Hierarchical Linear Models-Redux

Joseph Stevens, Ph.D., University of Oregon
(541) 346-2445, stevensj@uoregon.edu

Overview and resources

- Overview
- Listserv: http://www.jiscmail.ac.uk/lists/multilevel.html
- Web site and links: www.uoregon.edu/~stevensj/HLM-II
- Software:

HLM	MLWinN	Mplus	SAS
SPSS	R and S-Plus	WinBugs	

Workshop Overview

- Rationale for multilevel modeling
- Four examples as demonstrations of the power and flexibility of multilevel models
  - Achievement gap
  - Meta analysis
  - Longitudinal models of school effects
  - Interrupted time series
- Introduction to several technical issues as we discuss examples
- Lots of “how-to” information in last year’s workshop

Grouping and membership in particular units and clusters are important

For goodness sake, this is a huge field!
Why do we need to huddle like this all the time?

Hierarchical Data Structures

Many social and natural phenomena have a nested or clustered organization:

- Children within classrooms within schools
- Patients in a medical study grouped within doctors within different clinics
- Children within families within communities
- Employees within departments within business locations

Hierarchical Data Structures

More examples of nested or clustered organization:

- Children within peer groups within neighborhoods
- Respondents within interviewers or raters
- Effect sizes within studies within methods (meta-analysis)
- Multistage sampling
- Time of measurement within persons within organizations

Simpson's Paradox: Clustering Is Important

Well known paradox in which performance of individual groups is reversed when the groups are combined

	Quiz 1	Quiz 2
Gina	60.0%	10.0%
Sam	90.0%	30.0%

	Quiz 1	Quiz 2	Total
Gina	60 / 100	1 / 10	61 / 110
Sam	9 / 10	30 / 100	39 / 110

Simpson's Paradox: Other Examples

2006 US School study:

1975 Berkeley sex bias case:

- UCB sued for bias by women applying to grad school

"When the Oakies left Oklahoma and moved to California, it raised the IQ of both states."

– *Will Rogers*

First Example: Does Multilevel Modeling Matter?

- The Analysis of School Effects
  - Individual Level Analysis
  - Analysis of School Level Aggregates
  - Multilevel Analysis
- The Intraclass Correlation Coefficient (ICC)
- Fixed and Random Effects

Why Is Multilevel Analysis Needed?

- Nesting creates dependencies in the data
  - Dependencies violate the assumptions of traditional statistical models ("independence of error", "homogeneity of regression slopes")
  - Dependencies result in inaccurate statistical estimates
- Important to understand variation at different levels

Decisions About Multilevel Analysis

- Properly modeling multilevel structure often matters (and sometimes a lot)
- Partitioning variance at different levels is useful
  - tau and sigma ($\sigma^2_Y = \tau^2 + \sigma^2$)
  - policy & practice implications
- Correct coefficients and unbiased standard errors
- Cross-level interaction
- Understanding and modeling site or cluster variability

Example 1: Achievement Gap

Data Example from New Mexico State accountability system, 2001 reading data for grade 6 children, N = 5,544, j=36

Example used here examines relationship between ethnicity

First analysis considers all 5,544 students without taking school membership into account.

Second analysis considers the 36 schools without taking students into account.

Third analysis considers both the 5,544 students and the 36 schools using a multilevel modeling approach.

Disaggregated analysis (N = 5,544 students)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.389	.151	.151	36.128

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1287816	3	429272.082	328.890	.000[a]
	Residual	7230887	5540	1305.214		
	Total	8518703	5543			

a. Predictors: (Constant), OTHER, AMIND, HISP

b. Dependent Variable: READ01

Coefficients[a]

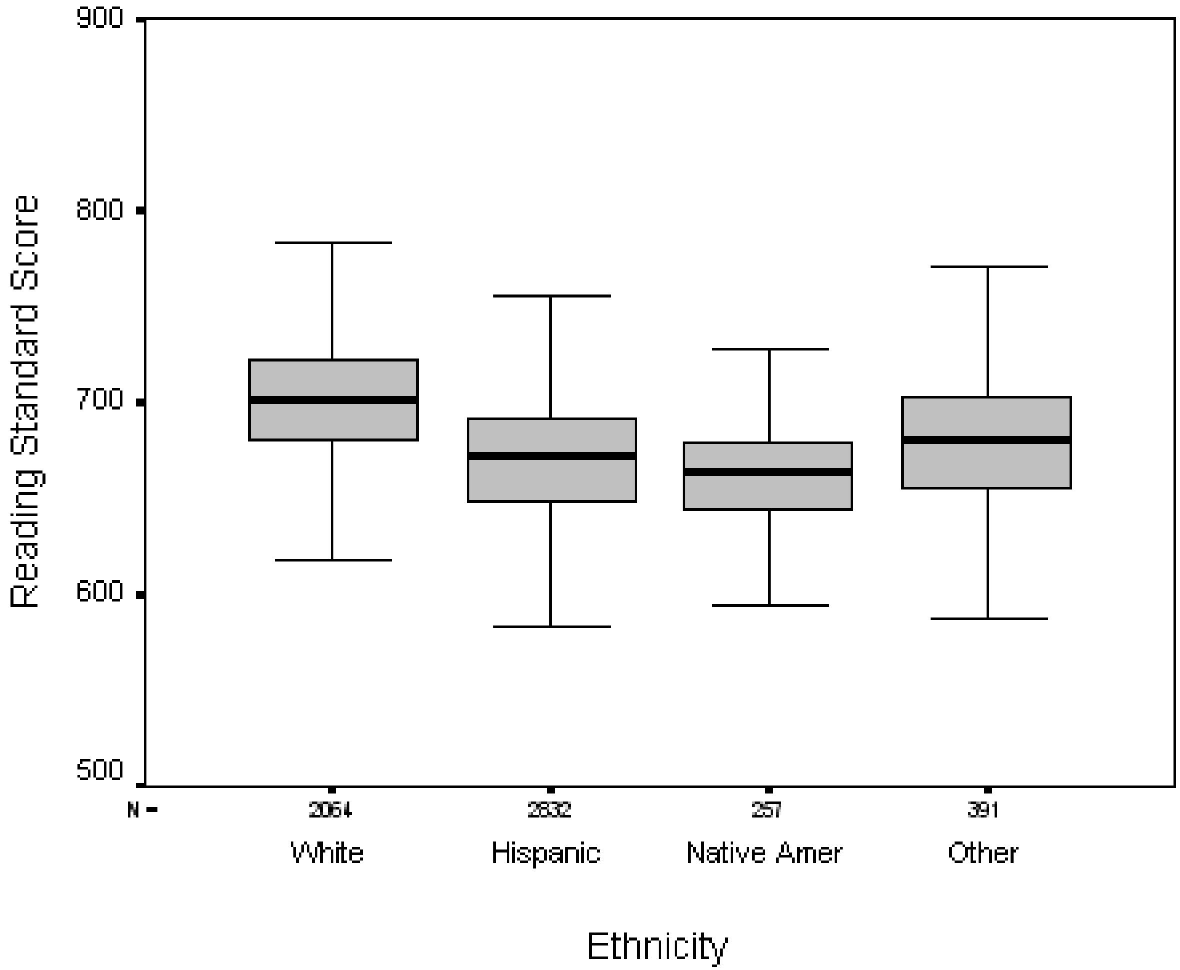
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	701.164	.795		881.726	.000
	HISP	-31.449	1.046	-.401	-30.078	.000
	AMIND	-38.740	2.390	-.208	-16.211	.000
	OTHER	-22.486	1.993	-.147	-11.285	.000

a. Dependent Variable: READ01

Disaggregated analysis (N = 5,544 students)

$$Y = 701.164 - 31.449(X_1) - 38.740(X_2) - 22.486(X_3) + r$$

Interpretation: White students average 6th grade reading performance is about 701 points; Hispanic students score on average 31 points less, American Indian students score on average 39 points less, and other ethnic categories of students score on average about 23 points less.


900
800
700
600
500
Reading Standard Score
N =
2064
2832
257
391
White
Hispanic
Native Amer
Other
Ethnicity

Another alternative is to analyze data at the aggregated group level

Participant (i)	Cluster (j)	Outcome (Y)	Predictor (X)
1	1	5	1
2	1	7	3
10	5	3	7

Cluster (j)	Outcome (Y)	Predictor (X)
1	6	2

The aggregated analysis considers the 36 middle schools without taking students into account.

Aggregated analysis (J = 36 schools)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.895	.801	.782	7.17838

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6630.867	3	2210.289	42.894	.000[a]
	Residual	1648.933	32	51.529		
	Total	8279.800	35			

a. Predictors: (Constant), OTHER, HISP, AMIND

b. Dependent Variable: READING

Coefficients[a]

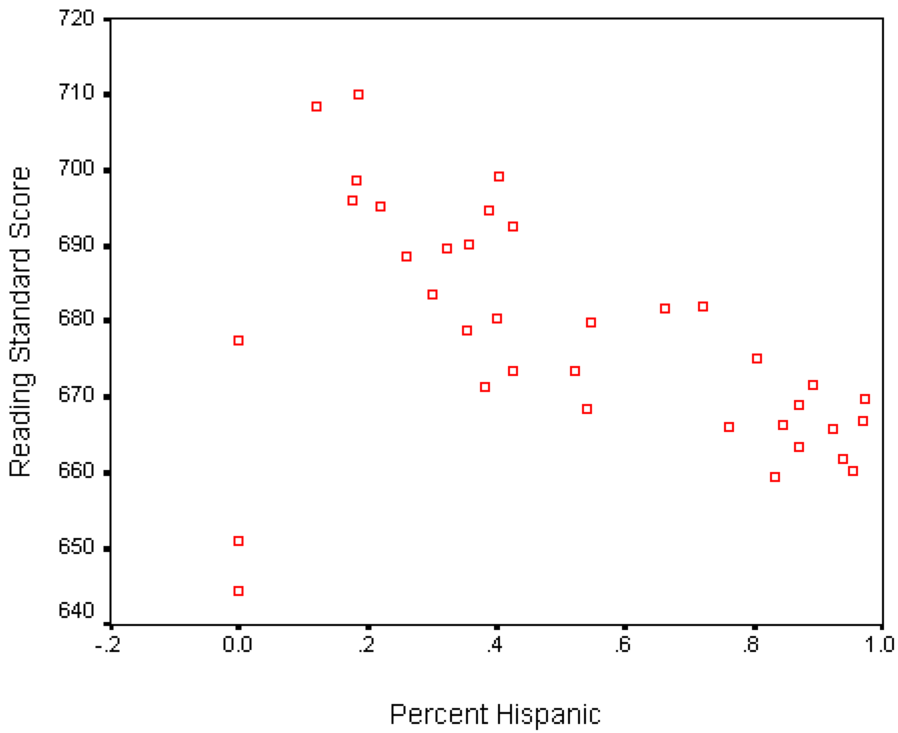
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	715.355	4.203		170.205	.000
	HISP	-50.789	5.095	-1.031	-9.969	.000
	AMIND	-60.006	6.155	-1.027	-9.750	.000
	OTHER	-70.699	21.540	-.305	-3.282	.002

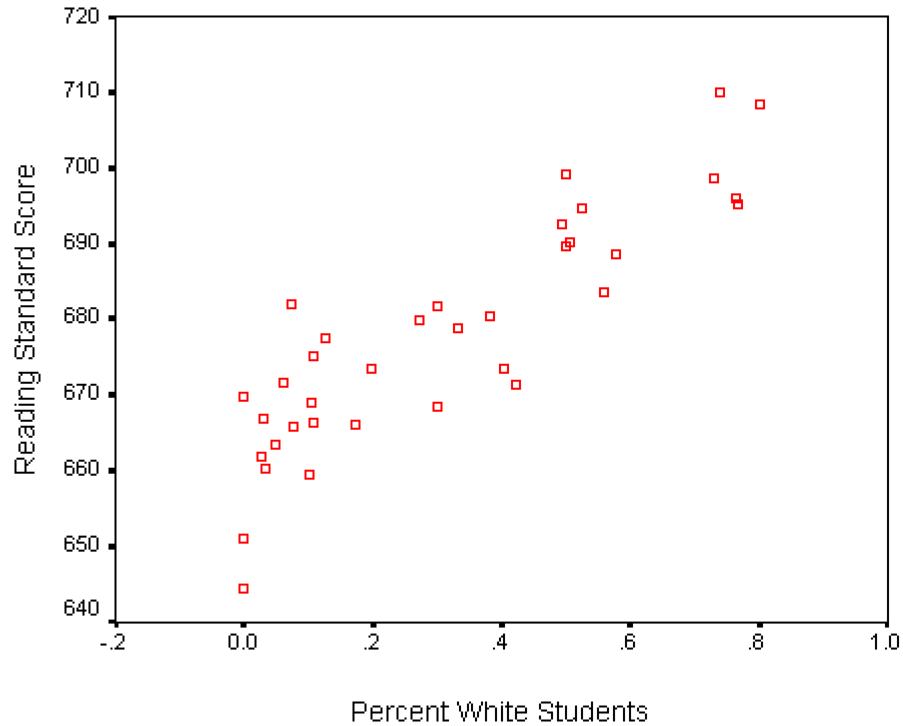
a. Dependent Variable: READING

Aggregated analysis (J = 36 schools)

$$Y = 715.355 - 50.789(X_1) - 60.006(X_2) - 70.699(X_3) + r$$

Interpretation: White students average 6th grade reading performance is about 715 points; Hispanic students score on average 51 points less, American Indian students score on average 60 points less, and other ethnic categories of students score on average about 71 points less.


720
710
700
690
680
670
660
650
640
Reading Standard Score
-.2
0.0
.2
.4
.6
.8
1.0
Percent Hispanic


720
710
700
690
680
670
660
650
640
Reading Standard Score
-.2
0.0
.2
.4
.6
.8
1.0
Percent White Students

Multilevel Models

- Unlike the two previous single-level regression models, multilevel modeling takes both levels (students and schools) into account simultaneously:

$$Y_{ij} = \beta_{0j} + \beta_1(X_1) + r_{ij} \qquad \text{Level 1}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad \text{Level 2}$$

$$\beta_1 = \gamma_{10} + u_{1j} \qquad \text{Level 2}$$

- Note that level 1 regression model parameters become outcomes at level 2

Multilevel Models

- Variance associated with the level 1 units (students) is partitioned from variance associated with level 2 units (schools)
- In essence, a different regression model is fit within each school
- Differences in model parameters (slopes and intercepts) can then be analyzed from one school to another
- A fundamental question in multilevel analysis is how much the outcome differs in relation to the level 2 grouping variable (e.g., schools); this relationship is estimated by the intraclass correlation coefficient (ICC)

Intraclass Correlation (ρ)

- The Intraclass Correlation Coefficient (ICC) measures the correlation between a grouping factor and an outcome measure
- In common notation there are 1 to J groups
- If participants do not differ from one group to another, then the ICC = 0
- As participants' outcome scores differ due to membership in a particular group, the ICC grows large

Intraclass Correlation Coefficient (ρ)

Total $\sigma^2_Y = \tau^2 + \sigma^2$

$$\text{ICC} = \frac{\text{between unit variance}}{\text{total variance}}$$

$$= \quad \tau^2 / (\tau^2 + \sigma^2)$$

When the ICC is 0, multilevel modeling is not needed and power is the same as a non-nested design.

Multilevel Analysis (N = 5,544 students nested in J = 36 schools)

Third analysis considers both the 5,544 students and the 36 schools using a multilevel modeling approach.

(with robust standard errors)

Within school variance = 226.091

Between school Variance = 1304.875

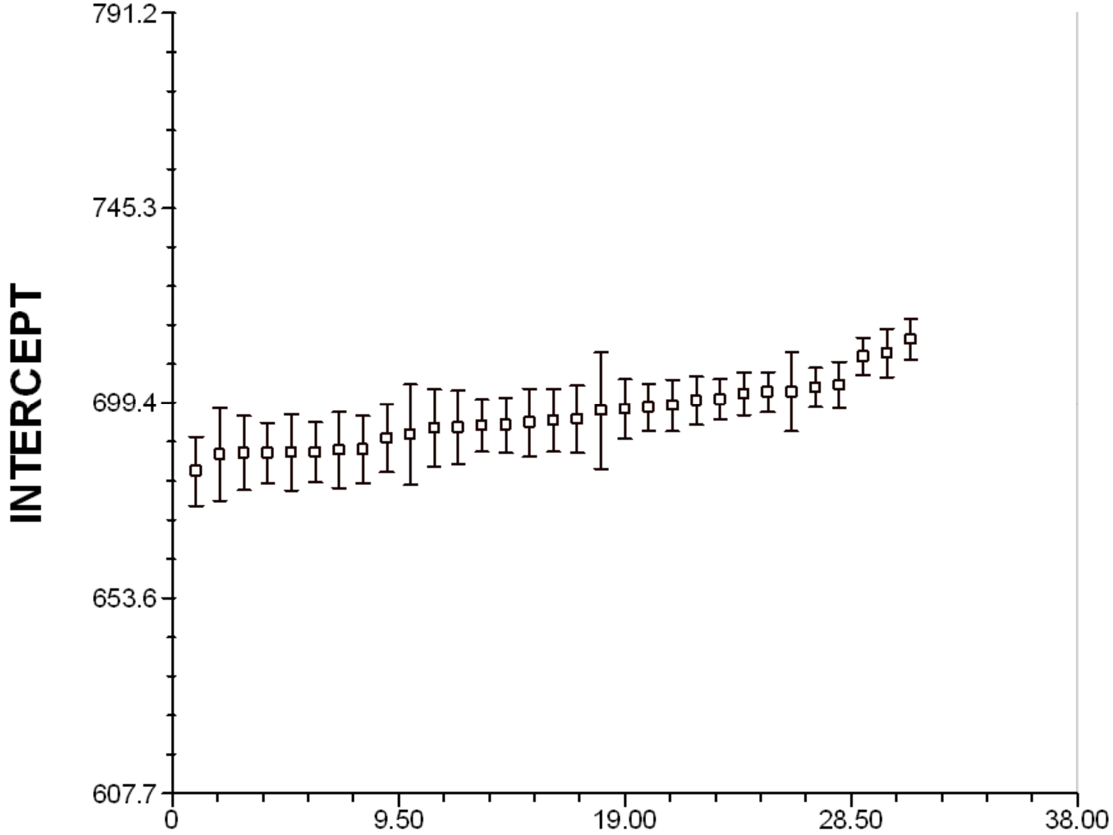
ICC = .148

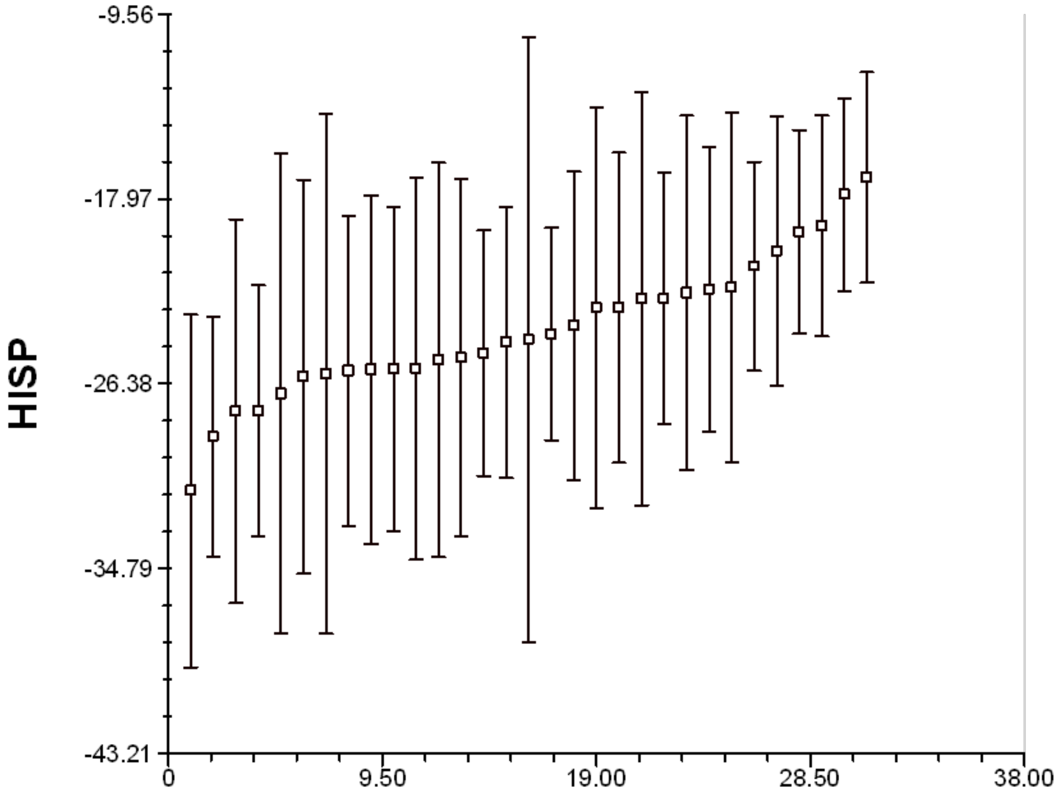
Fixed				prox. f.	P-value
For					
INTRCP				35	0.000
For H					
INTRCPT2, G10		2?.??????	?.?????	35	0.000
For AMIND slope, B2					
I					0.000
For					
I					0.000

Estimation of ICC an important result with policy implications, in and of itself. Over a large number of SER studies, ICC ranges from about 10-20%.

Fina

Ran					e	P-value
INTRCPT1,	U0	8.46948	71.73209	27	181.84069	0.000
HISP slope,	U1	4.95524	24.55441	27	39.63265	0.055
AMIND slope,	U2	6.70990	45.02271	27	25.33754	>.500
OTHER slope,	U3	7.03970	49.55731	27	36.06544	0.114
level-1,	R	34.97760	1223.43237			


791.2
745.3
699.4
653.6
607.7
INTERCEPT
0
9.50
19.00
28.50
38.00


HISP
-9.56
-17.97
-26.38
-34.79
-43.21
0
9.50
19.00
28.50
38.00

Comparing the Three Analyses

Model	R^2	F	b	SE	B	t
Disaggregated	.389	328.890				
Intercept			701.164	.795		
Hispanic			-31.449	1.046	-.401	-30.078
Amer. Indian			-38.740	2.390	-.208	-16.211
Other			-22.486	1.993	-.147	-11.285
Aggregated	.895	42.894				
Intercept			715.355	4.203		
Hispanic			-50.789	5.095	-1.031	-9.969
Amer. Indian			-60.006	6.155	-1.027	-9.750
Other			-70.699	21.540	-.305	-3.282
		χ^2				
Multilevel	Level 1 .156 Level 2 .697	379.686				
Intercept			695.412	1.722		
Hispanic			-24.109	1.497	-.308	-16.102
Amer. Indian			-28.703	2.733	-.154	-10.504
Other			-19.703	2.307	-.131	-8.541

Multilevel Model Specification

- Another important difference in the approaches is the greater flexibility of model specification in HLM
  - Multilevel models preserve information about individual differences (level 1 variance)
  - Multilevel models take groups into account and explicitly model group effects (level 2 variance)
  - Multilevel models allow for the examination of interactions between the two levels

Multilevel Model Specification

- In single level regression models, only fixed effects are possible for many parameters (all groups the same on many model parameters; i.e., homogeneity of regression slopes assumption)
- How to conceptualize and model group level variation?
- Do groups vary on the model parameters (fixed versus random effects)?
- Can group level information predict outcomes?

The Single-Level, Fixed Effects Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + r_i$$

- The parameters β_{kj} are considered fixed
  - One for all and all for one
  - Same values for all i and j; the single level model
- The r_i 's are random: $r_i \sim N(0, \sigma)$ and independent
- But what if the β_{kj} were random and variable?

Modeling variation at Level 2: Intercepts as Outcomes

$$Y_{ij} = \boxed{\beta_{0j}} + \beta_{1j}X_{1ij} + r_{ij}$$

$$\boxed{\beta_{0j} = \gamma_{00} + \gamma_{0j}W_j + u_{0j}}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

- Predictors (W's) at level 2 are used to model variation in intercepts between the j units

Modeling Variation at Level 2: Slopes as Outcomes

$$Y_{ij} = \beta_{0j} + \boxed{\beta_{1j}} X_{1ij} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{0j} W_j + u_{0j}$$

$$\boxed{\beta_{1j} = \gamma_{10} + \gamma_{1j} W_j + u_{1j}}$$

- Do slopes vary from one j unit to another?
- W's can be used to predict variation in slopes as well

Fixed vs. Random Effects

- Fixed Effects represent discrete, purposefully selected or existing values of a variable or factor
  - Fixed effects exert constant impact on DV
  - Random variability only occurs as a within subjects effect (level 1)
  - Should only generalize to particular fixed values used
- Random Effects represent more continuous or randomly sampled values of a variable or factor
  - Random effects exert variable impact on DV
  - Variability occurs at level 1 and level 2
  - Can study and model variability
  - Can generalize to population of values

Fixed vs. Random Effects?

- Use fixed effects if
  - The groups are regarded as unique entities
  - If group values are determined by researcher through design or manipulation
  - Small j (< 10); improves power
- Use random effects if
  - Groups regarded as a sample from a larger population
  - Researcher wishes to test effects of group level variables
  - Researcher wishes to understand group level differences
  - Small j (< 10); improves estimation

Variance Components Analysis

- VCA allows estimation of the size of random variance components
  - Important issue when unbalanced designs are used
  - Iterative procedures must be used (usually ML estimation)
- Allows significance testing of whether there is variation in the components (parameters) across units

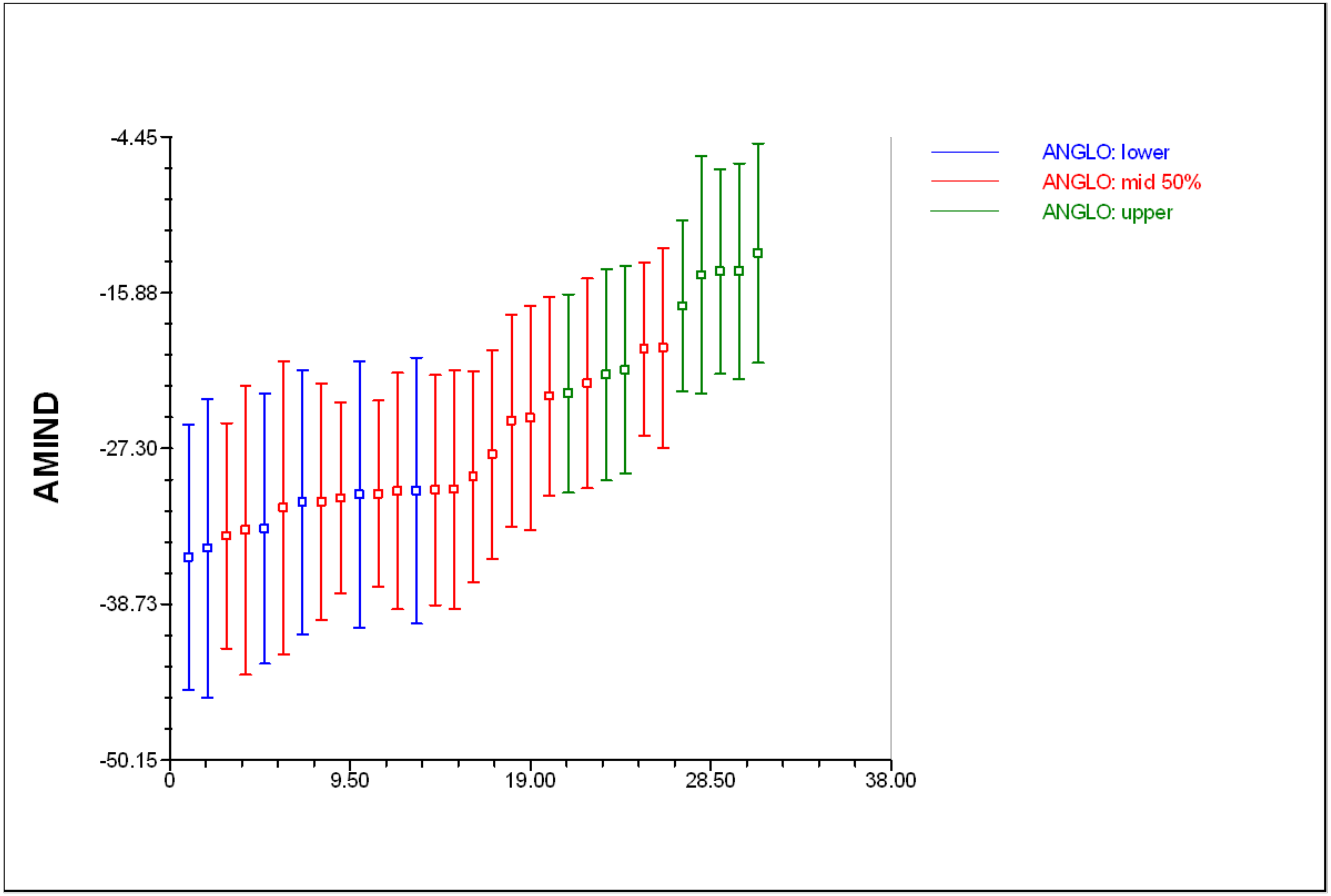
Achievement Gap Example Again

- Random effects allows parameters to vary across schools
- Introduces an entirely different set of research questions, for example:
  - Does the relationship between reading achievement and ethnic group differ from one school to another?
  - Can the differences in the ethnicity-reading achievement relationship be explained by characteristics of the schools?

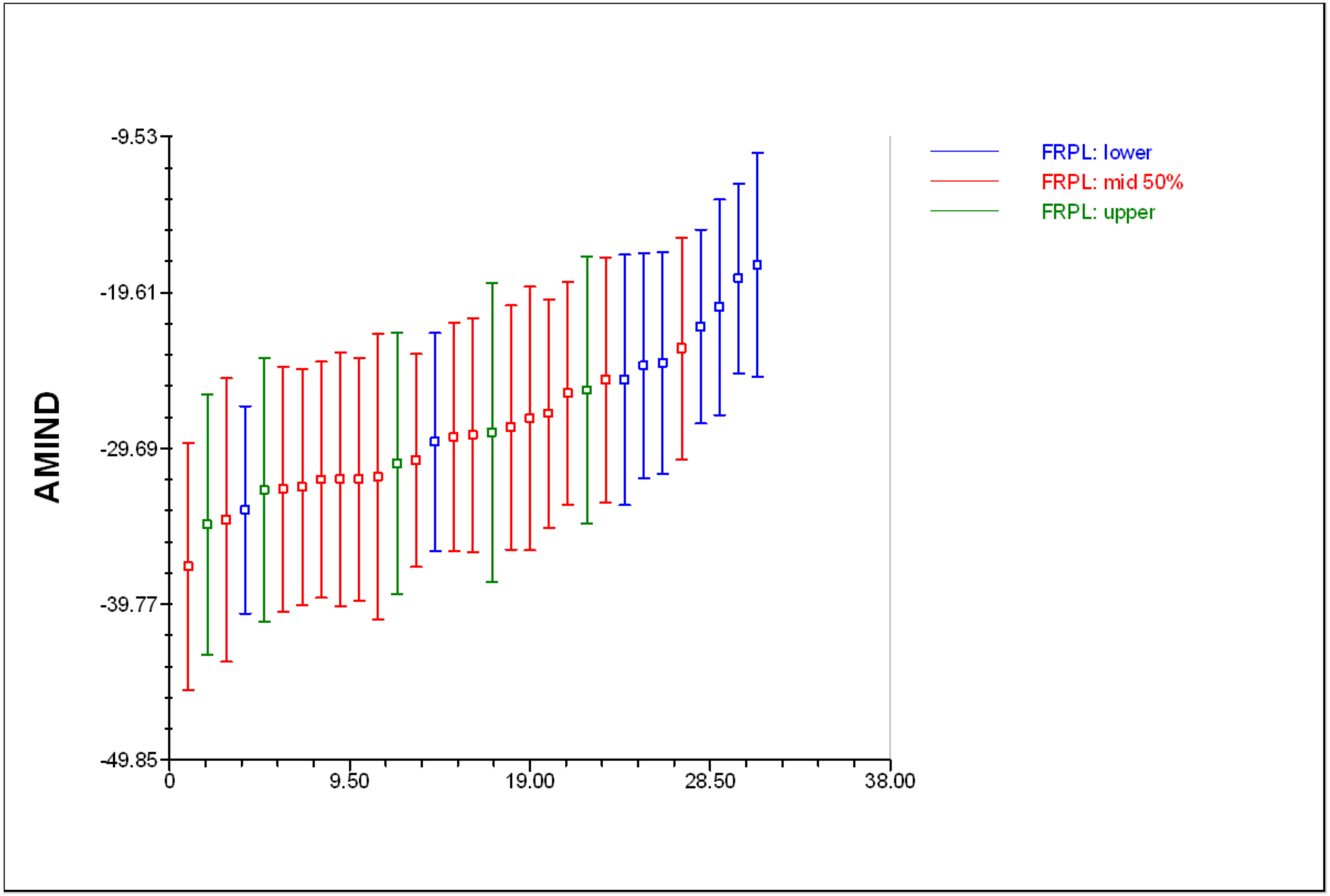
Multilevel Analysis with Level 2 Predictors

(N = 5, 544 students nested within J = 36 schools)

```
Final estimation of fixed effects
 (with robust standard errors)
 ----------------------------------------------------------------------------
                                          Standard             Approx.
    Fixed Effect             Coefficient   Error       T-ratio   d.f.     P-value
 ----------------------------------------------------------------------------
 For       INTRCPT1, B0
    INTRCPT2, G00             694.377894   1.590546   436.566        34     0.000
       ANGLO, G01              24.756614   5.684485     4.355        34     0.000
 For      HISP slope, B1
    INTRCPT2, G10             -22.994825   1.654343   -13.900        34     0.000
       ANGLO, G11               3.049102   5.680311     0.537        34     0.594
 For     AMIND slope, B2
    INTRCPT2, G20             -27.142110   2.525960   -10.745        34     0.000
       ANGLO, G21              23.113476   9.189661     2.515        34     0.017
 For     OTHER slope, B3
    INTRCPT2, G30             -20.440687   2.350818    -8.695        34     0.000
       ANGLO, G31              18.360940   9.255701     1.984        34     0.055
 ----------------------------------------------------------------------------
```



AMIND
-4.45
-15.88
-27.30
-38.73
-50.15
0
9.50
19.00
28.50
38.00
ANGLO: lower
ANGLO: mid 50%
ANGLO: upper



AMIND
-9.53
-19.61
-29.69
-39.77
-49.85
0
9.50
19.00
28.50
38.00
FRPL: lower
FRPL: mid 50%
FRPL: upper

Summary of Example 1 – Structure Matters!

- Correct statistical estimates
- ICC, separating parts from whole
- Understanding relations within and across levels

Example 2: Meta-Analysis

- Can estimation techniques used in HLM provide a more sophisticated way to synthesize quantitative results across studies?
- Example from Raudenbush & Bryk (2002)
  - Teacher expectancy ("the Pygmalion effect")
  - Contentious literature (see Wineburg, 1987; Rosenthal, 1987)
- Parameter Reliability
- Empirical Bayes Estimation

Statistical Estimation in HLM Models

- Estimation Methods
  - FML
  - RML
  - Empirical Bayes estimation
- Parameter estimation
  - Coefficients and standard errors
  - Variance Components
- Parameter reliability

Estimation Methods: Maximum Likelihood (ML)

- ML estimates model parameters by estimating a set of population parameters that maximize a likelihood function
- The likelihood function provides the probabilities of observing the sample data given particular parameter estimates
- ML methods produce parameters that maximize the probability of finding the observed sample data

Estimation Methods

RML – Restricted Maximum Likelihood, only the variance components are included in the likelihood function

FML – Full Maximum Likelihood, both the regression coefficients and the variance components are included in the likelihood function

Restricted: Sequentially estimates the fixed effects and then the variance components

Goodness of fit statistics (deviance tests) apply only to the random effects

RML only tests hypotheses about the VCs (and the models being compared must have identical fixed effects)

Full: Simultaneously estimate the fixed effects and the variance components.

Goodness of fit statistics apply to the entire model

(both fixed and random effects)

Check on software default

Estimation Methods

- RML expected to lead to better estimates especially when j is small
- FML has two advantages:
  - Computationally easier
  - With FML, overall chi-square statistic tests both regression coefficients and variance components, with RML only variance components are tested
  - Therefore if fixed portion of two models differ, must use FML for nested deviance tests

Computational Algorithms

- Several algorithms exist for existing HLM models:
  - Expectation-Maximization (EM)
  - Fisher scoring
  - Iterative Generalized Least Squares (IGLS)
  - Restricted IGLS (RIGLS)
- All are iterative search and evaluation procedures

Model Estimation

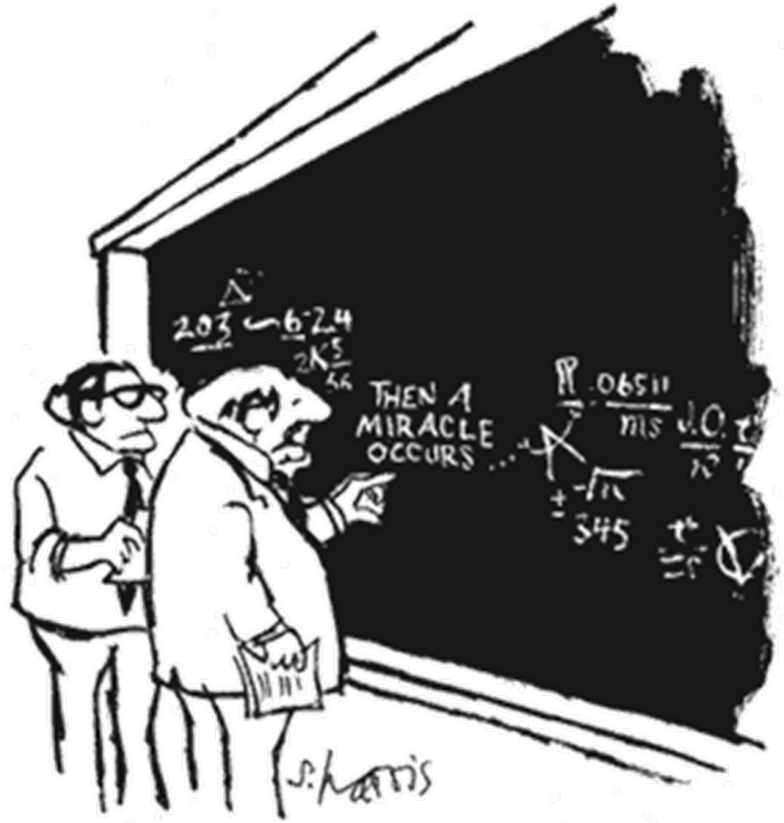
- Iterative estimation methods usually begin with a set of start values
- Start values are tentative values for the parameters in the model
  - Program begins with starting values (usually based on OLS regression at level 1)
  - Resulting parameter estimates are used as initial values for estimating the HLM model

Model Estimation

- Start values are used to solve model equations on first iteration
- This solution is used to compute initial model fit
- Next iteration involves search for better parameter values
- New values evaluated for fit, then a new set of parameter values tried
- When additional changes produce no appreciable improvement, iteration process terminates (convergence)
- Note that convergence and model fit are very different issues

Parameter estimation

- Coefficients and standard errors estimated through maximum likelihood procedures (usually)
  - The ratio of the parameter to its standard error produces a Wald test evaluated through comparison to the normal distribution (z)
  - In HLM software, a more conservative approach is used:
    - t-tests are used for significance testing
    - t-tests more accurate for fixed effects, small n, and nonnormal distributions)
- Standard errors
- Variance components



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

Parameter reliability

- Analogous to score reliability: ratio of true score variance to total variance (true score + error)
- In HLM, ratio of true parameter variance to total variability
- For exampl … pts, p … reliability, λ, is:

True variance of the sample means (estimated)

Variance of error of the sample means

$$\lambda_j = Var(\beta_{0j}) / Var(\bar{Y}_j) = \tau_{00}^2 / (\tau_{00}^2 + \sigma^2 / n_j)$$

True variance of the sample means (estimated)

Total variance of the sample means (observed)

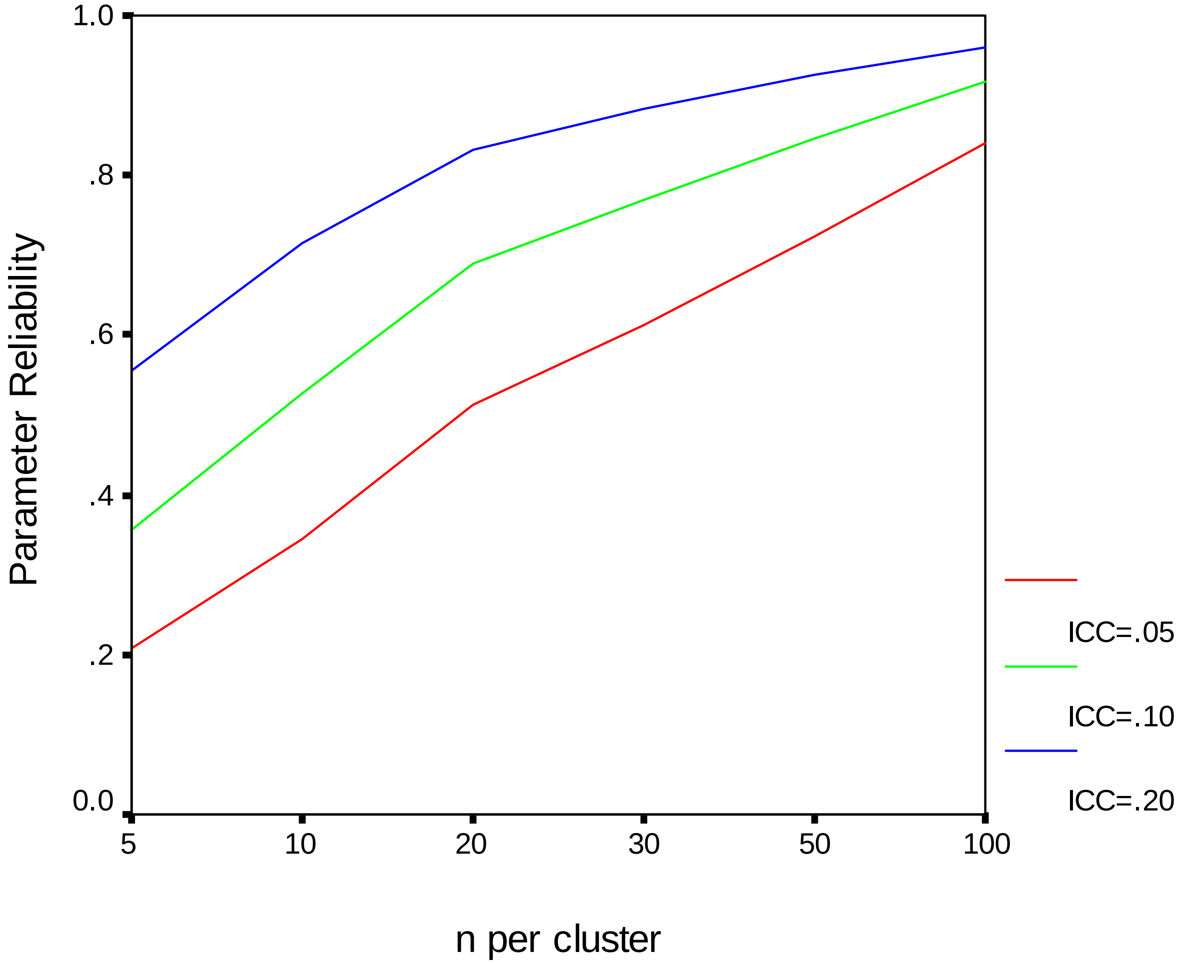
Parameter reliability

$$\lambda_j = \frac{n_j \rho_I}{1 + (n_j - 1)\rho_I}$$

	ICC (ρ_I)		
n_j	**.05**	**.10**	**.20**
5	.21	.36	.56
10	.34	.53	.71
20	.51	.69	.83
30	.61	.77	.88
50	.72	.85	.93
100	.84	.92	.96

Parameter reliability

$$\lambda_j = \frac{n_j \rho_I}{1 + (n_j - 1)\rho_I}$$



Predicting Group Effects

- It is often of interest to estimate the random group effects (β_{0j}, β_{1j})
- This is accomplished using Empirical Bayes (EB) estimation
- The basic idea of EB estimation is to predict group values using two kinds of information:
  - Group j data
  - Population data obtained from the estimation of the regression model

Empirical Bayes

- If information from only group j is used to estimate then we have the OLS estimate:

$$\beta_{0j} = \bar{Y}_j$$

- If information from only the population is used to estimate then the group is estimated from the grand mean:

$$\gamma_{00} = \bar{Y}_{..} = \sum_{j=1}^{N} \frac{n_j}{N} \bar{Y}_j$$

Empirical Bayes

- A third possibility is to c
  n
  on is a verage
  weighted by parameter re bility:

The smaller the reliability, the greater the weight of the grand mean

The larger the reliability, the greater the weight of the group mean

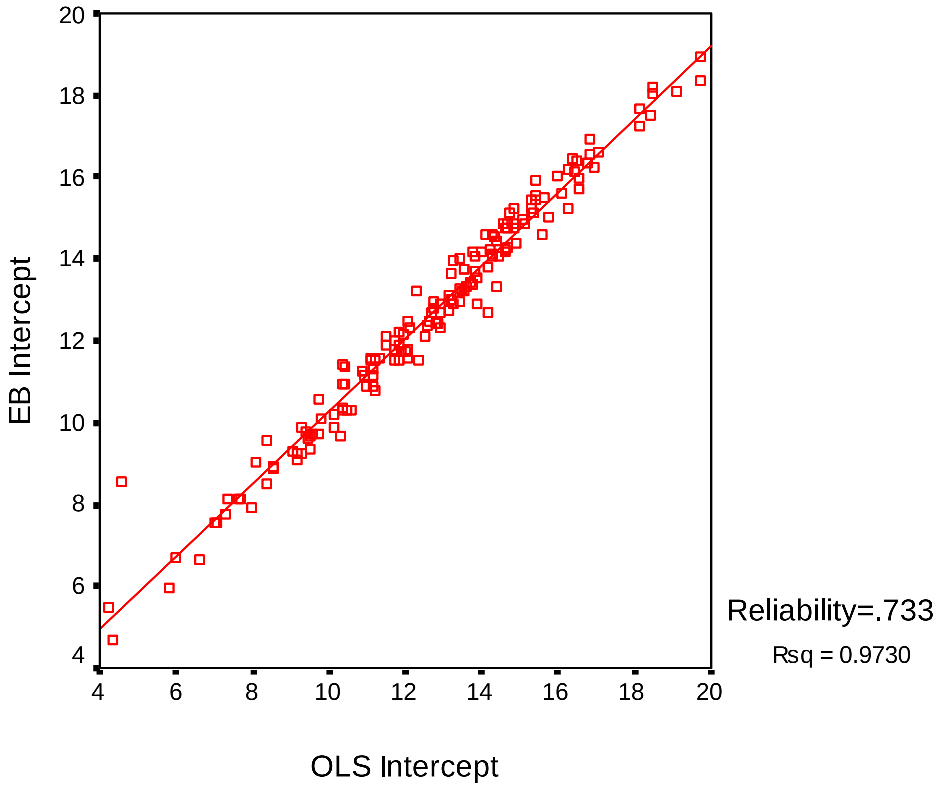
$$\beta_{0j}^{EB} = \lambda_j \beta_{0j} + (1 - \lambda_j)\,\gamma_{00}$$

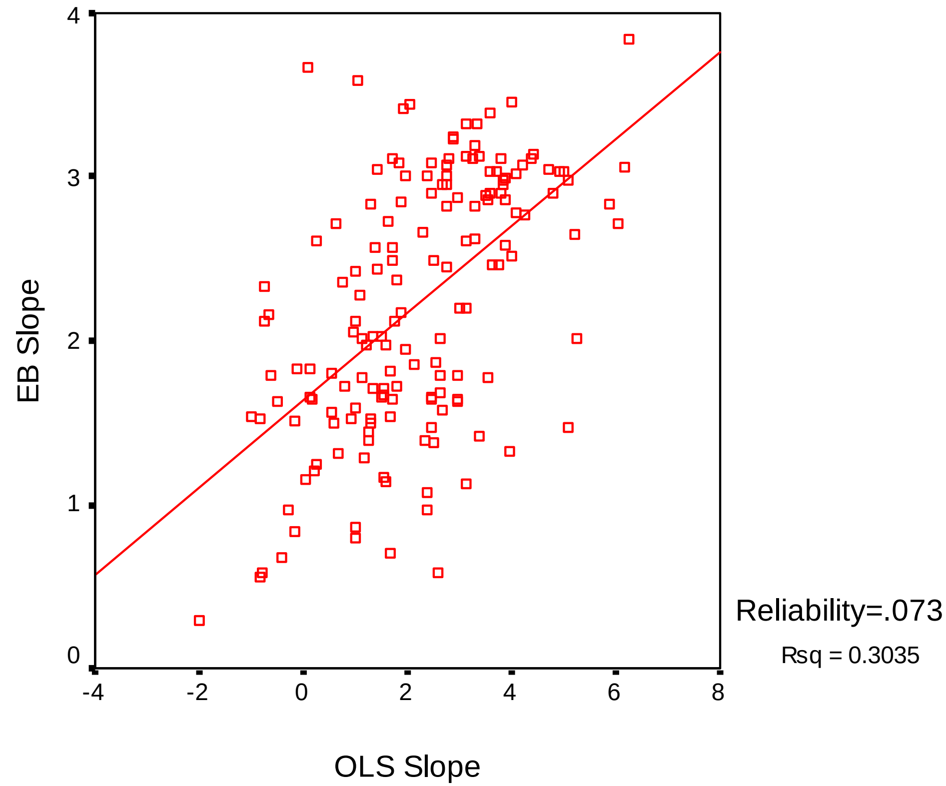
- This results in the "posterior means" or EB estimates

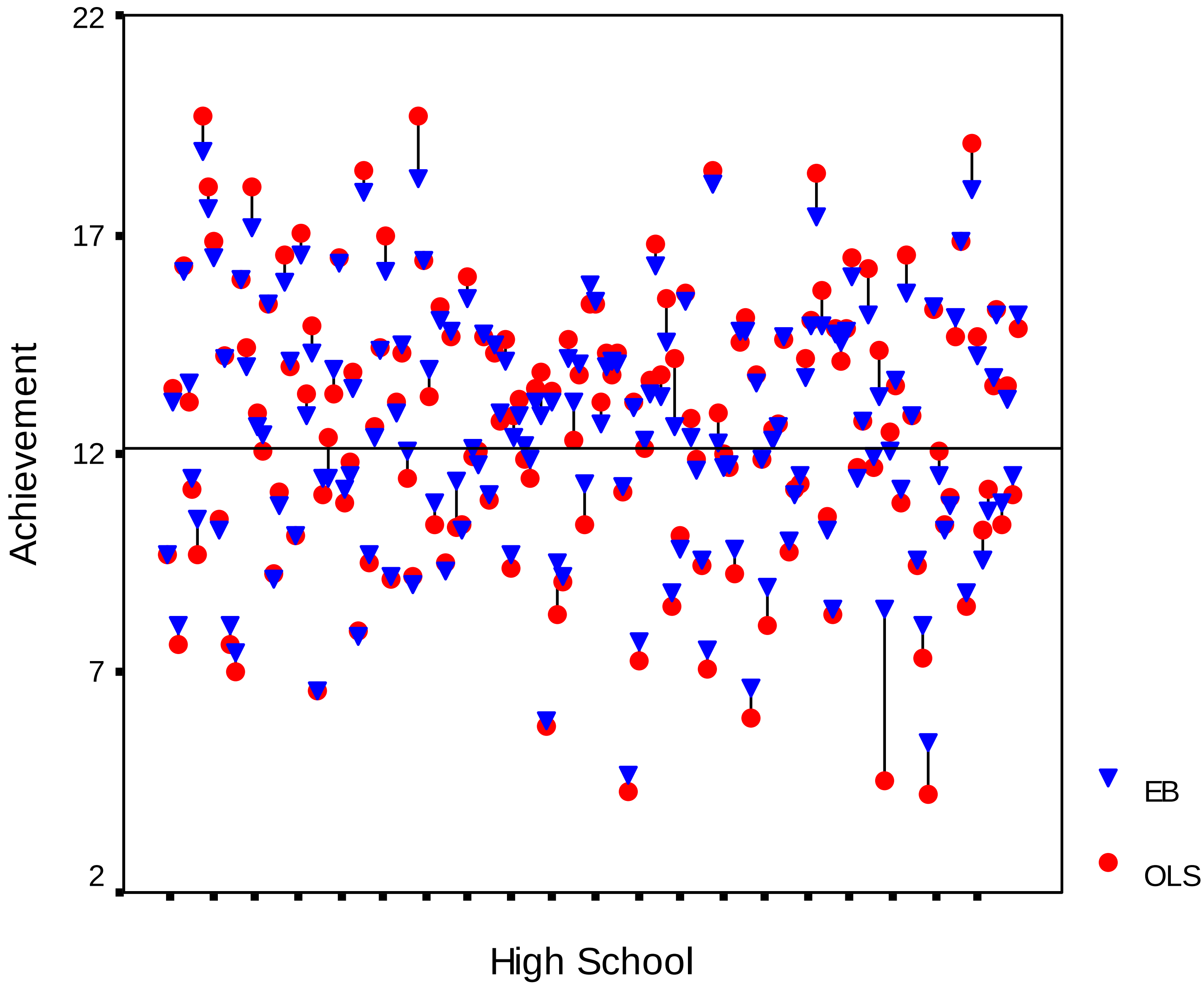
Bayesian Estimation

- Use of prior and posterior information improves estimation (depending on purpose)
- Estimates "shrink" toward the grand mean as shown in formula
- Amount of shrinkage depends on the "badness" of the unit estimate
  - Low reliability results in greater shrinkage (if $\lambda = 1$, there is no shrinkage; if $\lambda = 0$, shrinkage is complete, γ_{00})
  - Small n-size within a j unit results in greater shrinkage, "borrowing" from larger units

$$\beta_{0j}^{EB} = \lambda_j \beta_{0j} + (1 - \lambda_j)\gamma_{00}$$


EB Intercept
20
18
16
14
12
10
8
6
4
4
6
8
10
12
14
16
18
20
OLS Intercept
Reliability=.733
Rsq = 0.9730


EB Slope
4
3
2
1
0
-4
-2
0
2
4
6
8
OLS Slope
Reliability=.073
Rsq = 0.3035



22
17
12
7
2
Achievement
High School
EB
OLS

Intermission

"Frankly, Harold, you're beginning to bore everyone with your statistics."

Example 2: Meta-Analysis

- Can estimation techniques used in HLM provide a more sophisticated way to synthesize quantitative results across studies?
- Example from Raudenbush & Bryk (2002)
  - Teacher expectancy ("the Pygmalion effect")
  - Contentio
- Approach takes the standard error f effect size into account:

Note the effect of sample size on the standard error of the effect size

$$SE(d_j) = (\tau + V_j)^{1/2}, \text{ where } V_j = 1/\,(\mathrm{n}_j - 3)$$

Example 2: Meta-Analysis

- Term coined by Gene Glass in his 1976 AERA Presidential address
- An alternative to the traditional literature review
- Allows the reviewer to quantitatively combine and analyze the results from multiple studies
- Traditional literature review is based on the reviewer's analysis and synthesis of study themes or conclusions

What is Meta-Analysis (MA)?

- Meta-analysis
  - Collects empirical results from multiple studies
  - Expresses all results on a common scale, effect size
  - Can analyze covariates of effect size
  - Draws conclusions about the “overall” effect across studies no matter what the original study conclusions were
- Thus a MA becomes a research study on research studies, hence the term "meta"

Example 2: Meta-Analysis through HLM

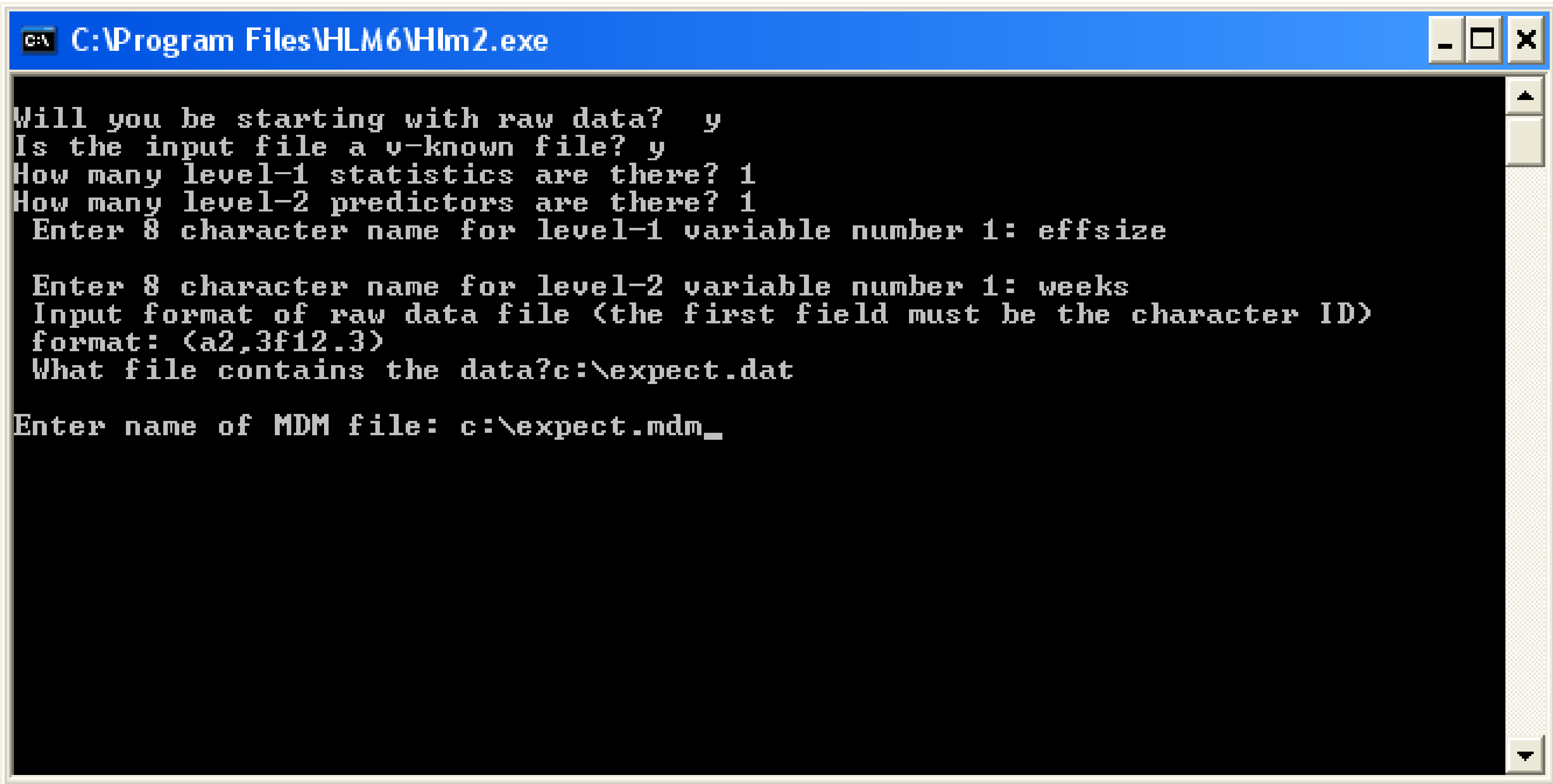
- Implemented through interactive DOS-based HLM programs rather than the Windows interface
- Involves estimation based on the observed variance-covariance matrix
- In this example, the v-c matrix is simply the study effect sizes and their standard errors
- Data file prepared with relevant variables (effect size, variance of effect size, predictors)
- Then an HLM ".mdm" file is created

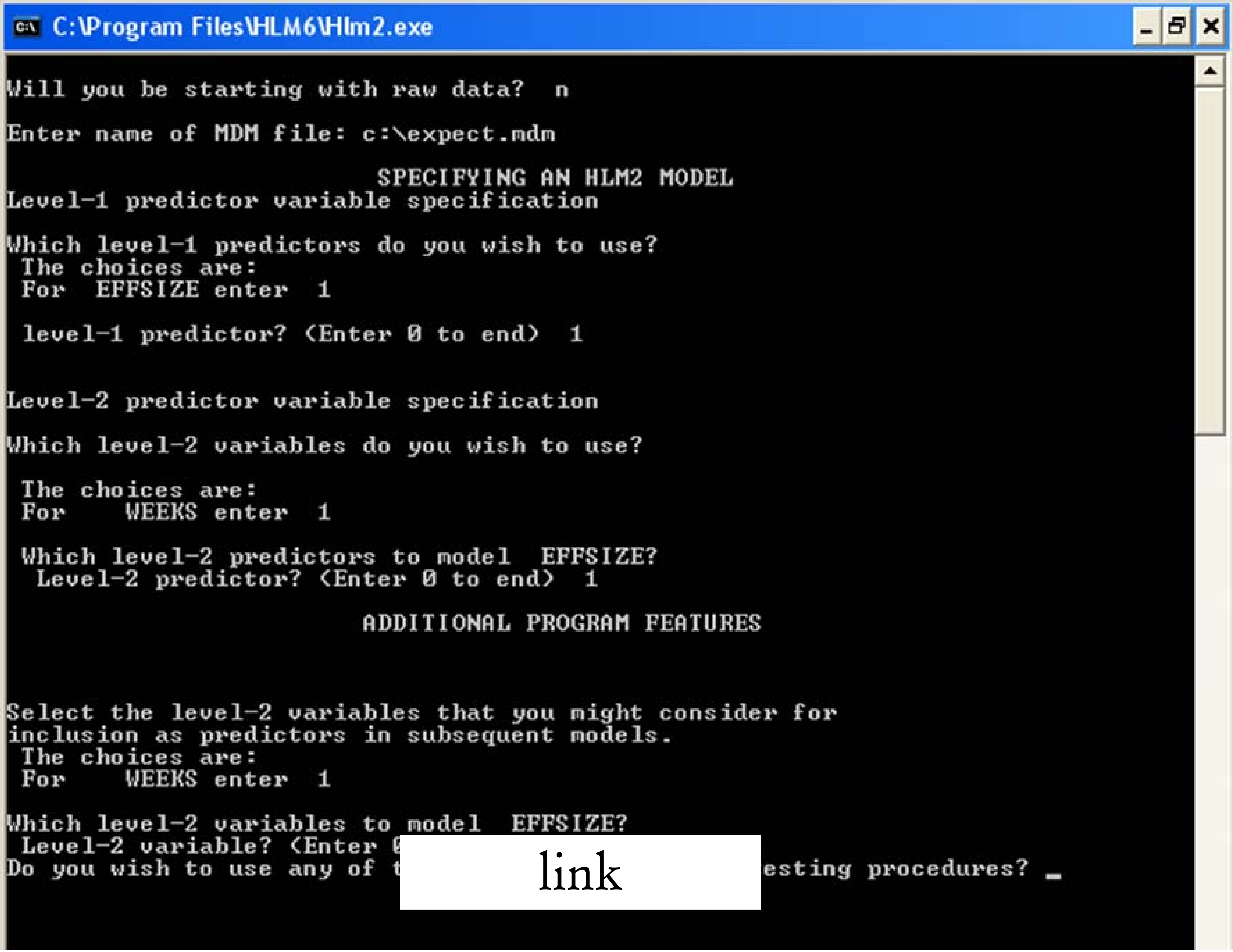
expect_data - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities S-PLUS Window Help

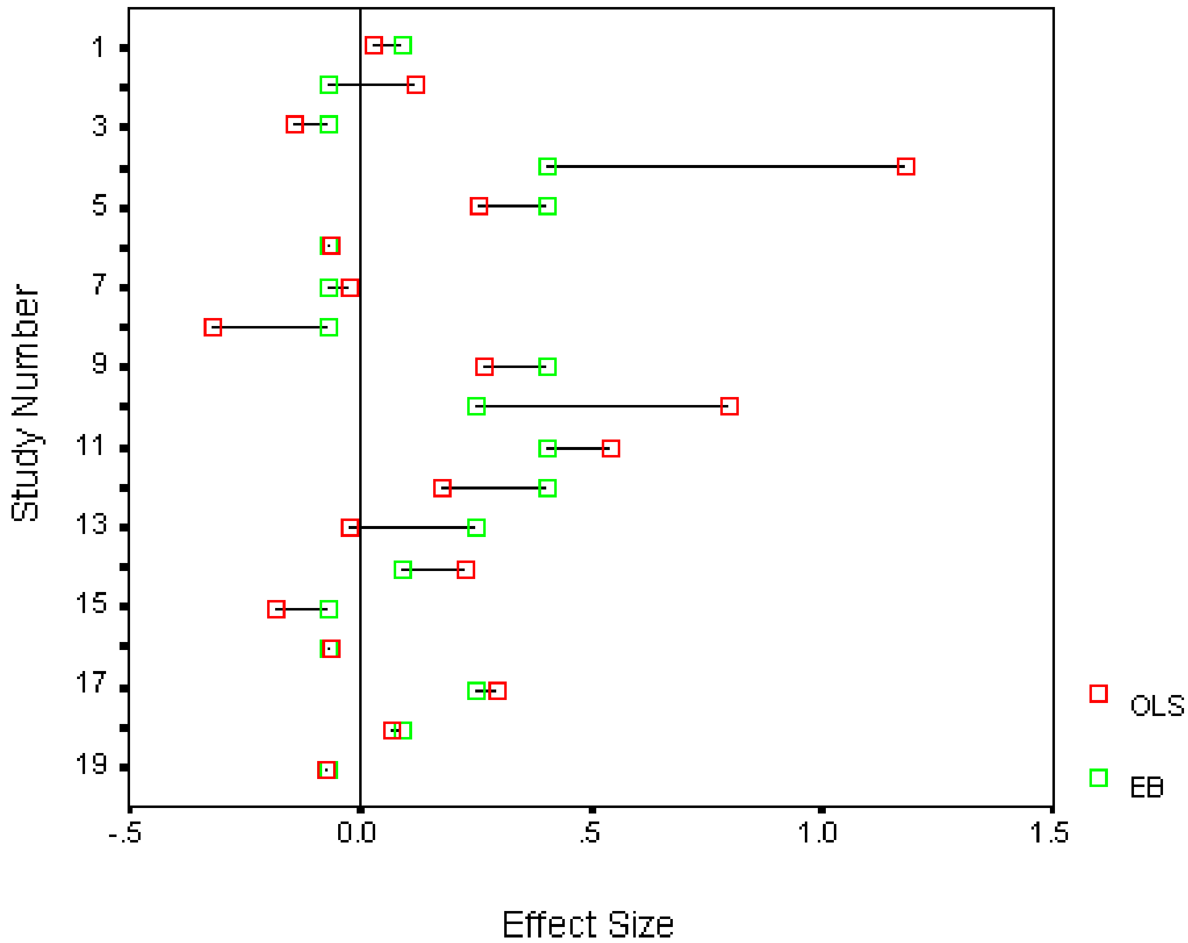
1 : studyid 1

	studyid	effsize	variance	weeks	var
1	1	0.030	.016	2	
2	2	0.120	.022	3	
3	3	-0.140	.028	3	
4	4	1.180	.139	0	
5	5	0.260	.136	0	
6	6	-0.060	.011	3	
7	7	-0.020	.011	3	
8	8	-0.320	.048	3	
9	9	0.270	.027	0	
10	10	0.800	.063	1	
11	11	0.540	.091	0	
12	12	0.180	.050	0	
13	13	-0.020	.084	1	
14	14	0.230	.084	2	
15	15	-0.180	.025	3	
16	16	-0.060	.028	3	
17	17	0.300	.019	1	
18	18	0.070	.009	2	
19	19	-0.070	.030	3	
20					
21					

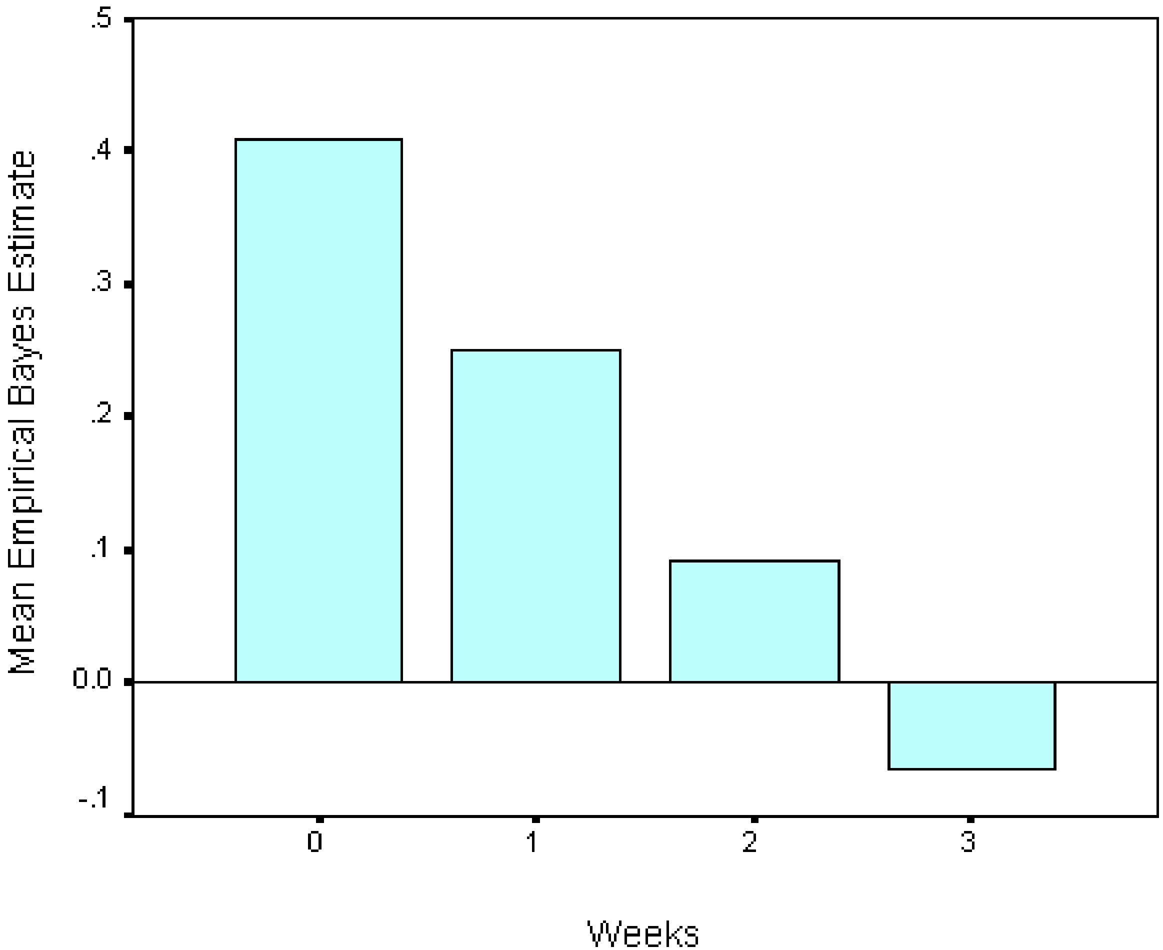

C:\Program Files\HLM6\Hlm2.exe
Will you be starting with raw data? y
Is the input file a v-known file? y
How many level-1 statistics are there? 1
How many level-2 predictors are there? 1
Enter 8 character name for level-1 variable number 1: effsize
Enter 8 character name for level-2 variable number 1: weeks
Input format of raw data file (the first field must be the character ID)
format: (a2,3f12.3)
What file contains the data?c:\expect.dat
Enter name of MDM file: c:\expect.mdm_


C:\Program Files\HLM6\Hlm2.exe
Will you be starting with raw data? n
Enter name of MDM file: c:\expect.mdm
SPECIFYING AN HLM2 MODEL
Level-1 predictor variable specification
Which level-1 predictors do you wish to use?
The choices are:
For EFFSIZE enter 1
level-1 predictor? (Enter 0 to end) 1
Level-2 predictor variable specification
Which level-2 variables do you wish to use?
The choices are:
For WEEKS enter 1
Which level-2 predictors to model EFFSIZE?
Level-2 predictor? (Enter 0 to end) 1
ADDITIONAL PROGRAM FEATURES
Select the level-2 variables that you might consider for
inclusion as predictors in subsequent models.
The choices are:
For WEEKS enter 1
Which level-2 variables to model EFFSIZE?
Level-2 variable? (Enter 0
Do you wish to use any of t
link
esting procedures?

The HLM analysis allows the use of Bayesian estimation methods to temper the estimates of study effect sizes

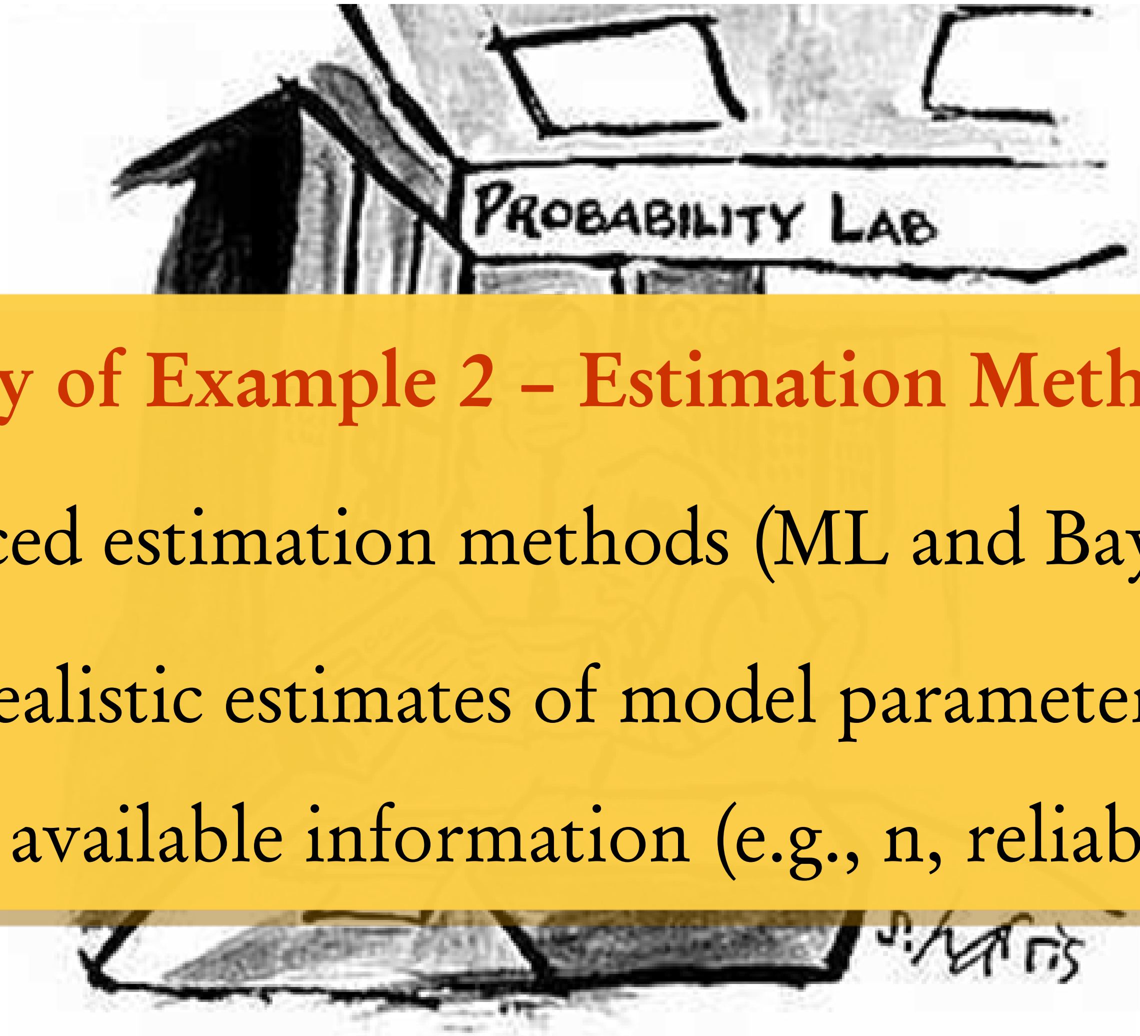


Use of a covariate to account for variation in study effect size: Teacher expectancy as a function of unfamiliarity



Summary of Example 2 – Estimation Methods

- Advanced estimation methods (ML and Bayesian)
- More realistic estimates of model parameters tempered by available information (e.g., n, reliability)



"According to this theory, it's strongly improbable that anything should ever happen anytime, anywhere."

Example 3: Longitudinal Models

- Growth models as an Alternative to NCLB Adequate Yearly Progress (AYP)
- HLM as a more flexible means to model repeated measures
  - Individual growth curves
  - Ability to model growth parameters

See Stevens (2005), Stevens & Zvoch (2006)

No Child Left Behind

- Purpose of legislation is to ensure the learning of all children
- Schools (and districts and states) judged on whether a sufficient proportion of students are learning each year
- Measure and report “Adequate Yearly Progress” (AYP) in each content area
- Disaggregation of results by ethnicity, economic advantage, disability, and ELL
- But does NCLB AYP validly reflect student learning?

No Child Left Behind

- NCLB and other recent federal mandates and programs place strong emphasis on "evidence based" or "scientifically based" research.

- Scientifically based research "...means research that involves the application of rigorous, systematic, and objective procedures to obtain reliable and valid knowledge relevant to education activities and programs" (NCLB, 2001)

No Child Left Behind

- However, NCLB methods appear to contradict the federal push for more rigorous, scientifically based evidence

- Collectively, NCLB regulations prescribe an unusual form of case study design that must be used to evaluate school effectiveness for AYP

NCLB accountability requirements impose a nonequivalent-groups, case study design for the evaluation of school effectiveness:

	Year 1	Year 2	Year 3
Group A (4th grade)	$X^{?}$ O_1		
Group B (4th grade)		$X^{?}$ O_2	
Group C (4th grade)			$X^{?}$ O_3

- $X^{?}$ is used to indicate unknown treatment implementation
- AYP in NCLB is a simple comparison of one O_t to a calculated target for improvement

How to Measure School Effectiveness?

- Estimating the impact a school has on students is a complex task; a problem in research or program evaluation design
- One of the most important challenges is separating "intake" to the school from "value added" by the school
- Raudenbush and Willms (1995) Type A and Type B effects or total causal effects vs. school effects
- Intake represents confounding pre-existing student differences as well as previous learning
- Intake also represents differences in group composition from school to school

The Analysis of Change

- Cross sectional comparisons do not likely measure change effectively/accurately
- Individual growth curve analysis an important tool for analyzing change
- HLM models are one mechanism for estimating growth curves
- Height analogy

Analogy: Measuring Physical Development

Measure Height

2004

Measuring Height, NCLB Method

2004

Actual Height

Height AYP for 2004

AYP defined by requiring 100% of children to be at least 6'0" by 2014 and projecting backwards to year in which height is first measured

All children must grow enough in each year to show AYP; all children must be tall by 2014

Get your facts first, and then you can distort them as much as you please.

- Mark Twain quoted by Rudyard Kipling in *From Sea to Shining Sea*

Measuring Height Using Longitudinal Methods

2004

Growth

2005

Longitudinal models using HLM

- Level 1 defined as repeated measurement occasions
- Levels 2 and 3 defined as higher levels in the nested structure
- For example, longitudinal analysis of student achievement:

  Level 1 = achievement scores at times 1 – t

  Level 2 = student characteristics

  Level 3 = school characteristics

Longitudinal models

- Three important advantages of the HLM approach to repeated measures:
  - Times of measurement can vary from one person to another
  - Data do not need to be complete on all measurement occasions
  - Growth parameters can be modeled at higher levels

HLM Longitudinal models

Level-1

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}(\text{time}) + e_{tij}$$

Level-2

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}(X_{ij}) + r_{0ij}$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11j}(X_{ij}) + r_{1ij}$$

Level-3

$$\beta_{00j} = \gamma_{000} + \gamma_{001}(W_{1j}) + u_{00j}$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101}(W_{1j}) + u_{10j}$$

Curvilinear Longitudinal models

Level-1

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}(time) + \boxed{\pi_{2ij}(time^2)} + e_{tij}$$

Level-2

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}(X_{ij}) + r_{0ij}$$
$$\pi_{1ij} = \beta_{10j} + \beta_{11j}(X_{ij}) + r_{1ij}$$
$$\boxed{\pi_{2ij} = \beta_{20j} + \beta_{21j}(X_{ij}) + r_{2ij}}$$

Level-3

$$\beta_{00j} = \gamma_{000} + \gamma_{001}(W_{1j}) + u_{00j}$$
$$\beta_{10j} = \gamma_{100} + \gamma_{101}(W_{1j}) + u_{10j}$$
$$\boxed{\beta_{20j} = \gamma_{200} + \gamma_{201}(W_{2j}) + u_{20j}}$$

Mathematics Achievement Predicted by Individual Characteristics

Fixed Effect	*Coefficient*	*SE*	*t*	*df*	*p*
School Mean Achievement, γ_{000}	663.54	1.28	513.86	241	< .001
White Student, γ_{010}	14.62	0.77	18.88	241	< .001
LEP, γ_{020}	-16.00	1.19	-13.50	241	< .001
Title 1 Student, γ_{030}	-11.10	1.44	-7.71	241	< .001
Special Education, γ_{040}	-33.09	1.88	-17.62	241	< .001
Modified Test, γ_{050}	-16.83	2.63	-6.40	241	< .001
Free Lunch Student, γ_{060}	-7.75	1.13	-6.85	241	< .001
Gender, γ_{070}	-1.21	0.59	-2.03	241	.042
School Linear Growth, γ_{100}	19.40	0.70	27.88	241	< .001
White Student, γ_{110}	-1.20	0.64	-1.86	241	.062
LEP, γ_{120}	0.70	1.13	0.60	241	.547
Title 1 Student, γ_{130}	-2.58	0.95	-2.72	241	.007
Special Education, γ_{140}	-2.16	1.67	-1.29	241	.196
Modified Test, γ_{150}	-2.43	2.47	-0.99	241	.325
Free Lunch Student, γ_{160}	-0.75	1.03	-0.73	241	.466
Gender, γ_{170}	-4.68	0.59	-7.98	241	< .001

Mathematics Achievement Predicted by Individual Characteristics (continued)

Fixed Effect	*Coefficient*	*SE*	*t*	*df*	*p*
School Curvilinear Growth, γ_{200}	-2.09	0.21	-9.78	241	< .001
White Student, γ_{210}	0.48	0.20	2.35	241	.019
LEP, γ_{220}	-0.10	0.36	-0.27	241	.790
Title 1 Student, γ_{230}	0.61	0.28	2.17	241	.030
Special Education, γ_{240}	0.61	0.50	1.22	241	.224
Modified Test, γ_{250}	-0.10	0.75	-0.14	241	.890
Free Lunch Student, γ_{260}	0.26	0.33	0.79	241	.427
Gender, γ_{270}	1.05	0.19	5.64	241	< .001

School Level Variance Component	*Level-1*	*Level-2*	*Variance Explained*
Mean Achievement, u_{00}	242.78	184.89	23.8%
Linear Growth, u_{10}	41.46	30.68	26.0%
Curvilinear Growth, u_{10}	2.94	2.60	11.6%

Mathematics Achievement Predicted by School Characteristics

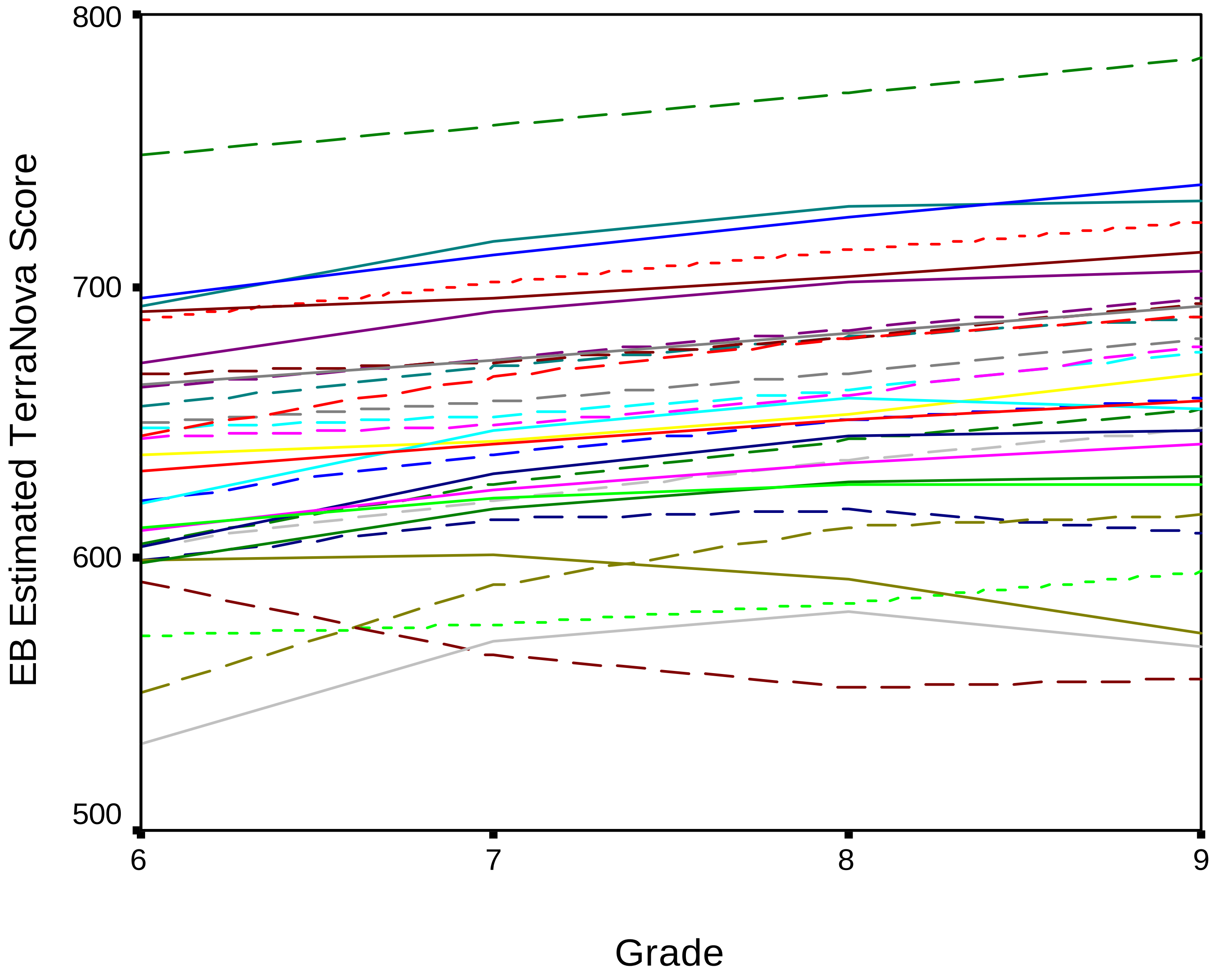
Fixed Effect	*Coefficient*	*SE*	*t*	*df*	*p*
School Mean Achievement, γ_{000}	662.53	1.07	620.80	237	< .001
Percent Bilingual Students, γ_{001}	4.19	4.00	1.05	237	.295
Percent LEP Students, γ_{0o2}	-0.99	4.56	-0.22	237	.828
Percent White Students, γ_{003}	19.55	3.72	5.25	237	< .001
Percent Free Lunch, γ_{004}	-5.29	3.18	-1.67	237	.096
School Mean Linear Growth, γ_{100}	19.18	0.71	26.87	237	< .001
Percent Bilingual Students, γ_{101}	-0.17	1.98	-0.09	237	.932
Percent LEP Students, γ_{102}	2.90	2.85	1.02	237	.309
Percent White Students, γ_{003}	3.51	2.74	1.28	237	.201
Percent Free Lunch, γ_{004}	-3.67	2.23	-1.65	237	.099

Mathematics Achievement Predicted by School Characteristics

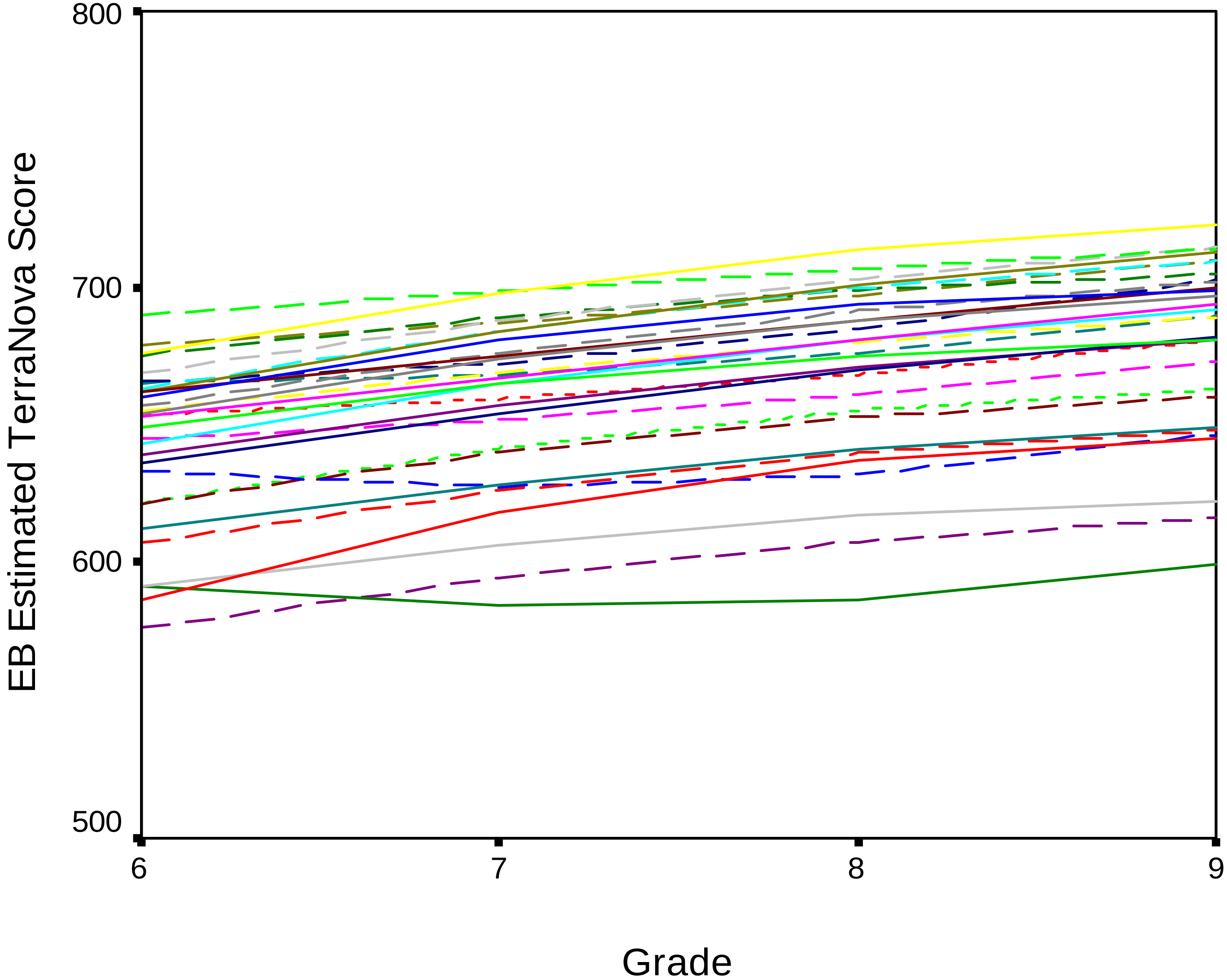
Fixed Effect	*Coefficient*	*SE*	*t*	*df*	*p*
School Curvilinear Growth, γ_{200}	-1.99	0.22	-9.10	237	$< .001$
Percent Bilingual Students, γ_{201}	-0.12	0.57	-0.21	237	.834
Percent LEP Students, γ_{202}	0.39	0.84	0.46	237	.643
Percent White Students, γ_{203}	-1.11	0.75	-1.48	237	.138
Percent Free Lunch, γ_{204}	-1.17	0.64	1.84	237	.065

School Level Variance Component	*Level-1*	*Level-2*	*Level-3*	*Variance Explained**
Mean Achievement, u_{00}	242.78	184.89	123.96	33.0%
Linear Growth, u_{10}	41.46	30.68	29.54	3.7%
Curvilinear Growth, u_{10}	2.94	2.60	2.49	4.2%

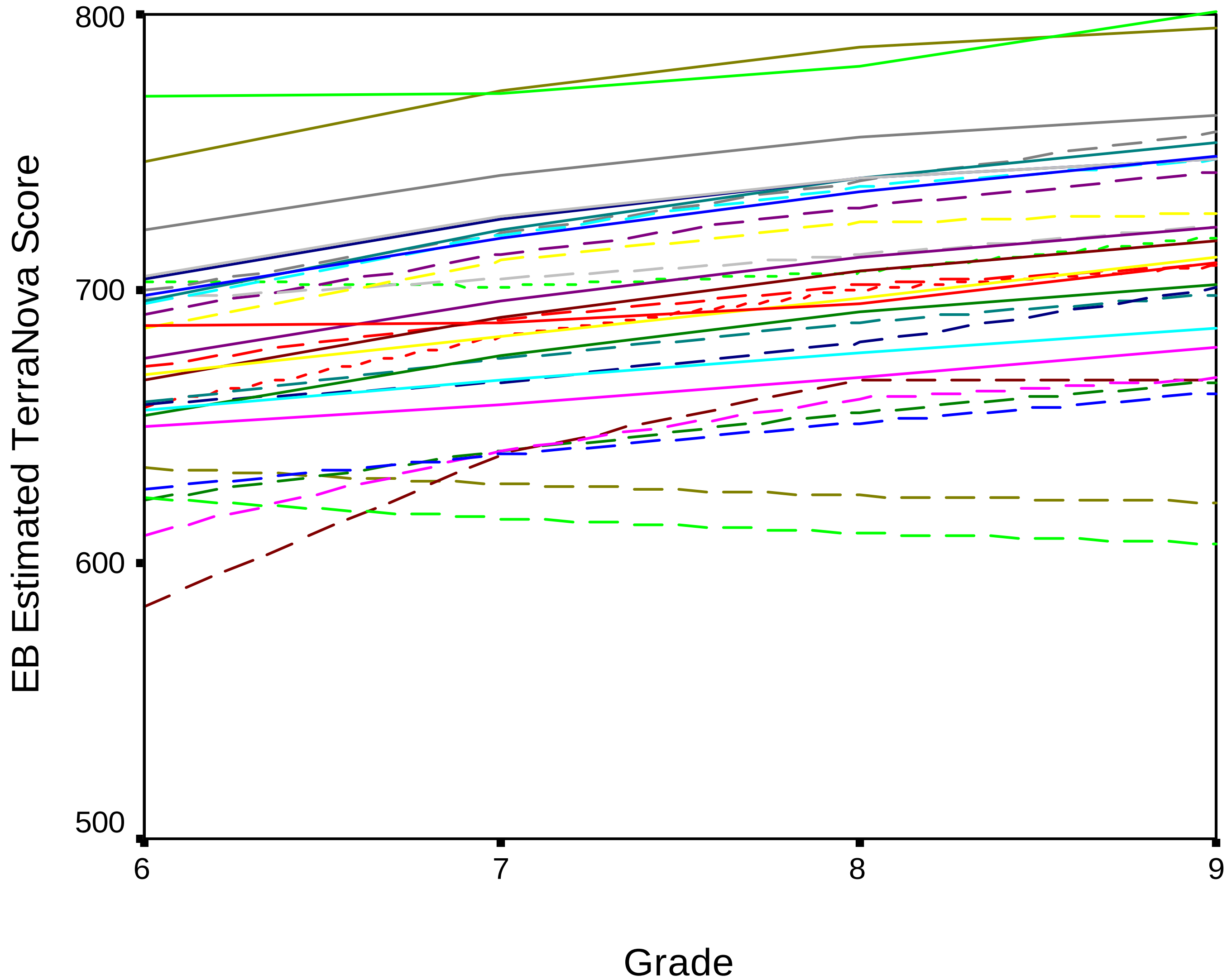
* Percent level 2 residual variance explained by level 3 model.



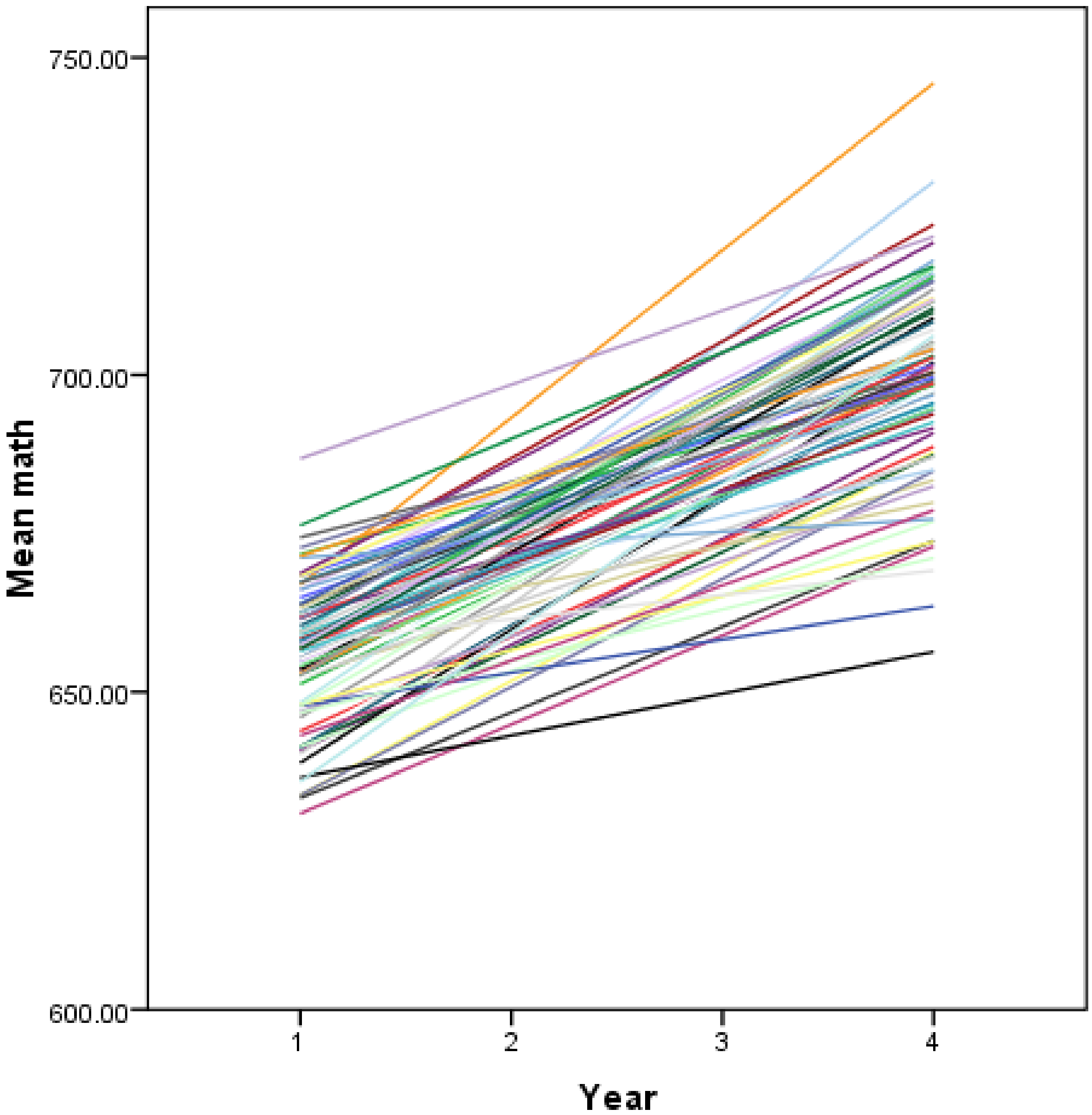
Hispanic Students



Native American Students



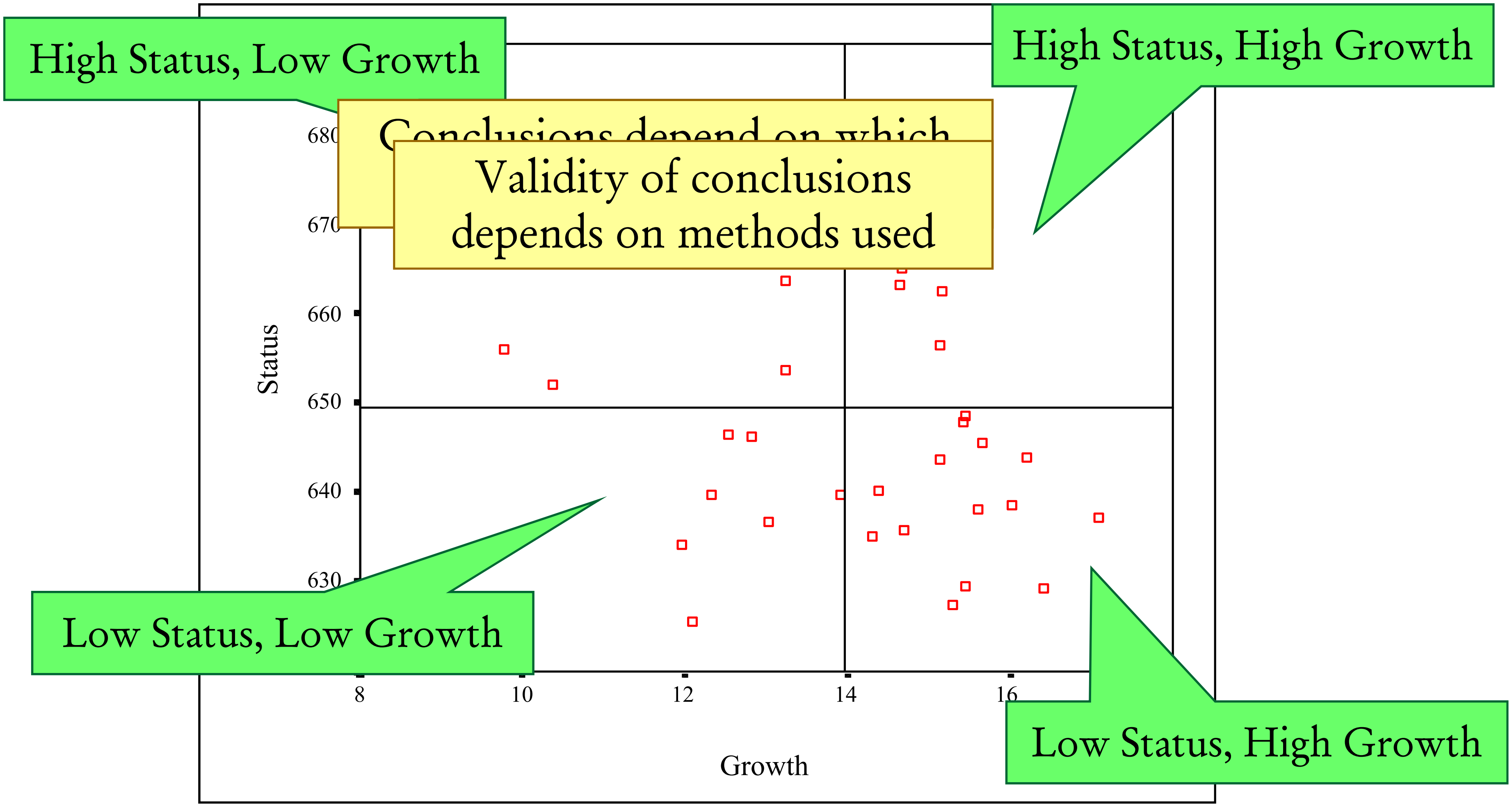
White Students

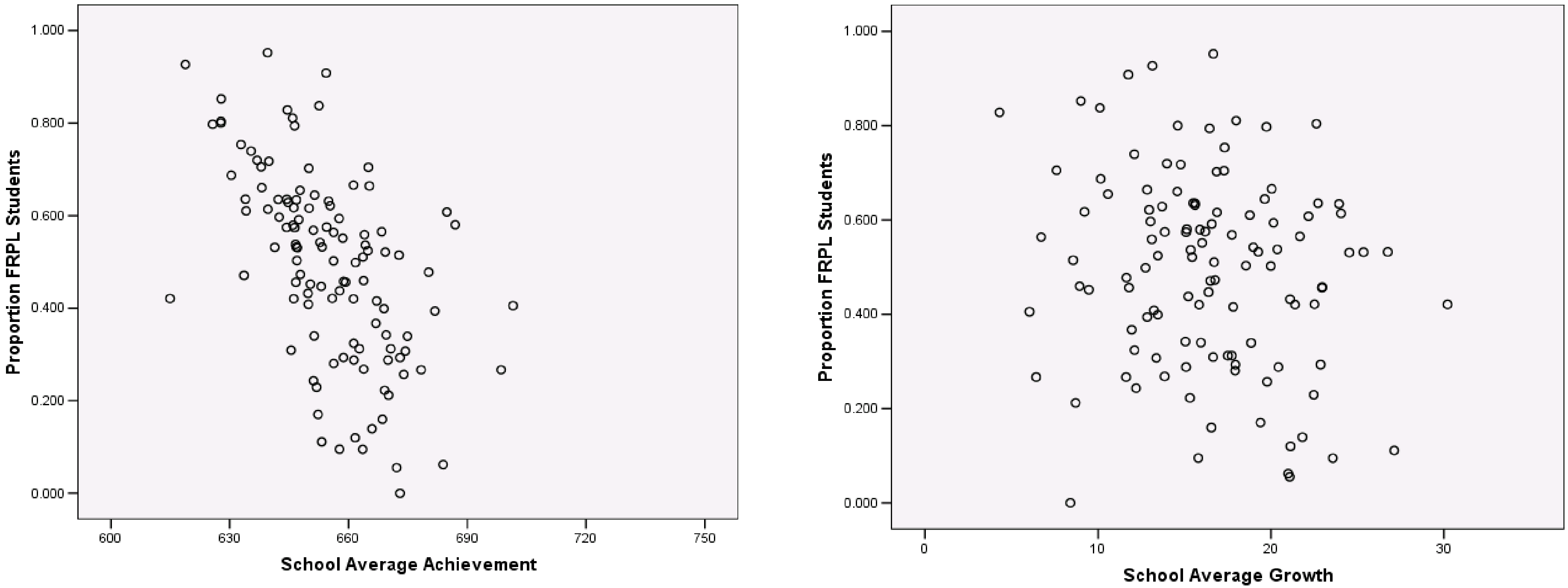

750.00
700.00
650.00
600.00
Mean math
1
2
3
4
Year

Achievement Status versus Achievement Growth

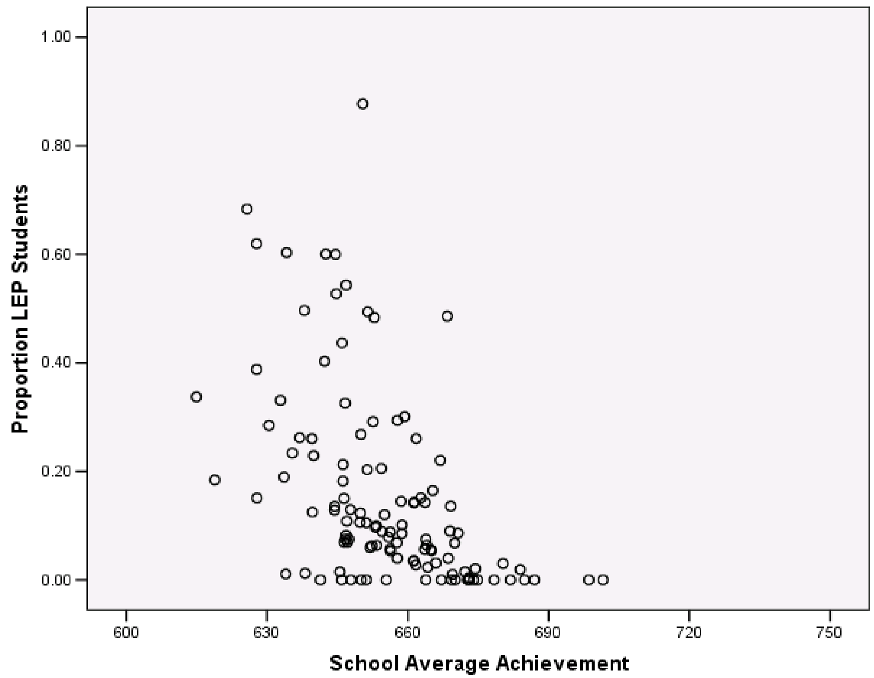
- Also important to note that the two research design approaches and the two parameters represent very different things
- In this example, the correlation between status and growth parameters was -. 378
  - Has important policy implications
  - Varies substantially across content, assessment, and state system

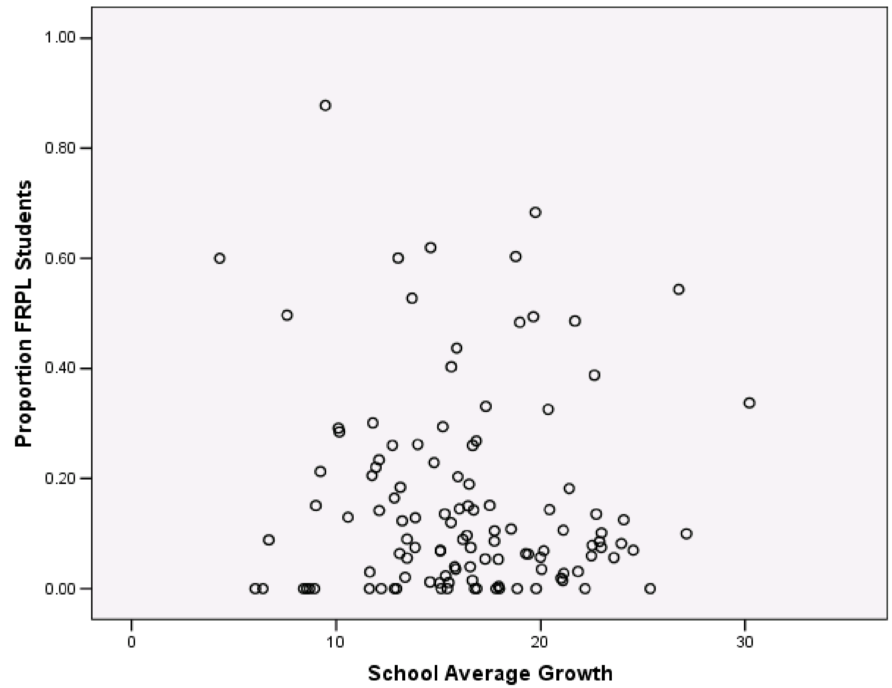
Inferences about School Performance



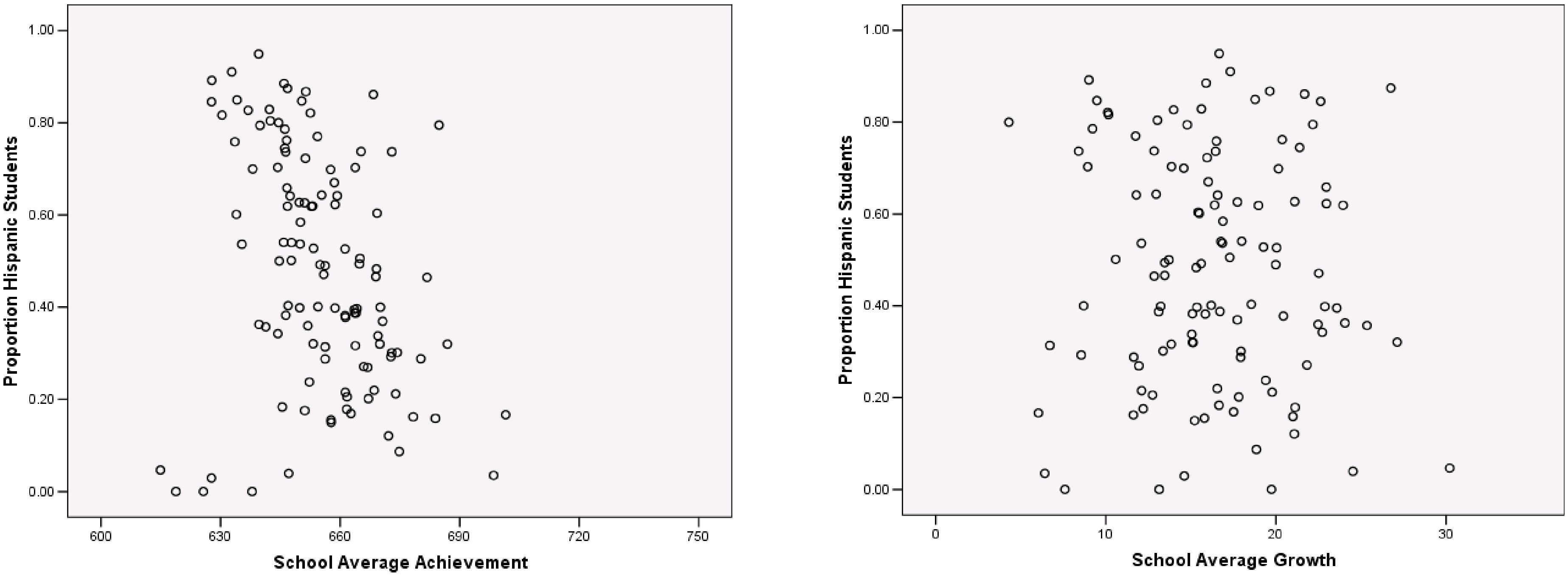


Relationships Between the Proportion of Free-Reduced Price Lunch (FRPL) in the School and Status (*r* = -.56) or Growth (*r* = -.17) in New Mexico Middle School Mathematics Achievement.

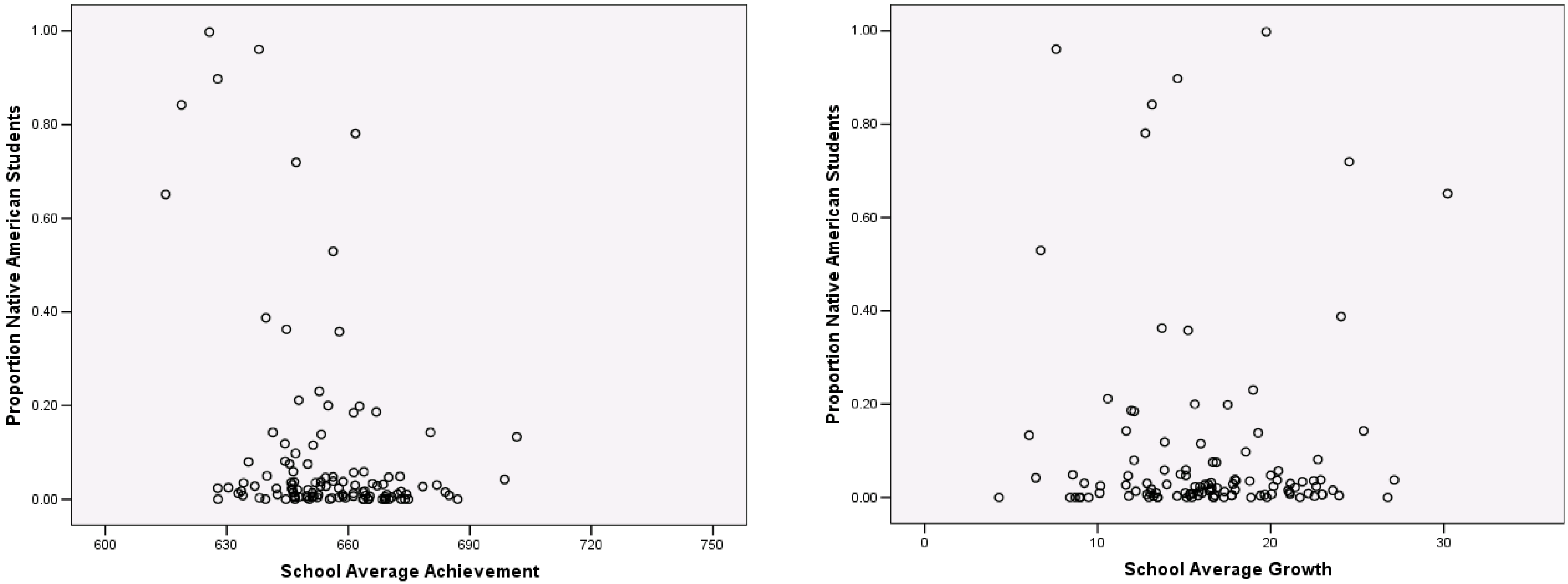




Relationships Between the Proportion of Limited English Proficient (LEP) Students in the School and Status (r = -.51) or Growth (r = -.06) in New Mexico Middle School Mathematics Achievement.

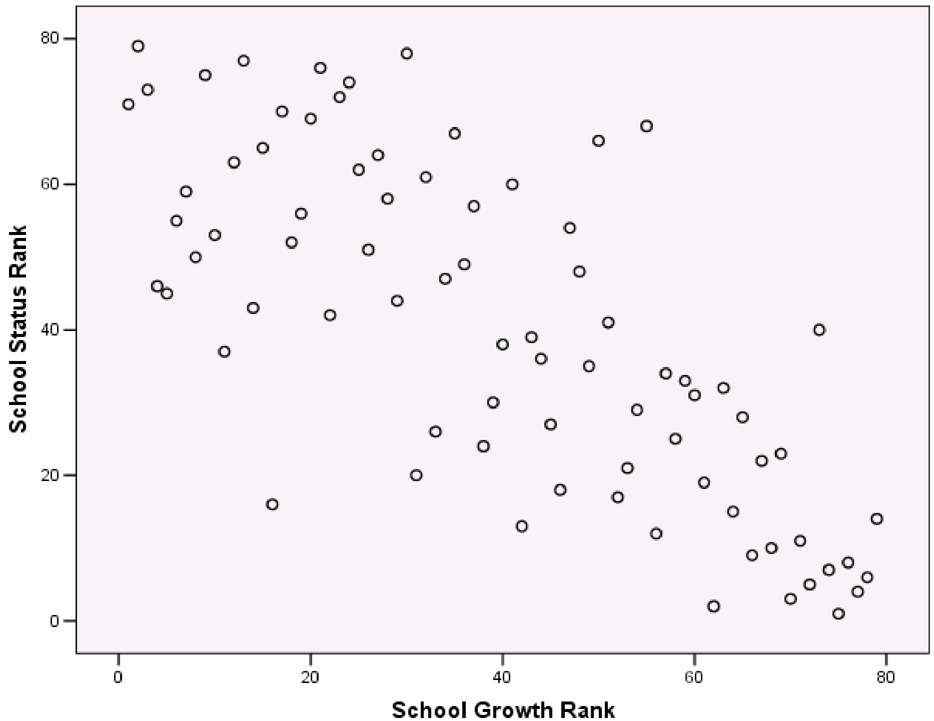


Relationships Between the Proportion of Hispanic Students in the School and Status (r = -.30) or Growth (r = -.05) in New Mexico Middle School Mathematics Achievement.



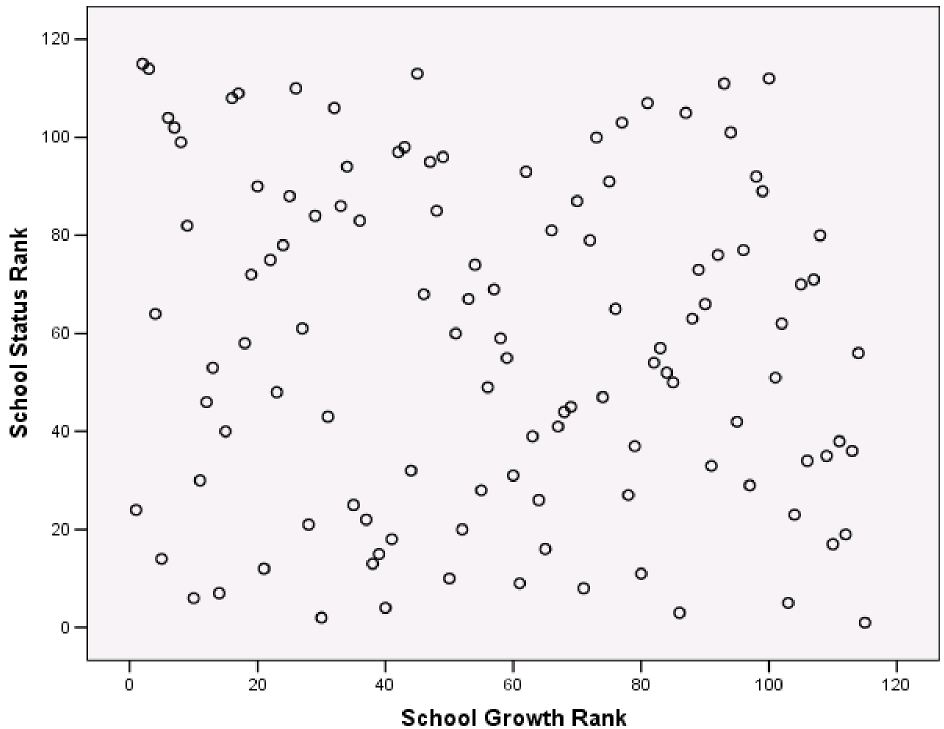
Relationships Between the Proportion of Native American Students in the School and Status (r = -.35) or Growth (r = -.02) in New Mexico Middle School Mathematics Achievement.

Classifying Schools using Status or Growth



Relationship Between Schools Ranked on Status and Schools Ranked on Growth (r = -.75) in New Mexico Elementary School Reading Achievement.

Classifying Schools using Status or Growth: Rankings can Differ Substantially



Relationship Between Schools Ranked on Status and Schools Ranked on Growth in New Mexico Middle School Mathematics Achievement (r = -.12).

Example 4: Interrupted Time Series

- Variation on longitudinal growth models presented earlier
- Flexible modeling of intervention effects over time
- In progress study of reading intervention in Bethel School District
  - Examine effects of time of intervention on reading performance
  - Examine effects of "dosage" of intervention on reading performance

Interrupted Time Series Designs: Change in Intercept

$$Y_{ij} = \pi_{0i} + \pi_{1i} Time_{ij} + \pi_{2i} Treatment_{ij} + \varepsilon_{ij}$$

Treatment is coded 0 or 1

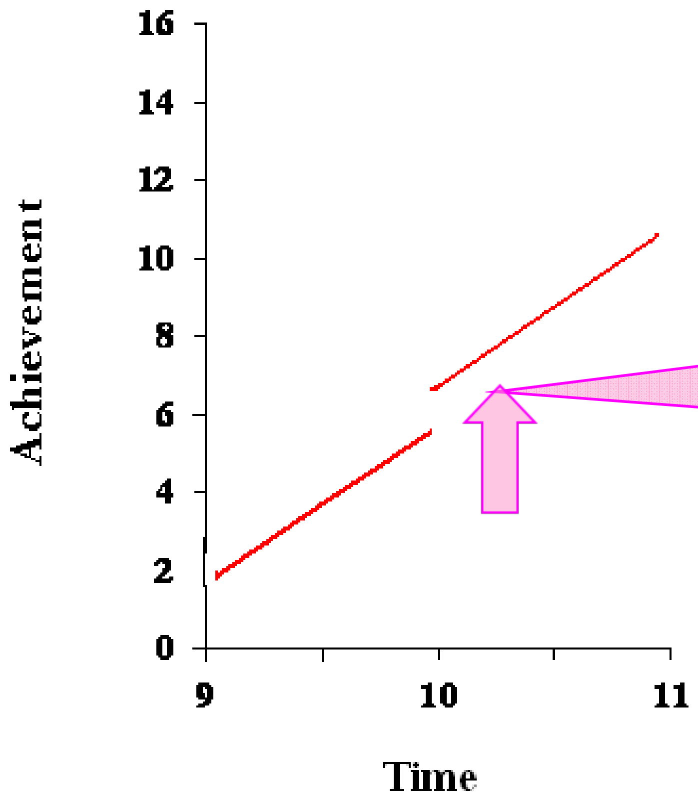
When Treatment = 0:

$$Y_{ij} = \pi_{0i} + \pi_{1i} Time_{ij} + \varepsilon_{ij}$$

When Treatment = 1:

$$Y_{ij} = (\pi_{0i} + \pi_{2i}) + \pi_{1i} Time_{ij} + \varepsilon_{ij}$$

Interrupted Time Series Designs



Treatment effect on level: $(\pi_{0i} + \pi_{2i})$

Interrupted Time Series Designs: Change in Slope

When Treatment = 1:

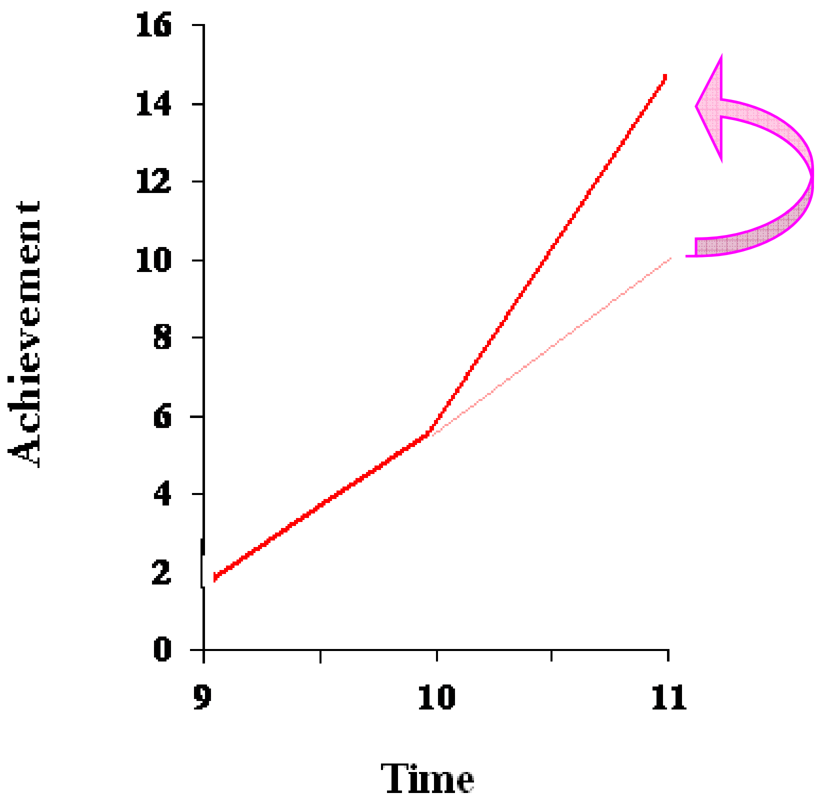
$$Y_{ij} = \pi_{0i} + \pi_{1i} Time_{ij} + \pi_{3i} TreatmentTime_{ij} + \varepsilon_{ij}$$

Treatment time expressed as 0's before treatment and time intervals post-treatment (i.e., 0, 0, 0, 1, 2, 3

When Treatment = 0:

$$Y_{ij} = \pi$$

Interrupted Time Series Designs



Treatment effect on slope: $(+\pi_{3i})$

Change in Intercept and Slope

$$Y_{ij} = \pi_{0i} + \pi_{1i}Time_{ij} + \pi_{2i}Treatment + \pi_{3i}TreatmentTime_{ij} + \varepsilon_{ij}$$

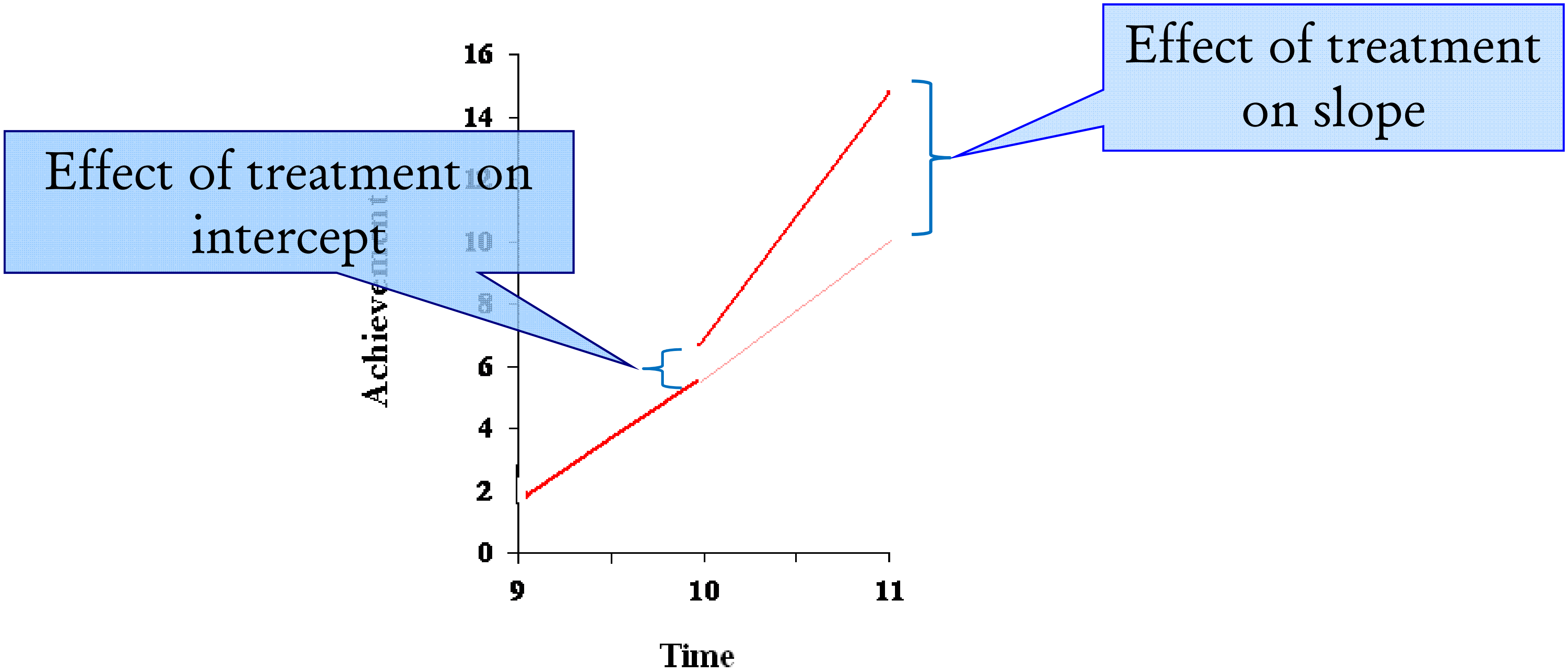
Effect of treatment on intercept

Effect of treatment on slope

When Treatment = 0:

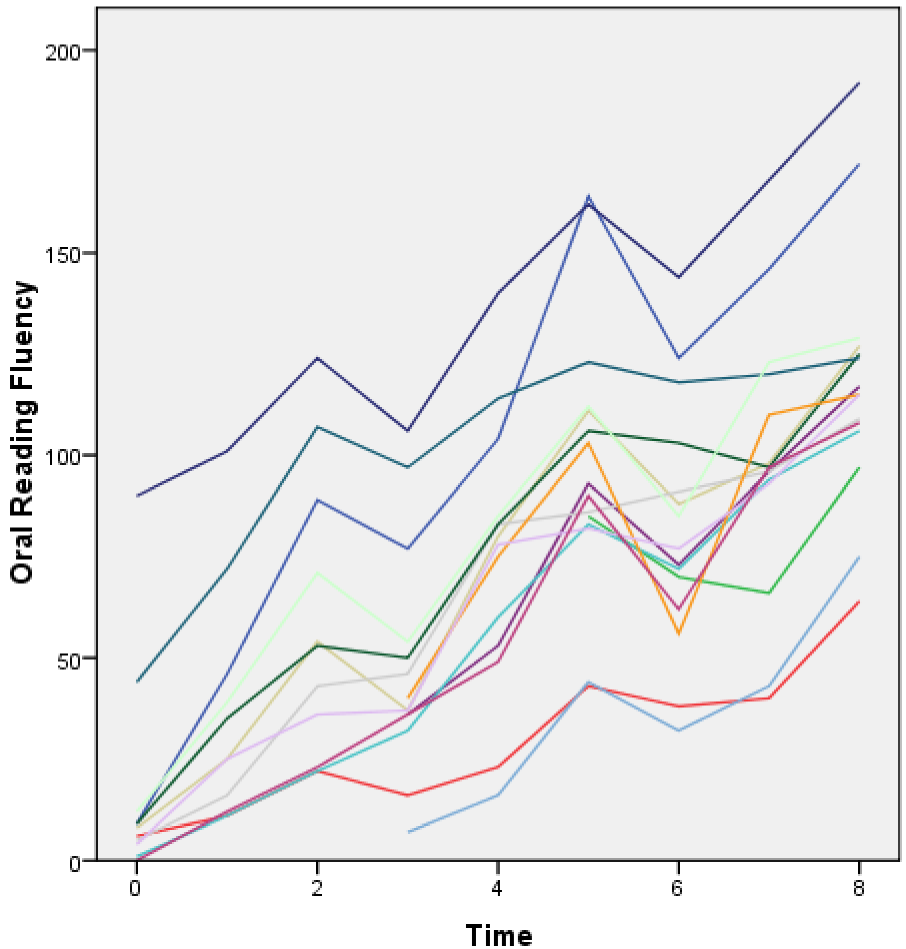
$$Y_{ij} = \pi_{0i} + \pi_{1i}Time_{ij} + \varepsilon_{ij}$$

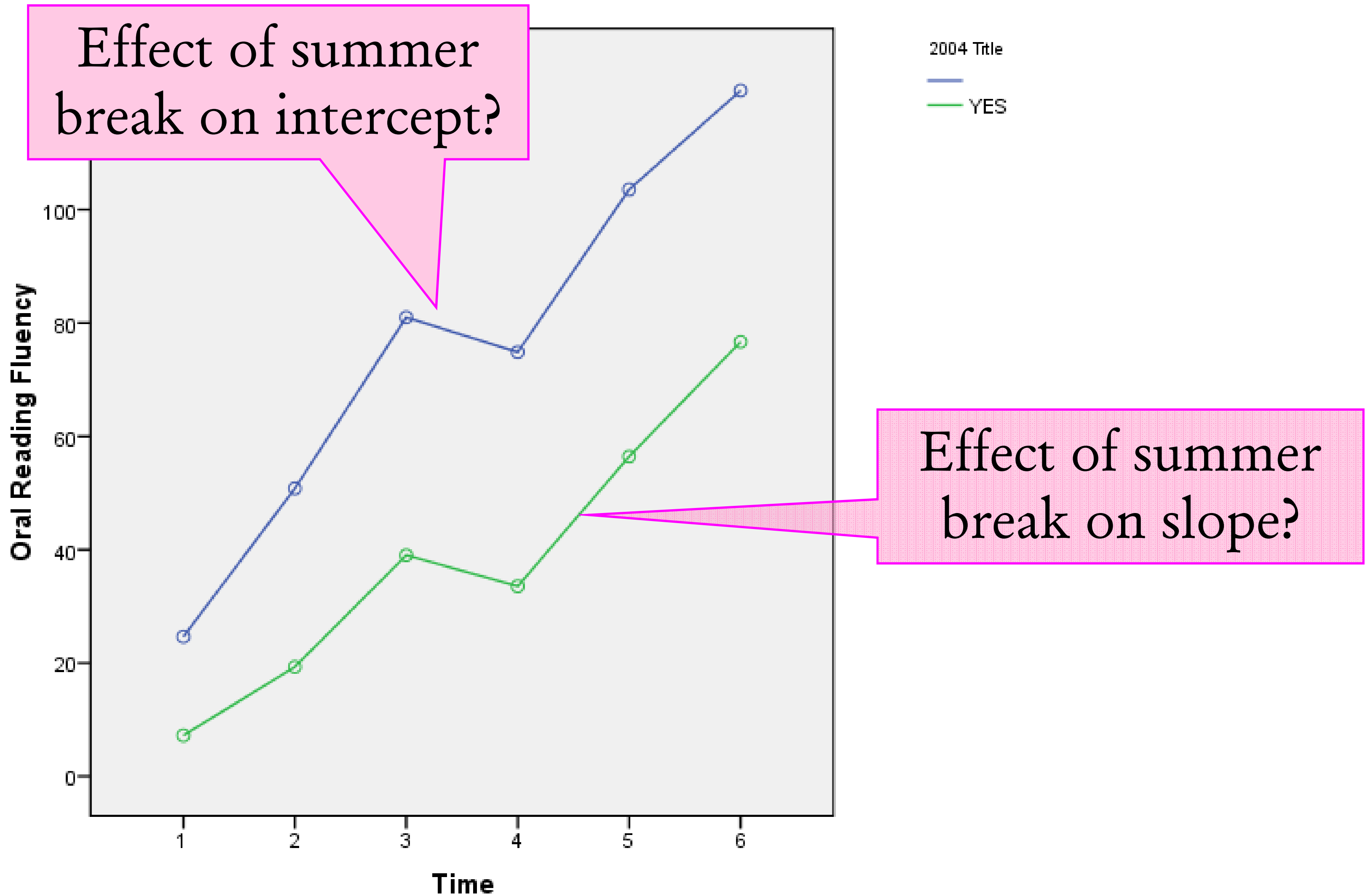
Interrupted Time Series Designs



Preliminary example modeling "What I did last summer"

- Prior to evaluating our treatment effects:
  - Examine nature of growth function
  - Explore the effect of summer drop in performance


Oral Reading Fluency
200
150
100
50
0
0
2
4
6
8
Time



Effect of summer break on intercept?
2004 Title
YES
100
80
60
40
20
0
Oral Reading Fluency
1
2
3
4
5
6
Time
Effect of summer break on slope?

Testing change in intercept and slope after summer break:

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
INTRCPT1, P0	11.934266	1.496115	7.977	6	0.000
TIME slope, P1	24.969961	0.588547	42.426	6	0.000
INTERCHA slope, P2	-25.886013	0.844983	-30.635	6	0.000
SLOPECHA slope, P3	-0.777770	0.899231	-0.865	6	0.421

Variation across students and schools?

Final estimation of level-1 and level-2 variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	R0	27.84145	775.14633	1362	12718.00625	0.000
TIME slope,	R1	10.65064	113.43621	1362	3795.84626	0.000
INTERCHA slope,	R2	4.62646	21.40417	1362	1508.33545	0.003
SLOPECHA slope,	R3	10.03375	100.67614	1362	2352.66736	0.000
level-1,	E	9.33343	87.11298			

Final estimation of level-3 variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2,	U00	3.49659	12.22617	6	31.18004	0.000
TIME/INTRCPT2,	U10	1.28195	1.64341	6	19.29324	0.004
INTERCHA/INTRCPT2,	U20	1.86312	3.47123	6	20.65449	0.002
SLOPECHA/INTRCPT2,	U30	2.15169	4.62977	6	39.43444	0.000

A last thought or two:

- Better modeling tools can expand the richness of research questions
- Better models allow more nuanced understanding of educational and social phenomena

Supposing is good, but finding out is better.

- *Mark Twain's Autobiography*

Bibliography

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park: Sage.

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

Goldstein, H. (1995). Multilevel Statistical Models (2nd ed.). London: Edward Arnold. Available in electronic form at http://www.arnoldpublishers.com/support/goldstein.htm.

Hedeker, D. (2004). An introduction to growth modeling. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage Publications.

Hedges, L. V., & Hedburg, E. C. (in press). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis*.

Hox, J. J. (1994). *Applied Multilevel Analysis*. Amsterdam: TT-Publikaties. Available in electronic form at http://www.ioe.ac.uk/multilevel/amaboek.pdf.

Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression*. Thousand Oaks: Sage.

Kreft, I. G., & de Leeuw, J. (2002). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Bibliography

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd Ed.). Orlando, FL: Harcourt Brace & Company.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.)*. Newbury Park: Sage.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., & Congdon, R.T. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Chicago, IL: Scientific Software International.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects, *Journal of Educational and Behavioral Statistics, 20* (4), 307-35.

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

Rosenthal, R. (1987). Pygmalion effects: Existence, magnitude, and social importance. *Educational Researcher, 16*, 37-41.

Rumberger, R.W., & Palardy, G. J. (2004). Multilevel models for school effectiveness research. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage Publications.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Seltzer, M. (2004). The use of hierarchical models in analyzing data from experiments and quasi-experiments conducted in field settings. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage Publications.

Bibliography

Spybrook, J., Raudenbush, S., & Liu, X.-f. (2006). *Optimal design for longitudinal and multilevel research: Documentation for the Optimal Design Software*. New York: William T. Grant Foundation.

Stevens, J. J., & Zvoch, K. (2006). Issues in the implementation of longitudinal growth models for student achievement. In R. Lissitz (Ed.), *Longitudinal and value added modeling of student performance*. Maple Grove, MN: JAM Press.

Stevens, J. J. (2005). The study of school effectiveness as a problem in research design. In R. Lissitz (Ed.), V*alue-added models in education: Theory and applications*. Maple Grove, MN: JAM Press.

Teddlie, C., & Reynolds, D. (2000). *The International handbook of school effectiveness research*. New York: Falmer Press.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations, *Development and Psychopathology, 10*, 395-426.

Wineburg, S. (1987). The self-fulfillment of the self-fulfilling prophecy. *Educational Researcher, 16*, 28-37.

Zvoch, K., & Stevens, J.J. (2006). Longitudinal effects of school context and practice on middle school mathematics achievement. *Journal of Educational Research, 99(6),* 347-356.

Zvoch, K., & Stevens, J. J. (2003). A multilevel, longitudinal analysis of middle school math and language achievement. *Educational Policy Analysis Archives, 11 (20).* (Available at: http://epaa.asu.edu/epaa/v11n20/).

Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. London: Falmer Press.