Chapter

The Study of School Effectiveness as a Problem in Research Design

Joseph Stevens
University of New Mexico

With the implementation of the No Child Left Behind (NCLB, 2001) legislation there is an increased focus on the evaluation of school effectiveness and the identification of both exemplary and failing schools. In order to accomplish the promise of this focus, methodologies must be brought to bear that can disentangle the impact a school has on its students from other influences on student learning. NCLB and other recent federal mandates and programs place strong emphasis on "evidence based" or "scientifically based" research to better understand educational programs and interventions that are effective in promoting student learning. 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). In some discussions, scientifically based research is equated with experiments using random assignment or randomized clinical trials. For example, criteria used by the federal What Works Clearinghouse (WWC, n.d.) for reviewing research studies assign higher ratings to randomized

clinical trials than quasi-experiments and appear to essentially exclude from consideration research that uses case study and other research designs. Throughout these recent discussions of scientifically based evidence, there is relatively little consideration of a wide array of other methods of experimental and statistical control.

The purpose of this chapter is to draw attention to certain research design issues inherent in the study of school effectiveness and to examine the way in which these issues relate to NCLB methods for evaluating school effectiveness using adequate yearly progress (AYP). It is also our intent to contrast those methods with alternative research designs, especially longitudinal designs, and offer preliminary evidence on the performance of alternative designs in controlling for covariates that may represent threats to internal validity.

Using common notation (Campbell and Stanley, 1967; Cook and Campbell, 1979), the research design often described as the "gold standard" for future educational research can be diagrammed as:

Treatment Group	R	X	O
Control Group	R		O

where R denotes random assignment to group, X is the delivery of the treatment or experimental manipulation, and O indicates a measurement occasion or observation. Random assignment to condition, of course, ensures that confounding factors are only randomly related to group membership as long as the sample size is large enough for randomization to be effective. An important alternative design is the randomized longitudinal design:

Treatment Group	R	O...O	X	O...O
Control Group	R	O...O		O...O

One strength of the longitudinal design is the use of multiple measurement occasions either before or after treatment. While these two designs are particularly strong ones, historically they have not been commonly applied in educational research and are particularly rare in School Effectiveness Research (SER; Teddlie, Reynolds, and Sammons, 2000).

In addition to the push for the use of randomized experimental designs, the NCLB requirements also prescribe a number of other methods and procedures that constrain the design of a state's accountability system. All states are required to test annually in grades 3-8, and test results must be translated into three or more achievement proficiency categories

(i.e., basic, proficient, advanced). Test results cannot be weighted or combined. This proscription prevents the use of statistical models such as regression that involve the computation of weighted linear composites. The percent of students that are proficient in each content category must be reported as well as the percent of students proficient in each of a number of disaggregated groups. Regulations also proscribe any adjustment of scores or the statistical consideration of other factors that might impact scores such as socio-economic status of the student. Under NCLB, schools, districts, and states must all demonstrate Adequate Yearly Progress (AYP) in each content area through the comparison of each year's unmodified percentage of students reaching proficiency or above with a progress standard that requires all students to be proficient by the year 2013-14.

However, these NCLB methods for evaluating school, district and state educational effectiveness appear to strongly contradict the simultaneous federal push for more rigorous, scientifically based evidence. Collectively, NCLB regulations impose a form of case study design for the evaluation of school effectiveness for AYP. The NCLB accountability requirements result in a single-group, 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

The diagram presents an example using fourth graders, but of course the process is repeated for all students tested in grades 3-8. In each year, unknown and unmonitored instructional treatments are delivered by teachers and schools as represented by X?. Most commonly, a single testing occasion occurs in each year as represented by each O. This annual observation is then compared to a performance expectation for AYP as defined by federally required calculations. Although there is a tendency to think that the NCLB design is measuring trends in performance over time, the method actually involves the comparison of each year's cohort to the AYP expectation in a single year rather than an actual evaluation of trend information.

From a research design perspective, there are few strengths in the design imposed by NCLB. This case study design does not employ a pretest, random assignment to condition, control group or other research design features that might control for threats to internal and external va-

lidity (see Campbell and Stanley, 1967; Cook and Campbell, 1979; Pedhazur and Schmelkin, 1991; Shadish, Cook, and Campbell, 2002). NCLB prohibitions on the treatment of data prevent the application of methods of statistical control as well. One potential strength of the design is the use of multiple cohorts over time. Replication may signal valid treatment effects when consistent results occur regardless of occasion or cohort studied. Weaknesses of the NCLB design are more apparent and include absence of pretest, no control group or random assignment to conditions, no control over treatment implementation, and no control over plausible confounding factors.

Absence of these research design features makes the NCLB case study particularly susceptible to the effects of pre-existing group differences or changes in group composition from one cohort to another. One of the greatest challenges in estimating school effectiveness is separating "intake" to the school from "value added" by the school. Willms and Raudenbush (1989) distinguish two types of "school effects". Type A school effects are defined as the total impact on a student of attending a particular school, including not only what we might call quality of schooling, but also school environment, milieu, community surrounding the school, quality of the teaching staff, etc. Note that this definition includes all factors associated with a particular school no matter what the source. Type B school effects represent a subset of the Type A effect and include only those influences or impacts of schooling that are directly attributable to school practice and policy. The importance of either effect depends on one's purpose in evaluating schools. For example, for parents clearly the most important issue is the school with the best Type A effect. That is, parents are interested in knowing in what school their child will achieve best no matter what the reason.

For the purpose of monitoring the impact and effectiveness of schools, however, it should also be clear that interest should be focused on the Type B effect (Raudenbush and Willms, 1995). That is, in schools with the same average student background and the same average school context and milieu, how effective are the practices and policies of the school being evaluated? If the methods and research design used for accountability evaluations do not control in some way for these effects, then schools can be "held accountable" for factors over which they have little or no influence.

These distinctions are particularly important in communities where the context and composition of students and schools vary greatly. School

composition and environment have been shown to have substantial effects on student outcomes over and above the effects associated with the individual student's ability and social class (Willms, 1986). For example, advantaged schools may have not only higher socio-economic status (SES), but may show differences in parental involvement, rate of disciplinary problems, school atmosphere, peer attitudes, characteristics of the teaching staff, or other demographic or context factors. These kinds of effects may vary from school to school and may vary over time (Willms and Raudenbush, 1989).

To evaluate school effectiveness, it is necessary to use methods and research designs that can measure those outcomes that can be attributed to school practices and policies; separating out rival influences on children's learning and development. Schools should not be held "accountable" for the impact of factors not within school control. The selective enrollment of students into schools results in intake differences that represent confounding pre-existing differences as well as differences in prior achievement before enrollment. Since enrollment is selective and nonrandom, intake composition of the student body also represents school to school differences that can masquerade as school effects or can interact with aspects of school policy and practice. For example, as Ballou, Sanders, and Wright (2004) point out, if low SES students disproportionately attend schools with a less prepared teaching staff, then SES effects may interact with teacher preparation differences and vice versa.

Value-added methods are one approach to disentangling the effects of schooling from other influences on student achievement. However, there are many variations subsumed under the term "value-added". Some models are cross-sectional research designs, using only one measurement occasion at a time as the modeled outcome. Others analyze residuals. Some models are gain score models that examine differences in scores between two measurement occasions. Discussion of these various models is not the purpose of the present chapter. However, we believe that each of these variations has important advantages and disadvantages that should be considered and evaluated in light of the purpose of evaluation in the accountability system.

Our focus in this chapter is on longitudinal methods that model three or more measurement occasions and that use a cohort of students individually matched over time to allow the estimation of individual student growth trajectories (Raudenbush, 2001). We believe that the use of such

longitudinal analyses provides an important research design advantage over a number of alternative designs. To realize that advantage, it is necessary to measure performance at multiple points in time. As cautioned by Rogosa (1995): "Two waves of data are better than one, but maybe not much better" (p. 744). Our focus is also on the estimation of student learning and school effects rather than teacher effects. Thus, we see our work as an application of methods from the analysis of change literature to issues in the study of school effectiveness and evaluation.

Although much of the language in NCLB suggests a focus on change phenomena (e.g., "yearly progress"), the research design required by NCLB is a cross-sectional design depending wholly on the evaluation and interpretation of assessment performance in a single year. This kind of comparison does not directly measure the change or progress of individual students but the performance of each year's cohort in comparison to the calculated goal for performance (AYP).

Barton and Coley (1998) observe that "average score trends and cohort growth tell us different things...it does appear to be important to look at *both* measures" (p. 15). Cross-sectional designs provide useful and important information on the level or status of performance. However, cross-sectional designs that study different groups of students at different points in time are an inadequate means to examine processes like learning, improvement, progress, or other aspects of change that are inherently longitudinal. Recent developments in methods for the analysis of change have made longitudinal growth curve modeling methods more accessible and tractable (Collins and Horn, 1991; Collins and Sayer, 2001; Duncan, and Duncan, 1995; Duncan, et al., 1999; Ferrer, et al., 2004; Gottman, 1995; Little, Schnabel, and Baumert, 2000; MacCallum, Kim, Malarkey, and Kiecolt-Glaser, 1997; Muthèn and Curran, 1997; Plewis, 1996; Raudenbush, 2001; Willett and Sayer, 1994).

True longitudinal analysis requires the tracking and measurement of the same individuals measured at multiple points in time. Nesselroade (1991) argues that in order to adequately study change, repeated measures on the same individuals are needed. Reynolds and Teddlie (2000) recommend that the study of school effects should be based on longitudinal data on individual children. In longitudinal models, three measurement occasions are a minimum for identifying linear trends and estimation of more complex curve forms require additional occasions, although there are a number of mixed, longitudinal/cross-sectional designs that may ease

the burden of data collection over time (see, for example, Willett, Singer, and Martin, 1998). In addition, it is important to recognize that true changes in school effectiveness take time and may not be adequately modeled by one or two measurement occasions or short time periods. Gray, et al (1995) observe that annual changes in school effectiveness are likely to be modest in size and spans of three to five years may be necessary for the identification of true school improvement.

Goldstein (1991), describing school effectiveness studies in Britain, stated that "...It is now recognised...that 'intake' achievement is the single most important factor affecting subsequent achievement, and that the only fair way to compare schools is on the basis of how much progress pupils make during their time at school" (p.14). One of the greatest advantages in using growth or change models is that they may be less susceptible to the influences of student background, intake characteristics, and other confounding factors. In a repeated measures design, students serve as their own controls. As a result, stable characteristics of the child are constant over time and cannot confound estimation of the growth curve. This potential research design advantage is an important one in nonexperimental studies where there is insufficient control over confounding influences.

Longitudinal designs have a number of other strengths in addition to control over stable intraindividual characteristics. The design can include replication over multiple cohorts, it may be less intrusive than more rigorous designs resulting in greater external and ecological validity, and it is a design that is more easily understood by stakeholders than more complex statistical models. Even though advanced statistical modeling may be involved, individual student growth curves are easily displayed and stakeholders understand well the idea of tracking an individual's progress over time. Perhaps one of the greatest strengths of the growth curve modeling approach is its focus on the fundamental interest of education and school effectiveness research: learning. Learning is, by definition, a change phenomenon that entails the adaptation and elaboration of cognitive structure over time. As such, learning is represented well by longitudinal designs like growth curve modeling.

Of course, there are also several weaknesses that may threaten the validity of inferences drawn from the kind of longitudinal designs applied in SER research including, in typical applications, absence of a control group or random assignment to conditions and no control over treatment implementation. In addition there may be little control over certain con-

founding conditions like historical influences, carryover effects and instrumentation. Another issue of importance in these designs is the attrition that is likely to occur over measurement occasions and the concomitant changes in group or school composition that may result. We explore attrition issues in Study II below by comparing two samples with different rates of exclusion and attrition.

Given that any research design has strengths and weaknesses and given the high stakes application of research designs in school effectiveness models, it is critical that there is an explicit accounting of how well each design performs, what weaknesses and strengths are likely to accrue with particular design choices, and acknowledgement of the way in which causal inferences must be tempered depending on the nature of the particular research design. This accounting should include evaluation of the construct and consequential validity of the methods and designs being applied (Messick, 1989; 1993; 1994; 1995). In addition, critical evaluation of the performance of measures of school effectiveness is required. Evidence is needed that demonstrates that the method legitimately captures the effects of school policy and practice and simultaneously an evaluation of the degree to which the method is relatively immune to the influences of construct irrelevant sources of variation. Shadish, Cook, and Campbell (2002) describe a process of "pattern matching" that involves the logical, theoretical consideration of the attributes and characteristics of a construct that should be present followed by a process of observation and matching of actual attributes and characteristics as a method of determining validity. This process may be particularly useful in the context of research designs with relatively low internal validity. Pattern matching evidence may also be bolstered through replication of results. Shafer (2001) argues for the careful planning of replications in field settings where there may be substantial opportunities to incorporate multiple classrooms or schools in the observational design. When results are consistent over replications, there is support for stronger inferences.

Examination of plausible rival hypotheses (Riechardt, 2000; Rindskopf, 2000) provides another mechanism for studying and validating alternative methods and research designs that may be especially useful when strong forms of experimental design are not feasible as in many school effectiveness research studies. The essence of the approach is the development of multiple working hypotheses in addition to the one preferred by the researcher that could also plausibly account for the kinds of

phenomena or relationships under study. These multiple hypotheses are included in the planned design of the research and then simultaneously tested along with the researcher's preferred hypothesis. The set of rival, competing hypotheses are evaluated in light of the observed data. By explicitly testing, and hopefully ruling out, rival hypotheses for the phenomenon of interest, internal validity can strengthened substantially even when strong forms of experimental design are not feasible.

This is the context for the two studies reported in this chapter. We were interested in examining the plausibility of specific alternative causal claims for the performance of several outcome measures of school effectiveness. New Mexico schools vary widely in composition and type of community from very advantaged, predominantly White student populations to schools that are poor, rural, and serve students who are predominantly limited English proficient and come from non-White cultures. We sought to gather evidence for how different classes of rival explanations or covariates were related to alternative measures of school performance. We hypothesized that measures of school practice and policy presumed to have an impact on student learning should show differential patterns of relationship with measures of school effectiveness to the extent that the measures validly estimate learning outcomes. Conversely, covariates best conceptualized as confounding factors should show opposite patterns of relationship with measures of school effectiveness to the extent that the measures control for extraneous influences. In previous studies that have investigated a variety of student background characteristics, associations have often been found with student or school mean achievement (e.g., Coleman, et al., 1966; Hanushek, 1986; Jencks, et al., 1972) but associations have been found less frequently between background characteristics and student growth rates (Stevens, 2000; Stone and Lane, 2003).

In order to examine issues surrounding differences in the evaluation of school effectiveness using several alternative outcome measures, we examined New Mexico state testing data in two separate studies. In the first study (Zvoch and Stevens, 2004), background characteristics and policy and practice variables were related to student achievement status and growth using multilevel, linear growth models. In the second study, multilevel, curvilinear growth models were used to examine the relationships between several covariates and student achievement status and growth. In the second study, we also assessed the plausibility of several rival hypotheses through the estimation of relationships between confounding variables and four measures of school effectiveness.

Study I

The purpose of Study I was to examine correlates of status and growth in mathematics achievement over a three year period. We had particular interest in whether correlates related differently to status versus growth outcome measures. Individual math achievement scores on the state mandated achievement test were used.

Methods

Data from middle school students and schools in a large urban school district located in the southwestern United States were analyzed. At the middle school level, the district has 24 schools that serve over 20,000 students in grades 6 through 8. All sixth, seventh, and eighth grade students were tested annually on a state mandated, norm-referenced achievement test, the TerraNova/CTBS5 Survey Plus (CTB/ McGraw-Hill, 1997). Achievement data from students who were in sixth grade in 1998-99, seventh grade in 1999-00, and eighth grade in 2000-01 were analyzed. As the intent of the study was to examine school effects on the achievement and growth of students, a sample was selected that consisted only of those students who remained in the same middle school during all three years of the study period. Nine hundred thirty students who transferred schools at least once during the three-year period were dropped from the working file, reducing the analytic sample to 5,168 (84.7% of the original sample).

Fifty-one percent of the sample was female (N = 2,622); forty-nine percent was male (N = 2,546). Forty-seven percent (N = 2,402) of the sample was Hispanic, 44% (N = 2,271) was Anglo, 3% (N = 152) was African American, 3% (N = 173) was Native American, and 2% (N = 99) was of Asian descent. Forty percent (N = 2,058) of the sample received a free or a reduced price lunch, 17% (N = 867) were classified as English language learners (ELL), and 17% (N = 887) were special education students. Exclusion of mobile students lowered the percentage of special education students and English Language Learners and raised the percentage of impoverished students relative to district averages by 1%, 3%, and 5%, respectively.

Dummy codes were used to classify individual students as female, non-Anglo, ELL, economically disadvantaged (i.e., free and reduced lunch recipients, FRPL), and special education students. A dummy code was also used to identify students who received a modified test administration (N = 808; 15%). Achievement data used in the study were student scores

on the TerraNova/CTBS5 Survey Plus, a standardized, norm referenced achievement test (CTB/McGraw-Hill, 1997). The mathematics composite score was used in the present study and is derived from the 31-item Mathematics and the 20-item Mathematics Computation subtests. A KR-20 reliability estimate of .86 was reported for the Mathematics subtest in the 6th, 7th, and 8th grade standardization samples. For Mathematics Computation, KR-20 was reported as .83 in grade 6, .80 in grade 7, and .85 in grade 8. For the Mathematics composite, KR-20 estimates were .91 in grade 6, .90 in grade 7, and .92 in grade 8 (CTB/McGraw-Hill, 1997).

Available school level measures used as predictors were the percent of students in each school who received a free or reduced price lunch (FRPL), mean educational attainment of the mathematics staff, and mathematics curriculum. Teacher educational attainment was computed as the approximate number of years required for degree completion plus any post-degree graduate credits (e.g., Bachelor's degree = 16 years, Master's degree + 15 credits = 18.5 years). Mathematics teachers had on average slightly less than a Master's degree. However, school variation in mean educational attainment indicated that in some schools mathematics teachers had on average slightly more than a Bachelor's degree, while in others mathematics teachers attained on average a Master's degree with 15 additional graduate course credits.

The second measure of school practice was based on the mathematics curricula that were delivered to students. Three of the math curricula (i.e., MATH Thematics, Mathematics in Context, and Connected Mathematics) were recently developed reform-based approaches for delivering mathematics instruction that emphasize problem-solving, higher-order thinking skills, hands-on interactive learning, and serve as an alternative to traditional instructional approaches (see Reys, et al., 2003; Schoenfield, 2002; Senk and Thompson, 2003). Mathematics programs were coded into two categories (traditional = 0, reform = 1). Nine of the 24 middle schools (38%) implemented one of the reform curricula, the remaining schools used a traditional approach to mathematics instruction during the study period.

Analytic Procedures

Multilevel modeling techniques were used to model and assess student and school growth trajectories. Three-level longitudinal models were estimated using the Hierarchical Linear Modeling (HLM) program, version 5.05 (Raudenbush, Bryk, Cheong, and Congdon, 2001). An uncondi-

tional three-level model was first used to estimate a mathematics growth trajectory for each middle school student, to partition the observed parameter variance into its within and between school components, and to estimate each middle school's mean achievement score and mean growth rate. Second, a conditional three-level model was used in order to regress the achievement outcomes on student and school characteristics. In both models, level-1 was composed of a longitudinal growth model that fitted a linear regression function to each individual student's mathematics achievement scores over the three years studied (grades 6, 7, and 8). Equation 1 specifies the level-1 model, where Y_{tij} is the outcome (i.e., mathematics achievement) at time t for student i in school j, π_{0ij} is the status of student ij at the end of 6th grade, π_{1ij} is the linear growth rate across grades 6-8 for student ij, and e_{tij} is a residual term representing unexplained variation from the latent growth trajectory.

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

At level-2, within-school variation in the status (π_{0ij}) and growth rate (π_{1ij}) of students was first modeled unconditionally in terms of the status and growth parameters of the student's school and student-level residuals. In the conditional model, individual characteristics were added to the equation. Equations 2a and 2b specify the form of the conditional level-2 model.

$$\pi_{0ij} = \beta_{00j} + \beta_{pij}(a_{Pij}) + r_{0ij} \quad (2a)$$
$$\pi_{1ij} = \beta_{10j} + \beta_{pij}(a_{Pij}) + r_{1ij} \quad (2b)$$

In equations 2a and 2b, within-school variation in the status and growth of students was modeled as a function of the status (β_{00j}) or growth (β_{10j}) of school j, the student characteristics (a_{Pij}) that were hypothesized to account for observed variation in the parameters of the student growth model, and respective student-level residual terms, r_{0ij} or r_{1ij} (Raudenbush and Bryk, 2002).

At level-3, between-school variation in the status and growth rate of schools was first modeled unconditionally in terms of the grand mean achievement or grand mean slope of schools and school-level residuals. School-level predictors were then added to the conditional three-level model. Equations 3a and 3b specify the form of the conditional level-3 model.

$$\beta_{00j} = \gamma_{000} + \gamma_{pqs}(W_{sj}) + u_{00j} \quad (3a)$$
$$\beta_{10j} = \gamma_{100} + \gamma_{pqs}(W_{sj}) + u_{10j} \quad (3b)$$

In equations 3a and 3b, between-school variation in the status and growth of schools was modeled as a function of grand mean achievement (γ_{000}) or the grand mean slope (γ_{100}), the school characteristics (W_{sj}) that were hypothesized to account for observed variation in the parameters of the school growth trajectory, and respective school-level residual terms, u_{00j} or u_{10j} (Raudenbush and Bryk, 2002).

Results

Table 1 presents the results of the three-level unconditional model. The first estimate presented, the grand mean, is the average 6th grade mathematics scale score for all students in the sample (γ_{000} = 648.96). The second estimate, the grand slope, is the average annual growth rate of the same students between the end of 6th grade and the end of 8th grade (γ_{100} = 17.64). Estimates of student and school-level parameter variance are presented next. Chi-square tests demonstrated that students and schools differed significantly in achievement levels and the rate of achievement growth. These test results indicate that there are individual differences from one student to another in mathematics achievement in grade 6 as well as in the rate of achievement growth throughout middle school. At the bottom of the table, the percentage of between-school variance in means and slopes is presented and shows that greater variation in student mathematics achievement and growth occurs within than between schools. However, the amount of between school variance in mathematics growth is relatively large and con-

Table 1

Study I: Three-Level Unconditional Model for Mathematics Achievement

Fixed Effect	*Coefficient*	*SE*	*t*
School Mean Achievement, γ_{000}	648.96	3.09	209.82***
School Mean Growth, γ_{100}	17.64	0.87	20.16***
Random Effect	*Variance Component*	*df*	χ^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.
*** $p < .001$

siderably greater than the amount of between school variance in mean mathematics achievement.

Results of the three-level conditional model are presented in Tables 2 and 3. Table 2 displays the within-school results. In Table 2, it can be seen that students from special student populations as well as students who received a modified test administration performed at a level that was significantly below their counterparts. The difference in achievement was approximately a quarter of a standard deviation for non-Anglo and economically disadvantaged students and approximately a half a standard deviation for English Language Learners, special education students, and students who received a modified test administration. A somewhat different pattern emerged when student growth in achievement was considered. In four of six comparisons, student background (and test administration status) was not statistically related to the rate at which students learned mathematics. On average, only female and ethnic minority students grew at a slower rate than their counterparts over the middle school years. Otherwise, the lack of relationship between economic, education, and English language status and growth in mathematics indicates that the initial

Table 2

Study I: Within-School Model Relating Individual Characteristics to Mathematics Achievement

Fixed Effect	*Coefficient*	*SE*	*t*
Individual Achievement, β_{00}	650.15	1.71	379.93***
Biological Sex, β_{01}	0.58	1.03	0.56
Minority Status, β_{02}	-9.11	0.87	-10.47***
Free Lunch Status, β_{03}	-9.63	1.44	-6.69***
LEP Status, β_{04}	-17.81	1.42	-12.52***
SPED Status, β_{05}	-20.72	3.82	-5.43***
Test Administration Status, β_{06}	-19.90	4.05	-4.92***
Individual Growth, β_{10}	17.91	0.79	22.63***
Biological Sex, β_{11}	-2.65	0.50	-4.35***
Minority Status, β_{12}	-1.56	0.46	-3.39**
Free Lunch Status, β_{13}	-0.41	0.50	-0.84
LEP Status, β_{14}	-0.50	0.82	-0.62
SPED Status, β_{15}	-1.69	1.66	-1.02
Test Administration Status, β_{16}	-2.98	2.32	-1.29
Variance Component	*Level-1*	*Level-2*	*Variance Explained*
Individual Achievement, r_{0ij}	1132.02	791.15	30.1%
Individual Growth, r_{1ij}	44.04	39.90	9.4%
School Mean Achievement, u_{00j}	222.42	63.04	71.7%
School Mean Growth, u_{10j}	17.01	14.02	17.5%

** $p < .01$, *** $p < .001$

achievement differences between these student groups remained constant over time.

At the bottom of Table 2, the percent reduction in the unexplained variance of the within and between components of the model are presented. A comparison of unconditional and conditional variance estimates reveals that individual background characteristics accounted for a small to moderate amount of the variation in student achievement outcomes and a moderate to large amount of the variation in school achievement outcomes. Student characteristics accounted for 31% of the variation in students' initial status in mathematics and 9% of the variation in students' mathematics growth and 72% of the variation in school mean achievement and 18% of the variation in school mean growth.

Table 3 presents the conditional between-school results. After adjustment for individual covariates, the percentage of free lunch recipients in 1998-99 was significantly related to the school mean achievement, but not to school mean growth. The opposite pattern of relationship between predictors and outcomes was observed for the measures of school practice. The mean educational level of the 6th grade mathematics staff in 1998-99 was not significantly related to school mean achievement, but its counterpart (staff mean attainment across the three study years) was significantly related to school growth rates. A similar pattern was observed with mathematics curricula in that the curriculum delivered to students

Table 3

Study I: Between-School Model Relating 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.
*$p < .05$,** $p < .01$, *** $p < .001$

was only associated with school mean growth and was not significantly related to school mean achievement.

Study I: Summary and Conclusions

Study I was designed to investigate the degree to which student and school characteristics relate to mathematics achievement outcomes and to demonstrate the differences between estimation of achievement level and achievement growth. Results indicated that students and schools differed significantly in achievement levels and growth rates. Results also showed that the variation within schools was greater than the variation between schools on both outcome measures. However, school-to-school differences in growth were greater than school-to-school differences in achievement level. Investigation of the source of the school-level achievement differences indicated that while the economic context of schools (FRPL) was the primary determinant of school achievement levels, that same factor was not significantly related to school growth rates. Examination of the relationship between aspects of school practice and school achievement outcomes revealed an opposite pattern. Educational level of the mathematics staff and the type of mathematics curricula implemented in the school were not significantly related to school achievement levels but were statistically significant predictors of growth in mathematics achievement. The patterns of association that were demonstrated in Study I suggest that conclusions drawn about relationships between student and school characteristics and student achievement outcomes may depend on the analytic model applied to the data.

Study II

The purpose of the second study was to further examine correlates of student status and growth in mathematics achievement but in this study curvilinear growth models were applied to student achievement data over a four-year period. The study also sought to examine differences that arise from evaluating schools using four different measures of school effectiveness: the state accountability rating system, a measure of student proficiency as directed by NCLB, and measures of status and growth estimated by multilevel growth modeling. One of our interests in examining the four outcome measures was to use a pattern matching strategy to investigate whether the four measures of school effectiveness showed differential relationships with covariates that can be conceptualized as potential confounding factors or rival explanations for school performance.

Specifically we hypothesized that growth parameters would not show relationships with stable characteristics of students that should be controlled by the repeated measures nature of the research design. Relationships might be found with covariates that were not stable individual characteristics or variables that represent other aspects of policy, practice, or context of schooling. Conversely, we hypothesized that the other three outcome measures would show relationships with covariates representing individual student characteristics.

Two analytic samples were used in the study. In order to make explicit any differences that might occur due to student mobility from one school to another or student drop-out, we applied a two-level HLM model to all students who took the state mandated TerraNova test in the school year 1999-00 regardless of school affiliation (referred to hereafter as the "student differences sample"). The second sample was composed of a subsample of these students who remained in the same middle school for at least two of the three years of middle school (referred to hereafter as the "school differences sample"). A three-level HLM model was applied to the data for the school differences sample.

Analytic Procedures

Multilevel modeling techniques comparable to those used in Study I were used to model and assess student and school growth trajectories in Study II. Two- and three-level longitudinal models were estimated using the Hierarchical Linear Modeling (HLM) program, version 5.05 (Raudenbush, Bryk, Cheong, and Congdon, 2001). For the student differences sample, a two-level unconditional model was first used to estimate a mathematics growth trajectory for each middle school student. Next, a conditional two-level model was run in order to regress the achievement outcomes on student characteristics. In both models, level-1 was composed of a longitudinal growth model that fitted a curvilinear regression function to each individual student's mathematics achievement scores over the four years studied (grades 6, 7, 8, and 9). Equation 4 specifies the level-1 model, where Y_{tij} is the outcome (i.e., mathematics achievement) at time t for student i in school j, π_{0ij} is the status of student ij at the end of 6th grade, π_{1ij} is the linear growth rate across grades 6-9 for student ij, π_{2ij} is the curvilinear growth rate across grades 6-9 for student ij, and e_{tij} is a residual term representing unexplained variation from the latent growth trajectory. Equations 5a-5c describe the level 2 unconditional model for the student differences sample. In the student differences sample, the

conditional model applied used the same form at level 1 as shown in equation 4 and added student level predictors at level 2 as illustrated in equations 7a-7c, where each a_{Pi} represents a student level 2 predictor.

In the school differences sample, levels 1 and 2 for both the unconditional and conditional models were the same as for the student differences sample. Analyses of the school differences sample also added a third level in the hierarchical models representing school effects as indicated by equations 6a-6c for the unconditional model and equations 8a-8c for the conditional model in which each school level predictor variable is represented by a W_{sj}.

Unconditional Models:

Level-1

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}(Grade) + \pi_{2ij}(Grade^2) + e_{tij} \quad (4)$$

Level-2

$$\pi_{0ij} = \beta_{00j} + r_{0ij} \quad (5a)$$
$$\pi_{1ij} = \beta_{p1j} + r_{1ij} \quad (5b)$$
$$\pi_{2ij} = \beta_{p1j} + r_{1ij} \quad (5c)$$

Level-3

$$\beta_{p0j} = \gamma_{000} + u_{00j} \quad (6a)$$
$$\beta_{p1j} = \gamma_{pq1} + u_{10j} \quad (6b)$$
$$\beta_{p1j} = \gamma_{pq1} + u_{10j} \quad (6c)$$

Conditional Models:

Level-2

$$\pi_{0ij} = \beta_{00j} + \beta_{pij}(a_{Pij}) + r_{0ij} \quad (7a)$$
$$\pi_{1ij} = \beta_{p1j} + \beta_{pij}(a_{Pij}) + r_{1ij} \quad (7b)$$
$$\pi_{2ij} = \beta_{p1j} + \beta_{pij}(a_{Pij}) + r_{1ij} \quad (7c)$$

Level-3

$$\beta_{p0j} = \gamma_{000} + \gamma_{pqs}(W_{sj}) + u_{00j} \quad (8a)$$
$$\beta_{p1j} = \gamma_{pq1} + \gamma_{pqs}(W_{sj}) + u_{10j} \quad (8b)$$
$$\beta_{p1j} = \gamma_{pq1} + \gamma_{pqs}(W_{sj}) + u_{10j} \quad (8c)$$

Student Differences Sample

In 1999-00, 23,469 sixth grade children took the state mandated TerraNova/CTBS5. This study includes the 23,296 sixth graders (99.3%) who took the mathematics subtest. These students were matched longitu-

dinally to 7th, 8th, and 9th grade records for the years 2001, 2002, and 2003. Eighty-five percent 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. Ethnic composition of the sample was 9,143 Hispanic students (53%) 1,693 Native American students (10%), and 6,377 White students (37%). Forty-nine percent of the students were female, 2,028 children (12%) were special education students, 1,450 (8%) received a modified test administration, and 2,335 (14%) students were classified as Limited English Proficient (LEP).

The TerraNova/CTBS5 mathematics subtest was used as the outcome measure. The publisher reports KR-20 estimates of reliability of .91 in grade 6, .90 in grade 7, .92 in grade 8, and .92 in grade 9 (CTB/McGraw-Hill, 1997). A lower-bound estimate of reliability for New Mexico students was found to be .89 for the mathematics subtest in 1999 (Stevens, 2001). The standardized scale used is a vertically equated developmental scale (CTB/McGraw-Hill, 1997).

In the conditional models, available variables that described student characteristics and background were used to examine the relationships between these variables and estimates of mathematics status and growth. The student level variables used at level 2 of the two-level HLM model were: gender (female), ethnicity (white), special education status, modified test administration, bilingual, limited English proficient (LEP), and stability in school. Forty-nine percent of the students were female, 34% were white, 15% were identified as special education students, 10% received a modified test administration, 10% were bilingual, and 16% were LEP. All student level predictors were dummy coded except the stability variable. The stability variable was coded 0 if the student changed middle school twice in three years, 1 if the student changed once in three years, and 2 if the student remained in the same school for all three years of middle school.

Results

An unconditional growth model was fit to the data to serve as a baseline comparison model. As can be seen in Table 4, average achievement was about 652 scale score points with a linear growth rate of almost 17 points per year and a curvilinear growth rate of about -1.5 points per year. Next a conditional model using all student level predictors was applied (see Table 5). This model provided significantly better fit than the unconditional growth model as indicated by reduction in model deviance, $\chi 2$ (24)

Table 4

Study II: Unconditional Model for Student Level Mathematics Achievement

Fixed Effect	Coefficient	SE	*df*	*t*
Mean Achievement, γ_{00}	651.75	0.32	23295	2153.81*
Linear Growth, γ_{10}	16.83	0.26	23295	63.84*
Curvilinear Growth, γ_{20}	-1.48	0.08	23295	-17.73*
Random Effect	**Variance Component**		*df*	$\chi 2$
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			

* $p < .001$

Table 5

Study II: Mathematics Achievement Predicted by Individual Characteristics

Fixed Effect	Coefficient	*SE*	*t*	*df*	*p*
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
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
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, r_{0ij}	1701.63	1181.41	30.6%
Linear Growth, r_{1ij}	291.05	278.43	4.3%
Curvilinear Growth, r_{2ij}	20.72	20.10	0.3%

= 9528.95, $p < .001$. For purposes of illustration, Figures 4, 5, and 6 show a random sample of growth curves for Hispanic, Native American, and White students. Reliabilities at Level 1 were .733 for the intercept, .186 for linear slope and .135 for curvilinear slope. All variance components were significant. The interrelationship between intercept or status and linear slope (τ_{01}) was -.378. As can be seen in Table 5, all variables except gender were significant predictors of mean achievement. In contrast, only four of the eight predictors were significantly related to either linear or curvilinear growth rates. For both growth parameters, stability and LEP status were associated with significantly higher linear growth rates with

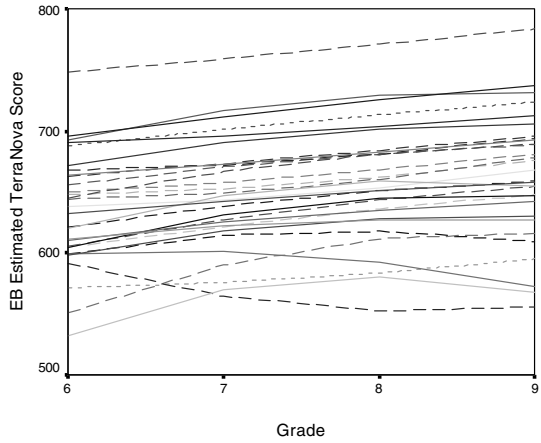

Figure 4. Sample Hispanic student growth trajectories

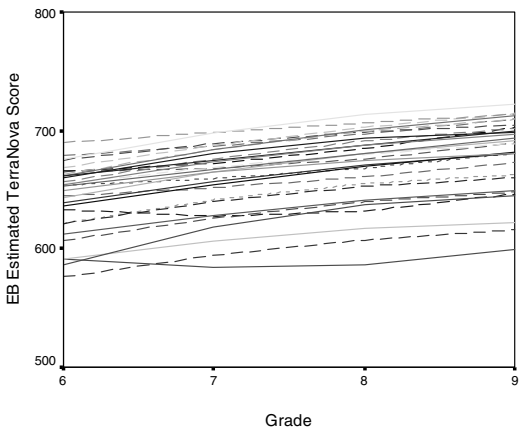

Figure 5. Sample Native American student growth trajectories

negative curvilinear components. Figure 7 shows a comparison of Non-LEP and LEP student growth curves.

Participation in a Title I program or being a female student was significantly associated with lower linear growth rates and positive curvilinear components. R^2 for the linear growth model at level 1 was .31 and for student level predictors at level 2, R^2 was .28. Examination of the variance components at the bottom of Table 5 shows that almost 31% of the variation in students' mean achievement was accounted for by the predictors used in the model but only small percentages of variation in student linear and curvilinear growth rates were accounted for with these predictors.

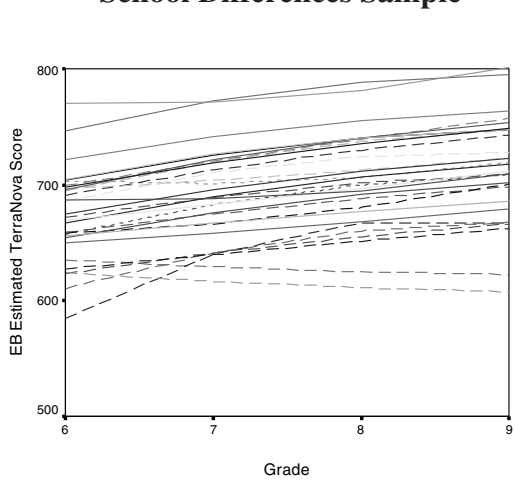

Figure 6. Sample White student growth trajectories

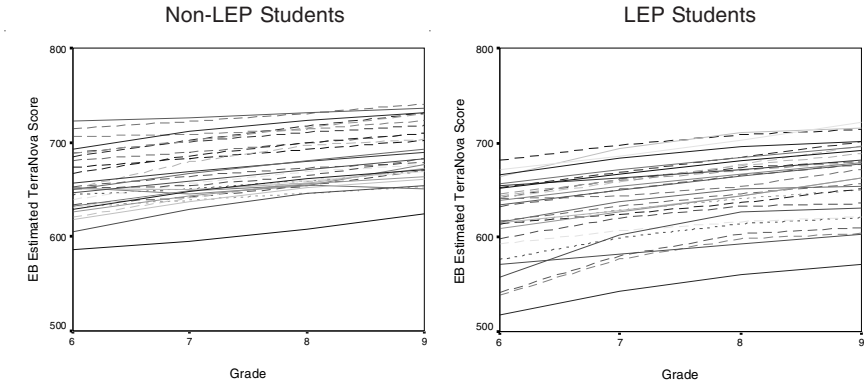

Figure 7. Estimated growth trajectories for Non-LEP and LEP students

In order to evaluate school level differences we also applied three level HLM models to a sample that included only those students who were present for testing in the same middle school for 2 or 3 years (17,596; 75.5% of student differences sample). Schools with less than five 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. These exclusions were designed to produce a sample of students who were available for an extended period of instruction to be impacted by school policy and practice and who had sufficient numbers per school to provide statistical estimates of status and growth. The resulting 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.

Table 6

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
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%

The models applied to this sample used the same student level predictors as the previous analyses with the student differences sample. At the school level, the following predictors were used: percent White students in the school (M = 32%), percent students in the school in a bilingual program (M = 15%), percent of students who were classified as LEP (M = 14%), and the percent of students who were receiving a free lunch (M = 53%). All school level predictors were grand mean centered.

Results

An unconditional linear growth model was first fit to the data to serve as a baseline. Next a conditional model using all student level predictors was applied (see Table 6). A deviance test showed that this model provided significantly better fit than the unconditional model, $\chi 2$ (315) = 6925.38, $p < .001$. The relationship of student mean achievement to linear growth was -0.274. Reliabilities at the student level were .664 for the intercept, .385 for linear slope and .354 for curvilinear slope.

All student level predictors were significantly related to mean achieve-

Table 7

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_{002}	-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
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

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_{20}	2.94	2.60	2.49	4.2%

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

ment. Only two of the seven student level predictors (gender and participation in a Title I program) were related to linear growth and only three of the seven (White ethnicity, gender, and participation in a Title I program) were significant predictors of curvilinear growth.

Next a conditional model was applied adding the four school level predictors (Bilingual, LEP, White, Special Education). This model also provided significantly better fit than the unconditional model, $\chi 2$ (327) = 7003.55, $p < .001$ (see Table 7). The relationship between mean achievement and linear growth was -0.480. Reliabilities at level 2 were .608 for the intercept, .378 for linear slope and .347 for curvilinear slope. After conditioning on the student level predictors, school level predictors were not significantly related to linear or curvilinear growth and only one of the four predictors (percent White students) was significantly related to school mean achievement. At level 2, R^2 for the student level predictors was.23. At the school level, R^2 was .24.

As described earlier, one of the purposes of Study II was to employ a pattern matching strategy to determine whether different outcome measures were more or less strongly associated with covariates that have the potential to confound the evaluation of school effectiveness. To this end, we examined the correlations between proficiency as defined under NCLB (percent proficient or above using state determined cutpoint), the state accountability rating of schools (a weighted combination of proficiency score, attendance, and dropout rates), and the HLM Empirical Bayes (EB) intercept and slope estimates arising from the models described above. Figures 8 and 9 show the relationships between school ethnicity and the four alternative measures of school effectiveness. Each figure shows individual middle school's bivariate performance on the outcome measure plotted against the school covariate value. As can be seen in the left and right panels of Figure 8, NCLB proficiency and the state school ratings are significantly correlated with the percentage of White students in the school ($R^2 = .48$, $p < .001$ and $R^2 = .40$, $p < .001$, respectively). Figure 9 shows the EB intercept and slope estimates which were also correlated significantly with percent of White students in the school ($R^2 = .27$, $p < .001$ and $R^2 = .06$, $p < .001$, respectively), although the size of relationship with slope estimates is small. Figures 10 and 11 show the relationships between the four outcome measures and the percentage of students in the school receiving free lunch and Figures 12 and 13 show the relationships with the percentage of LEP students in each school. Similar patterns of

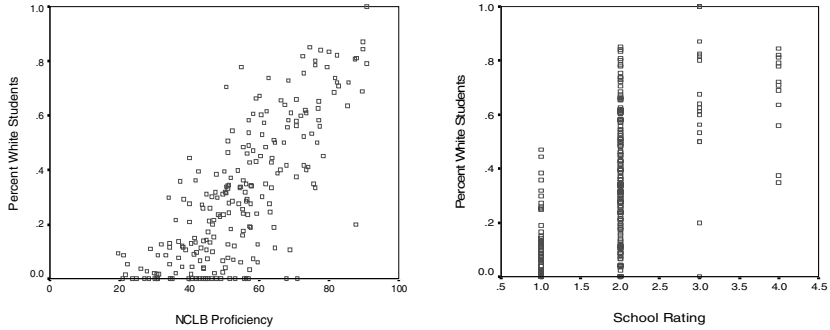

Figure 8. Relationship between percent White students in school and NCLB proficiency ($r^2 = .48$, $p < .001$) and state school accountability ratings ($r^2 = .40$, $p < .001$).

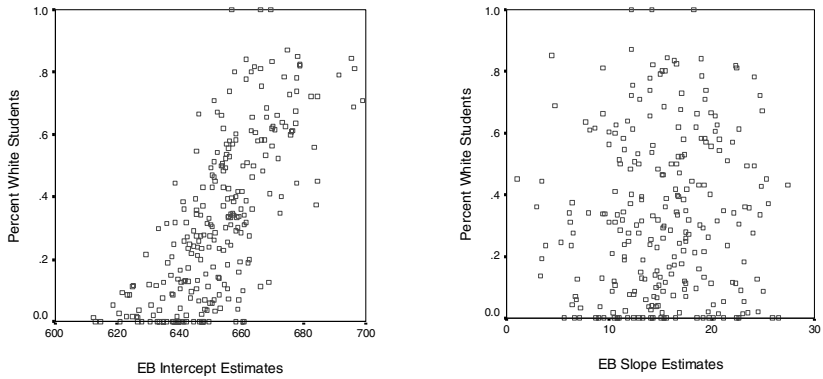

Figure 9. Relationship between percent White students in school and EB intercept estimates ($r^2 = .27$, $p < .001$) and EB slope estimates ($r^2 = .06$, $p < .001$).

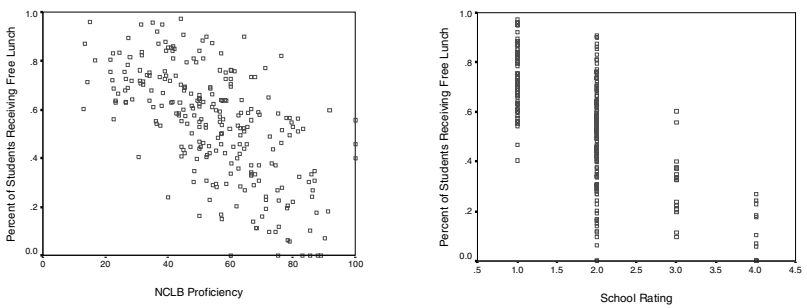

Figure 10. Relationship between percent of students receiving free lunch and NCLB proficiency ($r^2 = .38$, $p < .001$) and state school accountability ratings ($r^2 = .44$, $p < .001$).

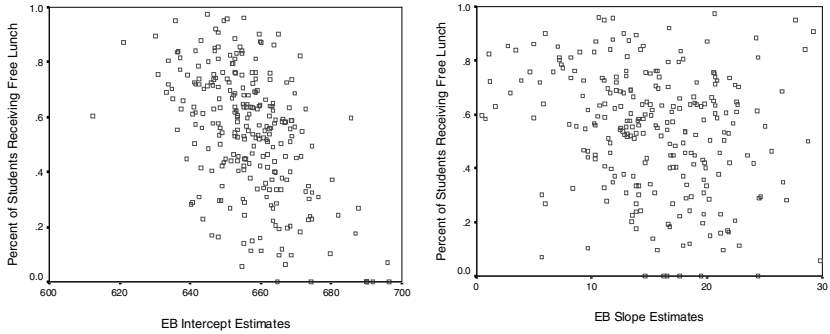

Figure 11. Relationship between percent of students receiving free lunch and EB intercept estimates ($r^2 = .26$, $p < .001$) and EB slope estimates ($r^2 = .06$, $p < .001$).

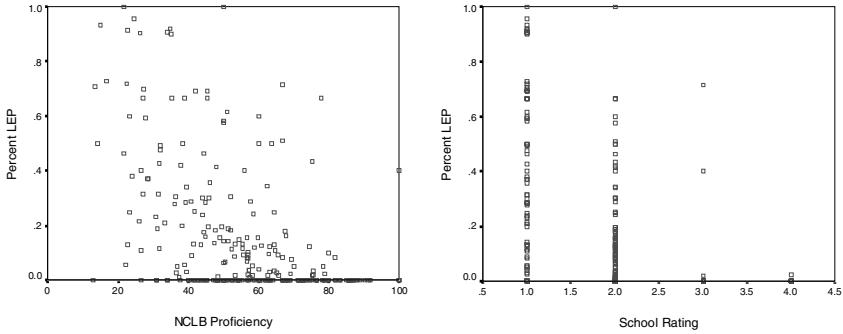

Figure 12. Relationship between percent of LEP students and NCLB proficiency ($r^2 = .24$, $p < .001$) and state school accountability ratings ($r^2 = .20$, $p < .001$).

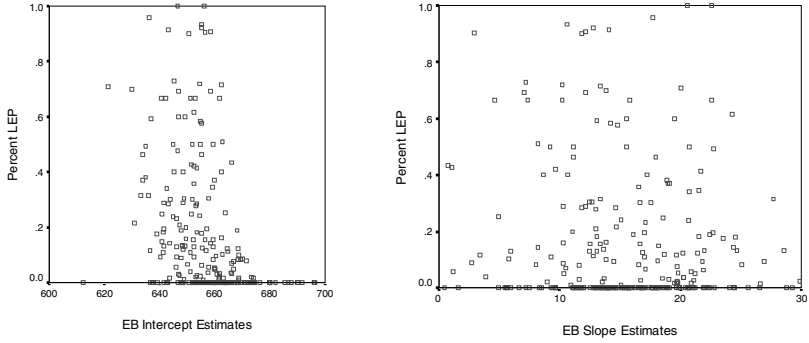

Figure 13. Relationship between percent of LEP students and EB intercept estimates ($r^2 = .13$, $p < .001$) and EB slope estimates ($r^2 = .01$, $p < .195$).

relationship occur in each of the figures with significant correlations between the covariate and NCLB proficiency and the state school rating. Somewhat smaller correlations are observed between each covariate and the EB intercept estimates and noticeably smaller relationships occurred with EB slope estimates (R^2 = .06 or smaller).

Study II: Summary and Conclusions

The results of Study II showed similar patterns to those for Study I. Application of multilevel growth models provided information on both status and growth of mathematics achievement and their relationships to student and school level predictors. As in Study I, covariate relationships differed depending on whether status or growth was examined. Almost all examined relationships of covariates with student status were statistically significant. A subset of covariates was significantly related to growth estimates. After conditioning on student level covariates, only one of the school level predictors was significantly related to school mean achievement and none of the covariates were significant predictors of school growth rates. Examination of patterns in the relationships between three covariates and school outcome measures showed significant relationships between school composition in the form of percent White students, percent free lunch students and percent LEP students and NCLB proficiency and state school accountability ratings. Somewhat smaller relationships were observed for EB intercept estimates and noticeably smaller relationships were observed for EB slope estimates.

Discussion

The purpose of this chapter was to draw attention to research design issues inherent in NCLB and alternative accountability measures for evaluating school effectiveness. Evaluation of the research design characteristics of accountability methods and outcome measures can provide important insights into their appropriate use and interpretation and can also validate the extent to which causal inferences can be drawn about the performance of school personnel, policy, or programs. While the use of more rigorous experimental methods and designs is recommended, it is likely that the nonexperimental, field study character of much educational evaluation and research including that conducted for high-stakes accountability purposes will remain commonplace. While the use of statistical methods of control are certainly to be recommended, we urge greater attention to opportunities for the use of research design tactics like replication, cross-validation, re-

peated measures, and the explicit examination of plausible rival hypotheses to validate and strengthen the inferential process.

NCLB regulations require the use and application of a posttest-only, case study design that provides little or no control over threats to internal and external validity. Unadjusted proficiency measures derived from a design like that required by NCLB are likely to be correlated with intake characteristics and differences in school composition and context. In Study II we evaluated the relationships between several measures of school composition and context and the state accountability ratings as well as a NCLB-type proficiency measure. These relationships were relatively large and statistically significant. As a result, it is not possible to rule out the rival hypotheses that these outcome measures are strongly influenced by construct irrelevant factors such as the schools composition in terms of language, poverty, and ethnicity. If so, then evaluations of school performance using these measures and methods may actually be better understood as estimates of school intake or Type A effects rather than an indication of the success or failure of school policy and practice. Since AYP depends on the successive comparison of different cohorts to a performance standard, differences in the composition of cohorts each year may also undermine AYP as a stable measure of progress (Linn and Haug, 2002). Without the application of experimental or statistical methods of control, such annual cohort fluctuations "…could swamp differences in instructional effects" (Baker and Linn, 1996).

Conditioned, EB estimates of school mean achievement showed lesser correlations with the measures of school composition and context in Study II. This is likely a result of the statistical adjustments that were used at the student level to take student characteristics into account.

In contrast to the results for the other outcome measures, EB growth estimates showed only small associations with the measures of school composition and context in Study II. We believe this evidence is suggestive of the intraindividual effect that occurs when true longitudinal designs are used and students serve as their own controls. The results also imply that, for the growth estimates, the rival hypotheses can largely be ruled out. The small associations observed indicate that construct irrelevant factors such as the school context of language, poverty, and ethnicity are unlikely contributors to the observed school growth estimates.

The patterns of evidence found in both Studies I and II also support the idea that the different outcome measures are differentially sensitive to

covariates representing policy, practice, social context, and student background. In Study I, teacher educational level and mathematics curricula were shown to have a relationship only with student rates of growth and not with mean achievement. The observed teacher effects on student growth are consistent with other recent studies that have identified the influence of teachers as one of the most important factors in promoting student achievement progress (Rivkin, Hanushek, and Kain, 2002; Rowan, et al., 2002; Wright, Horn, and Sanders, 1997) but are contrary to other studies and reviews that demonstrate mixed or negligible associations between teacher characteristics and student achievement levels (Hanushek, 1986; 1996). The results of Study I suggest that part of the discrepancy in past findings may depend on the kind of achievement outcome measure used. The same pattern of results was found for the effects of district mathematics programs. While there was no relationship of mathematics curricula to mean achievement, type of mathematics curricula was related to achievement growth.

These patterns were replicated in Study II. Significant relationships were found between school context and composition variables like ethnicity and poverty (FRPL) and student achievement levels. These findings are consistent with other studies of school effectiveness. Measures of school context tend to be strong predictors of average levels of achievement (e.g., Hauser, et al., 1976; Stevens, 2000; Stone and Lane, 2003; Willms, 1986). However, there have been few studies that have examined associations between school context and composition and student rates of growth. In the few studies that have reported on relationships between measures of school context and student rates of growth, little association between the economic context of schools and student growth rates has been found (Stevens, et al., 2000; Stone and Lane, 2003). Study II results also showed that several covariates that were significant predictors of mean achievement (ethnicity, poverty, special education, modified test administration) were not related to rates of growth while mobility, LEP, participation in Title I, and gender were related to growth rates. These results further underscore the differential sensitivity of status and growth to aspects of context, policy, and practice.

Although results of these studies are tentative, we believe they are consistent with an analysis of the research design characteristics of NCLB methods versus longitudinal growth models. Longitudinal research has been recognized as the “sine qua non of evaluation in nonexperimental

settings" (Marco, 1974). Tracking the achievement trajectory of individual students enables more precise estimation of school performance (Goldstein, 1997; Linn and Haug, 2002), enables evaluation of school mean achievement and school mean growth (Zvoch and Stevens, 2003), and serves as a robust means for assessing school policy impacts on changes in student achievement (Boyle and Willms, 2001). Furthermore, learning, the fundamental outcome of interest for the study and evaluation of school effectiveness, is a problem in the analysis of change that can best be addressed with longitudinal designs. One of the most attractive features of true longitudinal designs is the provision of some degree of control over stable characteristics of students. When true longitudinal designs are used, students serve as their own controls.

Another advantage in the application of growth models is that there are at least two parameters of substance: initial level of performance (intercept) and the rate of change (slope). Either parameter may be of interest to analysts, educators, and policy makers and the two parameters may interact with each other (Seltzer, Choi, and Thum, 2003). This is a crucial distinction since even an effective school may not be able to exert limitless influence over the absolute level of the child's functioning but can influence the learning of the child. In addition, taking the child's initial status into account may be important both pedagogically and in terms of evaluating growth or progress for accountability purposes.

While longitudinal designs provide great promise, they also have potential disadvantages. One of the most common concerns in the application of longitudinal models is the attrition of students that occurs over time. This is an area that deserves substantial additional attention since differential attrition may change school results substantially (see Zvoch and Stevens, in press). In Study II, there were significant differences in the characteristics of students between the student differences sample and the school differences sample that were not missing at random. It is probable that such sample differences introduce bias. However, there are almost no studies that examine the impact of these differences in school effectiveness research and many studies do not clearly report details on exclusions and missing data. While attrition is likely to be most severe the longer the longitudinal term of the study, these issues also apply to NCLB methods and value-added models that use one point in time or that use gain scores over two measurement occasions. It is also logical to assume that the source of the exclusion or attrition can determine the

direction of bias that results. For example, exclusion of special education students who often score at lower levels than other students is likely to produce an upward bias in the estimation of school effects. However, attrition that occurs from missing data on mobile students who never have the opportunity to learn at a particular school may actually result in more accurate estimation of school effects. Bias may also result from purposeful exclusions or attempts to influence the accountability process (e.g., Schemo and Fessenden, 2003).

Attention to additional research design issues can improve the quality of many approaches to the evaluation of school effectiveness including a variety of value-added methods. One such issue is the over-reliance on single outcome measures of student performance. Another weakness of much school effectiveness research is the failure to explicitly model "treatment variables". Evaluation of the success or failure of school policy and practice is flawed if direct measurement and monitoring of the delivery of instructional programs, interventions, and instruction itself is never included in analysis. Application of a wider array of methods and measures for control of extraneous influences is also worthy of pursuit.

One of our primary purposes in this chapter was to emphasize the idea that different research designs, measures, and methods of analysis and estimation are likely to provide different evaluations of school effectiveness and different ratings of schools in an accountability system. It is unlikely that any single method will be entirely successful as "the" method for evaluation of school effectiveness in all situations. The choice of the "best" design and method depends on the evaluative purpose being addressed. For example, models that are optimal for evaluating teacher effects may not be best for evaluating school effects—models that are optimal for assessing student mastery of a content standard may not work best to estimate student learning more generally. Part of the challenge in developing effective and valid methods is the empirical assessment of which methods and designs work best for particular accountability purposes and settings.

Notes

Supported in part by National Science Foundation grant NSF REC 0231774.

References

Baker, E. L., and Linn, R. L. (1996). Title I assessment in a new era. *The CRESST Line* (Spring, 1996), 1, 11.

Ballou, D., Sanders, W., and Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-66.

Barton, P., and Coley, R. (1998). *Growth in school: Achievement gains from the fourth to the eighth grade.* Princeton, NJ: Educational Testing Service.

Boyle, M. H., and Willms, J. D. (2001). Multilevel modeling of hierarchical data in developmental studies. *Journal of Child Psychology and Psychiatry, 42*, 141-162.

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

Bryk, A. S., and Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 159-204). San Diego: CA: Academic Press.

Campbell, D. T., and Stanley, J. C. (1963*). Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Coleman, J. S., Campbell, E. Q, Hobson, C. F., McPartland, J., Mood, A. H., Weinfield, R. D., and York, R. L. (1966). *Equality and educational opportunity*. Washington, DC: U.S. Government Printing Office.

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

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

Cook, T. D., and Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

CTB/McGraw-Hill (1997). *TerraNova Technical Bulletin 1*. Monterey, CA: Author.

Duncan, S. C., and Duncan, T. E. (1995). Modeling the processes of development via latent variable growth curve methodology. *Structural Equation Modeling: A Multidisciplinary Journal, 2*(3), 187-213.

Duncan, T. E., Duncan, S. C., Strycker, L. A., Li, F., and Alpert, A. (1999). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, NJ: Erlbaum.

Ferrer, E., Hamagami, F., and McArdle, J. J. (2004). Teacher's corner: Modeling latent growth curves with incomplete data using different types of structural equation modeling and multilevel software. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 452-483.

Goldstein, H. (1991). Better ways to compare schools? *Journal of Educational Statistics, 16*(2), 89-92.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8*, 369-95.

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

Gray, J., Jesson, D., Goldstein, H., Hedger, K., and Rasbash, J. (1995). A Multilevel Analysis of School Improvement: Changes in Schools' Performance over Time. *School Effectiveness and School Improvement, 6*(2), 97-114.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24,* 1141-1177.

Hanushek, E. A. (1996). A more complete picture of school resource policies. *Review of Educational Research, 66*, 397-409.

Hauser, R. M., Sewell, W. H., and Alwin, D. F. (1976). High school effects on achievement. In W. H. Sewell, R. M. Hauser, and D. L. Featherman (Eds.), *Schooling and achievement in American society* (pp. 309-341). New York: Academic Press.

Jencks, C. S., Smith, M., Ackland, H., Bane, M. J., Cohen, D., Ginter, H., Heyns, B., and Michelson, S. (1972). *Inequality: A reassessment of the effect of the family and schooling in America*. New York: Basic Books.

Linn, R. L., and Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24* (1), 29-36.

Little, T. D., Schnabel, K. U., and Baumert, J. (2000). *Modeling longitudinal and multilevel data: Practical issues, applied approaches and specific examples*. Mahwah, NJ: Erlbaum.

MacCallum, R. C., Kim, C., Malarkey, W. B., and Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research, 32*(3), 215-253.

Marco, G. L. (1974). A comparison of selected school effectiveness measures based on longitudinal data. *Journal of Educational Measurement, 11*, 225-234.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd Ed., pp. 13-103). New York: MacMillan.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett and W. C. Ward

(Eds.), *Construction versus choice in cognitive measurement* (pp. 61-74). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13-23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Muthèn, B. O., and Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2*(4), 371-402.

Nesselroade (1991). Interindividual differences in intraindividual change. In L. M. Collins and J. L. Horn (Eds.). *Best methods for the analysis of change* (pp.92-105). Washington, DC: American Psychological Association.

No Child Left Behind Act of 2001, Pub. L. No. 107-110 (2002).

Pedhazur, E. J., and Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated*

approach. Hillsdale, NJ: Erlbaum.

Plewis, I. (1996). Statistical methods for understanding cognitive growth: A review, a synthesis and an application. *British Journal of Mathematical and Statistical Psychology, 49*, 25-42.

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

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., and Congdon, R. T. (2001). *HLM 5: Hierarchical linear and nonlinear modeling*. Chicago: Scientific Software International.

Raudenbush, S. W., and Willms, J.D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*, 307-335.

Reichardt, C. S. (2000). A typology of strategies for ruling out threats to validity. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (Vol. 2, pp. 89-115). Thousand Oaks, CA: Sage.

Reynolds, and Teddlie (2000). The processes of school effectiveness. In C. Teddlie and D. Reynolds (Eds.). *The international handbook of school effectiveness research* (pp. 134-159). New York: Falmer Press.

Reys, R., Reys, B., Lapan, R., Holliday, G., and Wasman, D. (2003). Assessing the impact of standards-based middle grades mathematics curriculum material

on student achievement. *Journal for Research in Mathematics Education, 34(2),* 74-95.

Rindskopf, D. (2000). Plausible rival hypotheses in measurement, design, and scientific theory. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (Vol. 2, pp. 1-12). Thousand Oaks, CA: Sage.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2002). Teachers, schools, and academic achievement. National Bureau of Economic Research: Working Paper No. 6691.

Rogosa, D. (1995). Myths about longitudinal research. In J. M. Gottman (Ed.). *The analysis of change*. Mahwah, NJ: Erlbaum.

Rowan, B., Correnti, R., and Miller, R. J. (2002). What large scale, survey research tells us about teacher effects on student achievement: Insights from the *Prospects* study of elementary schools. *Teachers College Record, 104*, 1525-1567.

Schafer, W. D. (2001). Replication: A design principle for field research. *Practical Assessment, Research and Evaluation, 7*(15). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=15.

Schemo, D. J., and Fessenden, F. (2003, December 3). *Gains in Houston schools: How real are they?* New York Times.

Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing, and equity. *Educational Researcher, 31*(1), 13-25.

Seltzer, M., Choi, K., and Thum, Y. M. (2003). Examining relationships between where students start and how rapidly they progress: Using new developments in growth modeling to gain insight into the distribution of achievement within schools. *Educational Evaluation and Policy Analysis, 25*, 263-286.

Senk, S. L., and Thompson, D. R. (2003). *Standards-based school mathematics curricula: What are they? What do students learn?* Mahwah, NJ: Lawrence Erlbaum Associates.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Stevens, J. J. (2001). *Confirmatory Factor Analysis of the CTBS5/TerraNova*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Stevens, J. J., Estrada, S., and Parkes, J. (2000). *Measurement issues in the design of state accountability systems*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Stone, C. A., and Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16*, 1-26.

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

Teddlie, C., Reynolds, D., and Sammons, P. (2000). The methodology and scientific properties of school effectiveness research. In C. Teddlie, and D. Reynolds (Eds.), *The International handbook of school effectiveness research*. New York: Falmer Press.

What Works Clearinghouse. (n.d.). *WWC study review standards*. Retrieved December 20, 2004 from http://www.whatworks.ed.gov/reviewprocess/study_standards_final.pdf

Willett, J. B., and Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time, *Psychological Bulletin, 116(2)*, 363-381.

Willett, J. B., Singer, J. D., and 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.

Willms, J. D. (1986). Social class segregation and its relationship to pupils' examination results in Scotland. *American Sociological Review, 51*, 224-241.

Willms, J. D., and Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability, *Journal of Educational Measurement, 26*(3), 209-32.

Wright, S. W., Horn, S. P., and Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11,* 57-67.

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

Zvoch, K., and Stevens, J. J. (in press). Sample Exclusion and Student Attrition Effects in the Longitudinal Study of Middle School Mathematics Performance. *Educational Assessment*.

Zvoch, K., and Stevens, J. J. (2004). Longitudinal Effects of School Context and Practice on Mathematics Achievement. Manuscript submitted for publication.

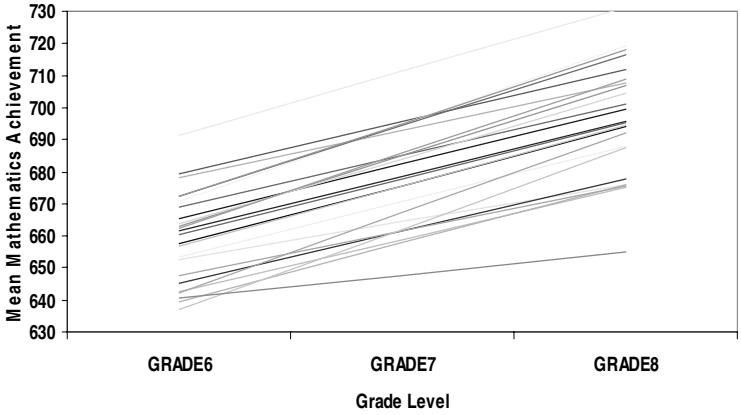

Figure 1. School mean linear growth trajectories

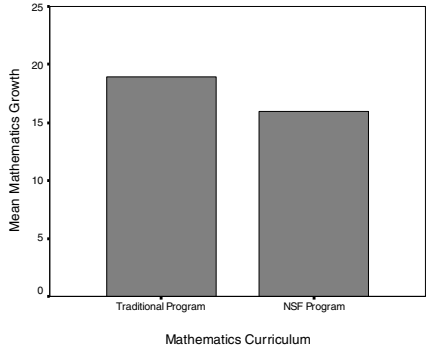

Figure 2. Mean mathematics growth by curriculum

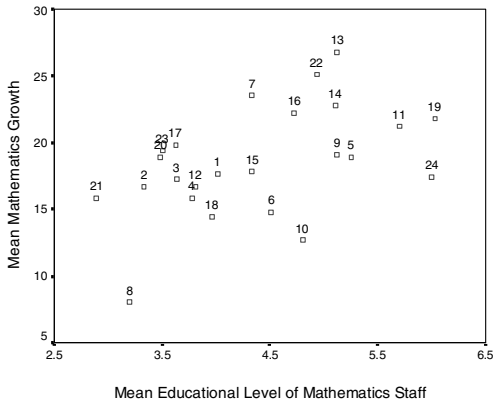

Figure 3. Mean mathematics growth as a function of school mean educational level of the mathematics staff