EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Volume 11 Number 20 July 8, 2003 ISSN 1068-2341

A Multilevel, Longitudinal Analysis of Middle School Math and Language Achievement

Keith Zvoch
Albuquerque (NM) Public Schools

Joseph J. Stevens
University of New Mexico

Citation: Zvoch, K. and Stevens, J. J. (July 8, 2003). A multilevel, longitudinal analysis of middle school math and language achievement. *Education Policy Analysis Archives, 11*(20). Retrieved [date] from http://epaa.asu.edu/epaa/v11n20/.

Abstract

The performance of schools in a large urban school district was examined using achievement data from a longitudinally matched cohort of middle school students. Schools were evaluated in terms of the mean achievement and mean growth of students in mathematics and language arts. Application of multilevel, longitudinal models to student achievement data revealed that 1) school performance varied across both outcome measures in both subject areas, 2) significant proportions of variation were associated with school-to-school differences in performance, 3) evaluations of school performance differed depending on whether school mean achievement or school mean growth in achievement was examined, and 4) school mean achievement was a weak predictor of school mean growth. These

results suggest that assessments of school performance depend on choices of how data are modeled and analyzed. In particular, the present study indicates that schools with low mean scores are not always “poor performing” schools. Use of student growth rates to evaluate school performance enables schools that would otherwise be deemed low performing to demonstrate positive effects on student achievement. Implications for state accountability systems are discussed.

With the enactment of the No Child Left Behind (NCLB; No Child Left Behind Act, 2002) legislation, states are now required to develop content-based standards in mathematics and reading or language arts and have tests that are linked to those standards in grades 3 through 8. The new legislation also requires states to set a proficiency standard for performance on those content-aligned tests. The proficiency standard will enable states to identify “probationary” schools, monitor their performance, and intervene if adequate yearly progress toward the standard does not occur. The increased emphasis that NCLB brings to the assessment of state content standards and the measurement of school effectiveness is intended to ensure that all students have access to an equitable and comprehensive education. However, the assessment of student performance and the measurement of school effectiveness are neither simple nor straightforward (Linn, 2000; Stevens, 2000). There are many complex issues involved in the development and implementation of accountability systems that are not acknowledged or considered in the fervor of political and public dialogue and policy discussion on educational reform. One issue of substantial importance is how the analytic methods used in an accountability system may impact the evaluation of school effectiveness. At its heart, the measurement of student learning and school effectiveness poses some challenges in research design that must be met if the effects of teachers and schools are to be validly estimated.

One of the most difficult challenges in evaluating school performance is to separate the effects of schooling from the intake characteristics of the students who attend the school (Raudenbush & Willms, 1995; Willms, 1992). The evaluative challenge stems from the manner in which students come to attend particular schools. Nonrandom selection processes sort families into neighborhoods and students into schools. The unequal distribution of student characteristics that follows tends to give schools with challenging intakes a competitive disadvantage in most accountability systems. Schools with disadvantaged intakes are at particular risk of unfavorable evaluation if the state accountability system fails to use statistical methods that properly account for student background and the hierarchical nature of school accountability data. State accountability models that use school means or medians as the primary or only indicator of school effectiveness are particularly problematic. Common practice is to aggregate student data to the level of the school in these models. However, if relevant data occur at different levels or for different sampling "units" as in the measurement of both students and schools, then aggregating data to the school level may be inappropriate. This issue is often referred to as a "unit of analysis" problem. In statistical terminology, students are nested within schools and analysis should incorporate the nested structure of the data into the design through the use of multilevel analysis methods (see Aitken & Longford, 1986; Brown, 1994; Burstein, 1980; Cronbach, 1976; Cronbach & Webb, 1975; Goldstein, 1988; Raudenbush & Willms, 1995). Yet, few state systems

appear to use multilevel methods. Accountability systems that fail to properly model the nested structure of data tend to confound school intake with school practice and policy and are probably biased in their estimation of school effects.

Another issue of importance in the measurement of school performance is the over-reliance in accountability systems on the comparison of successive cohorts of students as a measure of "change". States that use the successive cohort approach (e.g., the mean performance of 6th graders in 2001 is compared to the mean performance of 6th graders in 2002) to measure and evaluate school performance attempt to mitigate school differences in student intake by focusing on the year-to-year change in student achievement scores. The comparison of successive student cohorts enables states to evaluate schools in terms of proficiency gains instead of absolute performance levels. However, the use of different cohorts of students to measure school progress or school improvement is problematic for evaluative purposes. Recent investigations of the successive cohort approach demonstrate that estimates of year-to-year gains in proficiency are affected in large part by sampling variation, measurement error, and unique, non-persistent factors that are not associated with school size or school practice (Linn & Haug, 2002; Kane & Staiger, 2002). The lack of systematic variation in the successive cohort change score puts states at risk of assessing school performance on the basis of fluctuations in student cohorts or test administration conditions instead of actual changes in student performance (Linn & Haug, 2002).

Evidence that school performance cannot be estimated without bias when student test scores are aggregated at a single point in time or with precision when successive student cohorts are compared has led a number of authors to argue for the use of longitudinal analyses of individual student performance as a more direct and accurate estimate of school effects (Barton & Coley, 1998; Bryk & Raudenbush, 1988; Linn & Haug, 2002). For example, Goldstein (1991, p. 14), 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." Student progress can be measured by comparing year-to-year differences in individual performance, but the most appropriate methodology for measuring changes in student achievement is through estimation of individual growth trajectories by means of the multilevel model (Bryk & Raudenbush, 1992, 1987; Willett, 1988; Willms, 1992). In this approach, student test scores are linked across time. A regression function is then fit to the outcome data obtained on each student. The resulting growth trajectories index the rate at which students acquire certain academic competencies. A measure of school performance follows from averaging the individual growth trajectories within each school.

Multilevel, longitudinal analyses of individual student performance may allow the conceptualization of the most relevant and direct outcome measures of school effectiveness by facilitating estimation of the added benefit or "value" that students receive by attending a particular school (Boyle & Willms, 2001; Bryk & Raudenbush, 1988; Willms, 1992). The multilevel, longitudinal model facilitates value-added school performance estimates by providing a degree of control over a wealth of confounding factors that otherwise complicate the evaluation of school effectiveness. When longitudinal models are used, each student serves as his/her

own control for confounding factors that are stable characteristics of the student over time (Sanders & Horn, 1994; Stevens, 2000). Therefore, the confounding effects of factors like socio-economic status, limited English proficiency, and ethnic and cultural differences may be largely controlled through the application of a matched longitudinal design.

Despite the potential of using multilevel, longitudinal models to measure and evaluate school performance (Boyle & Willms, 2001; Teddlie & Reynolds, 2000), only a few reported studies have applied these models to student achievement data (e.g., Bryk & Raudenbush, 1988; Willms & Jacobsen, 1990). Given the lack of published examples, the purpose of the present study was to provide a demonstration of the use of multilevel, longitudinal models to estimate school effectiveness using a sample of middle school students. We were interested in examining the following research questions: 1) How does student achievement performance vary over a three year period? 2) How much of the variation in performance is associated with individual differences among the students and how much with differences from school to school? and, 3) How does the evaluation of school performance differ based on an examination of school mean achievement vs. an examination of school average rates of growth in achievement?

Method

Participants

Standardized test data from middle school students in a large urban school district located in the southwestern United States were analyzed in the present study. The school district that provided the data has over 100 schools and serves close to 90,000 students annually. The district has a diverse student body. In recent years, the student population has been approximately 46% Hispanic, 44% Anglo, 4% Native American, 3% African American, 2% Asian and 1% other. The district serves many students who are not fully English proficient. On average, twenty percent of the students who attend district schools are classified as Limited English Proficient (LEP). The district is also impacted by widespread poverty. In any given year, approximately 35% of the district's middle school students receive a free or a reduced price lunch.

At the middle school level, the school district has 24 schools that serve over 20,000 students in grades 6 through 8. All sixth, seventh, and eighth grade students are tested annually on a norm-referenced achievement test, the TerraNova/CTBS5 Survey Plus (CTB/McGraw-Hill, 1997). Approximately 6,500 students in each grade take the test each spring. 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 in the present study. Middle school students were used because they provided the only cohort on which three consecutive years of data were available. Mathematics and Language scores were used to provide a demonstration using core subject areas and those content areas required in the NCLB legislation. All students who completed an examination in all three study years were selected (N = 4,918). Since the purpose of the study was to examine school effects, 800 students who did not attend the same middle school in all three years were excluded resulting in a sample of 4,118 students. Any student who did not have a mathematics or

language composite score in all three study years was excluded as well as any student who received a modified test administration in any of the three years. This resulted in a final sample of 3,299 students nested within 24 middle schools.

The sample consisted of almost equal numbers of males and females. Fifty-one percent of the sample were female (*N* = 1,698); forty-nine percent were male (*N* = 1,601). Representation of ethnic groups was more variable. Forty-six percent (*N* = 1,524) of the sample were Anglo, 45% (*N* = 1,495) were Hispanic, 3% (*N* = 87) were African American, 3% (*N* = 86) were Native American, and 2% (N = 67) were of Asian descent. The ethnic background of 40 students was not identified. Thirty-five percent (*N* = 1,152) of the sample received a free or a reduced price lunch, 12% (*N* = 390) were classified as LEP, and 3% (*N* = 98) were special education students. In most respects, the backgrounds of students in the analytic sample were representative of the students who attend district middle schools. However, the exclusion of students who did not participate in all three test administrations, students who transferred schools, and students who received at least one modified test administration did lower the percentage of free lunch recipients and the percentage of LEP and special education students below district averages by 1%, 8%, and 18%, respectively. Nonetheless, the disproportionate exclusion of students from special populations did not affect the pattern of school mean achievement. Correlations between school mean achievement in grade 6 for the original and the analytic sample were .98 for mathematics and .99 for language.

Measures

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 Survey Plus is a test battery that spans grades 2 through 12. All test items are selected-response. The Survey Plus tests students in Reading, Language Arts, Mathematics, Science, Social Studies, Word Analysis, Vocabulary, Language Mechanics, Spelling, and Mathematics Computation. CTB/McGraw-Hill calculates an IRT derived score for each student in each subject area. CTB/McGraw Hill also provides each student with a weighted composite score in Reading, Mathematics, and Language.

Student scale scores on the Mathematics and Language composites were analyzed in the present study. The mathematics composite score is derived from the 31-item Mathematics and the 20-item Mathematics Computation subtests. According to the publisher, the Mathematics subtest measures a student's ability to apply grade appropriate mathematical concepts and procedures to a range of problem-solving situations. The Mathematics Computation subtest measures a student's ability to perform arithmetic operations on grade appropriate number types. CTB/McGraw-Hill reported a KR-20 reliability estimate of .86 for the Mathematics subtest in the 6th, 7th, and 8th grade norming samples. For Mathematics Computation, KR-20 was .83 in grade 6, .80 in grade 7, and .85 in grade 8. For the Mathematics composite, KR-20 was reported at .91 in grade 6, .90 in grade 7, and .92 in grade 8 (CTB/McGraw-Hill, 1997).

The Language composite was also derived from a weighted combination of two subtests. The 22-item Language subtest is intended to measure a student's ability to understand the structure and usage of words in standard written English. The

20-item Language Mechanics subtest is designed to measure a student's ability to edit and proofread standard written English. CTB/McGraw-Hill reported KR-20 reliability estimates of .86 in grade 6, .84 in grade 7, and .81 in the grade 8 norm groups. For Language Mechanics, KR-20 was reported as .84 in grades 6 and 7 and .85 in grade 8. For the Language composite, KR-20 was .91 in grades 6 and 7 and .90 in grade 8 (CTB/McGraw-Hill, 1997).

Analytic Procedures

Multilevel modeling techniques were used to estimate a mean achievement score and a mean growth trajectory for each school. The Hierarchical Linear Modeling (HLM) program, version 5.04 (Raudenbush, Bryk, Cheong, & Congdon, 2001) was used to estimate three-level longitudinal models. Level-1 was composed of a longitudinal growth model that fitted a linear regression function to each individual student's achievement scores over the three years studied (grades 6, 7, and 8). Equation 1 specifies the level-1 model:

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

As written, γ_{tij} is the outcome (i.e., mathematics or language achievement) at time t for student i in school j, π_{0ij} is the initial status of student ij (i.e., 6th grade performance), 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. Levels 2 and 3 in the HLM model estimate mean growth trajectories in terms of both initial status and growth rate across students (equations 2a and 2b) and across schools (equations 3a and 3b):

$$(2a) \qquad \pi_{0ij} = \beta_{00j} + r_{0ij}$$

$$(2b) \qquad \pi_{1ij} = \beta_{10j} + r_{1ij}$$

$$(3a) \qquad \beta_{00j} = \gamma_{000} + u_{00j}$$

$$(3b) \qquad \beta_{10j} = \gamma_{100} + u_{10j}$$

The initial status and growth of student achievement in equations 2a and 2b is conceived as a function of the school average achievement or school average slope and residual. Similarly, the initial status and growth by school in equations 3a and 3b is conceived as a function of the grand mean achievement or the grand mean slope and residual. Equations 3a and 3b were used to calculate estimates of school mean achievement and school mean growth reported in the present study.

Results

Model Assumptions

Visual examination of univariate frequency distributions and a check of summary statistics revealed that mathematics and language achievement scores were distributed normally (i.e., skew and kurtosis values < 1) in all three study years. A check of within-subject bivariate plots revealed linear relationships between

achievement scores across the three study years in both mathematics and language. After checking model assumptions, three SPSS data files were transferred to the HLM program for analysis. The Level-1 data file contained student and school identifiers, three years of student mathematics and language composite scale scores, and a field for year. This file contained 9,897 records (i.e., three records for each of 3,299 students). The Level-2 data file contained student and school identifiers (N = 3,299). The Level-3 data file contained only a school identifier (N = 24).

Mathematics

Table 1 presents the results of the three-level HLM model for mathematics. In the upper portion of the table, the results of the fixed effects regression model are presented. The first estimate shown, the grand mean (γ_{000}), is the intercept or the average 6th grade mathematics scale score for all students in the sample. The second estimate, the grand slope (γ_{100}), is the average yearly growth rate of the students. Thus, in this sample, the average mathematics score is estimated as 659.43 and on average, student mathematics achievement is expected to increase by 18.40 scale score points per year.

Table 1
Three-Level Unconditional Model for Mathematics Achievement

Fixed Effect	*Coefficient*	*SE*	*t*
School Mean Achievement, γ_{000}	659.43	2.97	222.20*
School Mean Growth, γ_{100}	18.40	0.93	19.86*
Random Effect	*Variance Component*	*df*	*Chi-square*
Individual Achievement, r_{0ij}	766.00	3275	8828.40*
Individual Growth, r_{1ij}	26.66	3275	3791.30*
Level-1 Error, e_{tij}	310.87		
School Mean Achievement, u_{00j}	203.03	23	704.70*
School Mean Growth, u_{10j}	19.11	23	359.31*
Level-1 Coefficient		*Percentage of Variation Between Schools*	

Individual Achievement, π_{0ij}	21.0
Individual Growth, π_{1ij}	41.8

Note. Results based on data from 3,299 students distributed across 24 middle schools.

** p < .001*

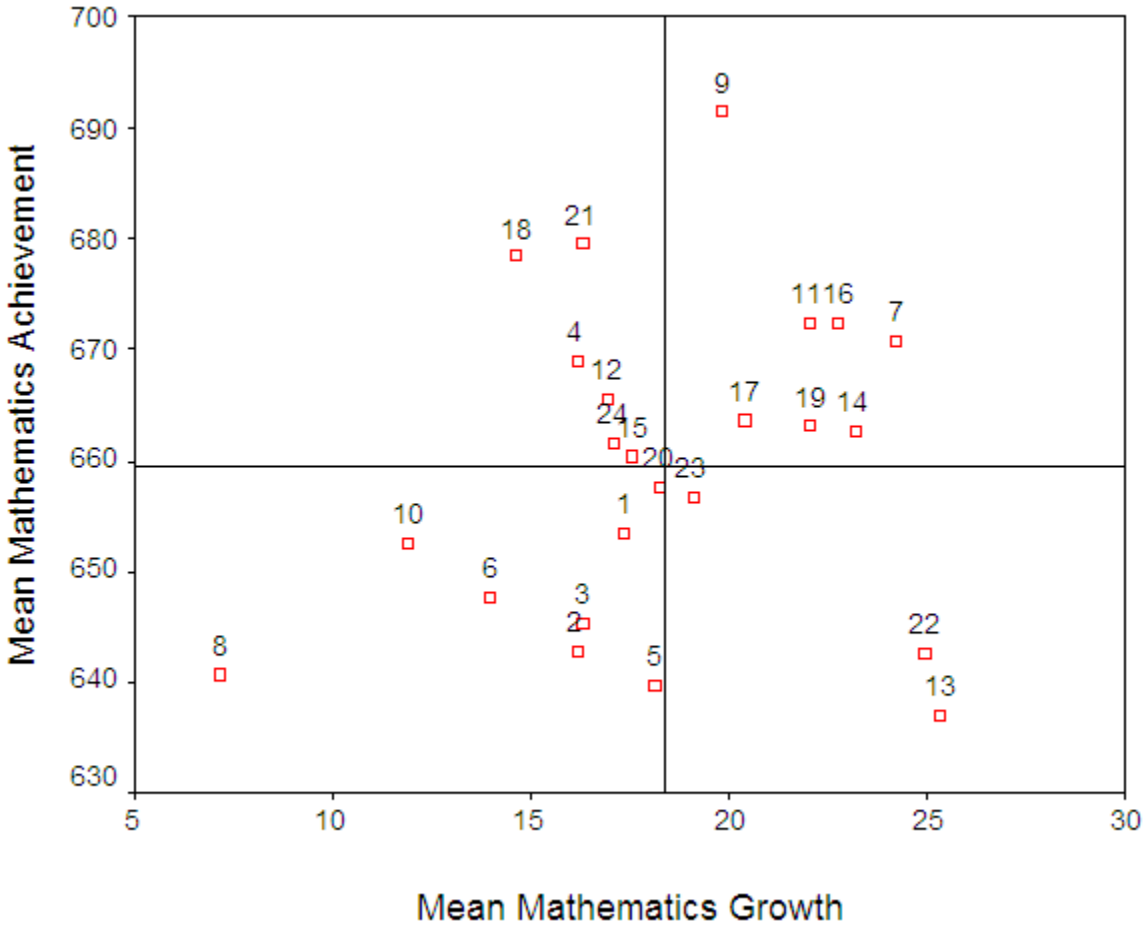

Figure 1. Relationship between mathematics achievement and mathematics growth by school.

Estimates of student and school-level parameter variance are presented next in Table 1. Chi-square tests of the hypotheses that students and schools differ in level and growth of mathematics achievement indicate that there was statistically significant variation across all parameters. Both students and schools differ significantly in initial achievement levels and the rate of achievement growth. This indicates that there are individual differences from one student to another in mathematics achievement initially in grade 6 as well as in the rate of growth in mathematics achievement throughout middle school. In addition, inspection of the variance components presented at the bottom of Table 1 show that the amount of between school variance in mathematics mean achievement (21.0%) and mean achievement growth (41.8%) is also relatively large and statistically significant.

Thus, over and above individual differences, there are systematic differences from one school to another in mean mathematics achievement initially in grade 6 and in the school average rate of growth in mathematics achievement from the 6th to the 8th grade.

To further illustrate these observed school level differences in mathematics achievement, Empirical Bayes (EB) estimates of the 24 middle-school mathematics mean achievement and mean growth rates are presented in the scatterplot in Figure 1 on the vertical and horizontal axes, respectively. The horizontal line in the interior of the figure represents the grand mean achievement in mathematics. The vertical line in the interior of the figure represents the grand mean growth in mathematics. The two grand mean reference lines are used to classify schools into four quadrants of school performance. The upper right quadrant contains schools with above average mean achievement in grade 6 and above average growth from grades 6 to 8. The lower right quadrant contains schools with below average mean scores, but above average growth. The two quadrants on the left side of the figure contain schools with below average growth and either high or low mean achievement. A number of interesting results can be seen in Figure 1. First, two schools (22 and 13) with low mean scores record the highest growth in the district. Strong growth is also evident in high scoring school 7. Also evident in Figure 1 is a school (8) with a low mean score and very poor mathematics growth. Relatively poor mathematics growth also occurs in two schools with high 6th grade mathematics achievement. Schools 21 and 18, second and third in 6th grade mean achievement, are noticeably below the district average in mathematics growth. Overall, a slight positive relationship exists between mathematics mean achievement and mathematics mean growth ($t_b = .14$). On average, schools with low mean scores record less growth than schools with high mean scores. Appendix A presents the individual school mean and school growth estimates on which Figure 1 is based.

Figure 2 displays the school estimates in growth trajectory form. Each line in Figure 2 shows the average mathematics achievement at one of the 24 middle schools. As can be seen, there is a good deal of variation from school-to-school both in initial status (i.e., grade 6 mean achievement) and in the average rates of growth over time. The variation in average mathematics growth rates is indicated in part by the number of crossing lines in the figure. Alternative line styles are used to highlight schools with exceptionally high or low growth rates. Schools with a high growth rate are represented by the broken dot pattern. Schools with a low growth rate are represented by the broken line pattern.

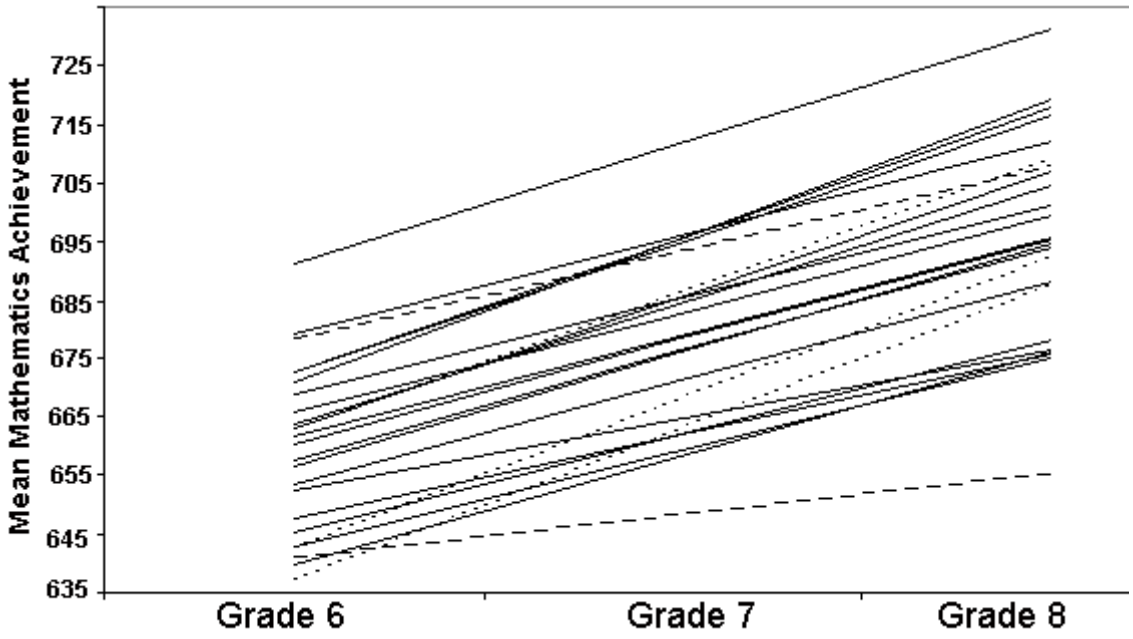

Figure 2. Mean mathematics achievement as a function of grade level and school location.

In Figure 2, the strong growth of two of the schools with low 6th grade achievement levels can be clearly seen. The school with the lowest 6th grade mean score (24th in rank) shows average achievement growth 1.4 times the overall average for mathematics growth. By 8th grade, the rank of the school has changed from 24th to 18th in average mathematics achievement. Similarly, the school that ranks 21st in mean performance in 6th grade, also shows average achievement growth 1.4 times the average and moves to a rank of 16th by the end of 8th grade. Strong growth is also apparent in some of the schools with high 6th grade mean scores. For example, the 10th ranked school in 6th grade mathematics achievement becomes the 6th ranked school in 8th grade mathematics achievement. Schools with lower than average mathematics growth are also readily apparent in Figure 2. The third ranked school in 6th grade mathematics performance falls to seventh ranked in 8th grade performance. In addition, the 22nd ranked school in 6th grade achievement not only becomes the lowest ranked school by the end of 8th grade, but by achieving at only 39% of the overall average for mathematics growth, also falls far behind the achievement level of all other middle schools in the district.

Language

The same three-level, longitudinal HLM model was applied to the language achievement scores of the same sample of students. Table 2 presents these results. As can be seen in Table 2, model results for language achievement were similar to those for mathematics achievement. Except for variation in student growth rates, all parameters of the three-level HLM model were statistically significant. The average language achievement for all students across the 24 middle schools was 661.43 in grade 6 and the average yearly growth in language achievement was 12.30 score points. Inspection of the variance components from the language model shows that while there is statistically significant individual

variation in students' initial language achievement in grade 6, individual language growth rates do not differ statistically. Table 2 also shows that school growth rates in language are less variable than school growth rates in mathematics. Of the variation that does exist in language growth rates, 84% was between school variance. Thus, in the case of language achievement, students differed in their average language achievement at grade 6 but showed rates of growth in language achievement that did not differ significantly. At the school level, there were statistically significant differences in average achievement at grade 6 and in average rates of growth in language achievement through grade 8.

Table 2
Three-Level Unconditional Model for Language Achievement

Fixed Effect	*Coefficient*	*SE*	*t*
School Mean Achievement, γ_{000}	661.43	2.58	256.55*
School Mean Growth, γ_{100}	12.30	0.45	27.44*
Random Effect	*Variance Component*	*df*	*Chi-square*
Individual Achievement, r_{0ij}	699.35	3275	8836.36*
Individual Growth, r_{1ij}	0.68	3275	3226.22+
Level-1 Error, e_{tij}	332.11		
School Mean Achievement, u_{00j}	151.66	23	588.34*
School Mean Growth, u_{10j}	3.51	23	91.84*
Level-1 Coefficient		*Percentage of Variation Between Schools*	
Individual Achievement, π_{0ij}		17.8	
Individual Growth, π_{1ij}		83.8	

Note. Results based on data from 3,299 students distributed across 24 middle schools.

* $p < .001$

Empirical Bayes estimates of the 24 middle-school language means and growth

rates are displayed in Figure 3. Instances of all four patterns of achievement described for the mathematics achievement results are also present in Figure 3. School 22 demonstrates high growth relative to its 6th grade mean language achievement while growth is low for school 8 in language as it was in mathematics. As with the mathematics results, school 18 again demonstrates low growth relative to 6th grade mean achievement while school 7 shows a high growth rate in language achievement. Overall, the relationship between language mean achievement and language mean growth is positive (t_b = .41). On average, schools with low language mean scores showed less growth than schools with high language mean scores. School language achievement means and growth rate estimates are presented in Appendix A.

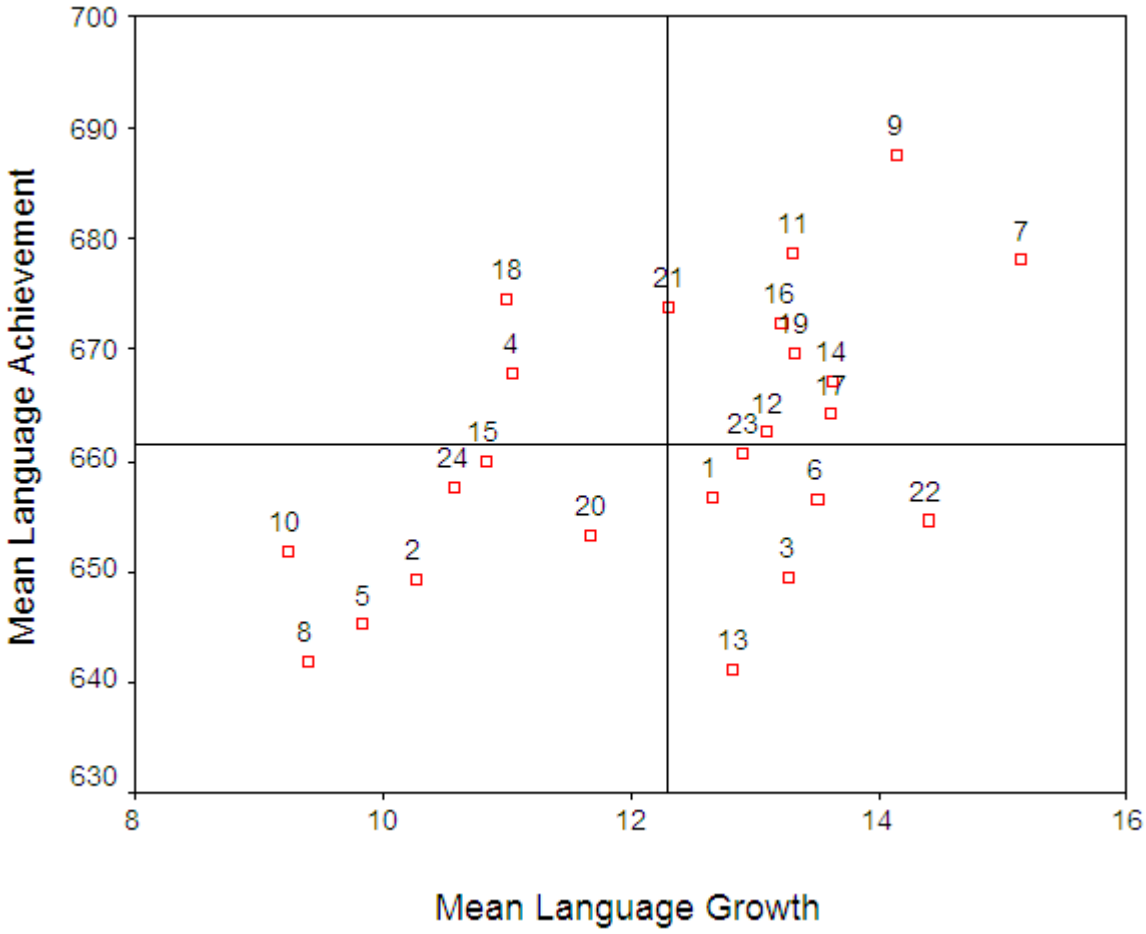

Figure 3. Relationship between language achievement and language growth by school.

Figure 4 displays these results in growth trajectory form. Alternative line styles are again used to represent schools with exceptionally high or low growth rates. The figure shows that, while middle schools differ substantially in mean language achievement at grade 6, the rate of language achievement growth is more similar across schools as evidenced by the parallel pattern of the growth trajectories. Relative to mathematics, fewer schools change rank position. Some possibly important differences in growth rate still exist however. The third ranked school in 6th grade language performance increased its relative standing over other schools in the district. Conversely, the 23rd ranked school in 6th grade language

performance becomes the lowest ranked school by the end of 8th grade.

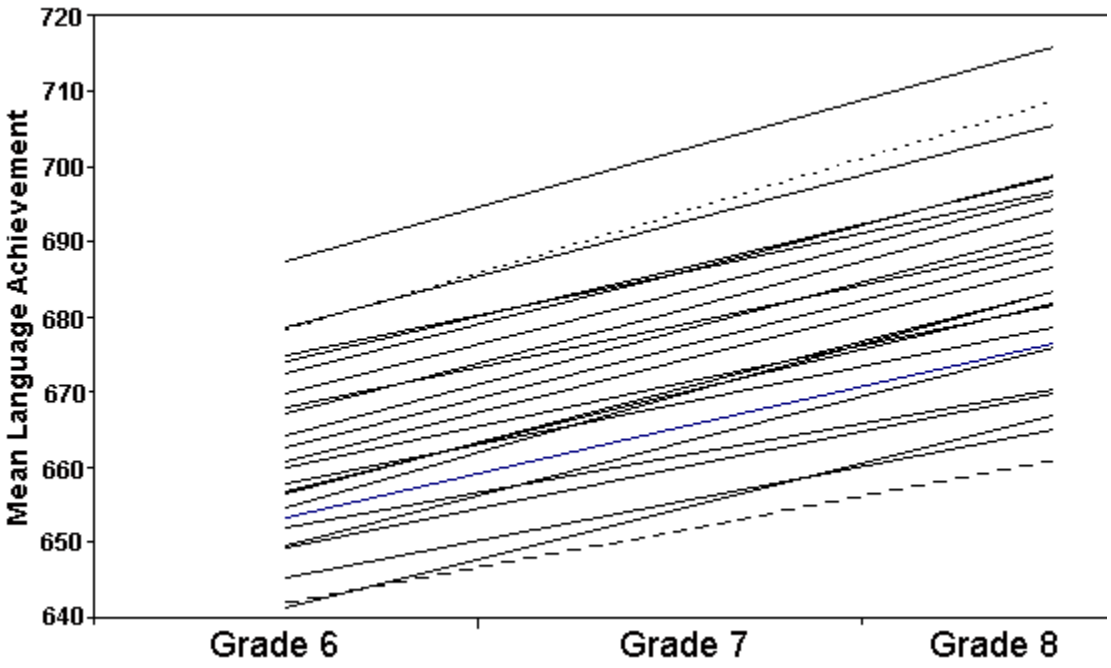

Figure 4. Mean language achievement as a function of grade level and school location.

Discussion

The purpose of the present study was to apply multilevel, longitudinal models in an analysis of school effectiveness in mathematics and language achievement and to demonstrate how assessments of school performance can differ based on how data are modeled and analyzed. The present study demonstrated that assessments of school performance varied across mathematics and language achievement measures. Estimates of the proportions of variance in achievement associated with individual students and with middle schools showed that significant proportions of variation were associated with school-to-school differences in performance. In the current sample, 21 percent of the unadjusted variation in mathematics achievement and 42 percent of the unadjusted variation in mathematics growth were attributable to between-school differences. For language, 18 percent of the unadjusted variation in school achievement means and 84 percent of the unadjusted variation in school growth trajectories were associated with school-to-school differences.

The present study also showed that evaluations of school performance differ depending on whether school mean achievement or school mean growth in achievement are examined. There was significant variation in the mean achievement of students in mathematics and language both from student-to-student and from school-to-school. The analyses also showed that there was significant variation in the rate of achievement growth from student-to-student and from school-to-school for mathematics and from school-to-school for language. Using results from the multilevel, longitudinal models, mean achievement and

mean growth were estimated for each middle school in mathematics and language. Evaluation of these estimates showed that the school mean level of performance was not strongly predictive of the school mean rate of growth. Correlations of school mean and school growth estimates were only .14 for mathematics and .41 for language. Inspection of Figures 1 and 3 showed that characterization of school performance is substantially different depending on whether mean achievement or mean growth is examined. In several cases, schools with low mean scores were not always “poor performing” schools. In fact, schools with low mean scores were in many cases the schools with the largest growth rates. Conversely, a high mean achievement score was not always a clear indicator of “good performance”. In several instances, schools with high mean scores had growth rates below the district average.

The demonstration that school performance can vary on the basis of the analytic model applied suggests that it is essential to use evaluative models that do not unfairly reward or penalize schools for factors that are beyond the control of school personnel (Hanushek & Raymond, 2001; Ladd, 2001). Current evaluative practice often falls short of this goal. The accountability systems now in use in many states apply evaluative methods that cannot separate school level variation from student level variation or validly disentangle school effects from factors that are outside the control of educational policy and practice at the school (Stevens, Estrada & Parkes, 2000). One of the most common approaches in state accountability systems is the use of the school mean as the only or the key component in the evaluation of school effectiveness. As an evaluative measure, the school mean has widespread appeal. School means are easily calculated and are readily understood. However, the school mean is also a biased indicator of school performance (Heck, 2000). School means reflect all influences on student performance, including those exogenous to the school (e.g., family background, prior achievement, community context). As a result, the school mean often provides a misleading picture of school performance. Schools with advantaged intakes tend to be evaluated more favorably than schools with disadvantaged intakes, regardless of the impact the school has on students over time (Stevens et al., 2000).

Another option for assessing school performance is available to those states or districts that collect comprehensive data on student background. School means can be adjusted on the basis on student characteristics, prior achievement levels, and community characteristics in an attempt to arrive at a mean value that isolates the contribution of school practice and policy (Clotfelter & Ladd, 1996; Raudenbush & Willms, 1995; Willms, 1992). However, these data are often difficult for states and districts to adequately and accurately collect and analyze. Adjusted school means also present states and districts with two unwanted concerns. One concern has to do with public response to having a lower standard of performance for certain special student populations. The second stems from the difficulty of having to convey the meaning of complex statistical adjustments to parents, teachers, and school administrators (Clotfelter & Ladd, 1996; Elmore et al., 1996).

An alternative to the adjusted school mean is a measure based on changes in students’ academic achievement over time. As a measure of school performance, school mean growth may offer a more tractable and accurate method of adjusting for socio-demographic characteristics by providing control over confounding influences associated with the stable characteristics of students (Haertel, 1999;

Lane & Stone, 2002; Stevens et al., 2000). Repeated measurement of individual students provides control over the background and intake characteristics that strongly impact the *level* at which a student performs by shifting the measurement process from indexing student performance at a single point in time to tracking the rate of pupil progress over time (Sanders & Horn, 1994). The calculation of student growth rates thus enable schools to be evaluated in terms of the gains students make instead of the level at which students start, thereby enabling a more valid comparison of schools that differ in the intake characteristics of their student bodies.

Despite the promise of using multilevel, longitudinal models to measure school performance, very few states have accountability systems that track individual students over time or use analytic methods that account for the hierarchical structure of accountability data (Council of Chief State School Officers, 2001; Education Week, 2001, 2002). However, the importance of basing an accountability system on an outcome measure that can be impacted by school practice and policy cannot be overstated. If school effectiveness is not evaluated in a way that actually reflects school practices and policies but instead reflects student intake characteristics, the state system for evaluating school performance can become a source for flawed decision-making and a target of criticism and possible litigation by disgruntled stakeholders (Parkes & Stevens, in press). Misguided assessment policy can thus stall attempts at constructive school-based change and effectively undermine the intent of the accountability system.

The present study demonstrated that assessments of school performance vary on the basis of the analytic methods applied to the data. Depending on whether schools were evaluated in terms of mean achievement or mean growth, assessments of school performance were shown to differ dramatically. In some cases, schools with low mean scores had high growth rates and schools with high mean scores had low growth rates. These results suggest that states should not rely on the school mean as the sole indicator of school effectiveness. Instead, consideration should be given to incorporating into school accountability systems measures that track student learning or growth over time. The importance of assessing student growth is further underscored by the amount of variation in growth rates that can be attributed to school-to-school differences. In the present study, school differences in mean growth were two times greater than school differences in mean achievement in mathematics and over four times greater than school differences in mean achievement in language. Identification of large amounts of school level variation in growth rates suggests that schools can have substantial influence on student achievement. Future research that examines the influence of school demographic factors, school climate, and school practice and policy on school growth trajectories will begin to facilitate our understanding of why some schools are more effective at promoting student growth in achievement than others.

References

Aitkin, M. & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A, 149*, 1-43.

Barton, P. & Coley, R. (1998). *Growth in school: Achievement gains from the fourth*

to the eighthgrade. Princeton, NJ: Educational Testing Service.

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

Brown, S. (1994). School effectiveness research and the evaluation of schools. *Evaluation and Research in Education, 8*, 55-68.

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

Bryk, A.S., & 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.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park: SAGE.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In D.C. Berliner (Ed.), *Review of research in education* (vol.8). Washington, DC: American Educational Research Association

Clotfelter, C.T., & Ladd, H.F. (1996). Recognizing and rewarding success in public schools. In H.F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 23-63). Washington, DC: The Brookings Institution.

Council of Chief State School Officers (2001).*Annual Survey: State Student Assessmenti Programs, Vols. 1 and 2 (1999-2000 data).*Washington, D. C.: Council of Chief State School Officers.

Cronbach, L.J. (1976). *Research on classrooms and schools: Formulation of questions, design, andanalysis*. An Occasional Paper, Stanford, CA: Stanford Evaluation Consortium.

Cronbach, L. & Webb, (1975). Between and within class effects in a reported aptitude by treatment interaction: Reanalysis of a study by G.L. Anderson. *Journal of Educational Psychology, 6,* 717-724.

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

Education Week. (2001). *Quality counts 2001: A better balance*. (Volume 20, Number 17, January 11, 2001). Bethesda, MD: Author.

Education Week. (2002). *Quality counts 2002: Building blocks for success*. (Volume 21, Number 17, January 10, 2002). Bethesda, MD: Author.

Elmore, R. F., Abelmann, C.H., & Fuhrman, S.H. (1996). The new accountability in state education reform: From process to performance. In H.F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 65-98). Washington, DC: The Brookings Institution.

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

Goldstein, H.I. (1988). Comparing schools. In H. Torrance (Ed.). *National assessment and testing: A research response*. London: BERA.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18* (4), 5-9.

Hanushek, E.A., & Raymond, M.E. (2001). The confusing world of educational accountability. *National Tax Journal, 54*, 365-384.

Heck, R.H. (2000). Examining the impact of school quality on school outcomes and improvement: A value-added approach. *Educational Administration Quarterly, 36*, 513-552.

Kane, T.J., & Staiger, D.O. (2002). Volatility in school test scores: Implications for test-based accountability systems. *Brookings Papers on Educational Policy, 1*, 235-283.

Ladd, H.F. (2001). School-based educational accountability systems: The promise and the pitfalls. *National Tax Journal, 54*, 385-400.

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21*(1), 23-30.

Linn, R.L. (2000). Assessments and accountability. *Educational Researcher, 29*, 4-16.

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

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

Parkes, J. & Stevens, J. J. (in press). Legal issues in school accountability systems. *Applied Measurement in Education*.

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

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

Sanders, W.L., & Horn, S.P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299-311.

Stevens, J. J. (2000). *Educational accountability systems: Issues and recommendations for New Mexico*(Technical Report). New Mexico State Department of Education.

Stevens, J. J., Estrada, S., & Parkes, J. (2000). *Measurementissues in the design of state accountabilitysystems*. Paper presented at the annual meeting of the American Educational Research Association,

Willett, J.B. (1988). Questions and answers in the measurement of change. In E. Rothkopf (Ed.), *Review of research in education 1988-89* (pp. 345-422). Washington: American Educational Research Association.

Willms, J.D. (1992). *Monitoring school performance: A guide for educators*. Washington, DC: FalmerPress.

Willms, J.D. & Jacobsen, S. (1990). Growth in mathematics skills during the intermediate years: Sex differences and school effects.*International Journal of Educational Research, 14*, 157-174.

About the Authors

Keith Zvoch
Albuquerque Public Schools
Albuquerque, NM

Email: zvoch@aps.edu

Keith Zvoch earned a doctorate in Quantitative Methods from the Educational Psychology Program at the University of New Mexico in 2001. He is currently a research scientist for the Albuquerque Public School District. Dr. Zvoch also teaches research methods and statistics at the University of New Mexico on an adjunct basis. His current research interest is the measurement and assessment of school effects.

Joseph J. Stevens
University of New Mexico
College of Education
Educational Psychology Program
Simpson Hall
Albuquerque, NM 87131

Email: jstevens@unm.edu

Joseph J. Stevens is a Professor of Educational Psychology at the University of New Mexico. Dr. Stevens' research concerns applications and validity of large-scale assessment systems, the evaluation of accountability systems and school effectiveness, and the assessment of cognitive diversity.

Appendix A
Sample Sizes and Empirical Bayes Achievement Mean and Slope Estimates by School and Content Area

	Mathematics	Language	

School	Mean	Slope	Mean	Slope	N
1	653.40	17.34	656.67	12.65	71
2	642.78	16.15	649.29	10.26	104
3	645.24	16.32	649.52	13.27	115
4	668.92	16.16	667.71	11.04	165
5	639.67	18.12	645.18	9.83	168
6	647.62	13.98	656.45	13.50	68
7	670.82	24.22	678.12	15.16	178
8	640.74	7.15	641.76	9.39	129
9	691.48	19.83	687.51	14.15	168
10	652.49	11.90	651.85	9.22	88
11	672.49	21.99	678.75	13.31	232
12	665.55	16.95	662.57	13.09	131
13	636.90	25.33	641.20	12.83	103
14	662.64	23.15	667.02	13.62	216
15	660.36	17.55	659.82	10.82	118
16	672.47	22.75	672.41	13.21	242
17	663.63	20.41	664.22	13.61	109
18	678.35	14.64	674.60	11.00	125
19	663.06	22.02	669.66	13.31	172
20	657.59	18.23	653.24	11.69	136
21	679.50	16.27	673.83	12.30	164
22	642.46	24.94	654.66	14.40	97
23	656.64	19.08	660.70	12.91	123
24	661.63	17.07	657.63	10.57	77

The World Wide Web address for the *Education Policy Analysis Archives* is
epaa.asu.edu

Editor: Gene V Glass, Arizona State University

Production Assistant: Chris Murrell, Arizona State University

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board

EPAA Spanish Language Editorial Board

EPAA is published by the Education Policy Studies Laboratory, Arizona State University