The Study of School Effectiveness as a Problem in Research Design

Joseph Stevens
University of New Mexico

Presented at the Conference on Value-Added Modeling, University of Maryland, College Park, October, 2004.

Contact Information:

120 Simpson Hall, University of New Mexico,
Albuquerque, NM 87131, 505-277-4203,
jstevens@unm.edu

Presentation available at:

http://www.unm.edu/~jstevens/papers/design.ppt

Supported in part by National Science Foundation grant NSF REC 0231774

Purpose

- Draw attention to certain research design issues inherent in the study of school effectiveness
- Examine the way in which these issues relate to NCLB methods for evaluating school effectiveness using percent of students proficient per school
- Contrast those methods with alternatives, especially longitudinal designs
- Offer preliminary evidence on the performance of alternatives in controlling for threats to internal validity

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)

- IES is currently funding initiatives to promote the use and application of scientifically based education research
- In some discussions, scientifically based research is equated with experiments using random assignment
- Less discussion about other methods of experimental and statistical control
- “Best” method depends on context and purpose of particular study or evaluation

Basic Randomized Experimental Design:

Treatment Group	R	X	O
Control Group	R		O

Randomized Longitudinal Design:

Treatment Group	R	O…O	X	O…O
Control Group	R	O…O		O…O

These designs are not commonly applied in many areas of educational research and are rare in School Effectiveness Research (SER; Teddlie, Reynolds, & Sammons, 2000)

Requirements of NCLB

- All states must implement high-stakes accountability systems, with annual testing in grades 3-8, and the definition of Basic, Proficient, and Advanced achievement levels (at least)
- Cannot weight or combine results
- Cannot consider factors or adjust scores
- Must measure “Adequate Yearly Progress” (AYP) in each content area through the report of unmodified percentage of students reaching proficiency or above

No Child Left Behind

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

- Collectively, NCLB regulations proscribe 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 subtraction of one O_t from another

Strengths and Weaknesses of the NCLB Case Study Design

Strengths:

- Use of multiple cohorts
- May have external and ecological validity

Weaknesses:

- Absence of pretest
- No control for pre-existing group differences or changes in group composition from cohort to cohort
- No control group, no control over treatment implementation
- No control over plausible confounding factors

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

Longitudinal Models

- Value-added, longitudinal models provide an alternative research design strategy in the evaluation of school effectiveness

- Common Design:

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

- Design often replicated over multiple cohorts (grades)
- As in the NCLB case study, treatment implementation is seldom monitored

Strengths of Longitudinal Designs

- Use of multiple cohorts
- May have external and ecological validity
- Focus on fundamental interest of SER: learning, a change phenomenon
- Tracking of individual student progress
- Provides control over individual differences between students; students serve as their own controls

Weaknesses of Longitudinal Designs

- No control group, no control over treatment implementation (in typical applications)
- No control over some other confounding influences (e.g., history)
- Attrition; changes in group composition
- Carryover effects; instrumentation

Evaluation of Alternative Designs

- Messick (1989) Validity and consequences of alternative approaches
- Pattern Matching (Shadish, Cook, & Campbell, 2002)
- Riechardt (2000) study of plausible threats to validity of treatment effects
- Context for our study: Ruling out specific, alternative causal claims

- NM schools vary drastically in composition and type of community; e.g., 90% White, 0% free lunch vs. 100% Native American and 90% free lunch vs. 95% Hispanic, 91% LEP and 88% free lunch
- One goal has been to examine the plausibility of confounding variables as factors affecting school performance
- Establish whether covariates that we wish to control for, like ethnicity, language, and poverty, are more strongly associated with some measures, models, or research designs than others

Study I: Zvoch & Stevens

- Study Purpose: To examine correlates of status and growth in mathematics achievement over a three year period.

- Individual math achievement scores on the TerraNova were used from a longitudinal sample of middle school students in the sixth grade in 1998-99, seventh grade in 1999-00, and eighth grade in 2000-01

Sample and Procedures

Study conducted in one urban school district in NM:

- 24 middle schools, over 20,000 students;
- 51% female, 49% male
- 47% Hispanic, 44% Anglo, 3% Native American, 3% African American, 2% Asian, and 1% Other
- 17% of students were classified as ELL
- 17% special education
- 40% of students receive a free or reduced price lunch

Sample and Procedures

- Data on individual and school characteristics were drawn from district files

- Students who were not at the same school all three years were excluded resulting in a sample of 5,168 students (84.7% of the original sample)

Student-Level Variables

- Sex: 0 = male, 1 = female (49%)
- Minority Status: 0 = non-minority, 1 = minority (54%)
- Free-Lunch Status: 0 = no free lunch, 1 = free lunch recipient (35%)
- LEP Status: 0 = English proficient, 1 = not English proficient (12%)
- SPED Status: 0 = general education, 1 = special education (13%)
- Test Administration: 0 = standardized, 1 = modified (15%)
- Measures: Mathematics Status and Growth (in scale score units)

School-Level Variables

- Percent Free-Lunch (*M* = .49, *SD* = .28)

- Mean Educational Level of Mathematics Staff (*M* = 17.61, *SD* = .58)

- Mathematics Curricula (0 = Traditional Program, 1 = NSF Reform Curricula, 9 of the 24 middle schools (38%)

Analytic Procedures
Multilevel Modeling (HLM 5)

Unconditional Model:

- **Level-1**

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

- **Level-2**

$$\pi_{0ij} = \beta_{00j} + r_{0ij}$$
$$\pi_{1ij} = \beta_{10j} + r_{1ij}$$

- **Level-3**

$$\beta_{00j} = \gamma_{000} + u_{00j}$$
$$\beta_{10j} = \gamma_{100} + u_{10j}$$

Conditional Models:

- **Level-1**

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

- **Level-2**

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}(Sex) + \beta_{02j}(Minority) + \beta_{03j}(Free\ Lunch) + \beta_{04j}(LEP) + \beta_{05j}(SPED) + \beta_{06j}(\text{Modified Test}) + r_{0ij}$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11j}(Sex) + \beta_{12j}(Minority) + \beta_{13j}(Free\ Lunch) + \beta_{14j}(LEP) + \beta_{15j}(SPED) + \beta_{06j}(\text{Modified Test}) + r_{1ij}$$

- **Level-3**

$$\beta_{00j} = \gamma_{000} + \gamma_{001}(\%\ Free\ Lunch) + \gamma_{002}(Teacher\ Education) + \gamma_{003}(Mathematics\ Curricula) + u_{00j}$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101}(\%\ Free\ Lunch) + \gamma_{102}(Teacher\ Education) + \gamma_{103}(Mathematics\ Curricula) + u_{10j}$$

Results

- The conditional model using individual and school level predictors provided significantly better fit than an unconditional model
- All student level predictors except gender were significantly related to mean achievement
- Only gender and minority status were significantly related to growth
- At the school level, only percent free lunch was related to school mean achievement, while teacher education and math curricula were related to school mean growth

Level-1 Unconditional Model for Mathematics Achievement

Fixed Effect	*Coefficient*	*SE*	*t*
School Mean Achievement, $_{000}$	648.96	3.09	209.82***
School Mean Growth, $_{100}$	17.64	0.87	20.16***

Random Effect	*Variance Component*	*df*	^{2}
Individual Achievement, r_{0ij}	1132.02	4548	22708.93***
Individual Growth, r_{1ij}	44.03	4548	5736.03***
Level-1 Error, e_{tij}	361.59		
School Mean Achievement, u_{00j}	222.42	23	855.44***
School Mean Growth, u_{10j}	17.01	23	347.30***

Level-1 Coefficient	*Percentage of Variation Between Schools*
Individual Achievement, π_{0ij}	16.4
Individual Growth, π_{1ij}	27.9

Note. Results based on data from 5,168 students distributed across 24 middle schools.

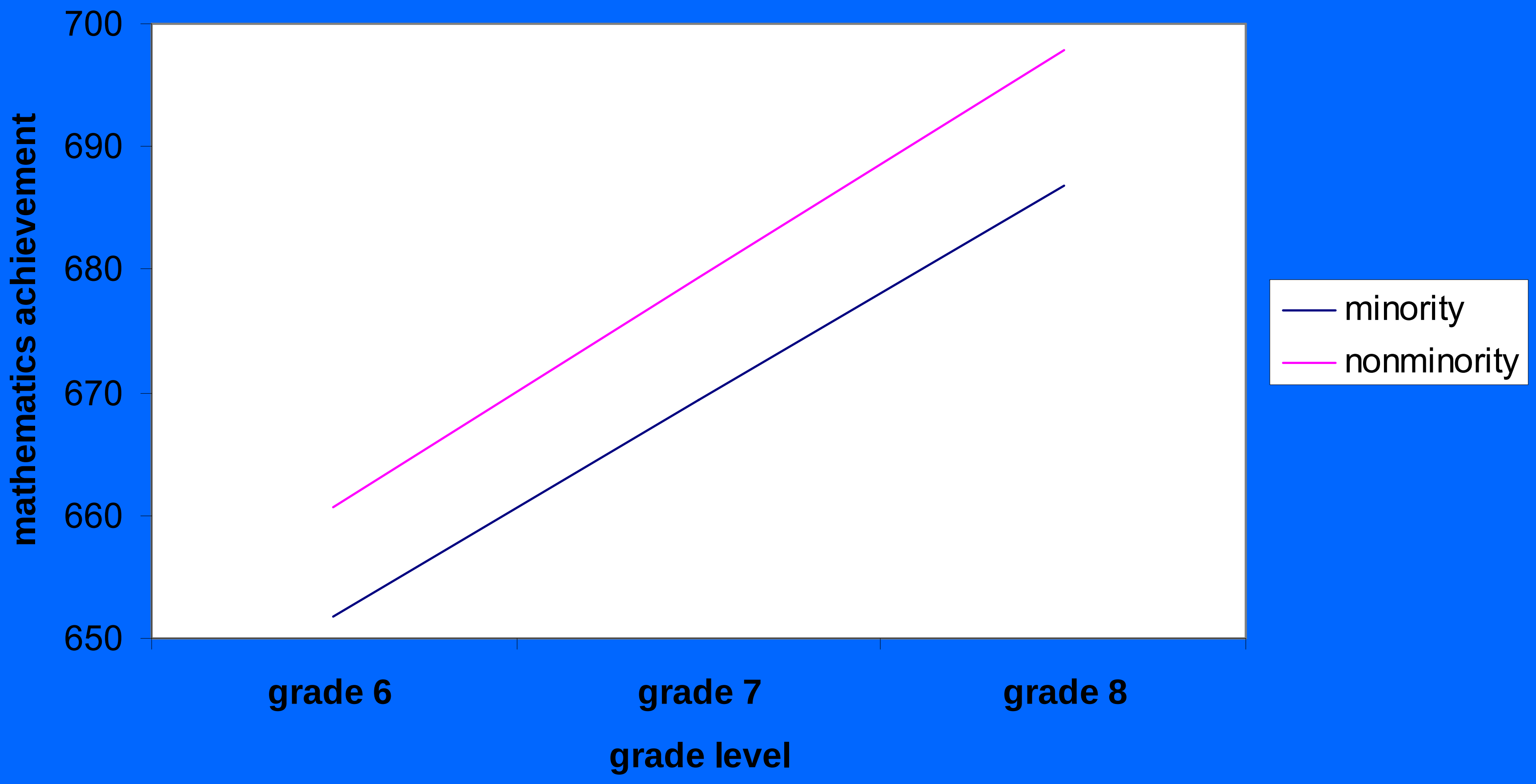
*** $p < .001$

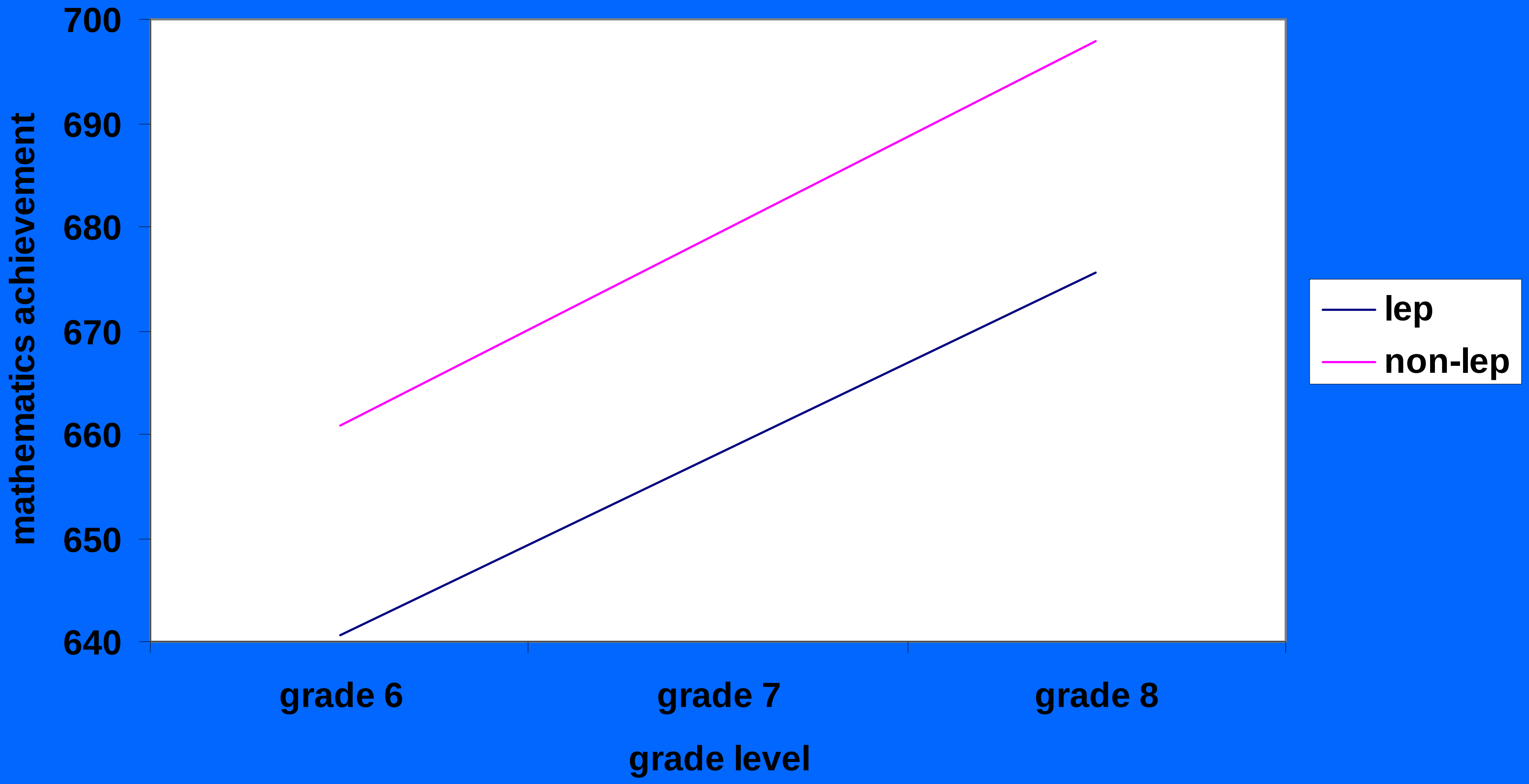
Level-2 Model Relating Individual Characteristics to Mathematics Achievement

Fixed Effect	*Coefficient*	*SE*	*t*
Individual Achievement, $_{00}$	650.15	1.71	379.93***
Gender, $_{01}$	0.58	1.03	0.56
Minority Status, $_{02}$	-9.11	0.87	10.47***
Free Lunch Status, $_{03}$	-9.63	1.44	-6.69***
LEP Status, $_{04}$	-17.81	1.42	-12.52***
SPED Status, $_{05}$	-20.72	3.82	-5.43***
Test Administration, $_{05}$	-19.90	4.05	-4.92***
Individual Growth, $_{10}$	17.91	0.79	22.63***
Gender, $_{11}$	-2.65	0.50	-4.35***
Minority Status, $_{12}$	-1.56	0.46	-3.39***
Free Lunch Status, $_{13}$	-0.41	0.50	-0.84
LEP Status, $_{14}$	-0.50	0.82	-0.62
SPED Status, $_{15}$	-1.69	1.66	-1.02
Test Administration, $_{16}$	-2.98	2.32	-1.29

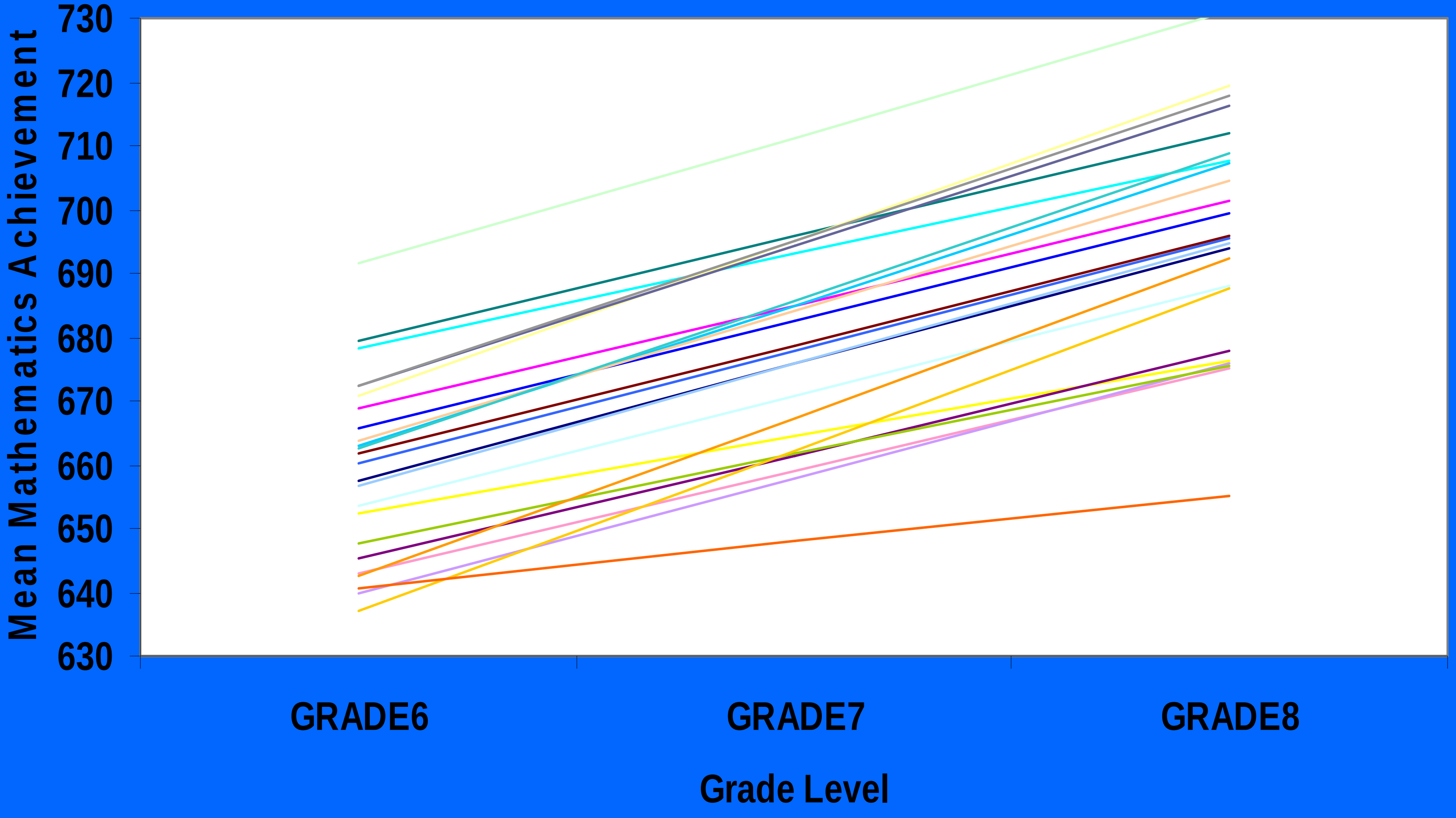
*Note. Random effects are not presented; *** p < .001*

Variance Component	*Level-1*	*Level-2*	*Variance Explained*
Individual Achievement, *r0ij*	1132.02	791.15	30.1%
Individual Growth, *r1ij*	44.04	39.90	9.4%
School Mean Achievement, *u00j*	222.42	63.04	71.7%
School Mean Growth, *u10j*	17.01	14.02	17.5%


700
690
680
670
660
650
mathematics achievement
grade 6
grade 7
grade 8
grade level
minority
nonminority


700
690
680
670
660
650
640
mathematics achievement
grade 6
grade 7
grade 8
grade level
lep
non-lep

School Level Results: Growth Trajectories

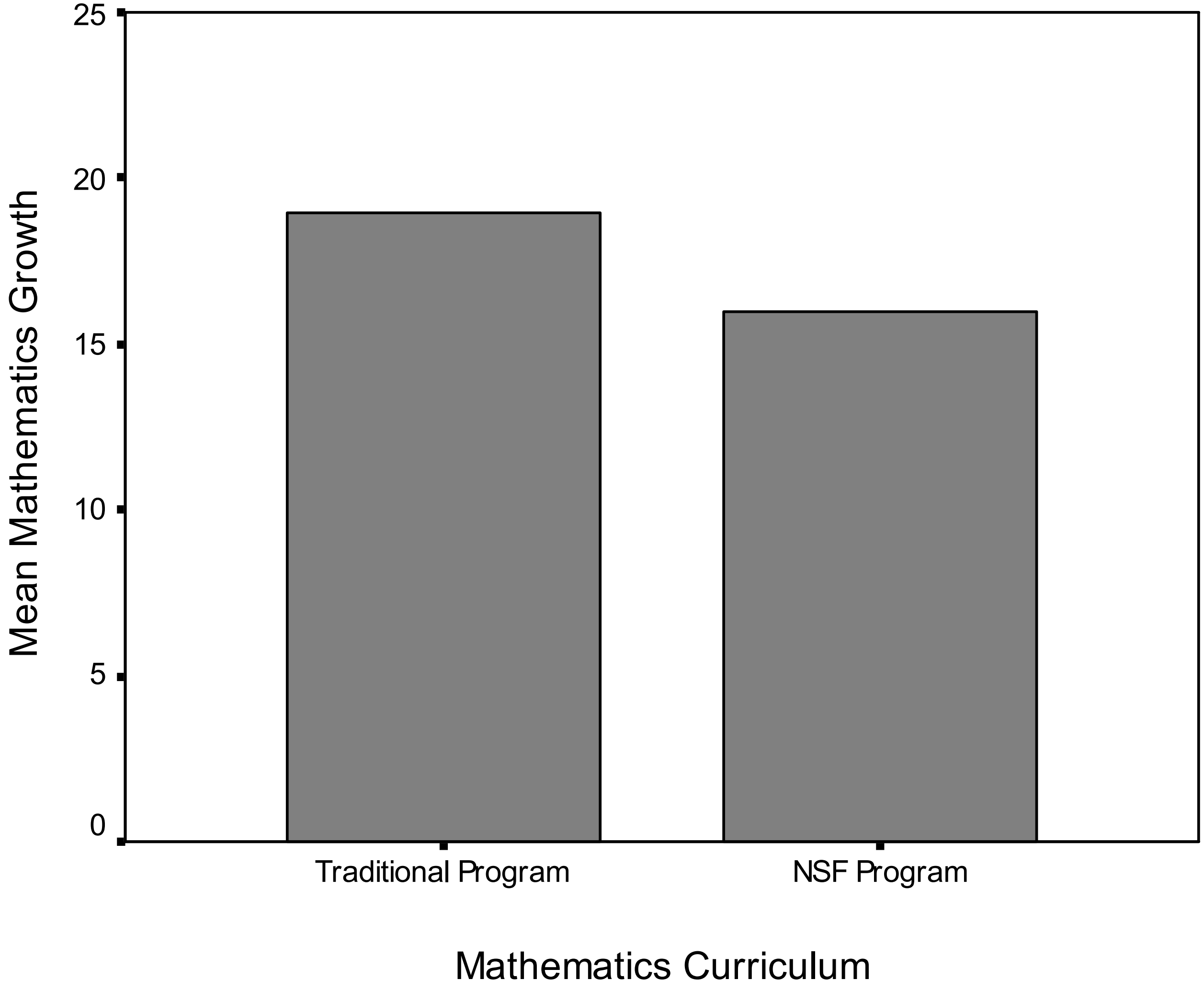


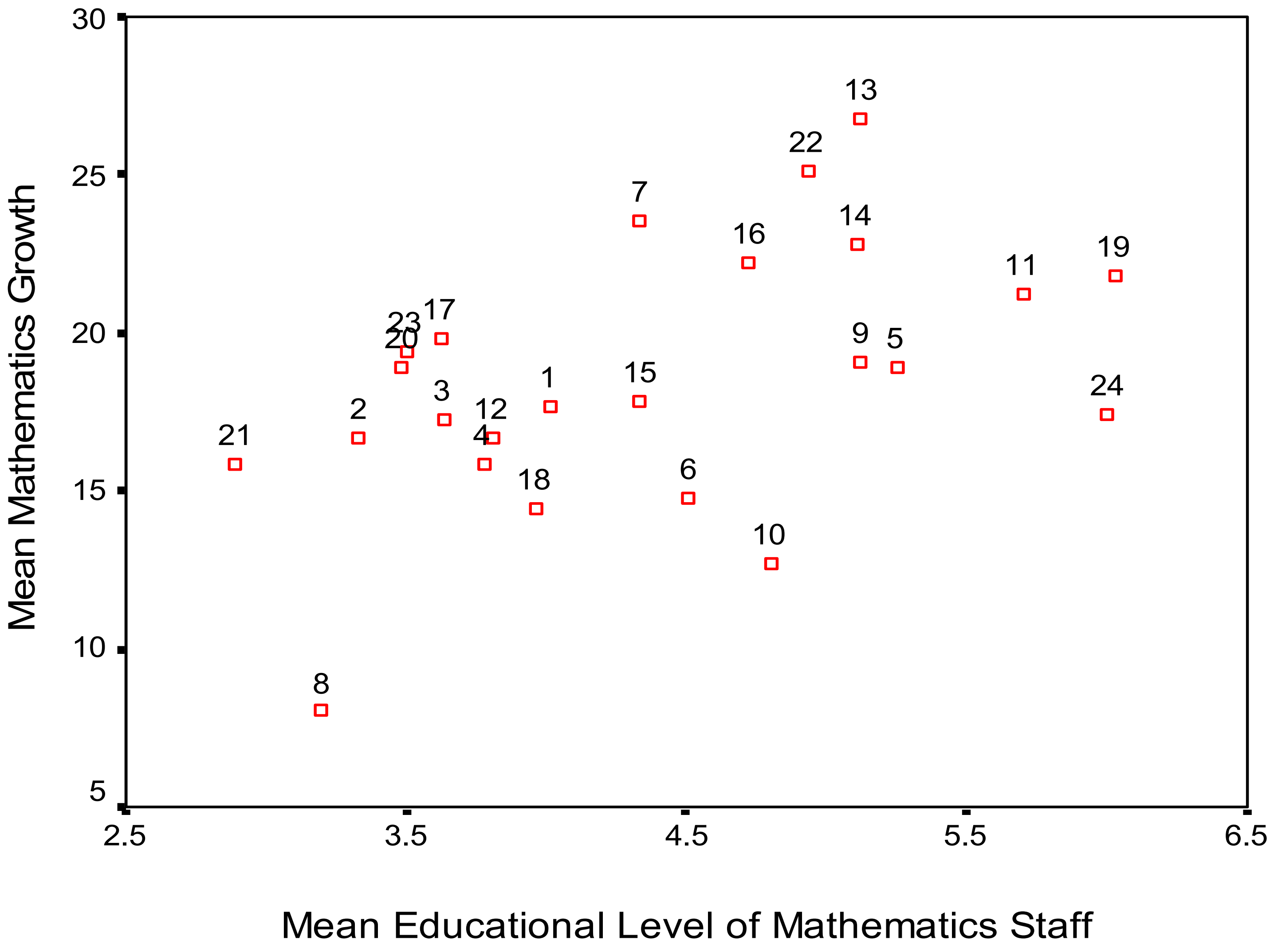
Level-3 Model Relating Individual and School Characteristics to Mathematics Achievement

Fixed Effect	*Coefficient*	*SE*	*t*
School Mean Achievement, γ_{000}	650.18	1.51	429.77***
Percent Free Lunch, γ_{001}	-0.22	0.05	-4.70***
Math Teacher Education, γ_{002}	0.93	2.00	0.47
Math Curricula, γ_{003}	-0.32	2.55	-0.13
School Mean Growth, γ_{100}	18.96	0.83	22.75***
Percent Free Lunch, γ_{101}	-0.02	0.02	-0.92
Math Teacher Education, γ_{102}	3.29	1.08	3.06**
Math Curricula, γ_{103}	-3.00	1.40	-2.14*

Variance Component	*Level-1*	*Level-2*	*Level-3*	*Variance Explained*
School Mean Achievement, u_{00j}	222.42	63.04	27.79	87.5%
School Mean Growth, u_{10j}	17.01	14.02	8.80	48.3%

Note. Results based on data from 5,168 students distributed across 24 middle schools.
* $p < .05$, ** $p < .01$, *** $p < .001$


Mean Mathematics Growth
25
20
15
10
5
0
Traditional Program
NSF Program
Mathematics Curriculum


Mean Mathematics Growth
30
25
20
15
10
5
2.5
3.5
4.5
5.5
6.5
Mean Educational Level of Mathematics Staff
13
22
7
16
14
19
11
23
17
20
9
5
1
15
3
12
2
24
21
4
18
6
10
8

Conclusions

- **Pattern of results differed depending on whether status scores or growth scores were examined**
- **For students:**

  1) Status scores reveal that minority students, LEP students, impoverished students, and special education students achieve at a level that is significantly lower than that of their counterparts

  2) Growth trajectories were relatively constant across the same groups over time

- **For schools:**

  1) Mean achievement (i.e., status) was strongly correlated with socio-demographics

  2) Mean growth, the rate at which students learn, was largely uncorrelated with socio-demographics but was related to two indicators of school policy and practice

Study II

- Study Purpose: To apply curvilinear growth models examining correlates of student status and growth in mathematics achievement over a four year period

- Two Analytic Samples: 1) a two-level student differences sample, and 2) a three-level school differences model

Methods

Student Differences Sample:

- 23,469 sixth grade children took the state mandated TerraNova in 1999-00
- Study includes the 23,296 sixth graders (99.3%) who took the mathematics subtest
- These students were matched longitudinally to 7th, 8th, and 9th grade records for the years 2001, 2002, and 2003
- 85.1% of the students were matched from 6th to 7th grade, 81.2% from 6th to 8th grade and 75.2% from 6th to 9th grade

Methods

Sample:

- Ethnic Composition:
  - 9,143 Hispanic students (53%)
  - 1,693 Native American students (10%)
  - 6,377 White students (37%)
- Gender: 50% female, 50% male

Methods

Sample:

- 2,028 children (12%) were special education students
- 1,450 (8%) received a modified test administration
- 2,335 (14%) students were classified as Limited English Proficient (LEP)

Methods

Measures:

- TerraNova mathematics subtest
- KR-20 .91 in grade 6, .90 in grade 7, .92 in grade 8, and .92 in grade 9 (CTB/McGraw-Hill, 1997)
- New Mexico lower-bound reliability estimate of .89 for the mathematics subtest in 1999 (Stevens, 2001)
- Standardized scale used is a vertically equated developmental scale

Analytic Procedures
Multilevel Modeling (HLM 5)

Unconditional Model:

- **Level-1**

$$Y_{tij} = \pi_{0ij} + e_{tij}$$

- **Level-2**

$$\pi_{0ij} = \beta_{00j} + r_{0ij}$$

Conditional Models:

- **Level-1**

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

- **Level-2**

$$\pi_{0ij} = \beta_{p0j} + \beta_{pij}(a_{Pij}) + r_{0ij}$$

$$\pi_{1ij} = \beta_{p1j} + \beta_{pij}(a_{Pij}) + r_{1ij}$$

$$\pi_{2ij} = \beta_{p1j} + \beta_{pij}(a_{Pij}) + r_{1ij}$$

Student-Level Variables (a_{Pij}):

- Gender: 0 = male, 1 = female (49%)
- Ethnicity: 0 = non-White, 1 = White (34%)
- Special Education Status (15%)
- Modified Test Administration (10%)
- Bilingual (10%)
- LEP Status (16%)
- Stability (0 = changed school twice in three years, 1 = changed once in three years, 2 = no change of middle school)

Results

- An unconditional Linear Growth Model was fit
- Next a conditional model using all student level predictors was applied. This model provided significantly better fit than the unconditional model, x^2 (24) = 9528.95, p < .001.
- τ_{01} was -0.378.
- Reliabilities at Level 1 were .733 for the intercept, .186 for linear slope and .135 for curvilinear slope; all variance components were significant.
- R^2 for the linear growth model was .31 and for student level predictors at level 2, R^2 was .28.

Unconditional Model for Student Level Mathematics Achievement

Fixed Effect	*Coefficient*	*SE*	*df*	*t*
School Mean Achievement, $_{00}$	651.75	0.32	23295	2153.81*
School Linear Growth, $_{10}$	16.83	0.26	23295	63.84*
School Curvilinear Growth, $_{20}$	-1.48	0.08	23295	-17.73*

Random Effect	*Variance Component*	*df*	^{2}
Individual Achievement, r_{0ij}	1701.63	19218	91,592.51*
Individual Linear Growth, r_{1ij}	291.05	19218	23,476.34*
Individual Curvilinear Growth, r_{2ij}	20.72	19218	22,207.73*
Level-1 Error, e_{tij}	449.64		

* $p < .001$

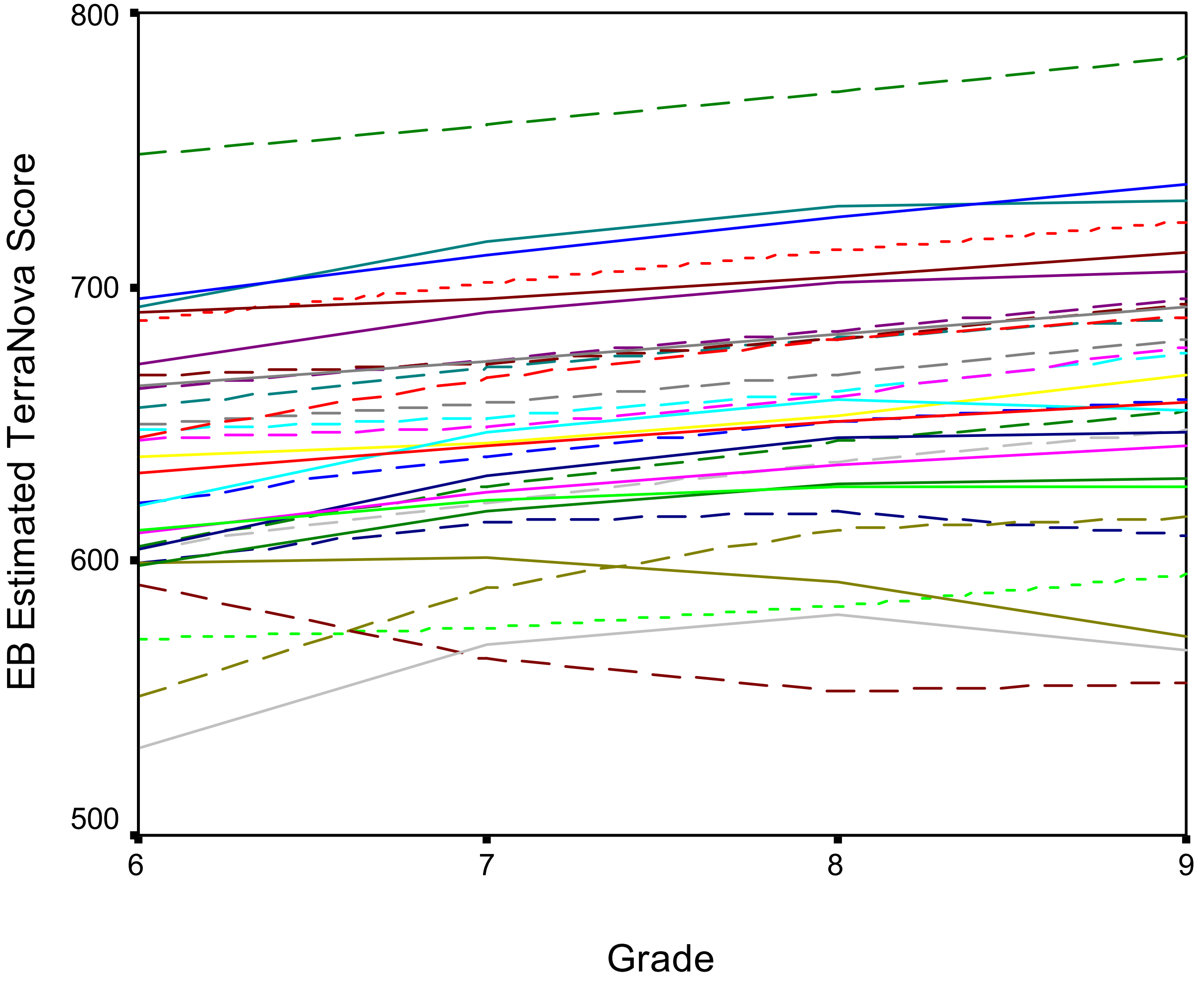
Mathematics Achievement Predicted by Individual Characteristics

Fixed Effect	*Coefficient*	*SE*	*t*	*df*	*p*
School Mean Achievement, γ_{00}	660.83	0.80	829.40	23287	< .001
White Student, γ_{01}	19.48	0.60	32.45	23287	< .001
Stability, γ_{02}	1.03	0.37	2.82	23287	.005
LEP, γ_{03}	-20.56	0.77	-26.74	23287	< .001
Title 1 Student, γ_{04}	-6.25	0.61	-10.31	23287	< .001
Special Education, γ_{05}	-32.50	1.38	-23.64	23287	< .001
Modified Test, γ_{06}	-14.66	1.67	-8.80	23287	< .001
Free Lunch Student, γ_{07}	-9.39	0.57	-16.55	23287	< .001
Gender, γ_{08}	-0.74	0.53	-1.39	23287	.164
School Linear Growth, γ_{10}	16.78	0.81	20.75	23287	< .001
White Student, γ_{11}	-0.18	0.58	-0.30	23287	.761
Stability, γ_{12}	2.83	0.40	7.13	23287	< .001
LEP, γ_{13}	4.15	0.84	4.96	23287	< .001
Title 1 Student, γ_{14}	-3.07	0.62	-4.96	23287	< .001
Special Education, γ_{15}	-1.19	1.42	-0.84	23287	.401
Modified Test, γ_{16}	-2.41	1.80	-1.33	23287	.183
Free Lunch Student, γ_{17}	-0.94	0.55	-1.71	23287	.088
Gender, γ_{18}	-5.24	0.52	-10.06	23287	< .001

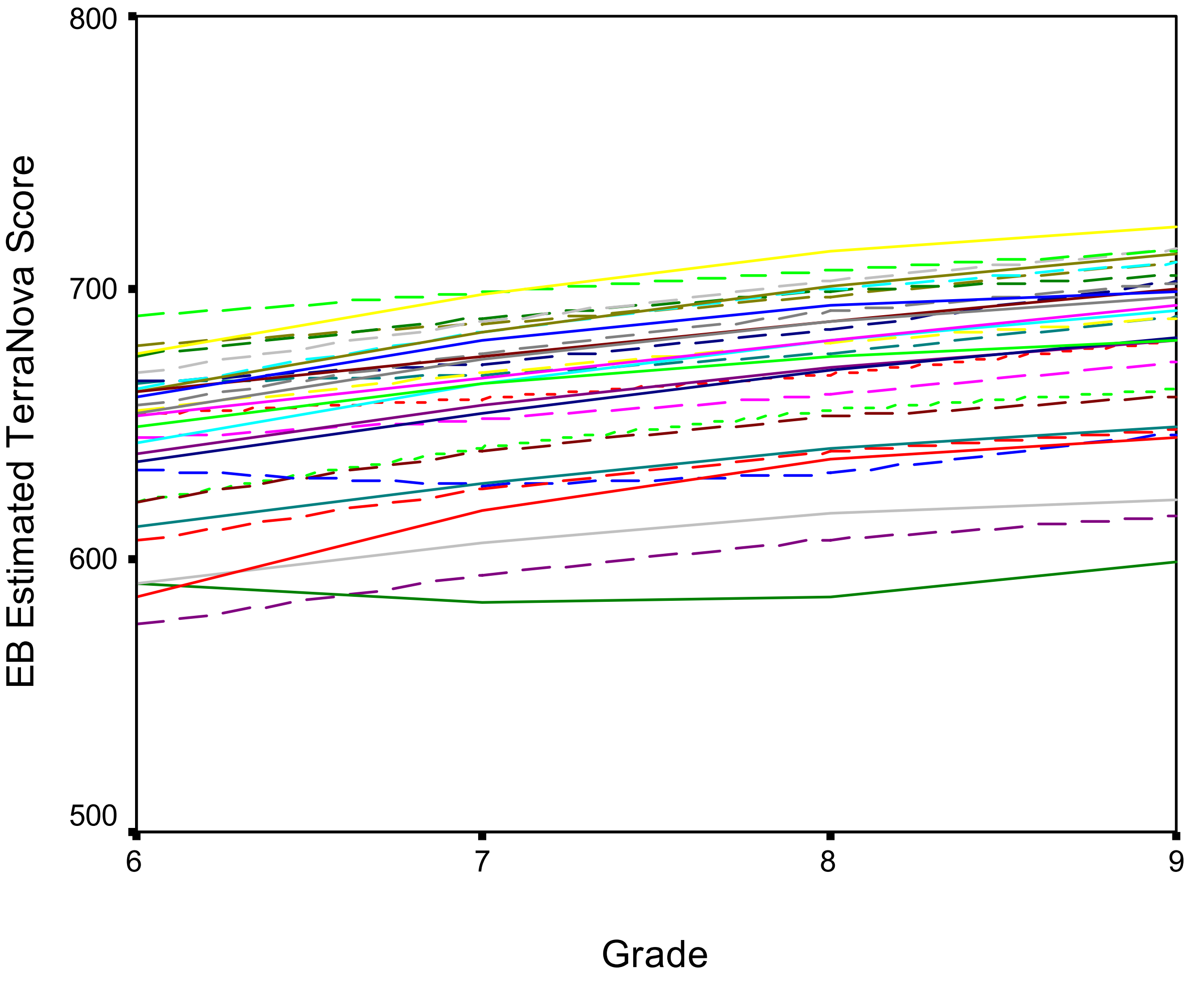
Mathematics Achievement Predicted by Individual Characteristics (continued)

Fixed Effect	*Coefficient*	*SE*	*t*	*df*	*p*
School Curvilinear Growth, γ_{20}	-1.46	0.26	-5.62	23287	< .001
White Student, γ_{21}	0.11	0.18	0.56	23287	.562
Stability, γ_{22}	-0.60	0.13	-4.66	23287	< .001
LEP, γ_{23}	-0.97	0.27	-3.62	23287	.001
Title 1 Student, γ_{24}	0.76	0.20	3.84	23287	< .001
Special Education, γ_{25}	0.25	0.44	0.58	23287	.565
Modified Test, γ_{26}	-0.21	0.57	-0.38	23287	.707
Free Lunch Student, γ_{27}	0.13	0.18	0.74	23287	.462
Gender, γ_{28}	1.20	0.17	7.23	23287	< .001

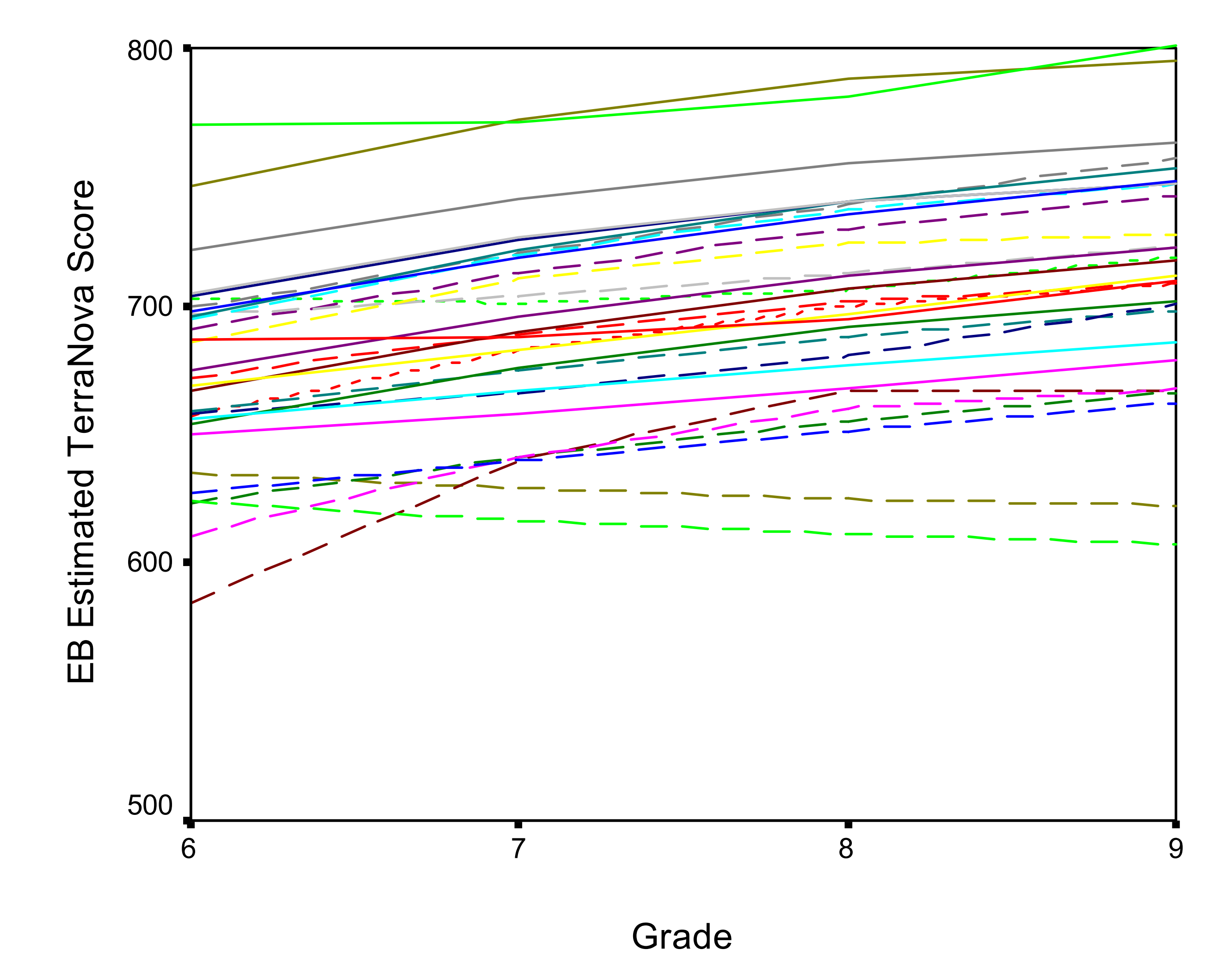
Variance Component	*Level-1*	*Level-2*	*Variance Explained*
Individual Achievement, *r0ij*	1701.63	1181.41	30.6%
Linear Growth, *r1ij*	291.05	278.43	4.3%
Curvilinear Growth, *r2ij*	20.72	20.10	0.3%



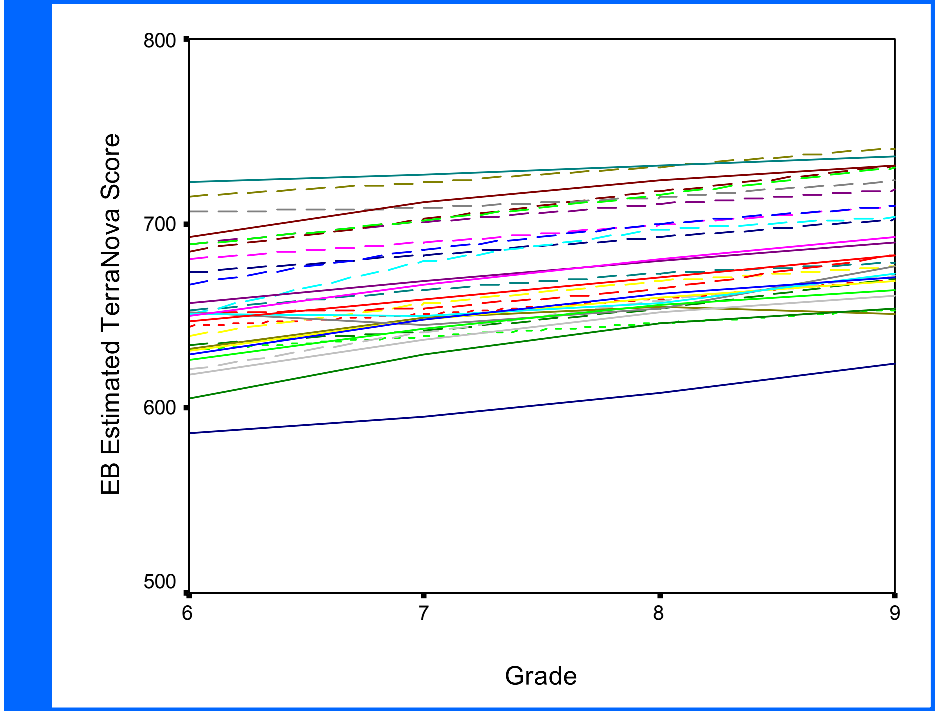
Hispanic Students

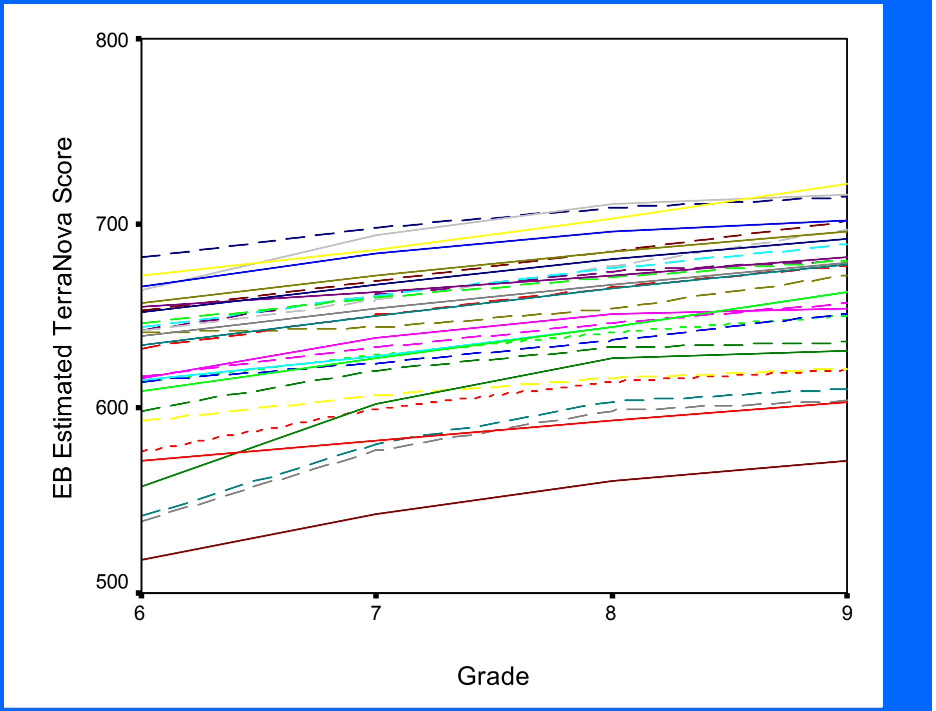


Native American Students



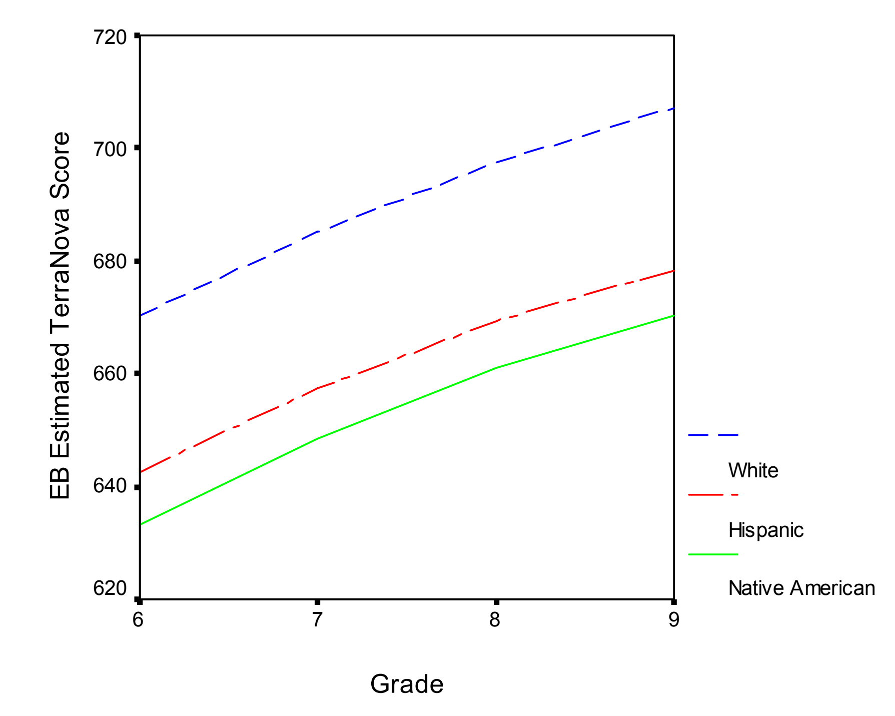
White Students



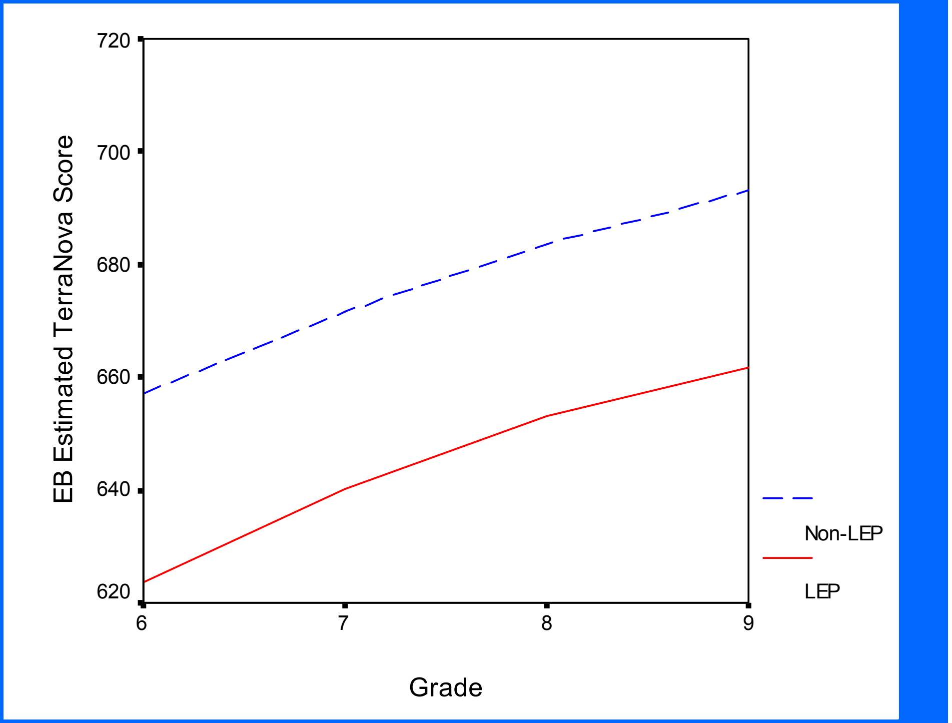


Non-LEP Students

LEP Students



Growth by Ethnicity



Growth by LEP Status

School Differences Sample

- In order to evaluate school level differences we applied three level HLM models to a sample that included only those students who were in their middle school for 2 or 3 years (17,596; 75.5% of students).
- Schools with less than 5 students were also excluded (13 schools with a total of 24 students), resulting in an analytic sample of 242 schools (94% of schools) with 17,572 students.
- This sample differs from the student differences sample in having about 1% more White and Hispanic, 1% fewer Native American, 1% fewer LEP and Special Education, and 2% fewer bilingual students.

Analytic Procedures
Multilevel Modeling (HLM 5)

Level-1

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

Level-2

$$\pi_{0ij} = \beta_{p0j} + \beta_{pij}(a_{Pij}) + r_{0ij}$$

$$\pi_{1ij} = \beta_{p1j} + \beta_{pij}(a_{Pij}) + r_{1ij}$$

$$\pi_{2ij} = \beta_{p1j} + \beta_{pij}(a_{Pij}) + r_{1ij}$$

Level-3

$$\beta_{p0j} = \gamma_{000} + \gamma_{pqs}(W_{sj}) + u_{00j}$$

$$\beta_{p1j} = \gamma_{pq1} + \gamma_{pqs}(W_{sj}) + u_{10j}$$

$$\beta_{p1j} = \gamma_{pq1} + \gamma_{pqs}(W_{sj}) + u_{10j}$$

- The models applied the same student level predictors as analyses with the student differences sample

- At the school level, the following predictors were used:
  - White students (*M* = 32%)
  - Students in a bilingual program (*M* = 15%)
  - Students who were classified as LEP (*M* = 14%)
  - Free Lunch (*M* = 53%)

Results

- An unconditional Linear Growth Model was fit

- Next a conditional model using all student level predictors was applied. This model provided significantly better fit than the unconditional model, χ^2 (315) = 6925.38, *p* < .001.

- τ_{01} was - 0.274.

- Reliabilities at the student level were .664 for the intercept, .385 for linear slope and .354 for curvilinear slope.

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%

Results

- A third conditional model was applied using the four school level predictors (Bilingual, LEP, White, Special Education). This model also provided significantly better fit than the unconditional model, χ^2 (327) = 7003.55, $p <$ .001.
- τ_{01} was -0.480.
- Reliabilities at level 2 were .608 for the intercept, .378 for linear slope and .347 for curvilinear slope.
- R^2 for the student level predictors was .23, at the school level R^2 was .24.

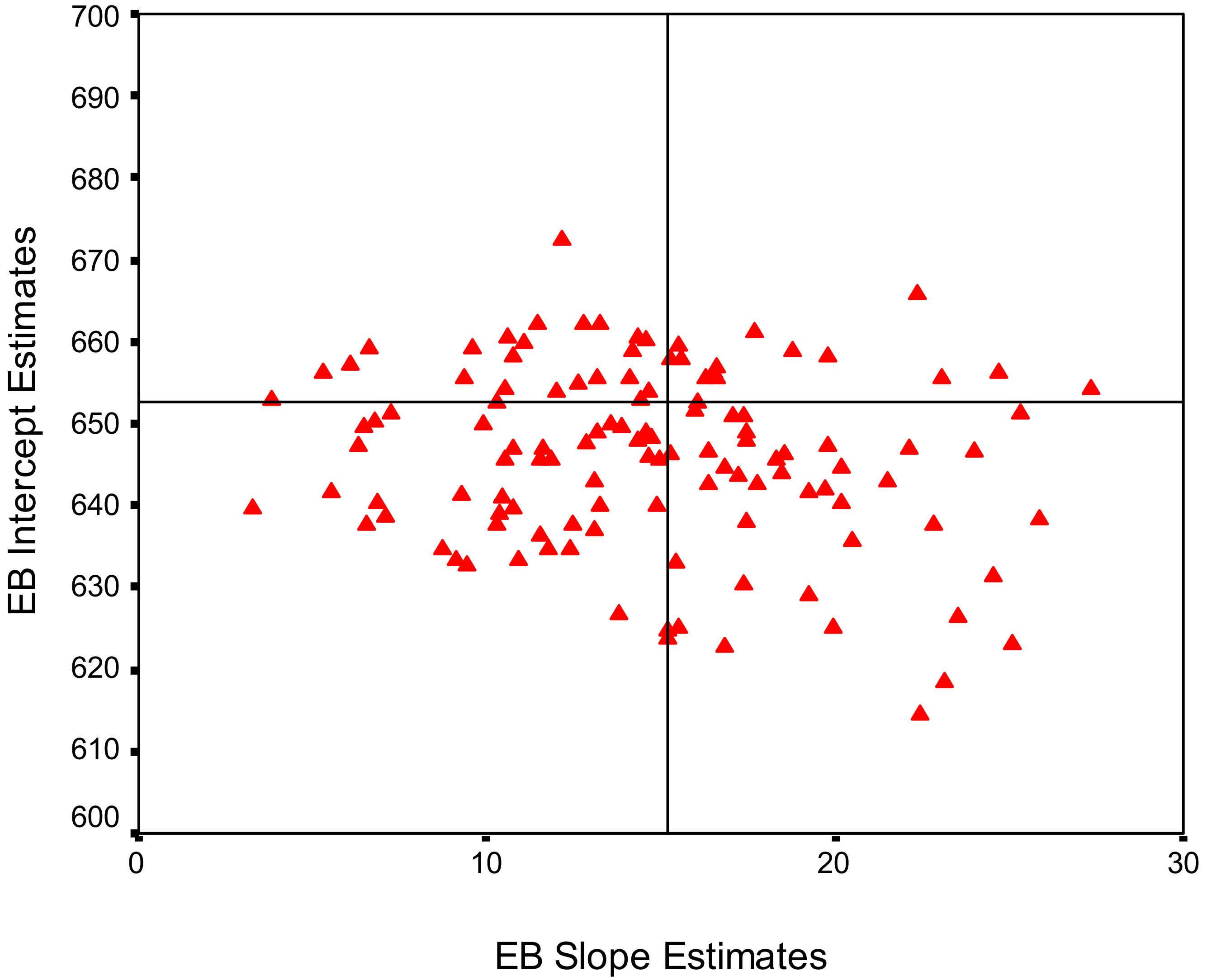
Mathematics Achievement Predicted by School Characteristics

Fixed Effect	*Coefficient*	*SE*	*t*	*df*	*p*
School Mean Achievement, $_{000}$	662.53	1.07	620.80	237	< .001
Percent Bilingual Students, $_{001}$	4.19	4.00	1.05	237	.295
Percent LEP Students, $_{0o2}$	-0.99	4.56	-0.22	237	.828
Percent White Students, $_{003}$	19.55	3.72	5.25	237	< .001
Percent Free Lunch, $_{004}$	-5.29	3.18	-1.67	237	.096
School Mean Linear Growth, $_{100}$	19.18	0.71	26.87	237	< .001
Percent Bilingual Students, $_{101}$	-0.17	1.98	-0.09	237	.932
Percent LEP Students, $_{102}$	2.90	2.85	1.02	237	.309
Percent White Students, $_{003}$	3.51	2.74	1.28	237	.201
Percent Free Lunch, $_{004}$	-3.67	2.23	-1.65	237	.099
School Curvilinear Growth, $_{200}$	-1.99	0.22	-9.10	237	< .001
Percent Bilingual Students, $_{201}$	-0.12	0.57	-0.21	237	.834
Percent LEP Students, $_{202}$	0.39	0.84	0.46	237	.643
Percent White Students, $_{203}$	-1.11	0.75	-1.48	237	.138
Percent Free Lunch, $_{204}$	-1.17	0.64	1.84	237	.065

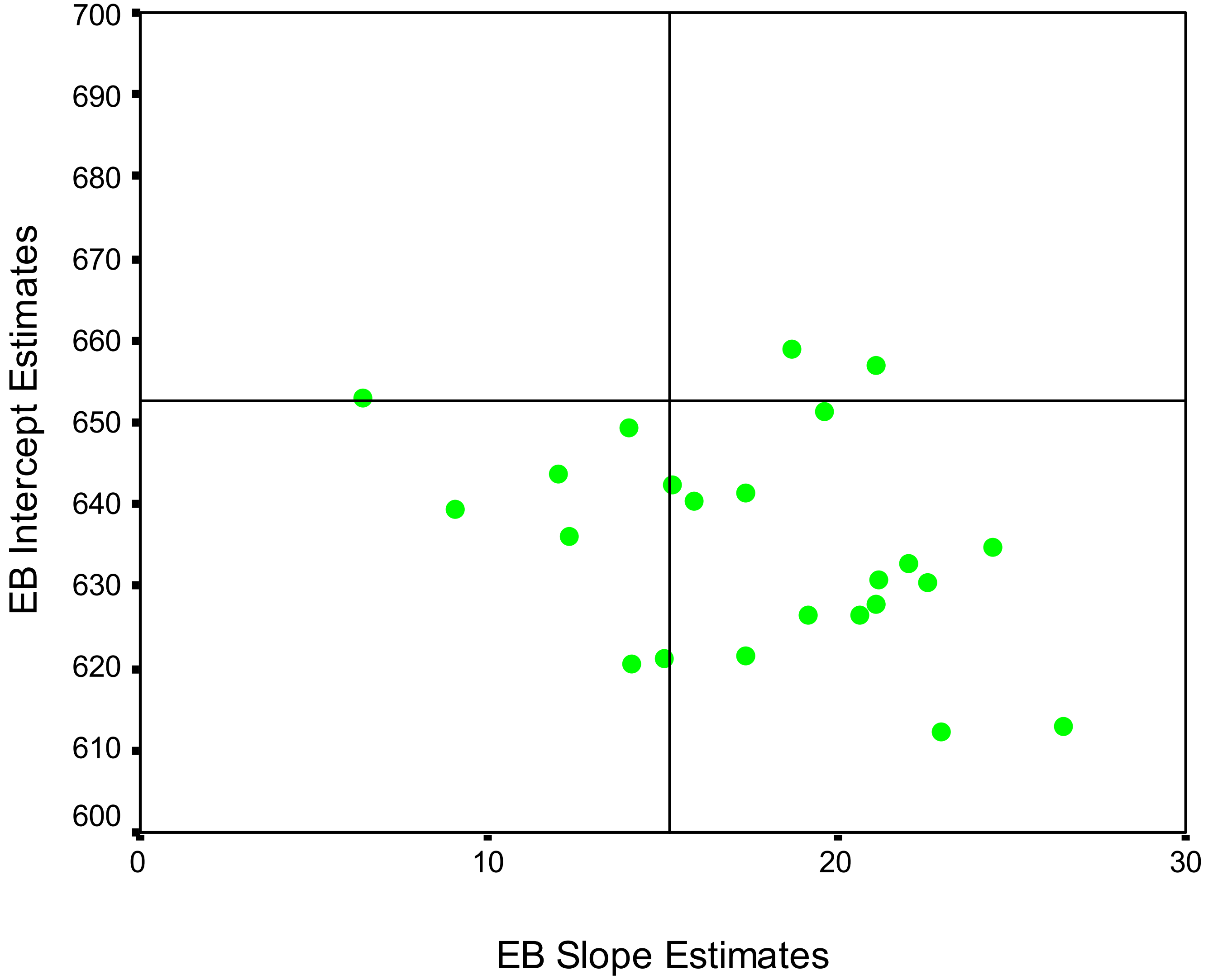
Note. Only school level results are presented for brevity, student results do not differ substantially from the previous model.

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%

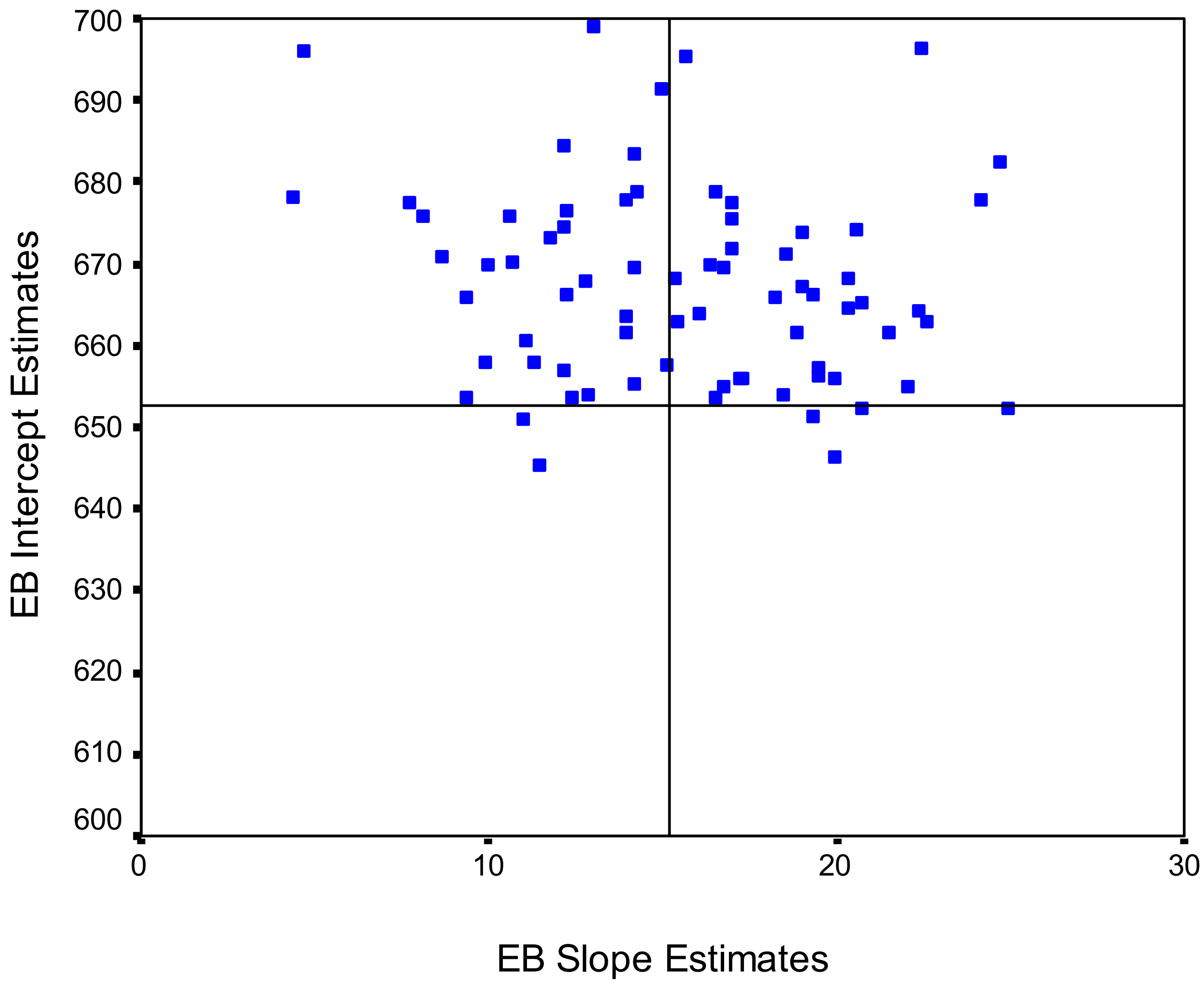
* Percent level 2 residual variance explained by level 3 model.



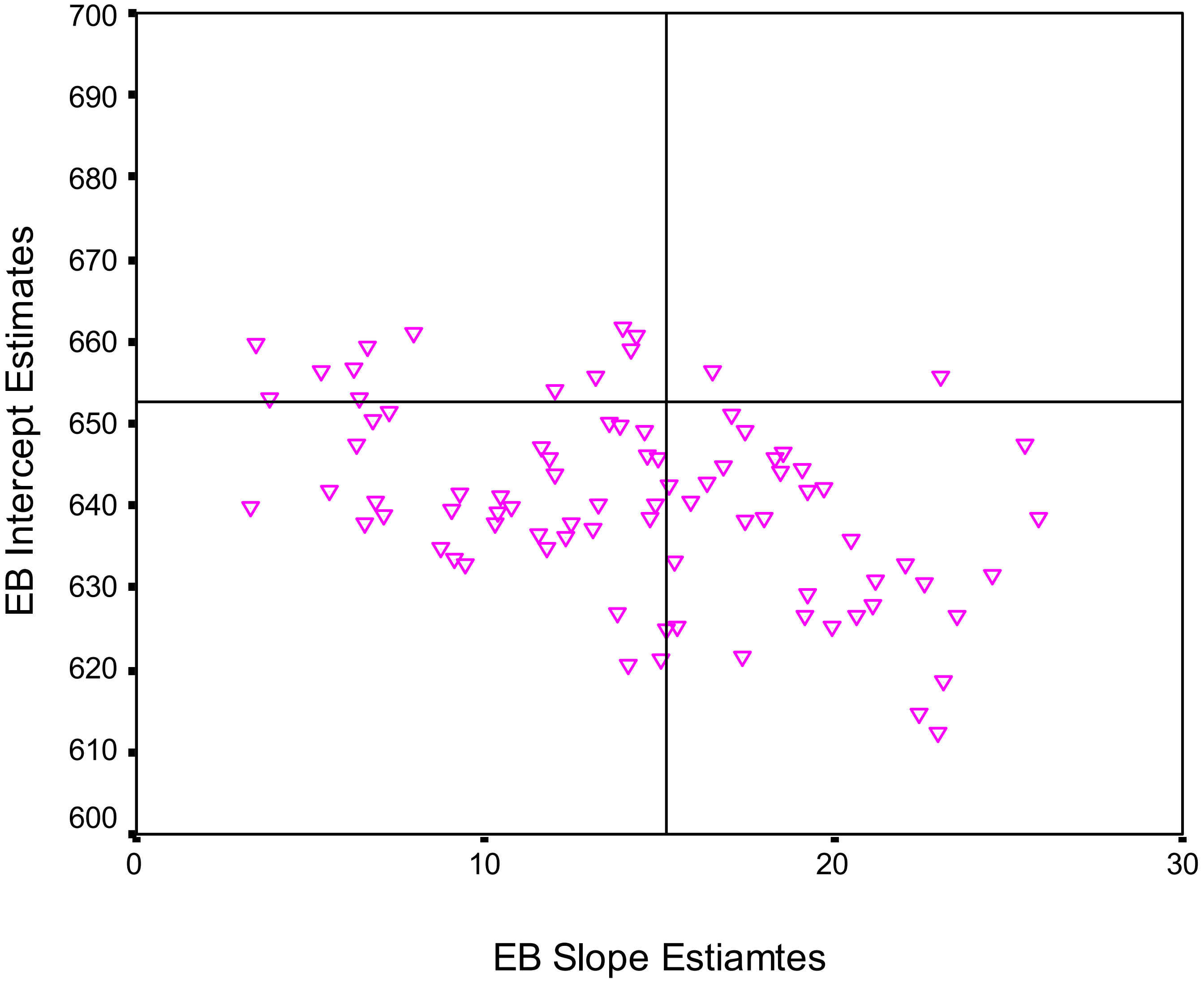
Schools with Majority Hispanic Enrollment



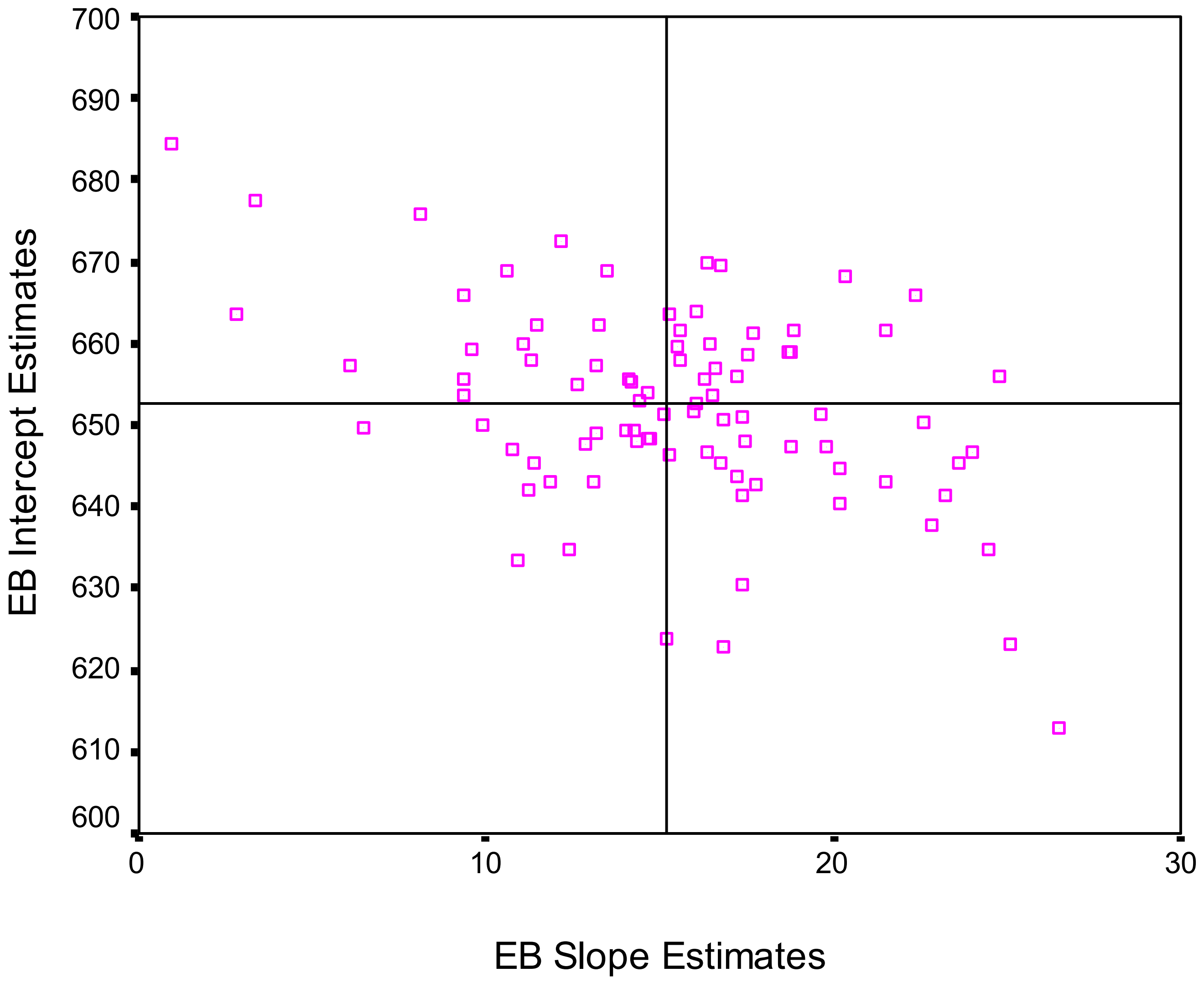
Schools with Majority Native American Enrollment



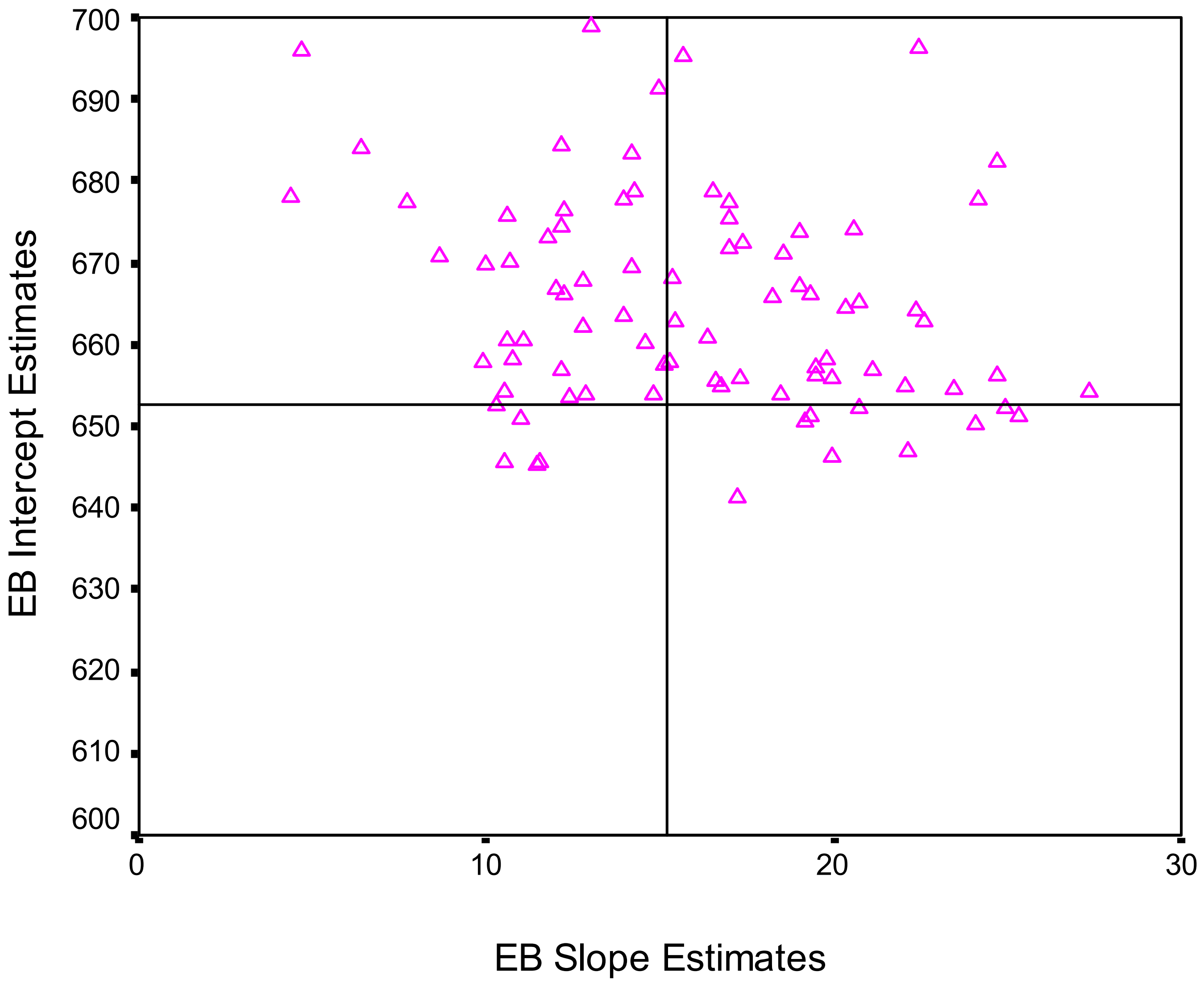
Schools with Majority White Enrollment



Lowest Third of Schools on Free Lunch



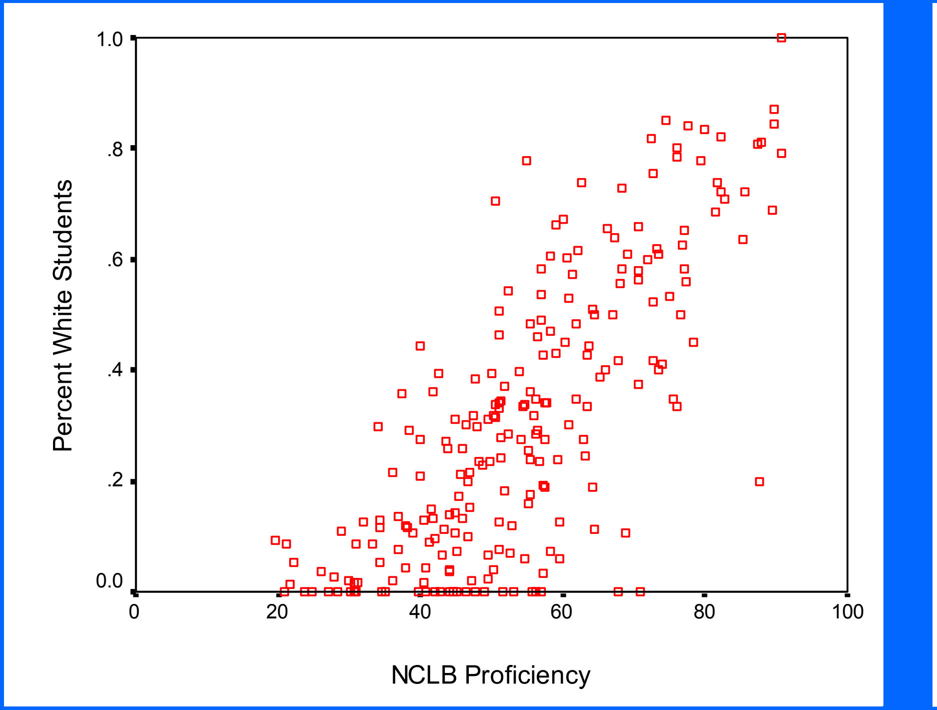
Middle Third of Schools on Free Lunch

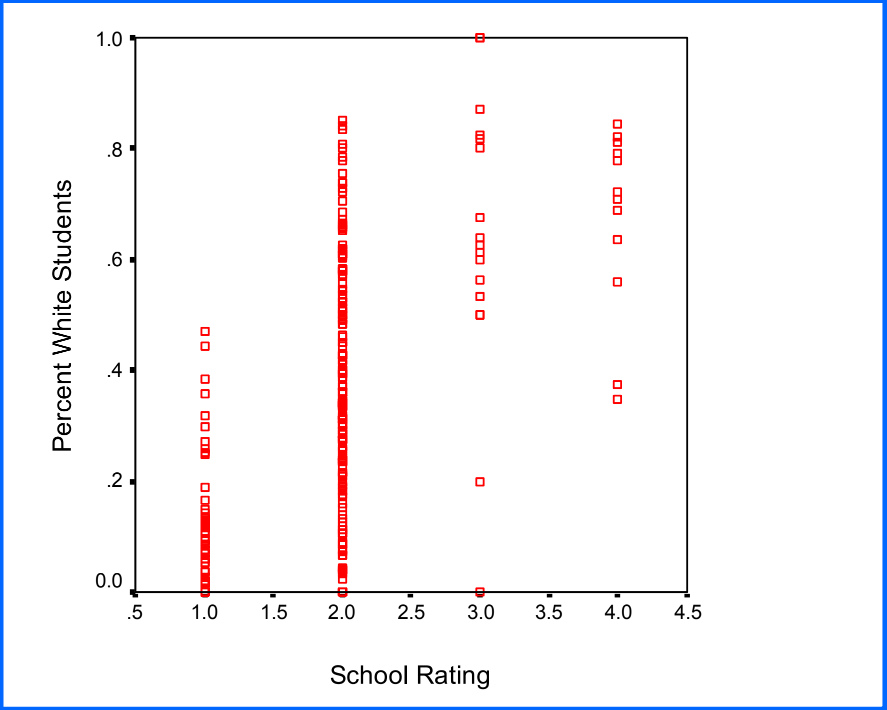


Highest Third of Schools on Free Lunch

Pattern matching: relationships between alternative measures of school effectiveness and confounding variables

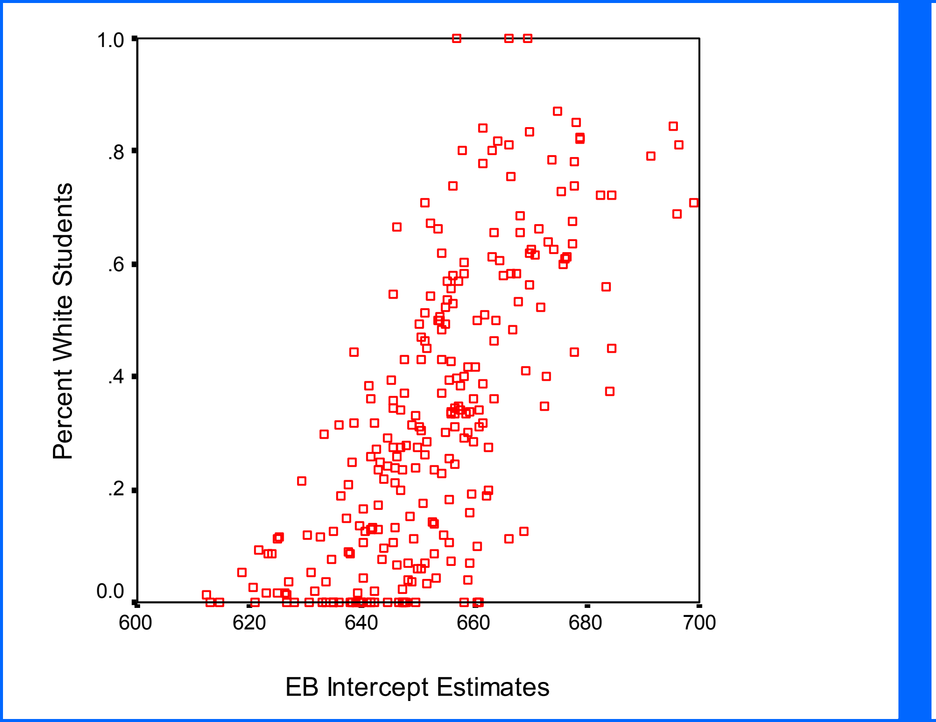
- NCLB Proficiency (percent proficient or above using state determined cutpoint)
- State rating of schools (weighted combination of proficiency score, attendance, dropout rates)
- HLM Empirical Bayes (EB) Intercept estimates
- HLM EB Slope estimates

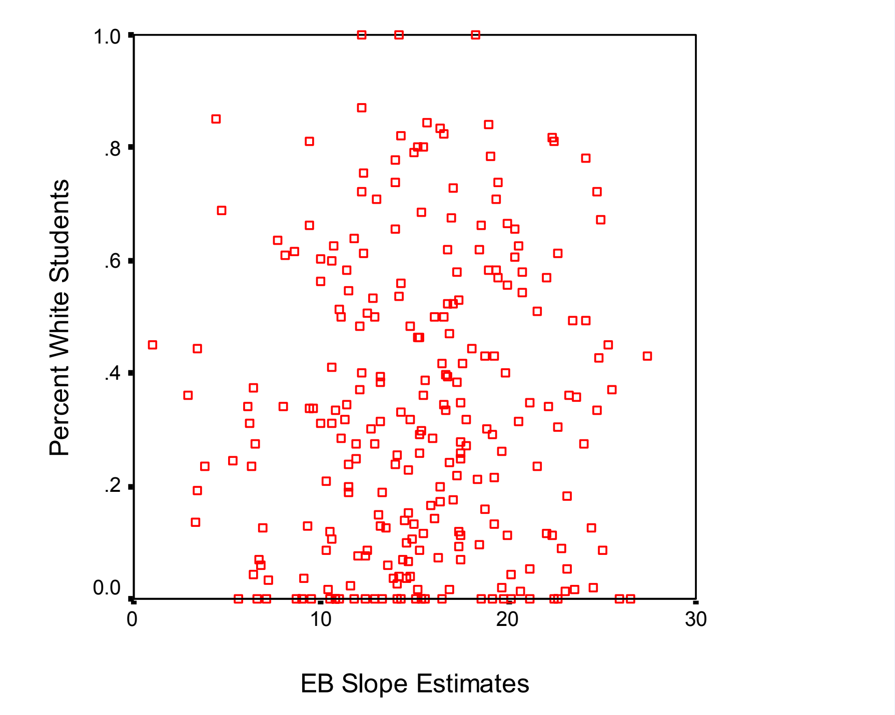




$r^2 = .48, p < .001$

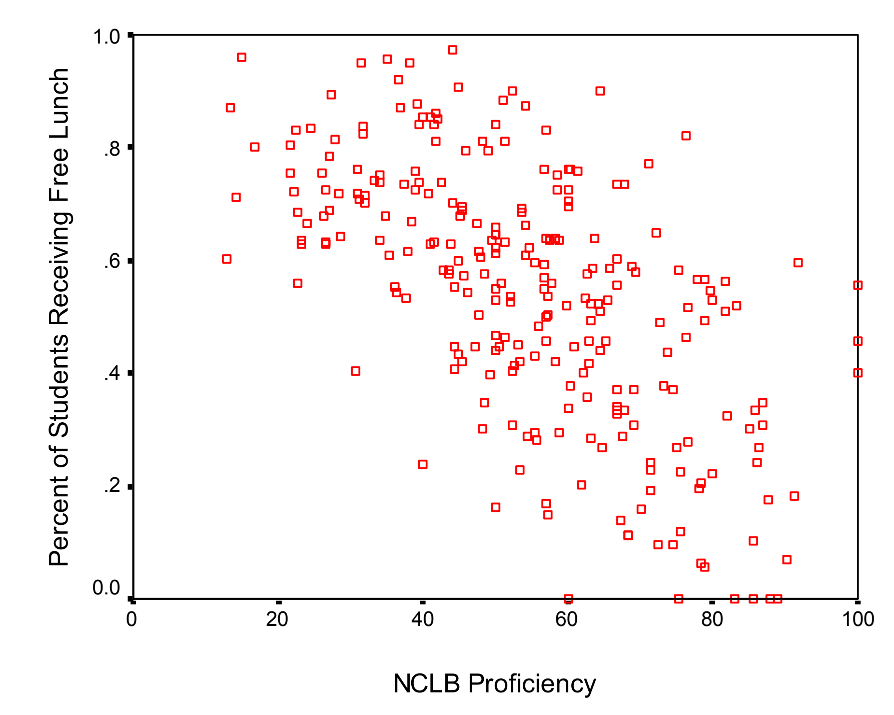
$r^2 = .40, p < .001$

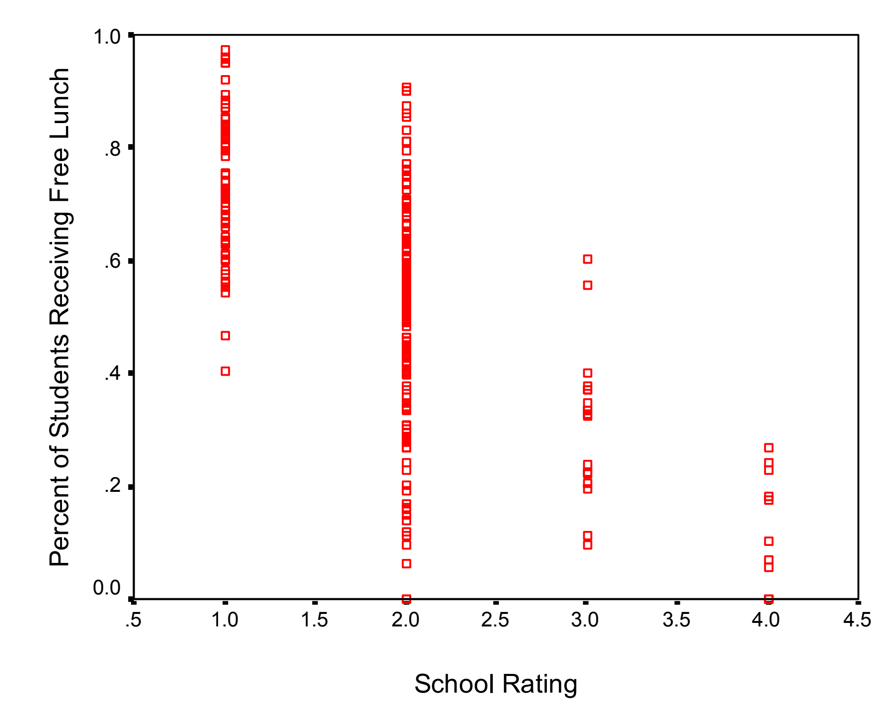




$r^2 = .27, p < .001$

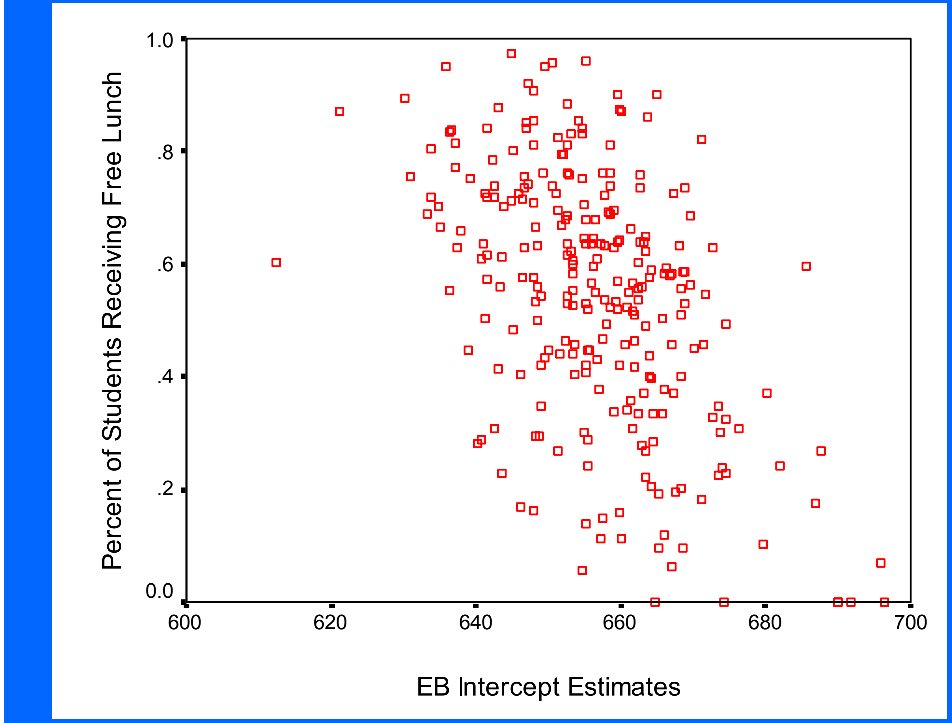
$r^2 = .06, p < .001$

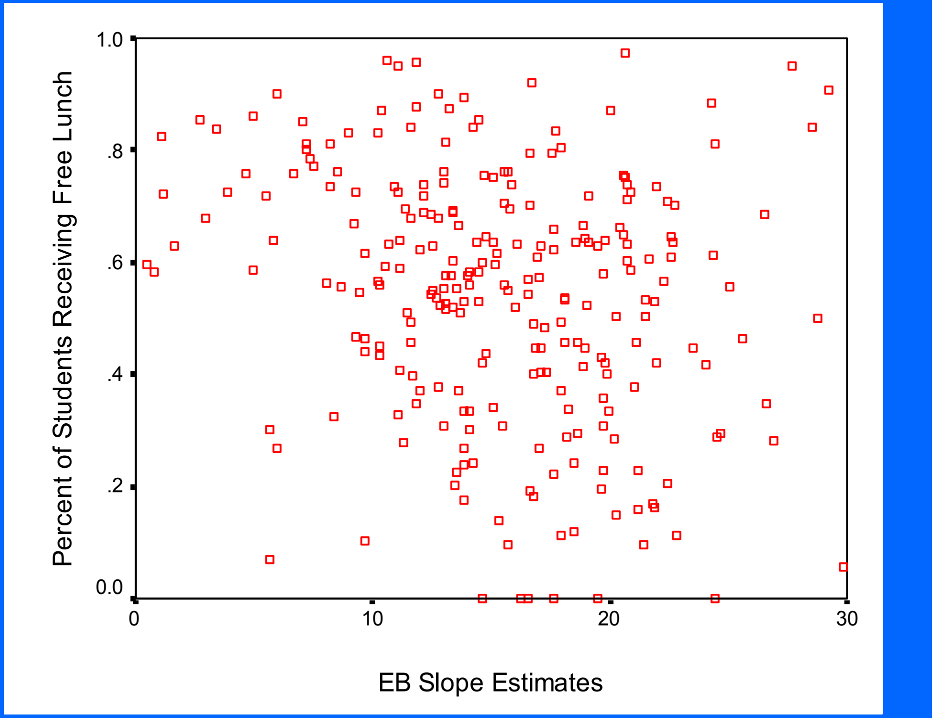




$r^2 = .38, p < .001$

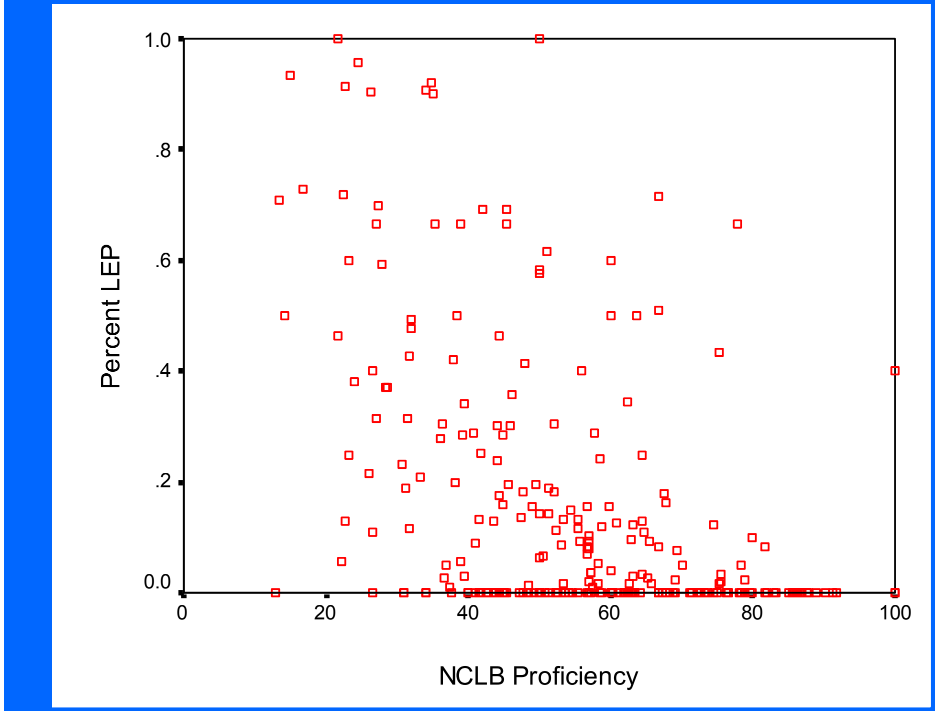
$r^2 = .44, p < .001$

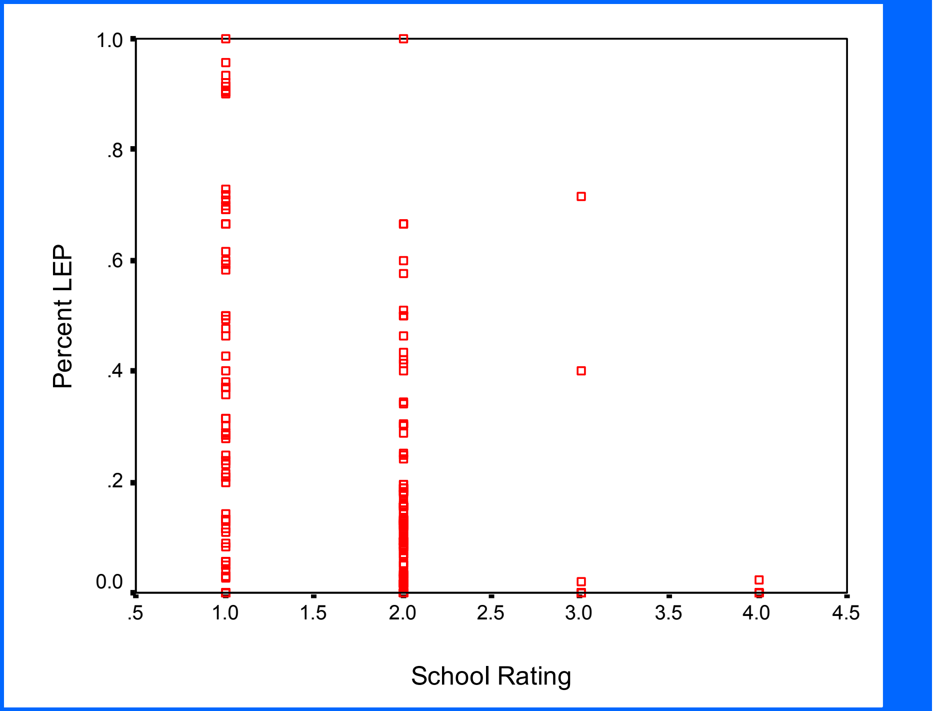




$r^2 = .26, p < .001$

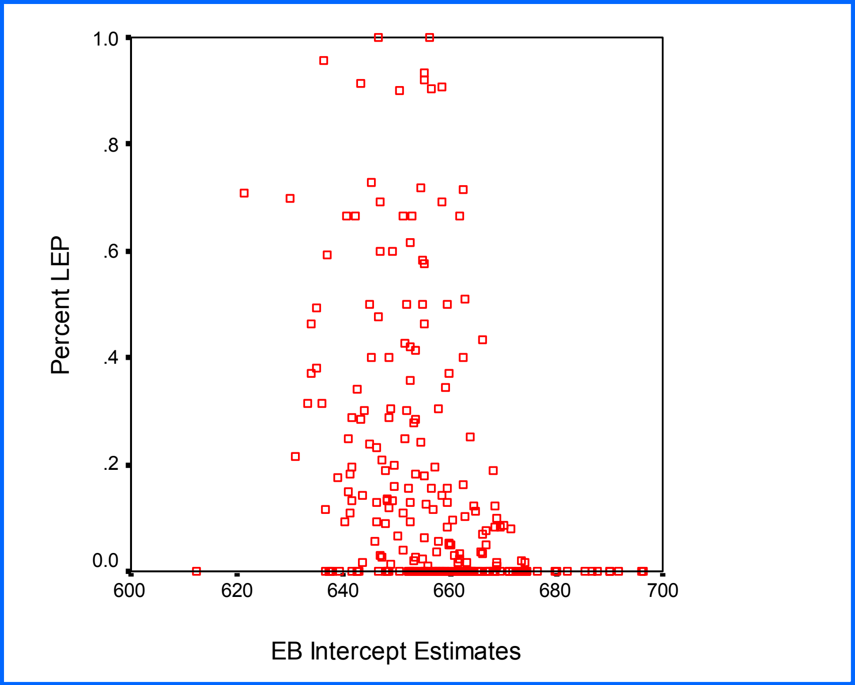
$r^2 = .06, p < .001$

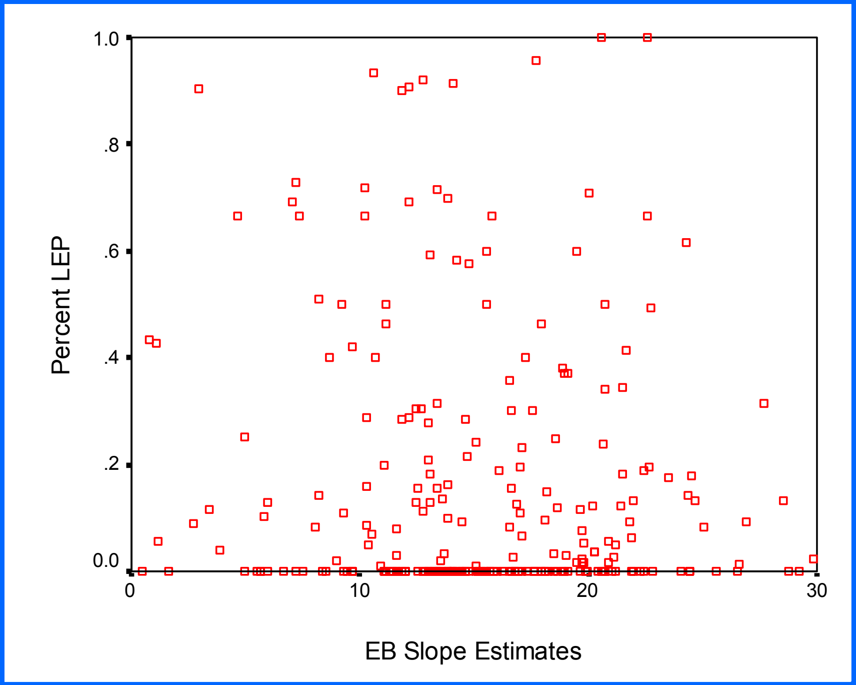




$r^2 = .24, p < .001$

$r^2 = .20, p < .001$





$r^2 = .13, p < .001$

$r^2 = .01, p = .195$

Conclusions – Study II

- All student level predictors were significantly related to mean achievement
- Only two of the seven student level predictors were related to linear growth; only three of the seven for curvilinear growth
- After conditioning on student level predictors, school level predictors were not significantly related to linear or curvilinear growth and only one of the four predictors was significantly related to school mean achievement
- A comparison of NCLB proficiency, NM School Rating, EB intercept estimates and EB slope estimates showed that slope estimates showed significantly smaller relationships with student majority status, LEP, and free lunch status

Summary and Conclusions

- Caveat – We have made no attempt to survey different designs being used across the states
- NCLB design provides little to no control over threats to internal validity
- Direct proficiency measures are likely to be correlated with intake characteristics and differences in school composition and context
- Differences from one cohort to another may undermine AYP as a stable measure of progress (Linn & Haug, 2002); such annual cohort fluctuations “could swamp differences in instructional effects” (Baker & Linn, 2002)

Summary and Conclusions

- Different designs, measures, and methods of analysis and estimation are likely to provide different evaluations of school effectiveness
- Choice of design and methods depends on evaluative purpose
- However, longitudinal research has been recognized as the “sine qua non of evaluation in nonexperimental settings” (Marco, 1974)
- Learning, the fundamental outcome of interest for school effectiveness, is a problem in the analysis of change, best addressed with longitudinal designs

Summary and Conclusions

- Longitudinal designs provide some degree of control over stable characteristics of students by using students as their own controls

- In our studies, socio-demographic factors that likely pose threats to the evaluation of school effectiveness are consistently related more strongly to status measures than growth measures

- In our studies, policy and practice is consistently related more strongly to growth measures than status measures

- Interaction of status and growth important as well

Summary and Conclusions

While longitudinal designs likely provide greater internal validity, a number of improvements can be made:

- Multiple outcome measures
- Explicit modeling of “treatment variables”
- Wider array of methods and measures for control of extraneous influences
- Need to pursue application of more rigorous designs

The Study of School Effectiveness as a Problem in Research Design

Contact Information:

120 Simpson Hall, University of New Mexico,
Albuquerque, NM 87131

505-277-4203, jstevens@unm.edu

Paper available at:

http://www.unm.edu/~jstevens/papers/design.ppt