approximately  $\frac{1}{10}$  to  $\frac{1}{4}$  of the sampling standard deviation.
- Based on this range, we chose values of  $\sigma_{vle}^2 = 2^2, 3^2, 4^2, 5^2$ .

## Two Sets of Test Scores Per Child

- At each replication we obtain a true score,  $Y_{ti}$ , and a score contaminated with linking error,  $Y_{ti}^*$ .

Student	$Y$	$Y^*$	Time
1	$Y_0$	$Y_0 + \eta_0$	0
1	$Y_1$	$Y_1 + \eta_1$	1
1	$Y_2$	$Y_2 + \eta_2$	2
1	$Y_3$	$Y_3 + \eta_3$	3

- Recall that  $Y_t = \theta_{g+1} + B$ . So adding in  $\eta_t$  is what would happen to the scaled score if the linking constant showed random departures over replications as illustrated in slide 14

## Analysis of True Scores

- Treat the entire vector of true scores,  $\mathbf{Y}_i = (Y_{0i}, \dots, Y_{3i})'$ , as the outcome variable in the following linear model:

$$Y_{ti} = \mu + \beta \cdot t + \epsilon_{ti}, \quad \epsilon_{ti} \sim \mathcal{N}(0, \mathbf{\Sigma}) \quad (12)$$

- Where  $t$  indexes time = (0, 1, 2, 3) and  $i$  indexes student
- The current approach captures the serial correlation in the student responses via  $\mathbf{\Sigma}$ , a block-diagonal matrix with covariances following a first-order autoregressive structure:

$$\mathbf{\Sigma}_i = \sigma_\epsilon^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \quad (13)$$

## Analysis of Contaminated Scores

- Treat the entire vector of true scores,  $\mathbf{Y}_i^* = (Y_{0i}^*, \dots, Y_{3i}^*)'$ , as the outcome variable in the following linear model:

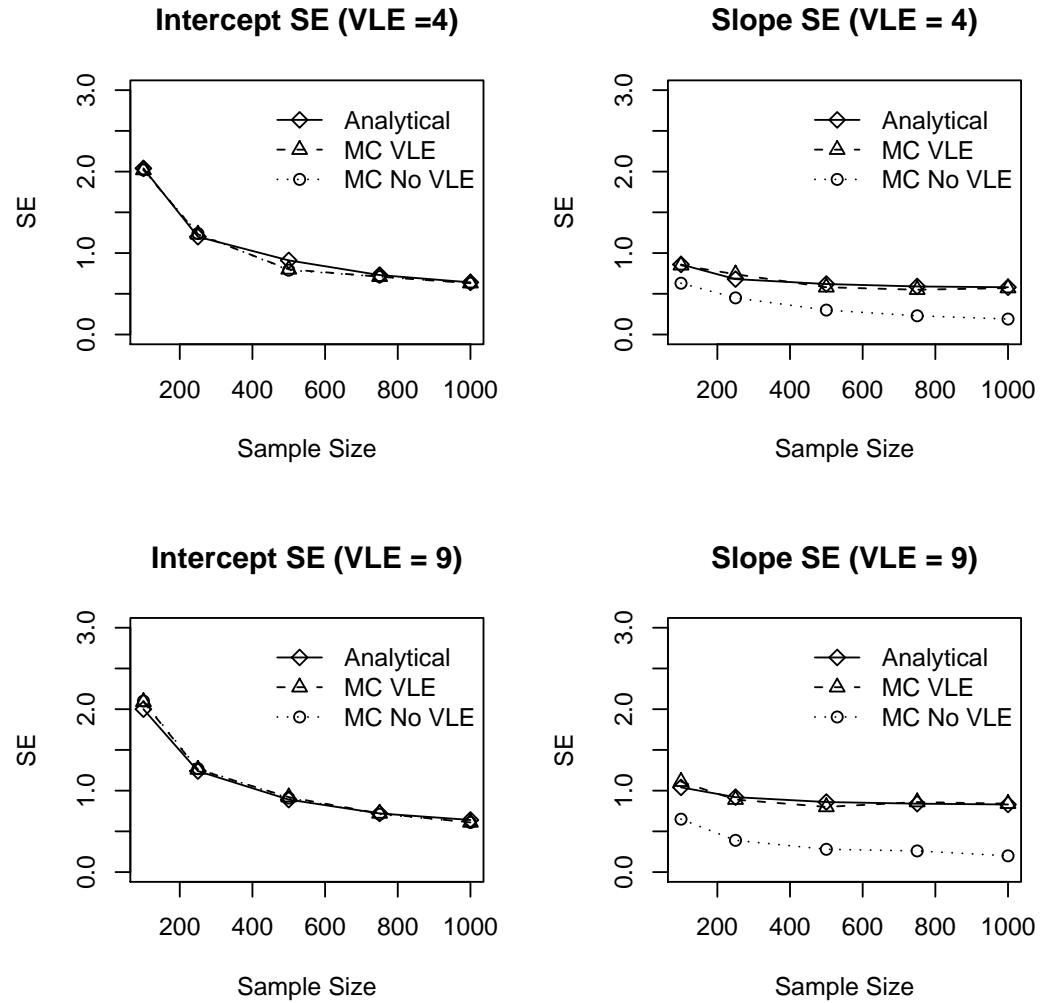
$$Y_{ti}^* = \mu + \beta \cdot t + \epsilon_{ti}, \quad \epsilon_{ti} \sim \mathcal{N}(0, \mathbf{\Phi}) \quad (14)$$

- Distributional assumptions are modified to account for the additional source of error in the data
- Specifically, the variance/covariance matrix is modified such that the linking error is correlated across students within each measurement occasion, but uncorrelated over time.
- The form of this matrix is in the paper, but requires more matrix operations than can be presented here

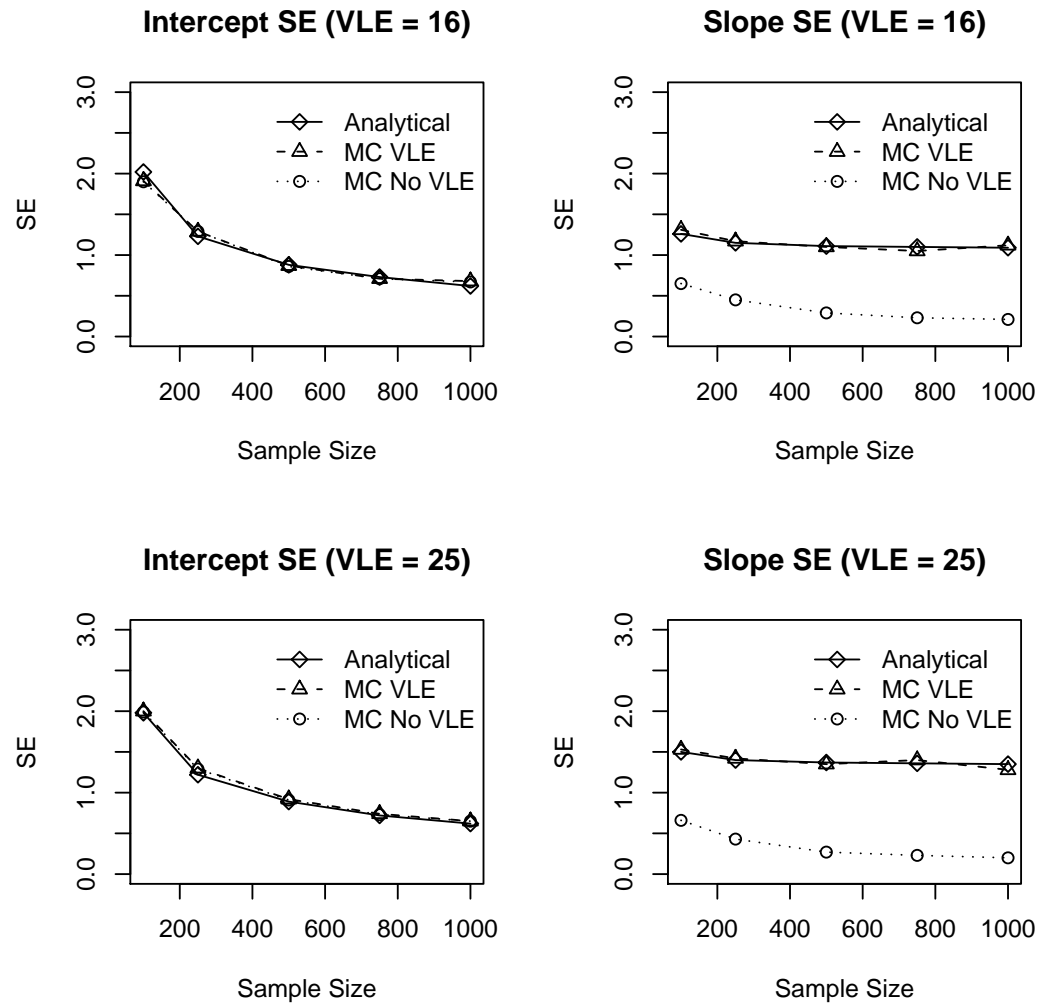
## Estimating Sampling Distributions

- This process is repeated 250 times at each sample size and at each value of  $\sigma_{vle}$ , resulting in 250 estimates of  $\mu$ , the intercept and  $\beta$ , the growth parameter.
- Taking the standard deviation of the 250 parameter estimates is the sampling distribution of the parameter
- For example, we obtain 250 estimates of  $\mu$  and  $\beta$  at the sample size of 500 and with  $\sigma_{vle} = 2$  and another 250 at the sample size of 500 and with  $\sigma_{vle} = 3$

# Results



# Results



## Implications of Equating Variance

- Standard errors of gain scores are underestimated. One may see average gain scores for schools or teachers bouncing around over time
- The psychometric challenge is to construct analytic formulae for estimating the variance of the linking constants
- Test developers must consider methods that minimize equating error
- Those measuring growth should be cognizant of this error variance and should incorporate estimates of the linking error into the estimation process
- Policymakers should be guarded against reacting to noise

## How Can This Be Studied?

- This question intersects with what a teacher effect is thought to be. Is it expected to change over time? If so, how much? Under what conditions?
- Establish baseline gains for each teacher over multiple years ignoring the linking variance. Which teachers are bouncing around?
- Re-estimate standard errors of these gains incorporating linking error. Identify which teachers are now bouncing around.
- Those that are no longer bouncing around may have been misclassified
- Obtain some qualitative data on these fluctuations. Some of these fluctuations are real, but which would result in misclassification?

## Literate Statistical Computing

- LSC is a method for combining statistical code within narrative documentation
- The result is a document where the statistical code is replaced by its output
- The benefit is that one can work in a single environment
- Additionally, routine tasks can be easily replicated using looping constructs
- In R, the **Sweave** library can be used with  $\text{\LaTeX}$  to create dynamic documents with variable information and graphics

## Variable Information Growth Reports

- Score reports are commonly clouded with too much information that is difficult for parents and teachers to understand
- They rarely combine visually pleasing graphical displays with simple, narrative language
- The challenge is to find ways to communicate the complexity of growth models with simple visual displays and narrative language

## Purpose of Variable Information Reports

- To express student growth using three facets of communication
  - Characterize statistical results using straightforward language
  - Provide data for those who best react to numbers
  - Provide visual displays for those who best react to visual displays
- Create visually pleasing displays to enhance understanding

## Technical Basis

- Control flow structures provide an avenue for development of documents with variable information
- This creates a very flexible environment for creating growth reports such that certain terms or even entire paragraphs can appear conditional on the statistical results
- As a result, one can easily program reports such that graphics and text are all created to be unique for each child
- This customizes reports such that complex statistical results can be characterized in simple ways and interpreted for the user
- Parents feel as though they are receiving reports created specifically for their child
- Both R and L<sup>A</sup>T<sub>E</sub>X have control flow structures that allow for conditional expressions to be evaluated

# Sample Growth Reports

Sample School

2004 to 2005 School Year

Prepared for Lisa Bergers

## Mathematics and Reading Growth Report

Prepared for Lisa Bergers

Dear Family:

Students in charter schools authorized by Central Michigan University take the Scantron Performance Series test twice each year; first in the fall and again in the spring. When students take the test twice, you will find it is helpful to examine how much they have improved from the beginning of Grade 4 to the end of Grade 4. This report describes Lisa's academic performance and how her performance compares to other Grade 4 students in schools authorized by Central Michigan University and to the national average. Growth is measured using scaled scores, a score designed for measuring student improvement. When students improve, their scaled scores will increase over time. Due to limitations in the reading test, students who start the year much higher than the average may not show as much improvement as students who start the year much lower than the average.

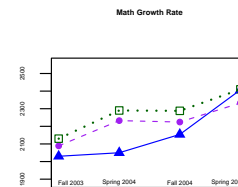
### 1. How Did Lisa Perform in Mathematics?

#### What Is Happening Here?

The chart to the right shows how Lisa is growing in math over time. The solid blue line (▲) is the growth rate for Lisa, the dashed purple line (●) is the average math growth rate for students in other schools authorized by Central Michigan University, and the dotted dark green line (□) is the national average.

#### What Do These Results Mean?

The data indicate that Lisa scored lower than the CMU average at the beginning of the year and grew at a faster rate than the CMU average. At the end of the year, Lisa's score was higher than the CMU average.



	Fall 2003	Spring 2004	Fall 2004	Spring 2005	Gain Score
Lisa	2030	2050	2154	2407	253
CMU Average	2089	2232	2224	2333	108
National Average	2130	2290	2288	2410	122

\* No Score Available

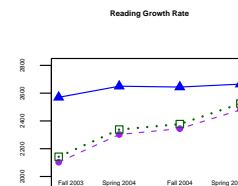
### 2. How Did Lisa Perform in Reading?

#### What Is Happening Here?

The chart to the right shows how Lisa is growing in reading over time. The solid blue line (▲) is the growth rate for Lisa, the dashed purple line (●) is the average reading growth rate for students in other schools authorized by Central Michigan University, and the dotted dark green line (□) is the national average.

#### What Do These Results Mean?

The data indicate that Lisa scored higher than the CMU average at the beginning of the year and grew at a slower rate than the CMU average. At the end of the year, Lisa's score was higher than the CMU average.



	Fall 2003	Spring 2004	Fall 2004	Spring 2005	Gain Score
Lisa	2570	2650	2644	2665	21
CMU Average	2103	2302	2344	2479	135
National Average	2142	2338	2377	2525	148

\* No Score Available

## Comparing Student Growth Trajectories to Standards-based Outcome Targets

- The key aspect of value-added modeling is to make inferences about teachers or school
- However, judging the adequacy of a student's growth is also of substantive interest
- How does one determine the adequacy of an observed growth rate?

## Comparing Student Growth Trajectories to Standards-based Outcome Targets

- One method is to compare the observed prior growth for an individual to the growth rate needed to be proficient by a certain point in time.
- The method answers the following question “if the student continues to grow at the same rate as has been previously observed, where will s/he be in year  $T$ ?”
- This model is based on two assumptions that I refer to as the *achievement* principle and the *timeline* principle.
- The Achievement principle denotes what level of achievement a student is expected to reach.
- The Timeline principle identifies the period of time in which the student has to achieve this outcome.

## Example

- The first step is to obtain an estimate of a student's growth rate
- The second is to estimate the average rate of change needed each year in order to be proficient at time  $T$ , the end of the timeline
- The third step is to compare the observed growth rate of student  $i$  to the expected growth rate for student  $i$
- All students are held accountable for same level of achievement

## REACH: A Contextual Student Level Model

- The following can be used to obtain the value needed for step 2

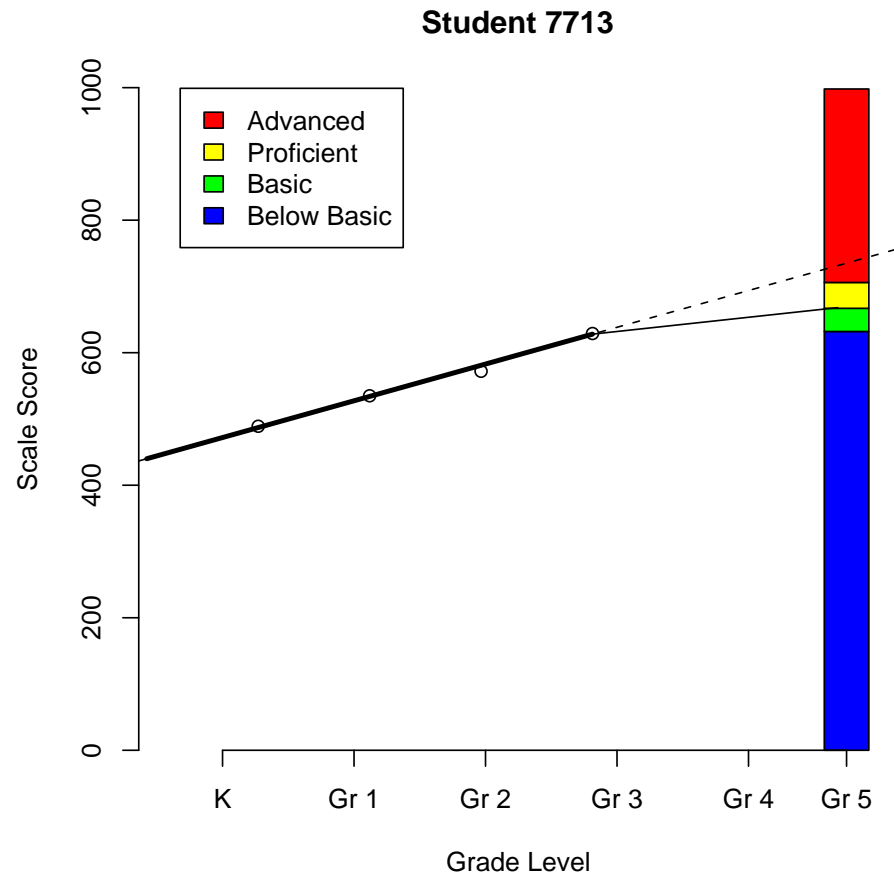
$$REACH_i = \frac{\lambda - y_{ti}}{T - \alpha_i} \quad (15)$$

- where  $\lambda$  is the lowerbound cut score for proficiency,  $y_{ti}$  is the current score for student  $i$  at time  $t$ ,  $T$  is the expected timeline, and  $\alpha_i$  is the current grade level of student  $i$ .
- The prior growth rate is then compared  $REACH_i$

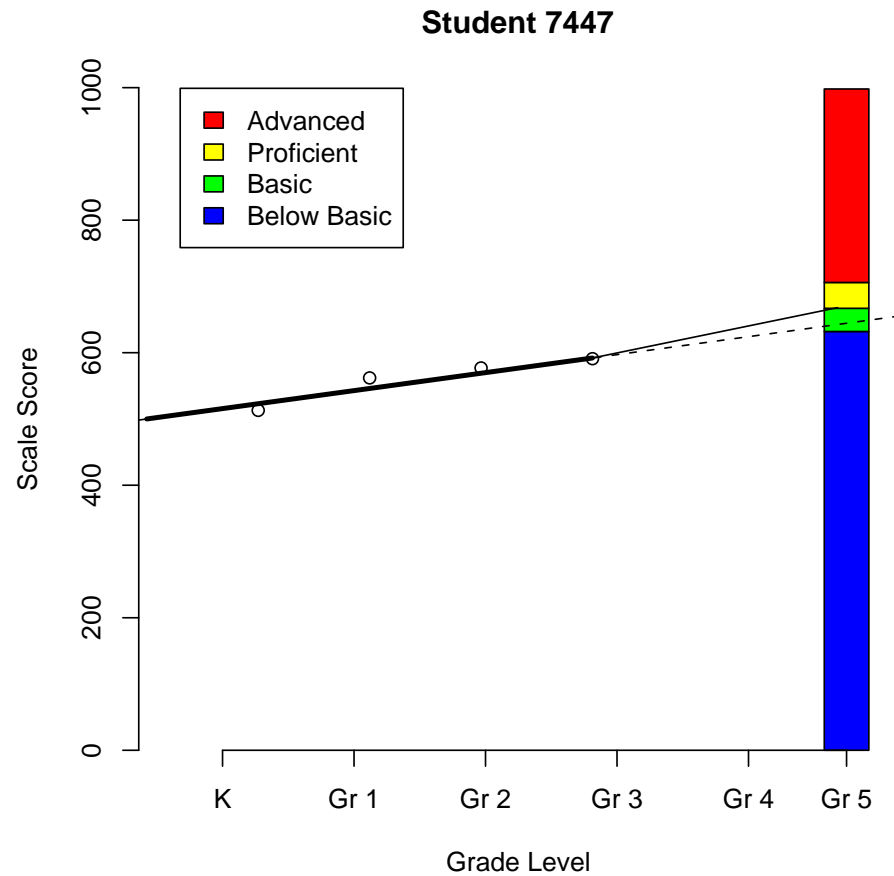
$$REACHRatio = \frac{Growth_i}{REACH_i} \quad (16)$$

- This ratio indicates whether the observed growth places this student “on track” to be proficient at time  $T$

# A Student “On Track” to Reach the Proficient Cutpoint



# A Student Unlikely to Reach the Proficient Cutpoint



## Software for Crossed Random Effects

- The `nlme` package in R has evolved significantly. Prior implementations were especially troublesome for large data files and models with crossed random effects
- Recent developments by Doug Bates has enhanced capacity by reliance on sparse matrix techniques
- The current version, `lmer` for linear-mixed effects revised is specifically designed for large data files with crossed random effects. The syntax has slightly changed.
- On the horizon is a new implementation of `lmer` that is extremely fast, and can be used on inexpensive machines to fit complex models with crossed random effects with teacher effects accumulating over time

## Some Timing Comparisons

- `egsingle` data file (7,230 obs, 1721 students nested in 60 schools)
- Old `nlme` took 30.14 seconds
- Same model specification in `lmer` took .86 seconds
- Star data file (26,796 obs, 11598 students, 1387 teachers) took 49.82 seconds
- DC data file (106,270 obs, 48,767 students, 3380 teachers, 168 schools) took 12 minutes