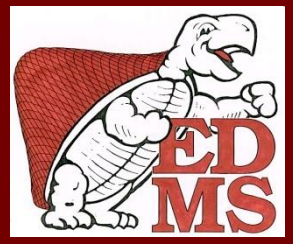
UNIVERSITY OF MARYLAND



Issues in Latent Growth Modeling with Longitudinal Public-Release Data

Ming Li, Jeffrey R. Harring & Laura M. Stapleton

Measurement, Statistics & Evaluation (EDMS)

University of Maryland

Modern Modeling Methods Conference

21 May 2014

Outline

- Introduction
- Longitudinal Data
- Latent Growth Modeling (LGM)
- Extensions to LGM
- Issues with Public-Release Data in LGM
- An Example Using ECLS-K Data Set
- Summary

Introduction

- For researchers interested in modeling longitudinal development of individuals using large-scale public-release data, it is important to understand the issues surrounding data collection and challenges of using these data to address questions about growth.

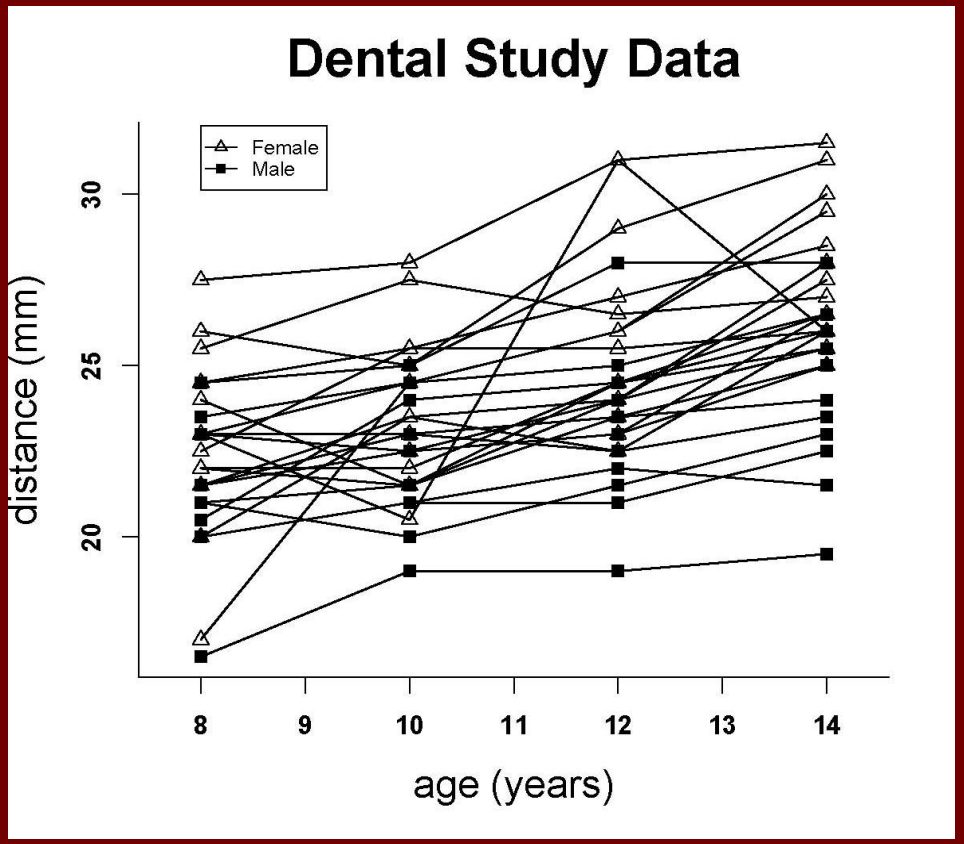
- Challenges of using public-release longitudinal data for growth analyses can be broadly grouped into four areas: construct measurement, modeling choices for growth, missing data, and sampling design accommodation.

Longitudinal Data

- **Repeated Measures**
  - **For each individual, repeated measurements of the same variables are made over time or other condition, which allows for the direct study of change**
  - **Observations within the same individual may be correlated, which requires special statistical models to account for the correlation of observations within subjects**

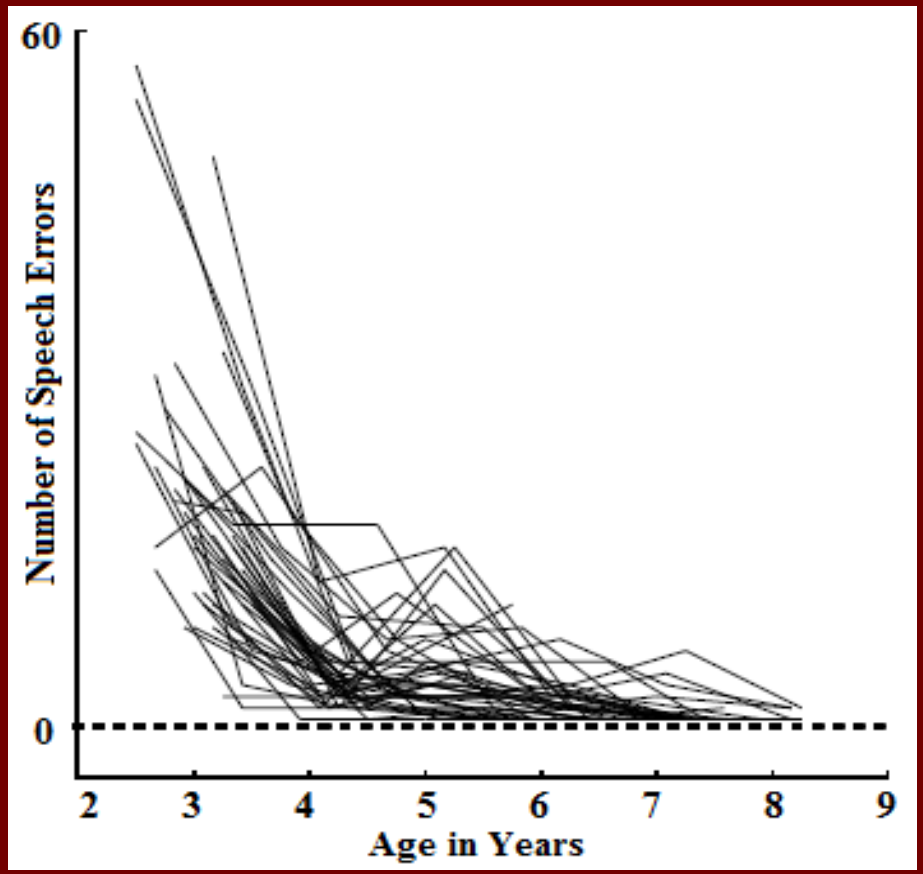
Longitudinal Data

- Potthoff and Roy (1964) conducted a study involving 27 children, 16 boys and 11 girls.
- On each child, the distance (mm) from the center of the pituitary to the pterygomaxillary fissure was made at ages 8, 10, 12, and 14 years of age.



Longitudinal Data

- Burchinal & Appelbaum (1991): Speech errors of 43 young children from ages 2-8 were recorded over 6 occasions.
- Ages at the time of testing differed for each child.



Latent Growth Modeling (LGM)

- Latent growth modeling (LGM), also called latent growth curve analysis, is a longitudinal analysis technique to estimate growth over a period of time.
- In LGM, the relative standing of an individual at each time is modeled as a function of an underlying growth process, with the best parameter values for that growth process being fitted to each individual.

Typical Research Questions in LGM

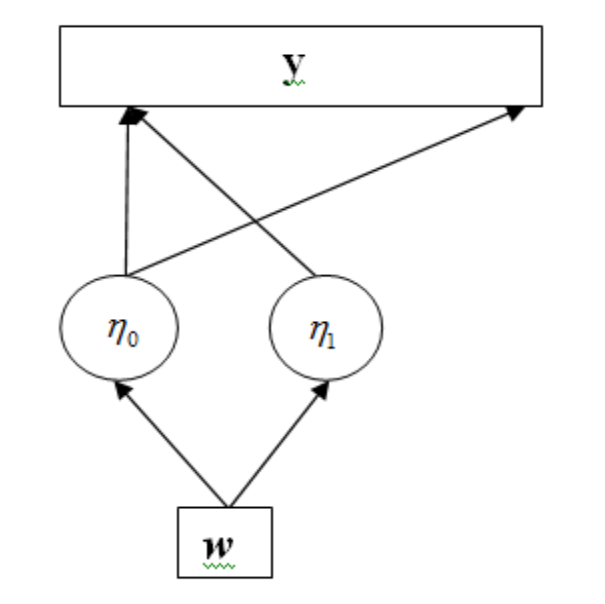
- What is the nature of change over time or condition? [e.g., What function best describes the change in alcohol consumption of students during college?]
- Do growth characteristics defining the change process vary across subjects? Within subjects? [e.g., Do students grow at different rates? How much does each student grow over time?]
- Correlation of the growth parameters [e.g., Do students with a higher score at the initial status grow slower or faster than those with a lower initial status?]

Typical Research Questions in LGM

- **What covariates/specific factors are associated with (help explain) individual variation in the growth characteristics? [e.g., Does the income trajectory differ for high school graduates depending on whether they attended college?]**
- **What are the consequences of individual variation in the change process? [e.g., Are higher average levels of alcohol use and more curvature in the growth form associated with higher levels of problem behavior?]**

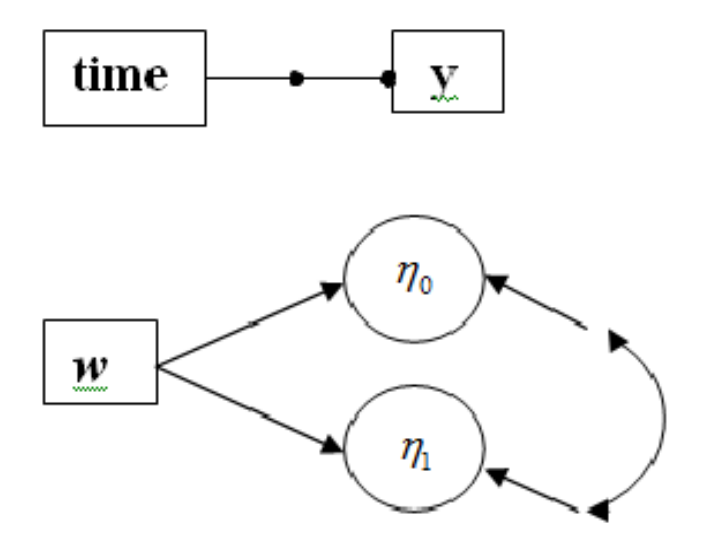
Two Approaches to LGM

- **Wide: Multivariate, 1-Level Approach**



- **SEM takes a multivariate approach to LGMs**

- **Long: Univariate, 2-Level Approach**



- **Multilevel takes a univariate approach to LGMs**

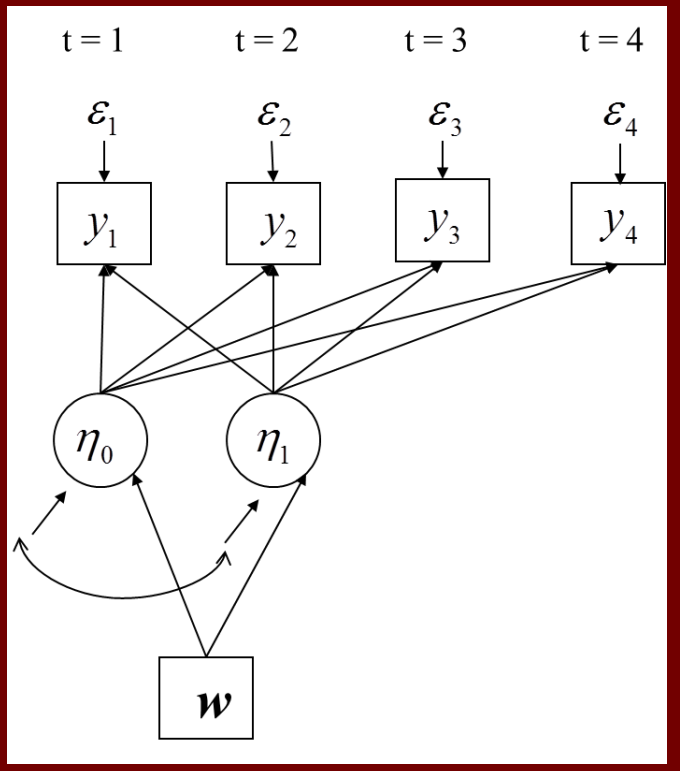
Two Approaches to LGM

- Latent growth modeling in SEM framework
  - An unconditional latent linear growth model:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} \eta_{0i} \\ \eta_{1i} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \end{pmatrix}$$

$$\mathbf{y}_i = \mathbf{\Lambda}_i \quad \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{\zeta}_i$$



  - Latent linear growth model with time-invariant covariate(s):

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{w}_i + \boldsymbol{\zeta}_i$$

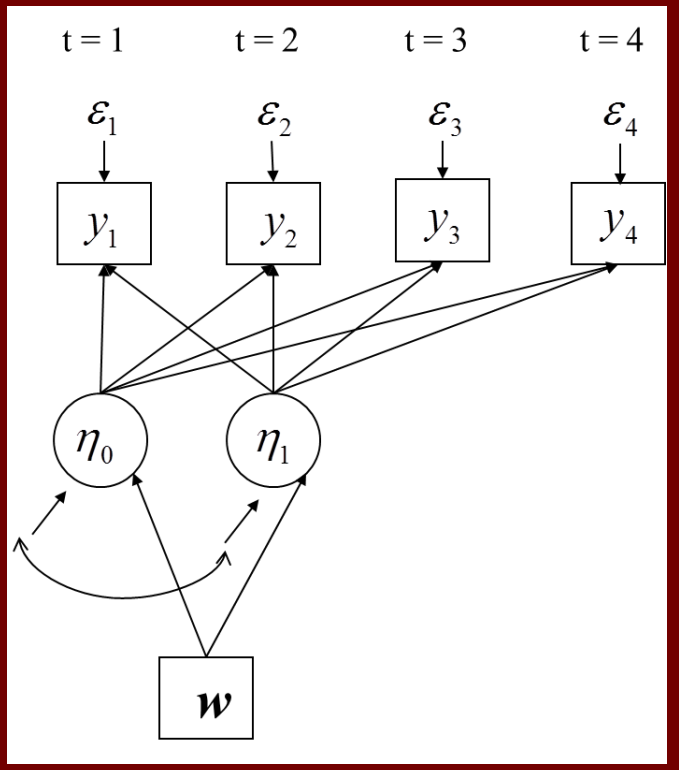
Two Approaches to LGM

- **Latent growth modeling in multilevel framework**
  - **An unconditional latent linear growth model:**

$$(1) \quad y_{ti} = \eta_{0i} + \eta_{1i} x_t + \varepsilon_{ti}$$

$$(2a) \quad \eta_{0i} = \alpha_0 + \zeta_{0i}$$

$$(2b) \quad \eta_{1i} = \alpha_1 + \zeta_{1i}$$



  - **Latent linear growth model with time-invariant covariate(s):**

$$(2a) \quad \eta_{0i} = \alpha_0 + \gamma_0 w_i + \zeta_{0i}$$

$$(2b) \quad \eta_{1i} = \alpha_1 + \gamma_1 w_i + \zeta_{1i}$$

Assumptions of LGM

$$(1) \quad y_{ti} = \eta_{0i} + \eta_{1i} x_t + \varepsilon_{ti}$$

$$(2a) \quad \eta_{0i} = \alpha_0 + \zeta_{0i}$$

$$(2b) \quad \eta_{1i} = \alpha_1 + \zeta_{1i}$$

Fixed Effects **Random Effects**

$$\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Theta}_i)$$

$$\boldsymbol{\Theta}_i = \sigma_j^2 \mathbf{I}_{n_i}$$

$$\boldsymbol{\zeta}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$$

$$Var(\boldsymbol{\zeta}_i) = \boldsymbol{\Psi} = \begin{pmatrix} \psi_{\eta_0} & \\ \psi_{\eta_1\eta_0} & \psi_{\eta_1} \end{pmatrix}$$

$$\text{cov}(\boldsymbol{\varepsilon}, \boldsymbol{\zeta}') = \mathbf{0}$$

Extensions – Nonlinear LGM

- **A very standard method of summarizing curvilinearity in the data is to fit higher-order polynomials (e.g., quadratic growth models)**

Stage 1 model:

$$y_{ti} = \eta_{0i} + \eta_{1i} x_t + \eta_{2i} x_t^2 + \varepsilon_{ti}$$

Stage 2 model:

$$\eta_{0i} = \alpha_0 + \zeta_{0i} \qquad \alpha_0 \text{ is average inital reaction time}$$

$$\eta_{1i} = \alpha_1 + \zeta_{1i} \qquad \alpha_1 \text{ uninterpretable}$$

$$\eta_{2i} = \alpha_2 + \zeta_{2i} \qquad \alpha_2 \text{ uninterpretable}$$

Extensions – Nonlinear LGM

- **Note: Often nonlinear functions can be tailored so that model parameters correspond to interesting facets of the change process**
- **Many different nonlinear growth functions have been used in the social and behavioral sciences including functions like the exponential, Gompertz, logistic, and Richard's Family of curves (Brown, 1993; Browne & du Toit, 1991; Grimm & Ram, 2009).**

Extensions – Nonlinear LGM

- A reparameterization of the quadratic model (Cudeck & du Toit, 2006)

$$y_{ti} = \alpha_{yi} - (\alpha_{yi} - \alpha_{0i})\left(\frac{x_t}{\alpha_x} - 1\right)^2 + \varepsilon_{ti}$$

- Parameter Interpretation

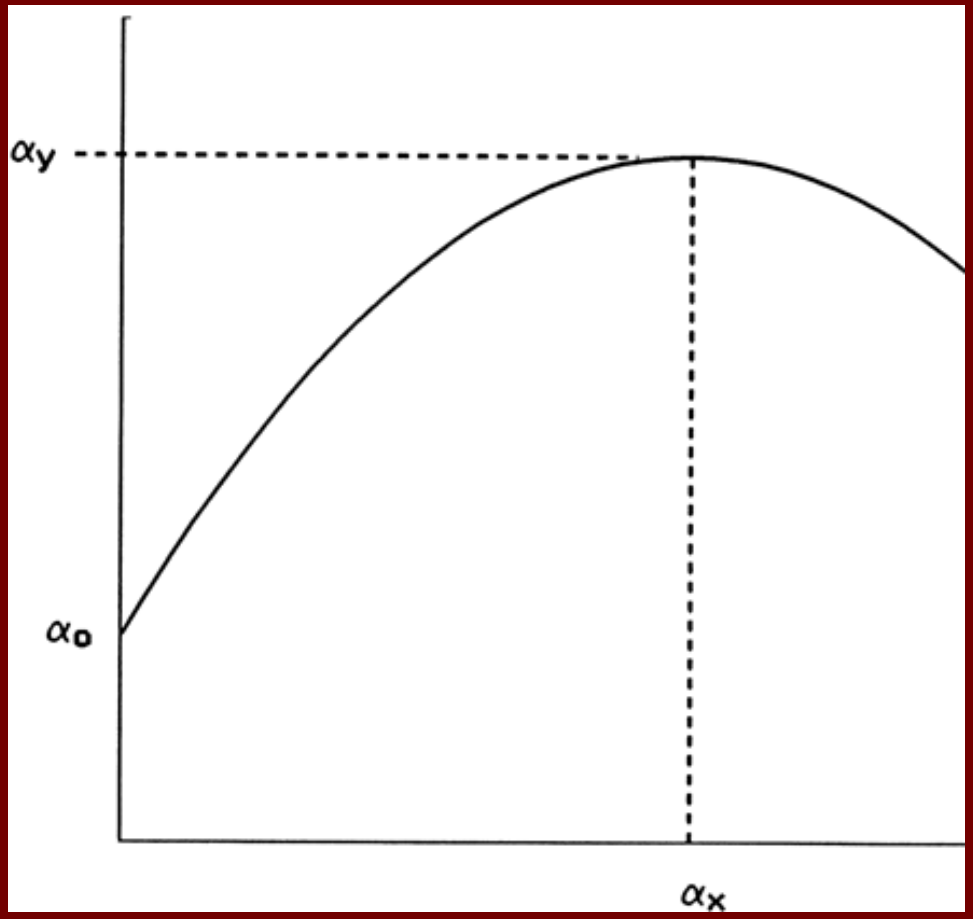
initial performance at $x_t = 0$

function maximizer

function maximum

New Quadratic

$$y_{ti} = \alpha_{yi} - (\alpha_{yi} - \alpha_{0i})\left(\frac{x_t}{\alpha_x} - 1\right)^2 + \varepsilon_{ti}$$



Other LGM Extensions

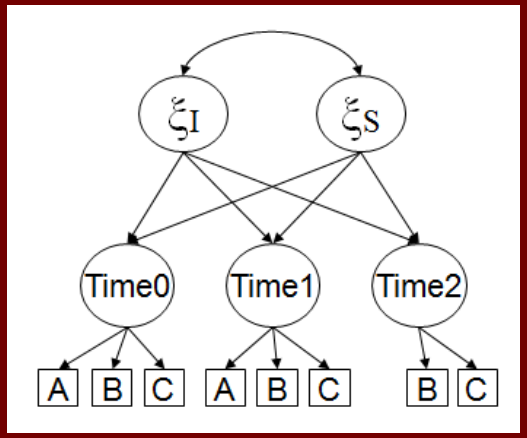
- Second-order LGMs allow examining change in latent—not measured—variables
- Multivariate LGM—modeling parallel processes
- Unique time scores for each individual
- Time-varying covariates

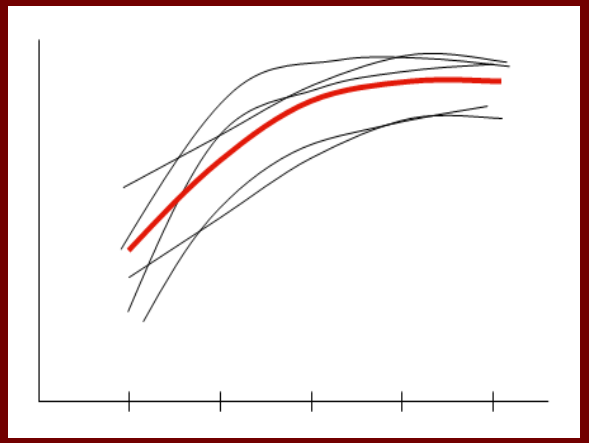
Issues with Public-Release Data in LGM

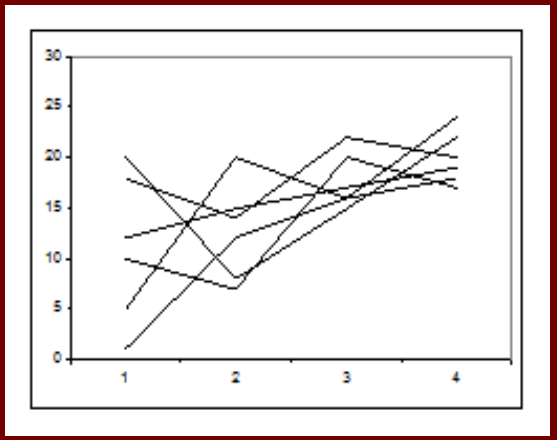
- Construct Measurement
- Functional Form of Model
- Missing Data
- Sampling Design

Issues Faced by the Applied Researcher

- Construct Measurement
  - Measurement consistency?
  - Ceiling or floor effects in instrument?
  - Differential reliability across wave?







Issues Faced by the Applied Researcher

- **Functional Form and Model Choice**
  - **Time points selected for data collection**
  - **Functional form of growth**
  - **Structures for between- and within-subject variability**

$$E[\mathbf{y}_i] = \boldsymbol{\Lambda} E[\boldsymbol{\eta}_i] = \boldsymbol{\Lambda}\boldsymbol{\alpha}$$

$$var(\mathbf{y}_i) = \boldsymbol{\Lambda}\,\mathrm{var}(\boldsymbol{\eta}_i) + \mathrm{var}(\boldsymbol{\varepsilon}_i)$$

$$\boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_i = \boldsymbol{\Sigma}_i$$

between-subject

within-subject

Issues Faced by the Applied Researcher

- **Missing Data**

–Item-level missingness

- **The individual participated in the wave of data collection but failed to provide an answer to a question.**

–Unit-level missingness

- **The individual did not respond during the wave of data collection (maybe as a result of planned missingness)**

–Other missingness is due to attrition from the study.

Issues Faced by the Applied Researcher

- **Ways to Address Missing Data (MAR)**

  - **For item-level missingness, use of multiple imputation or full information maximum likelihood estimation is reasonable (Hedeker & Gibbons, 2006).**

  - **For unit-level missingness, the analyst must decide whether to use such procedures or use the nonresponse adjusted panel weights provided on the dataset.**

Issues Faced by the Applied Researcher

- **Sampling Design**

–Unequal probabilities of selection should be accommodated to obtain unbiased parameter estimates and any clustering and/or stratification addressed to obtain appropriate measures of sampling variability for proper inference (Kish, 1965).

–For variance estimation, two techniques can be considered: linearization and replication. In addition, knowledge of the sampling design might allow for more complex research questions to be posed relating cluster characteristics to individual growth parameters.

ECLS-K

- Early Childhood Longitudinal Study – Kindergarten 1988
  - 3-stage sample (geographic areas, schools, students)
  - Stratification at first stages of sampling included region, metropolitan status, public/private
  - Stratified sampling of students by race/ethnicity, with oversampling of Asian/Pacific Islander students
  - Panel study
  - Data collected from parents, teachers, administrators and students

ECLS-K

- Assessments and Questionnaires Completed

—Math and Reading in early grades, Science added later

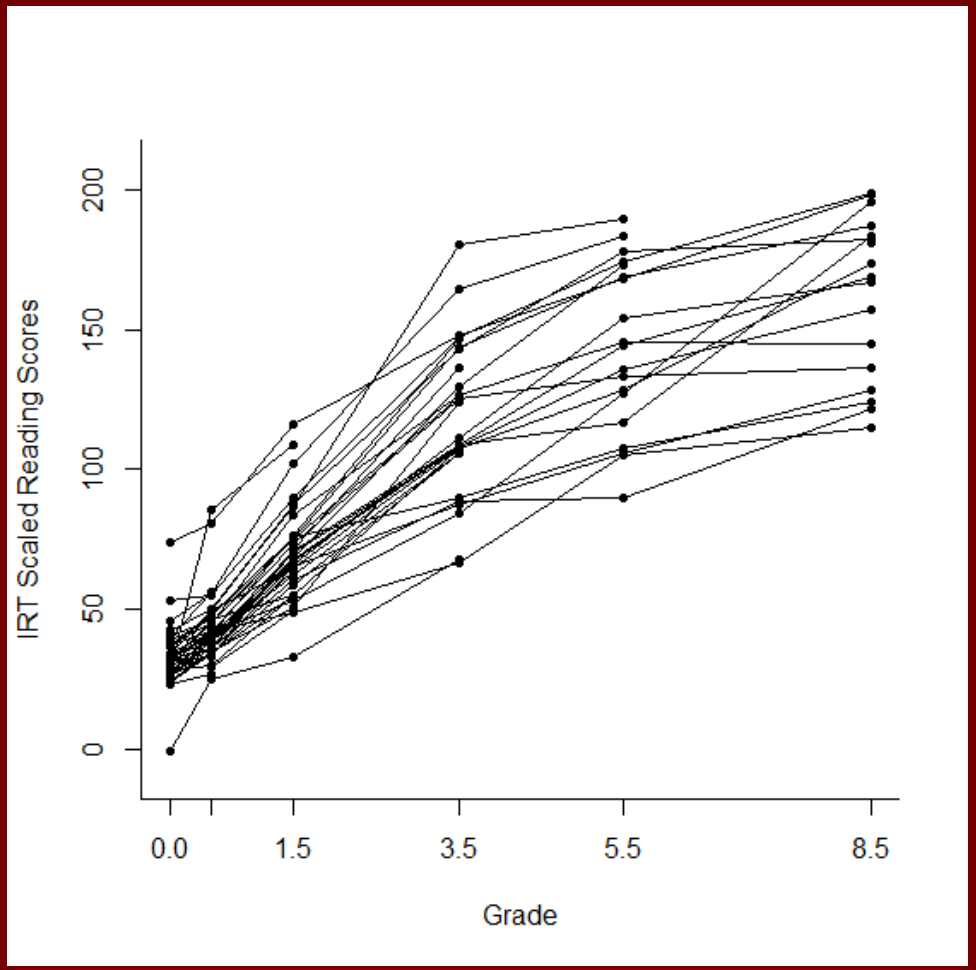
- fall kindergarten
- spring kindergarten
- fall 1st grade (20% sub-sample only)
- spring 1st grade
- spring 3rd grade
- spring 5th grade
- spring 8th grade

ECLS-K

- **Decisions to be Made:**
  - **What records/observations should be included in an analysis?**
  - **What panel weights should be used?**
  - **What method of sampling variance estimation should be used (Taylor Series or Replication)?**

Random Sample *n* = 50

- Individual profiles show a definite nonlinear trajectory
- Missing data is evident
- Variability appears to be growing over time



Exploration of Mean Function

Linear Function

$$y_{ij} = \eta_{0i} + \eta_{1i} t_j + \varepsilon_{ij}$$

Quadratic Function

$$y_{ij} = \eta_{0i} + \eta_{1i} t_j + \eta_{2i} t_j^2 + \varepsilon_{ij}$$

Exponential Function

$$y_{ij} = \eta_{1i} - (\eta_{1i} - \eta_{0i}) \exp\{-\eta_{2i} t_j\} + \varepsilon_{ij}$$

Logistic Function

$$y_{ij} = \frac{\eta_{0i}\eta_{1i}}{\eta_{1i} - (\eta_{1i} - \eta_{0i}) \exp\{-\eta_{2i} t_j\}} + \varepsilon_{ij}$$

Exploration of Mean Function

- **Oblivious Analysis (N=20,235):**

Model		# Param.	BIC	CFI	RMSEA	SRMR
Linear		6	811938.3	0.006	0.429	0.444
Quadratic		15	741629.3	0.901	0.179	0.077
Exponential		10	752134.8	0.767	0.231	0.139
Logistic		15	737588.9	0.952	0.124	0.051

Exploration of Mean Function

- **Panel Weight Analysis (N=7,754):**

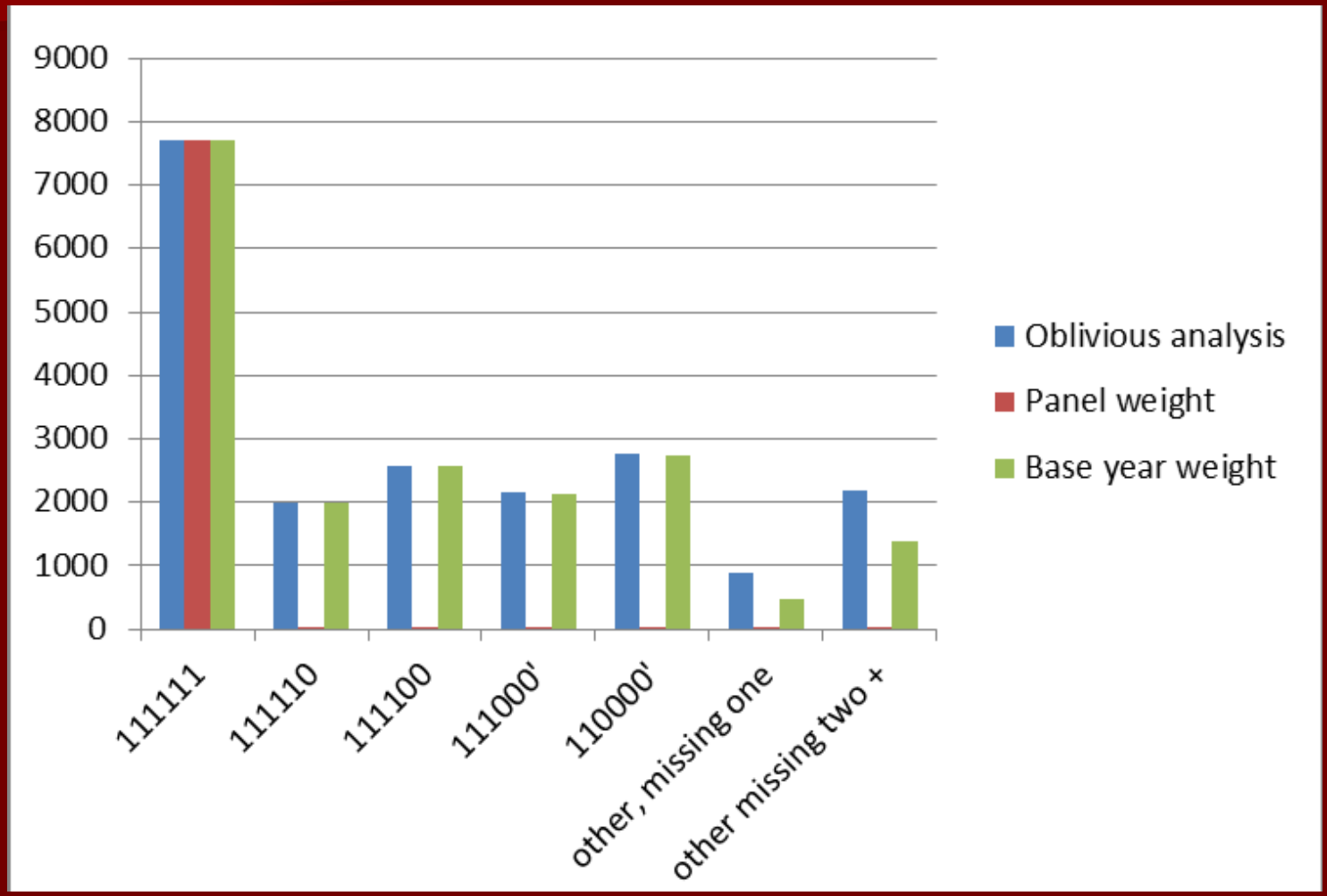
Model		# Param.	BIC	CFI	RMSEA	SRMR
Linear		6	434503.5	0.000	0.170	0.686
Quadratic		15	393406.4	0.800	0.093	0.103
Exponential		10	399590.4	0.657	0.103	0.230
Logistic		15	391330.1	0.895	0.068	0.054

Exploration of Mean Function

- **Base Year Weight Analysis (N=19,003):**

Model		# Param.	BIC	CFI	RMSEA	SRMR
Linear		6	763777.1	0.117	0.169	0.566
Quadratic		15	693213.6	0.789	0.110	0.121
Exponential		10	705703.3	0.659	0.117	0.206
Logistic		15	688358.2	0.951	0.053	0.089

Distribution of Analysis Observations By Missing Data Patterns



Conditional Logistic LGM

d_1 = 1 if RACE is Black/African American; 0 otherwise
d_2 = 1 if RACE is Hispanic; 0 otherwise
d_3 = 1 if RACE is Other; 0 otherwise
Reference group is White

η_{0i} : asymptotic behavior

η_{1i} : initial value $t_j = 0$

η_{2i} : rate parameter

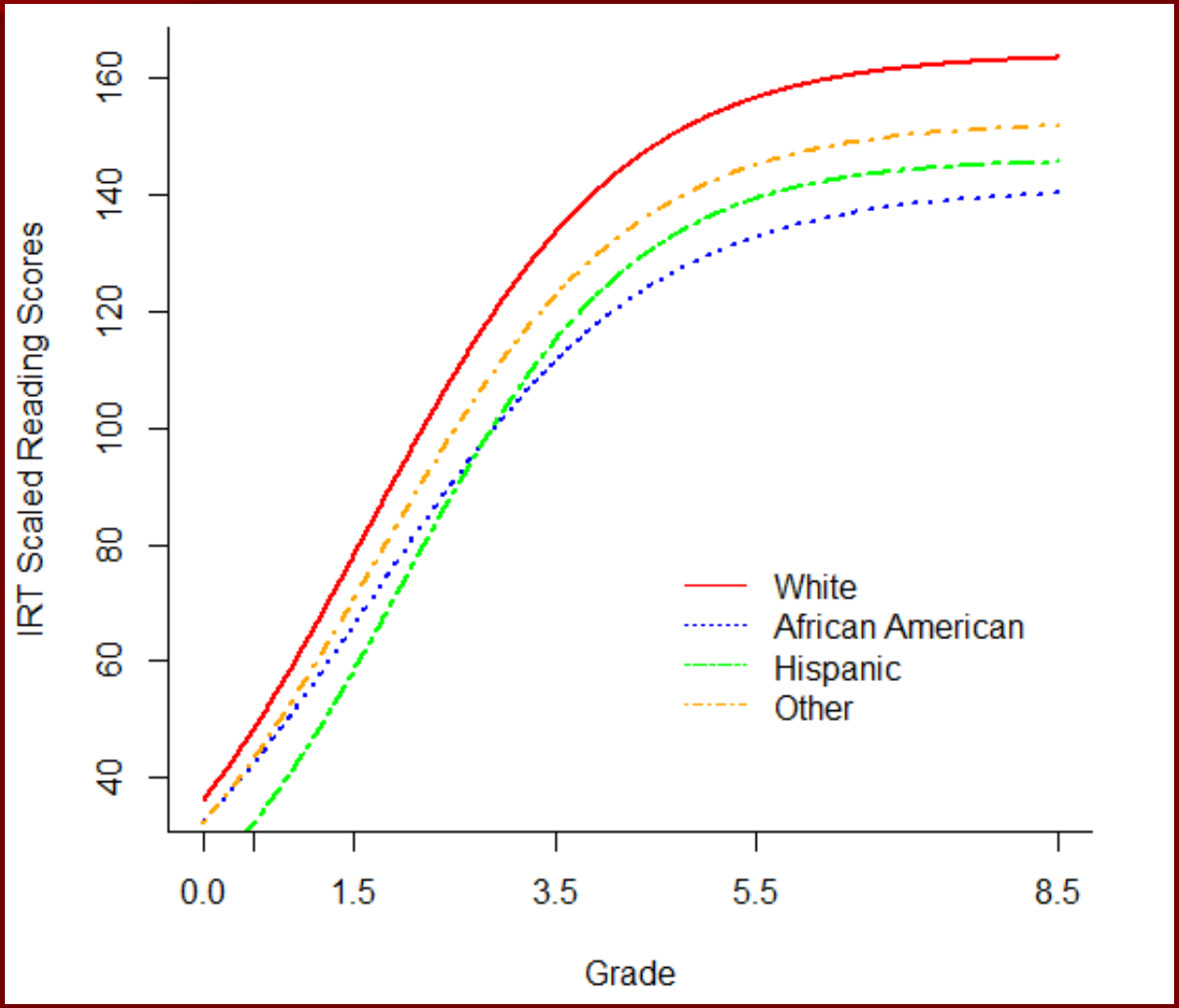
$$y_{ij} = \frac{\eta_{0i}\eta_{1i}}{\eta_{1i} - (\eta_{1i} - \eta_{0i})\exp\{-\eta_{2i}t_j\}} + \varepsilon_{ij}$$

$$\eta_{0i} = \alpha_0 + \gamma_{01}d_{1i} + \gamma_{02}d_{2i} + \gamma_{03}d_{3i} + \zeta_{0i}$$

$$\eta_{1i} = \alpha_1 + \gamma_{11}d_{1i} + \gamma_{12}d_{2i} + \gamma_{13}d_{3i} + \zeta_{1i}$$

$$\eta_{2i} = \alpha_2 + \gamma_{21}d_{1i} + \gamma_{22}d_{2i} + \gamma_{23}d_{3i} + \zeta_{2i}$$

Conditional Logistic LGM Under Base Year Weighting



Summary of Applied Example

- Analysis sample size differed greatly depending on weighting approach (panel vs. base year)
- Use of base year weighting assumes that missing data are MAR, conditional on model variables
- Modeled that individuals were assessed at the same time at each wave
- Standard errors were typically (and appropriately) larger for the weighted analyses with Taylor Series estimates of sampling variances

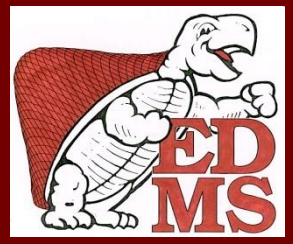
Summary of Applied Example

- Almost no differences in inference regarding parameters of interest
- In examining growth trajectories, no practical difference across weighting choices

Summary

- Appropriate research questions need to be defined in light of the available data
- Models need to be properly built to accommodate the data characteristics (e.g., time of data collection, measurement scale)
- Careful comparison of growth trajectories should be considered
- Parameter and variance estimation should address sampling design and missing data

UNIVERSITY OF MARYLAND



Thank You

for copies of these slides:

liming@umd.edu
harring@umd.edu
lstaplet@umd.edu