An Analysis of School District Data Using Value-added Methodology

CSE Report 622

Damian Betebenner
CRESST / University of Colorado at Boulder

March 2004

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.1: Comparative Analyses of Current Assessment and Accountability Systems. Strand 2: Outcomes of Different Accountability Designs.
Robert L. Linn, Project Director, CRESST/University of Colorado at Boulder

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

Abstract

This paper is an examination of student assessment data using multi-level analysis methods often referred to as "value-added" models. The analyses performed provide measures of "effectiveness" for both teachers and schools. The purpose of the paper is to examine the residuals derived from the model for teachers and schools (i.e., their "value-added" scores) and to examine their relation to demographic variables at the classroom and school level. Three models are examined: the basic variance components model, the random intercepts model, and a model including demographic covariates. Those interpreting the results should be aware of the likelihood of model misspecification in contexts where value-added measures are derived. These measures are a function of the model and much care must be taken to rule out the possibility of model misspecification.

Introduction

This study examines data from a large southern California school district using a variety of multi-level models generally referred to in the literature as *value-added models*. Within a given school system, value-added models purport to estimate the contribution of a given unit of analysis to the student score. In this way, an estimate of a given teacher's contribution to their students' scores can be derived. It is this facet of the model which has endeared it to those wishing to link test results to accountability (Rowan, Correnti, & Miller, 2002). The purpose of this study is to both use the results of the modeling and to examine the limitations of those results when applied to issues central to accountability initiatives.

The notion that there are determinants of academic achievement that are not directly attributable to the student (e.g., family, teachers, and schools) is almost a truism. The quintessential study of such factors is the Coleman report (1966). Among other things, this study demonstrated that quality of education is highly variable in the United States and brought to light the question of whether family factors (e.g., socio-economic status) were more important than school factors in determining academic performance. Hundreds of studies since have elaborated on this theme and have used increasingly sophisticated models combined with more complete data to better understand what the determinants of student achievement are. Value-added models applied to student assessment data are state of the art in terms of spelling out the determinants of student achievement (Rivkin, Hanushek, & Kain, 2002).

The primary use of value-added models today is to estimate effects associated with an aggregate unit employed in the model (e.g., the classroom unit or school unit). The first large-scale implementation of this is the Tennessee Value-Added Assessment System (TVAAS) developed by Dr. William Sanders. The TVAAS system employs mixed-model methodology[1] to estimate teacher and school level value-added measures. TVAAS employs two distinct models to analyze longitudinal student level data in a given large aggregate unit (usually

[1] The mixed-model methodology is a term which indicates that the statistical model under consideration contains both fixed and random effects. Multilevel models form a subclass of the class of all mixed-models and include the value-added models employed in the TVAAS system (Rowan et al., 2002).

a district) to quantify the effects associated with a given teacher or school on students' assessment test scores (Sanders, Saxton, & Horn, 1997; Wright, Horn, & Sanders, 1997).

Numerous studies have followed employing various models to determine the effectiveness of teachers or programs. One such study by Stone (2002) investigated the effectiveness of National Board for Professional Teaching Standards (NBPTS) certified teachers in Tennessee using the TVAAS value-added estimates. Stone found that those teachers who were NBPTS certified would not be considered excellent based upon the measures of excellence employed by the Chattanooga, TN use of the TVAAS teacher level value-added measures. It is in this way that assessment aligns well with accountability initiatives to determine effectiveness based upon value-added estimates for almost any conceivable predictor one wishes to consider.

The purpose of this study is to examine some of the methodology used to derive the value-added measures becoming so prominent in accountability initiatives. Criticism of whether these measures account for the demographic components of the students is of particular concern since accountability systems compare students, teachers, and schools across racial and socio-economic groups. This study will examine value-added scores in light of free and reduced lunch percentages at the class and school levels, and determine whether or not such measures can actually take account of the many exogenous factors that come together in classrooms across the United States today.

Data

Data for this study were supplied by a large southern California school district. The database supplied contains a variety of academic and demographic information for students and teachers in the 1997-98, 1998-99, and 1999-2000 school years. Figure 1 depicts the relational nature of the database. For a description of the variables provided in the database see page 23 of the Appendix.

Because analyses associated with the data centered on teacher and school effectiveness, only data where a student was linked with a teacher and school was examined. These data are in the StudentTeacherLink table. In this table there are 448,339 cases comprising 86,385 unique students. All student entries in StudentTeacherLink appear within the student demographic table, StudentDemo. This table was employed to derive free/reduced lunch percentages. A more complete description of the data set accompanies the description of variables in the Appendix.

Given the immense nature of the database, only a subset of all available data was employed to investigate teacher and school effectiveness. The analyses reported here utilized the data associated with elementary school teachers. These data are much more amenable to analysis given the fact that students are most often associated with a single teacher in elementary school making the association between that teacher and the students' achievement more plausible. School level analyses are possible at the elementary, middle, and high school levels. But only those at the elementary school level are reported here. The following is a summary of the data employed in this study.

1. We begin by separating StudentTeacherLink into two groups: elementary schools and middle/high schools. This is done by using the field `CrsID`, which has value 0 for elementary schools. The two tables formed are denoted StudentTeacherLinkElem and StudentTeacherLinkMidHigh.

2. In table StudentTeacherLinkElem there are 124,584 rows comprising 64,374 unique students and 2,633 unique teachers. In table StudentTeacherLinkMidHigh there are 323,755 rows comprising 34,385 unique students and 1,083 unique teachers.

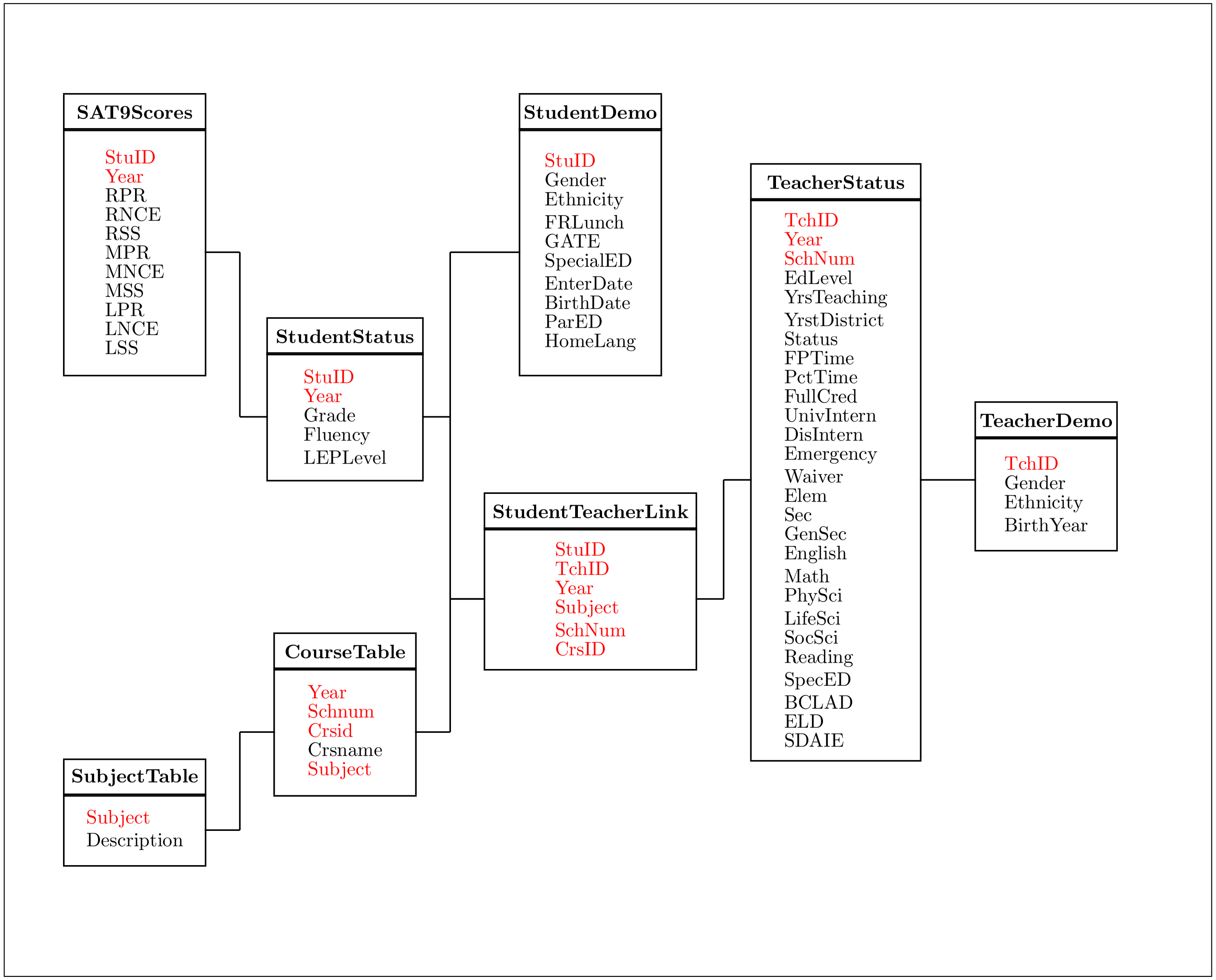

Figure 1: Tables, their variables, and relations for the school district data. Common variables between tables are shown in red

3. As is to be expected, a number of students appearing in the StudentTeacherLink file did not have test scores in the SAT9Scores file. Missing values were entered for those students once the tables were joined.

4. The number of distinct elementary schools within the file (distinct occurrences of `SchNum`) is 60. The number of distinct teachers in the files (distinct occurrences of `TchID`) is 2,633. The number of distinct students in the file is 64,374.

5. There are six cases in ElemBaseData where the scale score is 0. These six cases were changed to missing values.

6. There are 32 cases in StudentDemo where the ParEd lies outside the range of defined values (23 cases with 0, 1 case with 12, and 8 cases with 127). These cases were changed to missing. There are 64,565 missing cases.

7. There are 624 cases in StudentDemo where HomeLang lies outside the range of defined values (578 cases with 0, 46 cases with 127). These cases were changed to missing. There are 628 missing cases.

8. Analyses were conducted for the three subject matter tests (Language, Math, and Reading) both together and separately.

All data for this study was stored in its native relational format using the MySQL relational database server.

Method and Results

The Basic Variance Components Model

To derive measures of effectiveness at the teacher and school level from the data provided, this study employs multi-level models (also called hierarchical linear models and mixed-models) to derive estimates that can be used to measure the “value-added” of each teacher or school (Goldstein, 1995; Snijders & Bosker, 1999). These so called “value-added” models are not a single model but, more generally, a family of models where a residual estimate with respect to a random effect (e.g., the teacher or school) is used as a proxy for how much “value” the particular unit adds. More specifically, the ”value-added estimate” is a prediction of the random variable associated with the unit of analysis (e.g, the teacher or the school). For technical details about the calculation of these residuals see Snijder and Bosker (1999, p. 58).

Since the data provide up to three years of test data per student in three subject areas, the data can be represented as a multi-variate (referring to the three subject areas) repeated measures (referring to the three years) multilevel design. Moreover, given the data, a natural hierarchy places students within class (i.e., teacher) within school. It is this nesting structure that makes elementary school students much more amenable to multilevel modeling than middle and high school students who are assigned to multiple teachers. We first extracted the appropriate datasets with the relevant occasion, student, teacher, and school identifiers from the relational database and entered the data into the multi-level analysis program MLwiN (Rasbash, Browne, Goldstein, & Yang, 2000).

The analyses presented in this study fall into two broad categories: multivariate-multi-level analyses and univariate-multi-level analyses. The univariate/multivariate distinction refers to whether the model fit uses a single subject matter test for each student or all three subject matter tests (i.e., math, reading, and language). We believe that it is preferable from a theoretical standpoint to try and fit all the data within a single model so that possible test information on each student is included, but that there is always a question of practical significance, and the estimates derived separately may be “just as good” as those derived together.

The univariate and multivariate models were built using four-levels of hierarchy: *occasion* within *student* within *teacher* within *school*. Using the multivariate model allows for the variability present in the test scores of the students to be decomposed at each of the four levels. In its most basic form, the univariate expression of this model is represented as follows:[2]

$$X_{ijkl} = \beta_0 + f_{0l} + v_{0kl} + u_{0jkl} + e_{0ijkl} \tag{1}$$

And the multivariate form of the model has three equations, one for each exam:

$$X_{ijkl_{lang}} = \beta_{0_{lang}} + f_{0l_{lang}} + v_{0kl_{lang}} + u_{0jkl_{lang}} + e_{0ijkl_{lang}} \tag{2}$$
$$X_{ijkl_{math}} = \beta_{0_{math}} + f_{0l_{math}} + v_{0kl_{math}} + u_{0jkl_{math}} + e_{0ijkl_{math}} \tag{3}$$
$$X_{ijkl_{read}} = \beta_{0_{read}} + f_{0l_{read}} + v_{0kl_{read}} + u_{0jkl_{read}} + e_{0ijkl_{read}} \tag{4}$$

where i denotes *testing occasion*, j denotes *student*, k denotes *teacher*, and l denotes *school*. f_{ol}, v_{0kl}, u_{0jkl}, and e_{0ijkl} are the level 4 (occasion), 3 (student), 2 (teacher), and 1 (school) level residuals, respectively, and $\mathrm{var}(f_{0l}) = \sigma^2_{f0}$, $\mathrm{var}(v_{0kl}) = \sigma^2_{v0}$, $\mathrm{var}(u_{0jkl}) = \sigma^2_{u0}$, and $\mathrm{var}(e_{0ijkl}) = \sigma^2_{e0}$. Test scores, X_{ijkl}, were mean centered so that value-added estimates would be centered around 0.[3] Note that the multilevel analysis allows one to estimate the contribution of all combinations of occasion by student by teacher by school present in the dataset and thus calculate the contribution of each teacher or school separately from occasion.

Table 1 presents estimates using Model 1 with each of the language, math, and reading tests. This basic variance components model, used as a baseline model, shows how variability in the dependent variable is distributed across individual, teacher, and school levels. Particularly noteworthy is the fact that most of the variability is situated at the teacher level. Specifically, 35.1% of the variability of the language exam occurs at the teacher, k, level whereas 47.7% and 44.9% of the mathematics and reading exams, respectively, occurs at the teacher level.

Table 2 shows the multivariate estimates based upon the model expressed by Equations 2, 3, and 4. The only significant difference in the parameter estimates between Table 1 and Table 2 occurs with β_0 and σ^2_{v0} associated with the language exam. There is a particularly large difference in the estimation of the variance component which might be due to the

[2] This model is referred to variously in the literature as the empty model or the variance components model, since only variance components are estimated within it.

[3] For a more extensive discussion of variable centering in multilevel analysis see Snijder & Bosker (1999).

Parameter	Language	Mathematics	Reading
β_0	-10.715(2.182)	-9.141(2.132)	-9.067(2.548)
σ^2_{f0}	238.464(48.793)	232.231(50.563)	355.164(73.028)
σ^2_{v0}	687.496(24.411)	1,282.675(41.630)	1212.355(39.407)
σ^2_{u0}	531.954(16.619)	512.735(21.060)	575.796(18.558)
σ^2_{e0}	499.472(16.072)	662.694(20.679)	554.059(18.010)
school-level squared correlation	.122	.086	.131
teacher-level squared correlation	.351	.477	.449
student-level squared correlation	.272	.191	.213
$-2\log(likelihood)$	780,435.8	856,168.9	830,280.7

Table 1: Univariate parameter estimates and (standard errors) associated with Equation 1 on language, math and reading tests.

Parameter	Language	Mathematics	Reading
β_0	-4.679(2.116)	-9.292(2.156)	-8.291(2.587)
σ^2_{f0}	252.609(51.755)	227.452(49.354)	344.025(70.589)
σ^2_{v0}	876.424(28.770)	1277.096(41.393)	1229.010(39.865)
σ^2_{u0}	542.360(16.523)	514.600(20.767)	577.698(18.036)
σ^2_{e0}	504.809(15.993)	662.322(20.389)	552.651(17.500)
school-level squared correlation	.116	.085	.127
teacher-level squared correlation	.403	.476	.455
student-level squared correlation	.249	.192	.214
$-2\log(likelihood)$		2,321,005.000	

Table 2: Multivariate parameter estimates and (standard errors) for Equations 2, 3, and 4

fact that a correlation between the three exams is incorporated into the multivariate model and σ^2_{v0} for the mathematics and reading exams is much larger for those exams than for the language exam.

In addition, the multivariate model provides covariance estimates for the three subject tests at each level. These covariances combined with the variances from Table 4 can provide a correlation estimate that indicates the degree of association between randomly sampled units (i.e., schools, teachers, students, and occasions) in two of the subject areas. For example, at the student level, the estimated correlations between language-math, language-reading, and math-reading are 0.84, 0.93, and 0.75. At the teacher and school levels, all correlations exceeded 0.96. At the occasion level the correlations were approximately 0.5.

Though one could use the variance components model to derive "value-added" scores at the different model levels, it would be analogous to calculating residuals associated with a regression equation where the regression line was forced to have slope equal to zero (i.e., no covariate term, just an intercept). The variance components model is the simplest model one can test and forms a base line against which other nested models can be tested using a simple χ^2 statistic.

The Random Intercepts Model

The model from which value-added estimates for teachers and schools will be derived and examined is the model in which occasion is treated as a fixed effect. In the univariate case, the model can be expressed as the variance components model presented in Equation 1 plus the addition of dummy coded variables to account for the fixed occasions. This model is represented as:

$$X_{ijkl} = \beta_0 + \mu_1 d_{1ijkl} + \mu_2 d_{2ijkl} + f_{0l} + v_{0kl} + u_{0jkl} + e_{0ijkl}, \tag{5}$$

where d_{1ijkl} and d_{2ijkl} are variables coded 0 or 1 depending upon whether the student's test score is derived from testing occasion 1 (1998-99 in this case) or testing occasion 2 (1999-2000). Note that testing occasion 0 (1997-98) is implicitly derived by setting d_{1ijkl} and d_{2ijkl} equal to zero. Thus, μ_1 and μ_2 are parameters representing mean test scores for the 1997-98 and 1998-99 test administrations, respectively. Three similar equations express the multivariate model:

$$\begin{aligned} X_{ijkl_{lang}} &= \beta_{0_{lang}} + \mu_1 d_{1ijkl_{lang}} + \mu_2 d_{2ijkl_{lang}} + \\ &\quad f_{0l_{lang}} + v_{0kl_{lang}} + u_{0jkl_{lang}} + e_{0ijkl_{lang}} \qquad (6) \\ X_{ijkl_{math}} &= \beta_{0_{math}} + \mu_1 d_{1ijkl_{math}} + \mu_2 d_{2ijkl_{math}} + \\ &\quad f_{0l_{math}} + v_{0kl_{math}} + u_{0jkl_{math}} + e_{0ijkl_{math}} \qquad (7) \\ X_{ijkl_{read}} &= \beta_{0_{read}} + \mu_1 d_{1ijkl_{read}} + \mu_2 d_{2ijkl_{read}} + \\ &\quad f_{0l_{read}} + v_{0kl_{read}} + u_{0jkl_{read}} + e_{0ijkl_{read}} \qquad (8) \end{aligned}$$

In both the univariate and multivariate cases, this model is the classic model from repeated measures, the compound symmetry model, and is equivalent to the random intercepts model (Snijders & Bosker, 1999, p. 168). That is, the variance-covariance matrix calculated using the three occasions has constant variance on the diagonal and constant covariances in the off-diagonal positions. Roughly speaking, regression lines fit to the data can vary by intercept but share the same slope. Since there are no other explanatory variables in the

Fixed Effect	Language	Mathematics	Reading
β_0	-10.845(2.160)	-19.162(2.253)	-15.155(2.652)
μ_1	5.824(0.303)	9.118(0.320)	6.299(0.320)
μ_2	11.555(0.347)	16.136(0.364)	11.218(0.364)

Random Effect	Language	Mathematics	Reading
σ^2_{f0}	246.814(50.558)	250.707(54.636)	369.989(75.890)
σ^2_{v0}	700.138(24.890)	1395.156(45.116)	1284.257(41.346)
σ^2_{u0}	608.414(14.147)	648.865(16.830)	667.980(15.718)
σ^2_{e0}	411.544(13.383)	502.521(16.129)	451.072(14.936)
school-level squared correlation	.125	.090	.133
teacher-level squared correlation	.356	.499	.463
student-level squared correlation	.309	.232	.241
$-2\log(likelihood)$	779,368.6	854,296.0	829,369.8

Table 3: Univariate parameter estimates and (standard errors) associated with Equation 5 on language, math and reading tests.

model except those for the fixed occasions, the model represents a pure within-subjects design. Later models will include other fixed effects and covariates which will alter this aspect. Table 3 presents results for the univariate random intercepts model and Table 4 presents results for the random intercepts multivariate model.

A quick comparison between the result of using a full multivariate model versus the univariate analyses done separately indicates that there is not much difference between the estimates produced. In fact, at the school level, the correlation between the predicted residuals (i.e., sometimes referred to as value-added scores) derived using the univariate model and the multivariate model on the language test was 0.974. The correlation between the residuals derived via the two methods increased for the math and reading exams to 0.985 and 0.993. At the teacher level a comparison of the residuals is not so straight-forward. When estimated separately, some of the teachers had no valid scores for their students on a particular subject matter test. Hence, when univariate estimates were calculated these teachers were excluded. In the multivariate analysis, a number of these teachers get a predicted residual on a test that their students have no valid scores for. That is, their estimates are derived based upon how the teacher's students did on other subject matter tests.

When comparing the results with that of the basic variance components model, Tables 1 and 3 together with Tables 2 and 4 show little difference. Though the introduction of the fixed-effect for occasion decreases the likelihood statistic a significant amount (using a χ^2 statistic with 6 degrees of freedom), it is not clear that the decrease is practically significant. Moreover, the correlation between school level residuals derived from Models 1 and 5 was

Fixed Effect	Language	Mathematics	Reading
β_0	-17.459(2.228)	-19.123(2.214)	-16.438(2.599)
μ_1	6.007(0.299)	9.208(0.318)	6.729(0.313)
μ_2	11.219(0.337)	16.384(0.360)	12.090(0.355)
Random Effect	Language	Mathematics	Reading
σ^2_{f0}	260.258(53.796)	242.038(52.943)	354.290(73.297)
σ^2_{v0}	938.162(30.782)	1390.351(44.983)	1307.731(42.376)
σ^2_{u0}	618.722(14.102)	653.320(16.528)	674.843(15.135)
σ^2_{e0}	418.985(13.370)	499.600(15.823)	444.187(14.347)
school-level squared correlation	.116	.087	.127
teacher-level squared correlation	.420	.499	.470
student-level squared correlation	.277	.235	.235
$-2\log(\text{likelihood})$		2,319,069.000	

Table 4: Multivariate parameter estimates and (standard errors) for Equations 6, 7, and 8

.997.

To derive actual value-added effects for teachers and schools, predicted values associated with v_{0kl} and f_{0l} are calculated for each teacher or school. These predicted values are often times used as "value-added" estimates associated with a given teacher or school. It must be emphasized that these estimates are *model contingent* and are only as sound as the model on which they are based.

In addition to the broader concern that value-added estimates might be derived from a mis-specified model, these predicted values themselves are subject to error. Thus, in comparing value-added estimates for two different schools, for example, care must be exercised based upon uncertainty associated with the estimates. Figure 2 depicts the school-level residuals and their standard errors for each of the three exams. The standard error bars depicted are 1.4 times the length of the standard error. Thus, in using the figure to compare schools, one can infer that with 95 percent certainty that if the confidence bands for two schools don't overlap, then the residual estimates associated with each school are likely different (Goldstein & Healy, 1995). Thus, the top third (roughly 20) of schools can be reliably distinguished from the bottom third of schools based on their estimated value-added (residuals) scores in all three content areas.

A similar but much more cluttered picture is given at the teacher level in Figure 3. Figure 2 depicts one school with an especially large error band. The reason for the particularly large error band is that there were few valid scores available at the school ($n = 4$). Few students per teacher is also the reason for the large error bands seen in the three panels of Figure 3. Teachers with value-added estimates that are in the top 25% (roughly 500 teachers) can be reliably distinguished from teachers whose value-added estimates fall in the bottom quarter of the rank-order distribution in each of the three subjects.

The models looked at so far include no covariates or fixed effects other than those denoting test year. The last set of analyses will present a variety of results for models that include a number of predictors including student demographic variables and teacher characteristics. Of particular interest is the link between the "value-added" scores and the socio-economic status of the class/school as measured by free/reduced lunch.

Value-added Estimates and Additional Demographic Predictors

As mentioned previously, the predicted values associated with the random-effects residuals, the so called "value-added" estimates, are only as good as the model used and the scale over which the tests are vertically equated. In this section we investigate how the value-added estimates derived using Equation 5 are related to socio-economic status. Part of the reason for investigating this relationship is due to the suspicion that the value-added residuals might in fact carry with them some measure of poverty that is not controlled for with the blocking procedure employed with the multi-level model.

The most straight-forward analysis looking at "value-added" estimates with socio-economic status is to look at the correlation between these estimates and the percent of free/reduced lunch students at either the teacher or school level. The data provided by the district included demographic student data indicating whether the student was eligible for free/reduced lunch. Aggregates were compiled for each teacher and school.[4]

Figure 4 depicts free/reduced lunch percentages against value-added residuals on each of the three exams for both schools and teachers. As the six panels of the figures show, there is a negative correlation (near -0.6 for each subject at the school level and between -0.2 and -0.4 at the teacher level) between these residuals suggesting that those schools/teachers with a higher percentage of free/reduced lunch students generally receive lower "value-added" scores. What this implies is difficult to determine. One interpretation of the negative correlations observed in Figure 4 is that the "value-added" scores are confounded with the socio-economic status of the class/school. In other words, teachers at lower socio-economic schools are less likely to be certified and tend to have less experience than their peers at higher socio-economic schools. Students from lower socio-economic backgrounds may also

[4] In the data supplied by the district, all children present in the data file were either identified as free/reduced lunch or not. There were none with missing data. This level of coverage makes the data somewhat suspect given the nature of free/reduced lunch data identification.

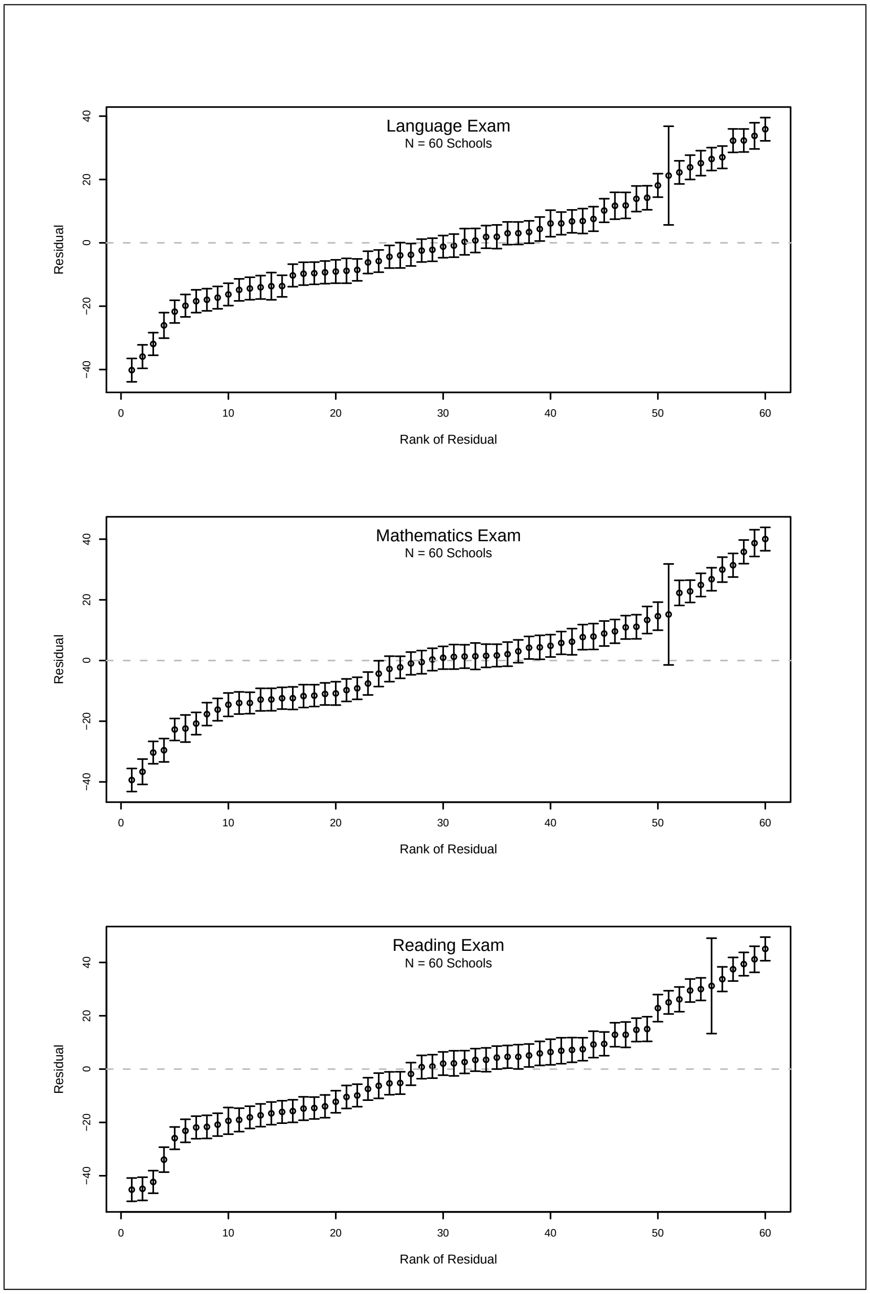

Figure 2: Value-added estimates (residuals) with associated standard error bars ordered by rank for schools

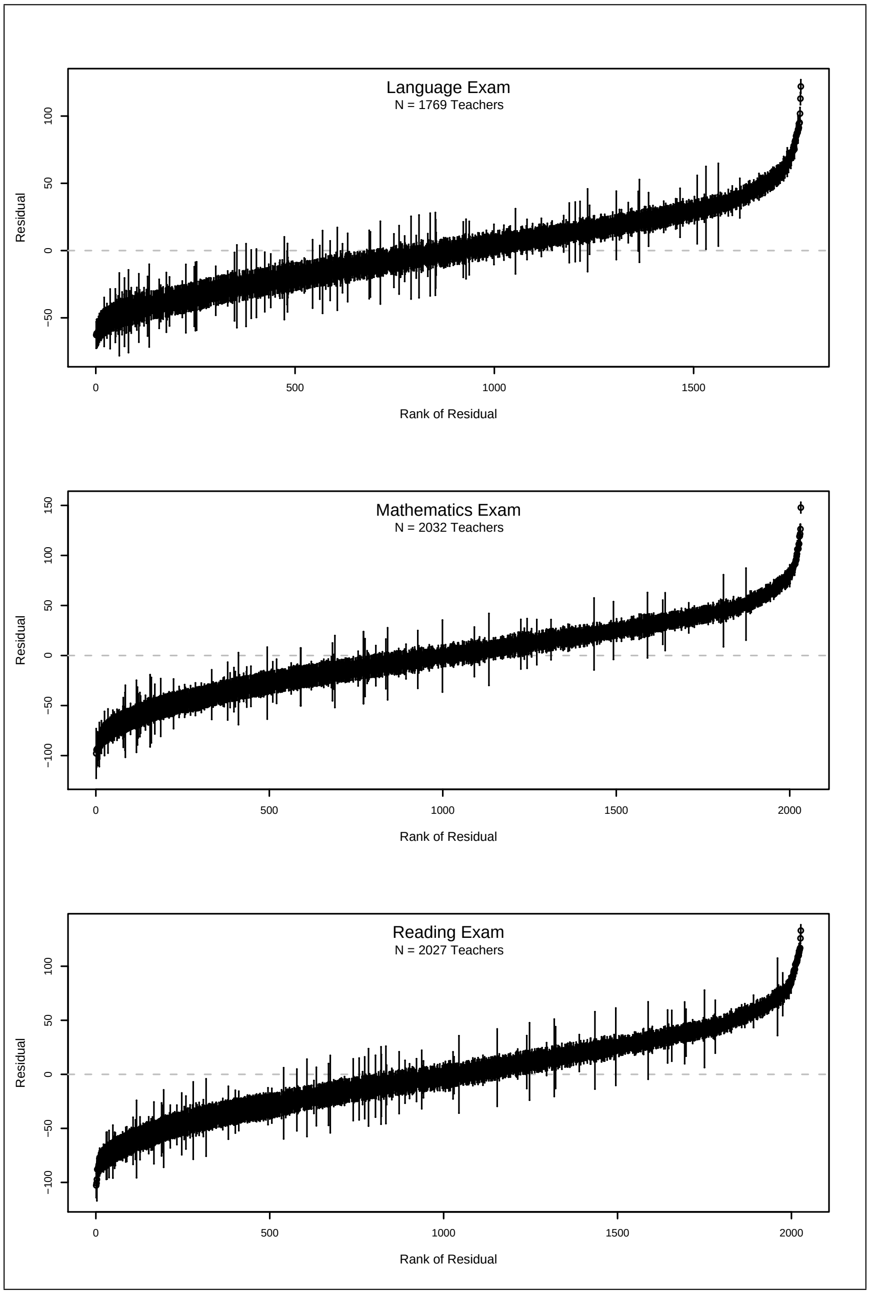

Figure 3: Value-added estimates (residuals) with associated standard error bars ordered by rank for teachers

receive less academic support outside of school than their peers from higher socio-economic backgrounds.

This result is consistent with what occurs when free/reduced lunch status is added to the multi-level model as a fixed effect. Adding free/reduced lunch status (FR) as a fixed-effect allows one to perform what is analogous to a simple t-test between those students indicated as receiving free/reduced lunch and those not. Using the multivariate formulation of the model specified in Equations 6, 7, and 8, the equations for the model incorporating free/reduced lunch are given by:

$$\begin{aligned} X_{ijkl_{lang}} &= \beta_{0_{lang}} + \mu_1 d_{1ijkl_{lang}} + \mu_2 d_{2ijkl_{lang}} + \\ &\quad f_{0l_{lang}} + v_{0kl_{lang}} + u_{0jkl_{lang}} + e_{0ijkl_{lang}} + \beta_{1_{lang}} FR \qquad (9) \\ X_{ijkl_{math}} &= \beta_{0_{math}} + \mu_1 d_{1ijkl_{math}} + \mu_2 d_{2ijkl_{math}} + \\ &\quad f_{0l_{math}} + v_{0kl_{math}} + u_{0jkl_{math}} + e_{0ijkl_{math}} + \beta_{1_{math}} FR \qquad (10) \\ X_{ijkl_{read}} &= \beta_{0_{read}} + \mu_1 d_{1ijkl_{read}} + \mu_2 d_{2ijkl_{read}} + \\ &\quad f_{0l_{read}} + v_{0kl_{read}} + u_{0jkl_{read}} + e_{0ijkl_{read}} + \beta_{1_{read}} FR \qquad (11) \end{aligned}$$

where FR denotes the free/reduced lunch status of each student. Table 5 provides the results associated with adding free/reduced lunch to the set of predictors. Not surprisingly, the addition of the dichotomous student level variable denoting eligibility for free/reduced lunch yields a coefficient that is significantly different than zero. That is, on all exams those students designated free/reduced lunch have lower achievement test score gains than those not designated. Specifically, the expected decrease in score is given by β_1. The greatest impact is on the reading test and the least impact is on the math test.

Additional fixed effects can be entered into the model so that the impact upon the scale score of the students by the fixed effect can be determined. The rich nature of the school district data allows for numerous student level and teacher level variables to be added to the basic random intercepts model to determine the effect of the variable on the outcome. Consider the model with the following variables added to those already present in Equations 9, 10, and 11.

Student Level Variables	Teacher Level Variables
Free/Reduced Lunch	Elementary Certification
Home Language	Temporary Teaching Waiver
Ethnicity	Years Teaching
Gifted and Talented Program	

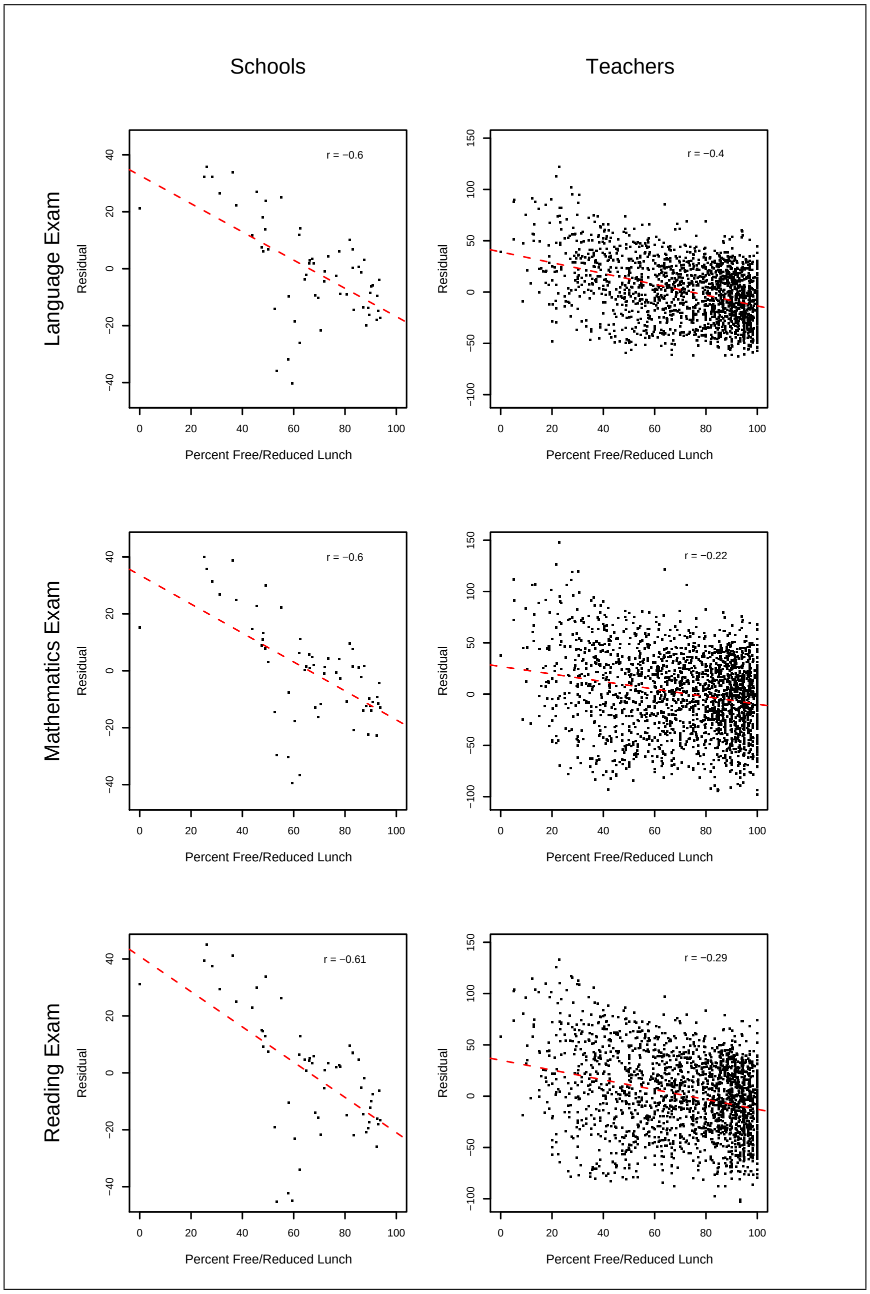

Figure 4: Value-added estimates (residuals) plotted against free/reduced lunch percentages for each test at both the school and teacher level

Fixed Effect	Language	Mathematics	Reading
β_0	-6.872(2.028)	-9.153(2.031)	-3.754(2.355)
μ_1	5.879(0.295)	9.088(0.314)	6.581(0.308)
μ_2	10.486(0.333)	15.707(0.357)	11.196(0.349)
β_1	-14.721(0.291)	-13.939(0.305)	-17.691(0.299)

Random Effect	Language	Mathematics	Reading
σ^2_{f0}	207.654(43.842)	193.525(43.808)	280.372(59.356)
σ^2_{v0}	920.807(30.192)	1383.349(44.728)	1290.535(41.783)
σ^2_{u0}	582.350(14.184)	620.164(16.588)	623.451(15.177)
σ^2_{e0}	424.507(13.523)	504.251(15.942)	450.263(14.486)
school-level squared correlation	.097	.072	.106
teacher-level squared correlation	.431	.512	.488
student-level squared correlation	.273	.230	.236
$-2\log(\text{likelihood})$		2,315,585.000	

Table 5: Multivariate parameter estimates and (standard errors) for Equations 9, 10, and 11

All the variables are dichotomous except Years Teaching. A number of the variables were dichotomized based upon nominal variables in the original data: Home Language (0 – English, 1 – non-English), Ethnicity (0 – White, 1 – non-white), Gifted and Talented Program (0 – No, 1 – Yes), Elementary Certification (0 – No, 1 – Yes), Temporary Teaching Waiver (0 – No, 1 – Yes). The results from the analysis are presented in Table 6.

The results are not too surprising. The inclusion of the variables reduces the amount that free/reduced lunch contributes to a student's score. Holding all other variables constant, the effect of elementary certification status appears to be positive for all subject areas with the largest contribution coming in Math. The contribution associated with being on waiver is not significant for any of the tests. The covariate years teaching yielded similar results across all three tests with there being a positive relationship between teacher experience and student test scores: for every year of teacher experience, students showed an expected gain of 0.3 points. Surprisingly, the dichotomous variable "homelanguage" showed a significant positive value associated with students whose home language was *not* English on the language and mathematics tests and a negative value on the reading tests. This appears to be in conflict with the coefficient for ethnicity which indicates that students who are white score roughly ten points higher than their non-white counterparts.

The estimated effects for student background variables suggest that teachers who teach poor and minority students would be at a relative disadvantage in comparison to their peers who teach white students from higher socio-economic backgrounds if value-added estimates ignored these student characteristics. Conversely, value-added estimates that ignored student

Fixed Effect	Language	Mathematics	Reading
β_0	-14.911(2.382)	-20.685(2.632)	-7.746(2.667)
μ_1	6.440(0.295)	9.902(0.316)	6.983(0.309)
μ_2	11.277(0.347)	16.653(0.379)	11.873(0.370)
free/reduced	-10.815(0.301)	-10.863(0.312)	-12.471(0.306)
elem certif	6.245(1.609)	7.983(1.823)	4.586(1.769)
waiver	-3.099(1.814)	-0.588(2.013)	-1.797(1.971)
yrs teaching	0.301(0.065)	0.278(0.080)	0.297(0.077)
homelang	2.015(0.282)	7.889(0.295)	-1.346(0.290)
ethnicity	-10.094(0.362)	-11.009(0.376)	-12.432(0.369)
gifted	40.548(0.478)	44.270(0.506)	42.898(0.496)

Random Effect	Language	Mathematics	Reading
σ^2_{f0}	137.986(30.353)	146.054(34.392)	173.143(38.860)
σ^2_{v0}	764.366(25.478)	1210.096(39.689)	1106.213(36.354)
σ^2_{u0}	504.767(13.816)	526.914(15.953)	529.523(14.748)
σ^2_{e0}	413.814(13.269)	485.771(15.435)	439.576(14.197)
school-level squared correlation	.076	.061	.077
teacher-level squared correlation	.420	.511	.492
student-level square correlation	.277	.222	.236
$-2\log(\text{likelihood})$		2,230,271.000	

Table 6: Multivariate parameter estimates and (standard errors) for model including all predictors

background characteristics would give a relative advantage to teachers teaching students in the gifted and talented program compared to teachers who taught students who were not in the gifted and talented program. It is also the case that teachers who lack elementary certification, are teaching with a temporary waiver, have less teaching experience, and are less likely to have high value-added scores than more experienced teachers who have elementary certification. Overall, the results in Table 6 suggest that value-added estimates that ignore student background characteristics and teacher characteristics may be unfair.

Conclusions

This report presents an examination of data from a large southern California school district using a family of value-added models. The analyses contained in this report have used a variety of multi-level models to test the extent to which the models can be used to assess change in student scores that is attributable to teachers, to assess the extent to which teachers can be compared relative to the amount of "value-added" those teachers provide to students, and to assess how covariates impact the "value-added" scores attributable to teachers. The major findings are as follows:

- Using students scores on several different tests across several different years as the dependent variable, most of the variability observed is attributable to teachers. Specifically, 35.1%, 47.7%, and 44.9% student score variability in language, mathematics, and reading, respectively, is attributable to the teacher level.

- Value-added estimates derived for schools and teachers with standard-error bands around those estimates demonstrate that one can reliably distinguish roughly the top third from the bottom third of schools. Similarly, one can reliably distinguish approximately the top quarter from the bottom quarter of teachers using their value-added estimates.

- Value-added estimates were impacted by demographic variables of students, classes, and teachers. Most significantly, the socio-economic/racial makeup of a class impacted the value-added estimates derived for teachers. Other characteristics such as teacher elementary certification also had an impact.

Given these findings, it is important that both researchers and practitioners who begin to utilize value-added estimates to determine systemic efficacy pay attention to model-specification as part of understanding the reliability of these estimates to the system they purport to measure. To do otherwise in the proliferating high stakes environments nationwide could lead to consequences that may not improve the quality of education for students.

References

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity* (Tech. Rep.). Washington, D.C.: U.S. Government Printing Office.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). New York: Edward Arnold.

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, Series A*, *158*(1), 175–177.

Rasbash, J., Browne, W., Goldstein, H., & Yang, M. (2000). *A user's guide to MLwiN* [Computer program, version 1.10.0007]. London: Institute of Education.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2002). *Teachers, schools and academic achievement.* article pre-print.

Rowan, B., Correnti, R., & 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* (Tech. Rep.). Philadelphia, PA: Consortium for Policy Research in Education (CPRE).

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press, Inc.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis.* Thousand Oaks, CA: Sage Publications.

Stone, J. E. (2002, May). *The value-added achievement gains of nbpts-certified teachers in tennessee: A brief report.* (Available on-line at http://www.education-consumers.com/briefs/stoneNBPTS.shtm)

Wright, S. P., Horn, S. P., & 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.

Appendix

Variable Description

Table Name: CourseTable

Field Name	Description
Year	1 – 1997-1998, 2 – 1998-1999, 3 – 1999-2000
SchNum	3-digit school identifier
CrsID	4-digit course identification number
CrsName	Name of Course
Subject	Code identifying subject area (see table Subject Table for descriptions)

Table Name: SAT9Scores

Field Name	Description
StuID	Unique 9-digit student identification number
Year	Year of administration with 1 – 1997-1998, 2 – 1998-1999, 3 – 1999-2000
RPR	National Percentile Rank in Total Reading
RNCE	Normal Curve Equivalent in Total Reading
RSS	Scaled Score in Total Reading
MPR	National Percentile Rank in Total Math
MNCE	Normal Curve Equivalent in Total Math
MSS	Scaled Score in Total Math
LPR	National Percentile Rank in Total Language
LNCE	Normal Curve Equivalent in Total Language
LSS	Scaled Score in Total Language

- stuid—98,151 unique entries. There are 232 entries that appear in the SAT9Scores table that do not appear in the StudentDemo table. There are 21,970 unique entries that appear in the SAT9Scores table that do not appear in the StudentTeacherLink table.

Table Name: StudentDemo

Field Name	Description
StuID	Unique 9-digit student identification number
Gender	M – Male, F – Female
Ethnicity	A – Asian, B – Black, F – Filipino, H – Hispanic, I – Native American, P – Pacific Islander, W – White
FRLunch	Free or Reduced Lunch Status: 1 – Qualifies for free/reduced lunch, 0 – does not qualify for free/reduced lunch
GATE	1 – Identified as gifted, 0 – Not identified as gifted
SpecialED	1 – Special Ed student, 0 – Not Special Ed
EnterDate	Date that student entered the district
BirthDate	Date of birth of student
ParEd	Parent Education: 1 – Not HS Grad, 2 – HS Grad, 3 – Some College, 4 – College Grad, 5 – Postgraduate Training
HomeLang	2-digit code for language spoken at home: 1 – English, 2 – Spanish, 3 – Vietnamese, 4 – Filipino, 5 – Cantonese, 6 – Korean, 7 – Hmong, 8 – Khmer, 9 – Armenian, 10 – Russian, 11 – Lao, 12 – Other
Ethnicitynum	Recoded version of Ethnicity: 1 – Asian, 2 – Black, 3 – Filipino, 4 – Hispanic, 5 – Native American, 6 – Pacific Islander, 7 – White
Ethnicitynumdi	Dichotomous version of Ethnicitynum: 0 – White, 1 – Other
Homelangdi	Dichotomous version of Homelang: 0 – English, 1 – Other

- stuid—121,088 unique entries. There are 34,703 entries that appear in the StudentDemo table that do not appear in the StudentTeacherLink table. There are 23,169 entries that appear in the StudentDemo table that do not appear in the SAT9Scores table.

Table Name: StudentStatus

Field Name	Description
StuID	Unique 9-digit student identification number
Year	1 – 1997-1998, 2 – 1998-1999, 3 – 1999-2000
Grade	2-digit grade level 00 through 12
Fluency	1 – ELL (English Language Learner), 0 – Fluent in English
LEPLevel	IDEA levels for ELL Students: A through F and M where A denotes lowest fluency, F denotes highest fluency, and M denote attainment of oral fluency

Table Name: StudentTeacherLink

Field Name	Description
StuID	Unique 9-digit student identification number
TchID	Unique 9-digit teach identification number
Year	1 – 1997-1998, 2 – 1998-1999, 3 – 1999-2000
Subject	Code for subject area. See table SubjectTable for descriptions
SchNum	3-digit school identifier
CrsId	4-digit course ID

- stuid—448,339 cases comprising 86,385 unique entries. All entries appear within the StudentDemo table. Thus the students in the StudentTeacherLink table are a proper subset of those appearing in the StudentDemo table. There are 10,204 entries that appear in the StudentTeacherLink table that do not appear in the SAT9Scores table.

- tchid—3,665 unique entries. There are 110 unique entries that appear within the StudentTeacherLink table that do not appear in the TeacherStatus table. There are 281 unique entries that appear within the StudentTeacherLink table that do not appear within the TeacherDemo table.

Table Name: SubjectTable

Field Name	Description
Subject	Alphanumeric code for subject area
Description	Description of subject area

Table Name: TeacherDemo

Field Name	Description
TchID	Unique 9-digit teacher identification number
Gender	M – Male, F – Female
Ethnicity	A – Asian, B – Black, F – Filipini, H – Hispanic, I – Native American, P – Pacific Islander, W – White
BirthYear	Year of Birth

- tchid—4,599 unique entries. There is 1 unique entry that appears within the TeacherDemo table that does not appear in the TeacherStatus table. There are 1,215 unique entries that appear within the TeacherDemo table that do not appear within the Student-TeacherLink table.

Table Name: TeacherStatus

Field Name	Description
TchID	Unique 9-digit teacher identification number
Year	1 – 1997-1998, 2 – 1998-1999, 3 – 1999-2000
SchNum	3-digit school identifier for school that teacher worked at this year
EdLevel	1 – Doctorate, 2 – Master's + 30, 3 – Master's, 4 – Bachelor's + 30, 5 – Bachelor's, 6 – Less Than Bachelor's
YrsTeaching	Years Teaching
YrstDistrict	Years in District
Status	T – Tenured, P – Probationary, L – Long-term substitute, O – Other
FPTime	F – Full time, P – Part time
PctTime	Percent of time teaching
FullCred	Full credentialed: 1 – Yes, 0 – No
UnivIntern	University Intern: 1 – Yes, 0 – No
DistIntern	District intern: 1 – Yes, 0 – No
Emergency	Emergency credential: 1 – Yes, 0 – No
Waiver	Teaching on a waiver: 1 – Yes, 0 – No
Elem	Elementary multiple subject credential: 1 – Yes, 0 – No
Sec	Secondary single subject credential: 1 – Yes, 0 – No
GenSec	Secondary multiple subject credential: 1 – Yes, 0 – No
English	Secondary English certification: 1 – Yes, 0 – No
Math	Secondary Math certification: 1 – Yes, 0 – No
PhySci	Secondary Physical Science certification: 1 – Yes, 0 – No
LifeSci	Secondary Life Science certification: 1 – Yes, 0 – No
SocSci	Secondary Social Studies certification: 1 – Yes, 0 – No
Reading	Reading Specialist: 1 – Yes, 0 – No
BCLAD	Bilingual Education Certified: 1 – Yes, 0 – No
ELD	English Language Development Certified: 1 – Yes, 0 – No
SDAIE	Specially Designed Academic Instruction in English Certification: 1 – Yes, 0 – No
SpecEd	Special Education Specialist: 1 – Yes, 0 – No
Statusnum	Recoded version of Status: 0 – T, 1 – P, 2 – L, 3 – O

- tchid—5,544 unique entries. There are 1,989 unique entries that appear within the TeacherStatus table that do not appear in the StudentTeacherLink table. There are

946 unique entries that appear within the TeacherStatus table that do not appear in the TeacherDemo table.

- schnum—117 unique entries

Table Name: TeacherDemo

Field Name	Description
TchID	Unique 9-digit teacher identification number
Gender	M – Male, F – Female
Ethnicity	A – Asian, B – Black, F – Filipino, H – Hispanic, I – Native American, P – Pacific Islander, W – White
BirthYear	Year of Birth