TECHNICAL APPENDIX A • FEBRUARY 2010

# Gaining Ground in the Middle Grades: Why Some Schools Do Better

## A Large-Scale Study of Middle Grades Practices and Student Outcomes

EdSource®
Clarifying Complex Education Issues®

520 San Antonio Rd, Suite 200, Mountain View, CA 94040-1217 | 650/917-9481
Fax: 650/917-9482 | edsource@edsource.org
www.edsource.org | www.ed-data.org

# Gaining Ground in the Middle Grades: Why Some Schools Do Better

## TECHNICAL APPENDIX A

## TABLE OF CONTENTS

PAGE

# Gaining Ground in the Middle Grades: Why Some Schools Do Better

## *APPENDIX A - Research Methodology and Analyses*

### Overview

*This section provides the overview of the study including how the sample was selected, the outcome variables used, and analysis steps.*

### Constructing the Survey Data File

*This section describes how the principal, teacher, and superintendent survey data were entered, data cleaning and recoding, and statistical reliability of the survey items.*

### Constructing Composite Independent Variables (Subdomains)

*This section describes the conceptual and technical development of subdomain composite variables from individual survey items that measure various schooling practice areas.*

### Constructing Longitudinal Outcome Variables

*This section describes the use of the special longitudinal data file obtained from the California Department of Education (CDE) to develop longitudinal outcome variables that controlled for past student performance.*

### Constructing Data Files for Analysis

*This section describes how both the longitudinal and cross-sectional data were utilized in the analyses in addition to listing the control variables used in the study.*

### Specifying Predictor Pools

*This section describes the tools developed to effectively map the survey items into subdomains, and the subdomains into domains.*

### Regression Analyses

*This section describes the primary analytic technique—regression analysis—and lists the steps taken for the analysis.*

### Statistical Comparisons Across Study Domains

*This section describes the statistical methodology used in comparing domains.*

# Gaining Ground in the Middle Grades: Why Some Schools Do Better

## *APPENDIX A - Research Methodology and Analyses*

### *Overview*

Statistical analyses for the Middle Grades Study were carried out primarily by the Principal Data Analyst, Jesse Levin, who is a Senior Research Scientist at American Institutes for Research (AIR). Overall responsibility for planning and coordinating the analyses rested with—Senior Technical Director—Edward Haertel, who is a professor in the School of Education at Stanford University. Levin and Haertel were ably assisted by Ben Webman and other EdSource staff members. The project team met approximately twice per month from December 2008 through January 2010, with more frequent meetings as needed.

This was a complicated study, using over 1,000 variables derived from three separate surveys (of principals, teachers, and superintendents) to predict school-level outcomes on seven different California Standards Tests (CSTs). As described in one PowerPoint presentation, the study required analysis of over 1,000,000 teacher item responses, over 100,000 principal item responses, and nearly 30,000 superintendent item responses. Over 400 distinct regression models were examined. Specification of all these analyses required over 6,300 lines of statistical programming. Over 20,000 variables were created at various points in the process. The school CST score means serving as outcome variables were derived from the test scores of over 200,000 students.

California public schools with both 7th and 8th grade students served as the primary sampling units and as the unit of analysis. These included both middle schools and K-8 elementary schools. The sample was further restricted to schools within two bands (the 20th-35th and 70th-85th percentiles) of the California Department of Education School Characteristics Index (SCI), a composite of demographic variables indicating the degree of educational challenge each school confronts.[1] Of the 528 schools in this target sample, 133 were eliminated because their school districts declined to participate (typically citing time pressures and uncertainties due to the current funding climate in California), or a school had closed or consolidated with another. Of the 395 schools contacted, 303 provided both teacher and principal data used in the study. Of these schools, 244 also had corresponding surveys that were completed by the superintendent presiding over their district or, in the case of charter schools, the chief administrator of the charter management organization. Within each participating school, all regular mathematics and/or English language arts (ELA) teachers of 6th, 7th, or 8th grade students were surveyed.

The main outcome variables used were school-level means of the CSTs in English Language Arts for grades 6 through 8 (ELA6, ELA7, and ELA8), Mathematics for grades 6 and 7, and for grade 8 General Mathematics and Algebra I (Math6, Math7, Math8Gen, and Math8Alg). Analyses were based solely upon school-level data from students taking a CST without modifications. That is, no use was made of data from the California Modified Assessment (CMA), Standards-based Tests in Spanish (STS), or California Alternate Performance Assessment (CAPA). As described in the body of this report, schools were recruited from two demographic bands defined by the 2006-07 SCI. The combined set of all schools in both the 20th to 35th and 70th to 85th SCI percentile bands is referred to as the pooled sample. While most analyses used this pooled sample, some analyses used only the 20th-35th band or the 70th-85th band schools. Only a subset of the schools served students in the 6th

[1] A report on the construction of the SCI is available at http://www.cde.ca.gov/ta/ac/ap/documents/tdgreport0400.pdf. Details of the 2007 SCI (used for sample selection) are available at http://www.cde.ca.gov/ta/ac/ap/documents/tdgreport0708.pdf.

grade, and not all schools offered General Mathematics and/or Algebra I at the 8th grade level. In addition, superintendent surveys were received from only a subset of the school districts with participating schools. For these reasons, the numbers of schools included varied across analyses according to outcome being analyzed and available survey data, as shown in Figure A-1.

**Figure A-1: Numbers of Schools Included in Analyses for Each Outcome, by Sample and Surveys**

| | Analyses Using Principal and Teacher Surveys | | | Analyses Using Principal, Teacher, and Superintendent Surveys | | |
|---|---|---|---|---|---|---|
| | **Pooled** | **20-35** | **70-85** | **Pooled** | **20-35** | **70-85** |
| **ELA6** | 220 | 102 | 118 | 169 | 86 | 83 |
| **ELA7** | 303 | 144 | 159 | 244 | 125 | 119 |
| **ELA8** | 303 | 144 | 159 | 244 | 125 | 119 |
| **Math6** | 220 | 102 | 118 | 169 | 86 | 83 |
| **Math7** | 303 | 144 | 159 | 244 | 125 | 119 |
| **Math8Gen** | 252 | 109 | 143 | 204 | 94 | 110 |
| **Math8Alg** | 298 | 141 | 157 | 242 | 125 | 117 |

In addition to unadjusted spring 2009 testing outcomes (the seven cross-sectional outcome variables), test scores from 2006, 2007, 2008, and 2009, linked at the individual student level, were used to construct mean *residualized* 2009 CST scores for each school (the seven growth outcome variables).[2] Thus, there were fourteen main outcome variables. The numbers of schools included were the same for the growth outcomes as for the corresponding cross-sectional outcomes.

Because the number of potential explanatory variables was large relative to the number of schools surveyed (i.e., over 1,000 variables and only 303 schools), a disciplined approach was required in planning an analysis that could distill as much information as possible from the multitude of variables available. To do this, items were grouped into clusters on substantive grounds, with no reliance on information concerning their statistical associations with outcomes. Composite variables (subdomains) were created from these item clusters. Analyses using the subdomain variables as explanatory variables for all outcomes were run in parallel, following a common analysis plan.[3]

Stata Version 10 was used for all statistical analyses. In addition, sophisticated Excel workbooks were developed to facilitate communication between the conceptual and technical groups that comprised the research team. These Excel workbooks codified the grouping of survey items into subdomains, the labeling of those subdomains, and the specification of subsets of subdomains that were to be considered for possible inclusion in each predictive equation. Each of these Excel workbooks contained one or more worksheets designed to be easily imported into Stata, which were used to drive the analyses. The final construction of tables summarizing pairwise significance tests comparing the explanatory power of the ten domains for the various outcomes was also carried out using Excel.

Major steps in the data analysis were: 1) constructing the survey data file; 2) developing composite independent variables (subdomains) organized into ten practice domains; 3) creating longitudinal outcome variables; 4) assembling these pieces and constructing data files for analysis; 5) specifying

[2] That is, we estimated the portion of school-level CST scores for 2009 that could not be explained by prior year scores (2006, 2007, and 2008). For a detailed discussion of how this was done, see the section, Constructing Longitudinal Outcome Variables, that follows.

[3] To minimize finding significant results simply by chance, our analysis plan did not allow for any a priori selection of survey items to include in subdomains based on whether they were significantly correlated with outcomes.

pools of potential predictive variables for inclusion in each analysis; 6) performing the regression analyses; and, 7) making statistical comparisons across domains to determine their relative predictive power. In addition, decision rules were developed to guide the synthesis and interpretation of findings from this very complicated study. The remainder of this appendix section describes these steps in turn.

### *Constructing the Survey Data File*

Completed surveys were received at EdSource, logged, and shipped to WestEd for keyed data entry under the supervision of John Bosma. A random sample of surveys was keyed twice for purposes of data verification. Data were provided to EdSource in the form of Excel files. The principal, teacher, and superintendent data files received from Bosma were in excellent shape, but of course some data cleaning remained. The first step in data cleaning was to examine all alpha responses to numeric variables and recode those that could be recoded (e.g., a written-in response of "about 20" was recoded to 20). For questions requiring a numeric response, any remaining alpha responses (those that could not be recoded) were treated as missing. Next, all responses to write-in variables (e.g., responses to questions of the form "Other (please specify) ______________") were examined for possible recoding to similar or identical options provided in that question's preceding list of options.

Next, data were imported into Stata using the application StatTransfer. All files were then checked for out-of-range responses. Decision rules were also developed for treating non-scalable responses such as "Does Not Apply" or "Don't Know" as informative for some items or simply as missing for others. Items were recoded as necessary so that the signs of expected correlations to outcomes were all positive. In a few cases, single items were replaced by a set of several binary variables. In a few other cases, for complex, multi-part items, distributions of response pattern frequencies were tabulated and codes were created for each high-frequency pattern. Rules were developed for trimming out-of-range numerical responses (e.g., for instructional minutes per day). Internal consistency checks enabled correction and/or imputation for missing responses, or in some cases for deletion of responses where skip patterns were not observed. On "Check All That Apply" items, rules were developed for distinguishing blanks signifying "Does Not Apply" from omitted responses. A very few items were dropped because they appeared to have been misconstrued by significant numbers of respondents.

The data cleaning and recoding described above was an iterative process, entailing both logical analysis and substantive decisions. Throughout, all decisions were captured in Stata programs specifying the creation of recoded variables. At the same time as the full research team worked through recoding decisions, we also reviewed the placement of derived variables and retained survey item response variables in subdomains and the mapping of subdomains to domains. Where a substantive case could be made for multiple placements of an item into two or more subdomains, the variable was flagged for future review, so that final item placement could be informed by the empirical relationship of the ambiguous item to the other items already placed in each of the candidate subdomains. The final result was a set of Stata programs that processed the raw survey data to produce clean survey data files for further analysis. Cleaned data files included surveys from 303 principals, from 3,752 ELA and Math teachers, and from 157 superintendents (who represented 244 of the 303 total schools in the sample).

One additional step was carried out for teacher surveys. For each recoded item, intra-class correlations for teachers within schools and corresponding reliabilities of school means were calculated. Any item with a reliability statistic less than 0.25 was dropped from further consideration. The 39 items dropped represented just over 13% of all the teacher survey items. Note that these checks were performed using all teacher surveys. As described below, various subsets of teacher surveys within each school were used for some analyses. Some retained teacher items may have fallen below the 0.25 reliability threshold for teacher subsets (e.g., those who reported teaching 8th grade Algebra).

### *Constructing Composite Independent Variables (Subdomains)*

The surveys developed for teachers, principals, and superintendents were based on prior research and theory, as well as careful consideration of the current policy context for middle grades schooling in California. Once data were collected, the same theory and practice considerations guided the grouping of items into one of the following ten substantive domains:

Domain A – A positive, safe, engaging school environment.
Domain B – An intense, school-wide focus on improving academic outcomes.
Domain C – School organization of time and instruction.
Domain D – Coherent and aligned standards-based instruction and curricula.
Domain E – Extensive use of data to improve instruction and student learning.
Domain F – Early and proactive academic interventions.
Domain G – Attention to student transitions.
Domain H – Teacher competencies, evaluation, and support.
Domain I – Principal leadership and competencies.
Domain J – Superintendent leadership and district support.

Within each of these domains, in general each item was included in exactly one subdomain, with the exception that some items included in one of Domains A through H could also be included in Domains I or J. Domains A through H dealt with various aspects of schooling policy and practice. Domains I and J dealt with the role of the principal and with the role of the superintendent and/or district leadership, respectively. If an item referred to the principal's or superintendent's role with respect to a domain-specific policy or practice, then the item might appear in a subdomain within Domains A through H and in another within Domains I or J.

The subdomains constructed from teacher, principal, and superintendent surveys were all school-level variables. Each subdomain was made up of one or more items from one of the three surveys (i.e., there were teacher, principal, and superintendent subdomains). Some subdomains included only a single item, but every effort was made to group multiple items into subdomains where possible. Because procedures differed slightly for subdomains created from the principal and superintendent versus teacher survey items, they are described separately.

Principal subdomains were constructed as follows. First, each item was standardized to mean zero and standard deviation (SD) 1. Second, for each principal, a straight (unweighted) average of all non-missing standardized values was calculated. Finally, the resulting averages were restandardized to mean zero and standard deviation 1 across all principals. Superintendent subdomains were constructed in exactly the same way as principal subdomains. As noted, the 157 superintendents responding oversaw districts containing 244 of the 303 schools in our sample. Where a superintendent's responses pertained to two or more schools (i.e., where a superintendent's district included two or more schools in our sample), the same superintendent subdomain responses were used for each of those schools.[4]

Creation of teacher subdomains was more complex. As part of data cleaning and recoding, teacher responses had already been recoded to quantitative variables so that averages up to the school level would result in meaningful values. In addition, as noted, items that failed to discriminate adequately among schools (i.e., whose school-mean reliability was less than 0.25) had been dropped. The next step was to average each teacher survey item up to the school level. However, not all teachers' responses were relevant to all analyses. A 6th grade ELA teacher's responses concerning some aspect of classroom practice might not be relevant in modeling predictors of Algebra I for 8th graders, for example. For this reason, 16 distinct versions of each school-level average were created, each making use of item responses from different (sub)populations of teachers, as shown in Figure A-2.

**Figure A-2: Versions of School-Level Average Teacher Survey Items**

| Version | Inclusion Criterion |
|---|---|
| 1 | All Teachers |
| 2 | All ELA Teachers |
| 3 | All Math Teachers |
| 4 | 6th Grade ELA Teachers |
| 5 | 7th Grade ELA Teachers |
| 6 | 8th Grade ELA Teachers |
| 7 | 6th Grade Math Teachers |
| 8 | 7th Grade Math Teachers |
| 9 | 8th Grade Gen Math Teachers |
| 10 | 8th Grade Algebra Teachers |
| 11 | 6th Grade Teachers |
| 12 | 7th Grade Teachers |
| 13 | 8th Grade Teachers |
| 14 | All Teachers with English Learners in Their Classrooms |
| 15 | 6th and 7th Grade ELA Teachers (Versions 4 and 5 Combined) |
| 16 | 6th and 7th Grade Math Teachers (Versions 7 and 8 Combined) |

Sixteen versions of each teacher subdomain were then constructed by aggregating responses from each of the subsets of teachers to the school level, standardizing the school-level items to mean zero and standard deviation 1, and averaging across items to construct subdomains as was done for the principal

[4] Sample sizes were not sufficient for meaningful hierarchical modeling.

and superintendent surveys. There were far fewer missing teacher item responses at the school level because for each item, school-level averages had been created from all non-missing responses. However, where all teachers in some version subset within a school omitted an item response, an average was taken across the remaining (non-missing) items within that subdomain, exactly as for the principal and superintendent surveys.

Checks of internal consistency and of dimensionality were run for all subdomains including two or more items. These checks were performed for the principal and superintendent subdomains and for version 1 (the “All teachers” version) of the teacher subdomains. For subdomains with just two items, there was a simple check that the two items were positively correlated. Cases where the correlation was less than 0.40 were scrutinized closely. If, on substantive grounds, it appeared that the two items belonged together, then the subdomain was kept as is. If there was a substantively interesting difference between the two items, then the subdomain was divided into two new single-item subdomains. Negatively correlated item pairs were sometimes separated into two new single-item subdomains. In other cases, one or both items were recoded, resulting in a positive association. For subdomains with more than two items, internal consistency reliabilities were examined, and a principal component analysis was run. If the reliability was very low or if more than one factor was extracted according to the default criterion (i.e., the number of eigenvalues being greater than 1.00), the subdomain was closely scrutinized and in some cases divided into two or more subdomains.

The missing data imputation described earlier in this section pertained to missing responses for a subset of the items within a subdomain. Where there was only one item in a subdomain, or where all items in the subdomain were missing, these procedures would still result in a final missing value calculated for the subdomain. In order to make best use of the limited sample of schools available, and in light of the large number of potential predictors being considered, remaining missing values were replaced with the mean of the non-missing responses for that variable. This was required much more frequently for the principal and superintendent subdomains than the teacher subdomains because teacher subdomains would be missing only if all teachers in a version subset within a school omitted all items included in a given subdomain. In all cases where mean replacement was required, an additional variable referred to as an *imputation flag* was created. This was a binary variable, specific to a given subdomain, taking on the value of 1 for cases where missing values for that subdomain were replaced with the subdomain mean across all non-missing cases, and 0 for the remaining schools (where means replacement was not required). In a subsequent stage of the analysis, described below, these dummy variables were entered into the regression together with the corresponding subdomain. If the imputation flag dummy showed a statistically significant effect for a particular outcome variable, then the corresponding subdomain was dropped from the analysis for that outcome. No further use was made of imputation flags following these checks.

## *Constructing Longitudinal Outcome Variables*

We requested a set of special files from the California Department of Education (CDE) to permit the construction of growth-outcome variables corresponding to each cross-sectional CST outcome variable. Preliminary files using 2008 CSTs as the final year were obtained first and were used only to refine procedures and test Stata program code. Final files, using 2009 CSTs as the final year, were obtained as soon as a reasonably clean (near-final) version of the 2009 data became available within the CDE. All final analyses are based on 2009 CSTs, which match the cross-sectional outcomes and represent the end-of-year student outcomes corresponding to the year in which the teacher, principal, and superintendent survey data were collected. These files included a record for each student in the 6th, 7th, or 8th grades in one of our participating schools in the final year (2008 for the preliminary file, 2009 for the final file). There were no student identification indicators included in the data. Variables in each file included CST scores in ELA and math for the final year and up to three preceding years, as available, together with indicators permitting the derivation of each student's grade-level in school each of these years as well as the particular ELA and math CSTs taken each year. In addition to a stringent nondisclosure agreement, confidentiality of individual students' responses was ensured by the addition of a small random number to each CST scale score (random data perturbation). The variance of the random numbers was small enough, relative to the variance of the CST scale scores, that school-level means of residualized scores were virtually unaffected.

For each outcome (e.g., CST 8th Grade General Mathematics), multiple linear regression was used to predict the 2009 score from all prior year scores. Only prior math scores were used to predict math scores, and only prior ELA scores were used to predict ELA scores. Due to missing data patterns and in a few cases due to grade retention, as many as 10 to 20 or so prior test score patterns might occur for a given 2009 CST outcome. Student records with rare patterns (i.e., patterns with fewer than 200 students across the entire sample of participating schools) were dropped. Also, any student who did not have a CST score for the immediately prior year (2008) was dropped.

For the remaining patterns, multiple regression was used to predict the 2009 score from all available prior year scores. For patterns including missing data for one or two years, these regressions could be run using either of two groups of students. For example, consider the pattern "2006-missing, 2007-missing, 2008-ELA5, and 2009-ELA6" for students who took the grade 6 ELA CST in 2009 and the grade 5 ELA CST in 2008, but for whom no earlier ELA CST scores were available. A regression could be run using only the students with this pattern. Alternatively, a regression predicting the 2009 grade 6 ELA CST using only the 2008 grade 5 ELA CST could be run using all students with these two tests for these two years, including those with patterns like "2006-ELA3, 2007-ELA4, 2008-ELA5, and 2009-ELA6" or "2006-ELA3, 2007-missing, 2008-ELA5, and 2009-ELA6."

Thus, two regressions were run for each pattern—one pattern-specific and the other as inclusive as possible. (These two regressions were identical for patterns with no missing values for prior year test score predictors.) In each case where these models differed, the choice between them represented a trade-off between potentially greater bias and increased precision. The two regressions were compared, and where there was a statistically significant difference in the coefficients for one or more prior year test score predictors, the pattern-specific model was used. Where there was not a statistically significant difference, the more precise model based on the larger sample size was used. The goal was to predict each included student's 2009 ELA CST and math CST scores as precisely as possible using all available prior year data. Once final models and regressions were determined, each student's predicted 2009 scores were subtracted from the corresponding observed 2009 scores and the resulting residuals, pooled across patterns, were averaged up to the school level. These aggregated residuals served as the "Growth-Outcome" dependent variables.

### *Constructing Data Files for Analysis*

The special files we received from the CDE enabled calculation of 2009 cross-sectional mean scores in cases where fewer than ten students in a school had taken a given grade-specific or, in the case of math, grade-and-course-specific CST. (On the CDE website, school-level results based on fewer than ten students are suppressed.) A small number of otherwise missing outcome values for small schools was retrieved in this manner.

In addition to these cross-sectional outcome variables, the longitudinal growth outcome variables already described, and the subdomain variables derived from the three surveys, a set of baseline demographic variables was specified for each analysis. After some preliminary investigations to resolve the details of variable selection and coding, the final set of baseline demographic variables was determined, as shown in Figure A-3. Choices among alternative sets of variables were based on patterns of collinearity and adjusted $R^2$ statistics in regressions predicting the cross-sectional outcome variables. As the sample of schools was bimodal with respect to SCI (i.e., schools fell within the 20th-35th or 70th-85th SCI band), a key question was whether the relationship between schooling practices (as measured by subdomain variables) and CST achievement differed between the lower and higher SCI band schools. To do this, a *high-SCI band* indicator (representing schools in the 70th-85th SCI band) was interacted with the subdomain variables. In turn, the high-SCI band indicator (main effect) itself was included even though it added little to the predictive equation because in any regression including interaction terms, the corresponding main effects should also be included. Thus, because the final regressions included some interactions of subdomains with SCI band (indicating different regression slopes for 20th-35th versus 70th-85th SCI percentile band schools), the SCI band "main effect" was required. Demographic variables were obtained from the CDE website. In all cases, the most recent data available were employed. As shown in Figure A-3, certain demographic variables were constructed by collapsing across available categories.

**Figure A-3: Demographic (or "Baseline") Variables in Final Regression Models**

| **Description** | **Source** |
|---|---|
| **STUDENT CHARACTERISTICS - GENERAL** | |
| Percent Economically Disadvantaged | STAR 2009 (Spring 2009) |
| Percent [English Learner (EL) + Redesignated Fluent-English-Proficient (RFEP)] | STAR 2009 (Spring 2009) |
| **STUDENT CHARACTERISTICS - ETHNICITY** | |
| Percent African American | STAR 2009 (Spring 2009) |
| Percent Asian | STAR 2009 (Spring 2009) |
| Percent Filipino | STAR 2009 (Spring 2009) |
| Percent Hispanic | STAR 2009 (Spring 2009) |
| *Percent White (includes Percent White, Pacific Islander/Hawaiian Native, American Indian/Alaskan Native)* | *STAR 2009 (Spring 2009)* |
| **STUDENT CHARACTERISTICS - PARENT EDUCATION** | |
| Percent Parental Education - College Graduate Plus Graduate School | STAR 2009 (Spring 2009) |
| Percent Parental Education - High School Graduate Plus Some College | STAR 2009 (Spring 2009) |
| *Percent Parental Education - Less than High School Graduate* | *STAR 2009 (Spring 2009)* |
| **SCHOOL CHARACTERISTICS - GRADE CONFIGURATION** | |
| Grade Configuration - K-8th Grade | EdSource Principal Survey |
| Grade Configuration - 7th-8th Grade | EdSource Principal Survey |
| *Grade Configuration – 6th-8th & Other* | *EdSource Principal Survey* |
| **SCHOOL CHARACTERISTICS - MATH COURSE TAKING** | |
| Proportion of 7th graders taking Algebra I instead of the 7th grade math test[a] | STAR 2009 (Spring 2009) |
| Proportion of 8th graders taking Algebra I instead of another 8th grade math test[b] | STAR 2009 (Spring 2009) |
| Proportion of 8th graders taking Geometry instead of another 8th grade math test[b] | STAR 2009 (Spring 2009) |
| **SCHOOL CHARACTERISTICS - GENERAL** | |
| Percentage of students counted as part of school enrollment in October 2008 CBEDS and has been continuously enrolled since that date | 2009 Growth API |
| Cohort Size - Average Grade Enrollment - [(Percent of enrollments in grades 7 and 8 multiplied by school enrollment) divided by 2] | 2009 Growth API |
| Indicator of SCI Band (20th -35th and 70th-85th Percentile in 2007) | 2007 Base API |
| For selected categorical demographic variables, the (omitted) reference categories (e.g., percent white non-Hispanic/other, percent of students whose parents are not high school graduates, or grade 6-8 school configuration) are shown in grey table rows. | |

[a] Included as control variable only for the following outcome variables: Grade 7 General Math, Grade 8 General Math, and Grade 8 Algebra I.

[b] Included as control variable only for the following outcome variables: Grade 8 General Math and Grade 8 Algebra I.

### *Specifying Predictor Pools*

Four Excel workbooks were constructed to document the construction and inclusion of variables for the various analyses. The largest of these (called the Domain Development Tool or DDT) defined the mappings of all individual survey items into subdomains as well as the labels for those subdomains. The DDT allowed for provisional placement of an item in more than one subdomain. After factor analyses and reliability checks, each item was almost always included in only one subdomain, with the exception that items could be included in one of Domains A through H and also in Domain I or Domain J, as explained earlier. The other three Excel workbooks specified which teacher, principal, and superintendent subdomains were to be considered for inclusion in models predicting each of the CST scores. For example, subdomains pertaining to ELA-specific practices might be excluded from models predicting math outcomes. In addition, the teacher workbook specified which of the 16 versions of each subdomain were to be used in a given analysis (see Figure A-2). These Excel tables were imported into Stata, enabling efficient code generation and minimization of errors. Use of these tables also enabled revisions to decision rules for subdomain construction and inclusion with only very minor updates to the Stata code itself.

### *Regression Analyses*

The main result of the regression analyses was a set of 840 distinct regression models. These included models limited to teacher and principal survey predictors as well as models using the teacher, principal, and superintendent surveys, for each of the ten domains, for each of the seven CST outcomes (ELA6, ELA7, ELA8, Math6, Math7, Math8Gen, and Math8Alg), for cross-sectional versus longitudinal growth versions of the outcomes, and for the pooled versus the 20th-35th and 70th-85th percentile SCI band school samples. Exhibit 1 provides an illustration of the total number of regression models estimated for the study.

As explained above, several regression runs were required to generate each of these 840 final models. After reaching final models using only the principal and teacher subdomains, additional steps were followed to augment these models with superintendent subdomain variables. This division of the analysis into two major stages enabled the most efficient possible use of the data given that superintendent surveys were available for only a subset of the schools (see Figure A-1).

The steps leading to these 840 models were as follows. Note that each step is repeated for the 70 combinations defined by the seven CSTs and the ten domains. This entire process was carried out first using the seven cross-sectional outcomes and then using the seven longitudinal growth outcomes.

1. Use the pooled sample of schools to run each of the 70 regressions defined by the seven (cross-sectional or longitudinal) outcomes serving as dependent variables and ten domain-specific sets of subdomain variables and corresponding missing value indicators for imputed subdomain observations serving as explanatory variables (in addition to the baseline variables listed in Figure A-3). For all missing value imputation indicators that prove significant with a p-value less than or equal to 0.05, drop both the imputation indicator and corresponding imputed subdomain from further consideration. Repeat until no imputation indicators are significant with a p-value less than or equal to 0.05. Drop all missing value imputation indicators.

2. Using the final regression specifications from Step 1, evaluate the variance inflation factor (VIF) for each subdomain variable and drop all variables with a VIF greater than 10 from further consideration.

3. Using the 20th-35th SCI band school sample, run the final regressions specified in Step 2, and construct lists of all (retained) subdomains with coefficients significant with a p-value less than or equal to 0.05.

4. Repeat Step 3 using the 70th-85th SCI band school sample.

5. Using the baseline variables and the retained set of subdomains from Step 2 for the pooled sample of schools, perform a forward stepwise regression procedure locking in all baseline variables but allowing principal and teacher subdomains designated for possible inclusion to enter equations. The required significance level for the principal and teacher subdomain variables to be retained in models is set to p-value less than or equal to 0.10. Construct lists of all subdomains that were retained in these regressions.

6. Make lists of subdomains identified as significant (retained) in Steps 3, 4, and 5 above, and create interaction variables of these with the high-SCI band indicator (representing schools in the 70th-85th SCI band). Using the pooled sample of schools, enter all baseline variables as well as all retained subdomain variables from Steps 3, 4, and 5 and corresponding high-SCI band interactions into regression models. Perform check of VIFs and drop any subdomains with a VIF greater than 10 from further consideration.

7. Using the lists of principal and teacher subdomain variables and interactions retained in Step 6, run final principal plus teacher models by performing a forward stepwise regression procedure that locks in all baseline variables and principal plus teacher subdomains, but allows corresponding high-SCI band interactions designated for possible inclusion to enter equations. The required significance level for the principal and teacher subdomain/high-SCI band interactions to be retained in models is set to p-value less than or equal to 0.10.

8. Rerun the final principal plus teacher models from Step 7 with one additional dummy variable representing availability of superintendent survey data to check whether outcomes for schools with superintendent data systematically differed from those without. (There was no case in which this dummy variable proved to be statistically significant.)

9. Using the 20th-35th SCI band school sample, enter all variables retained in the final principal plus teacher subdomain regression models in Step 7, plus all superintendent subdomains designated for possible inclusion. Construct lists of all superintendent subdomains with coefficients significant with a p-value less than or equal to 0.05.

10. Repeat Step 9 for the 70th-85th SCI band.

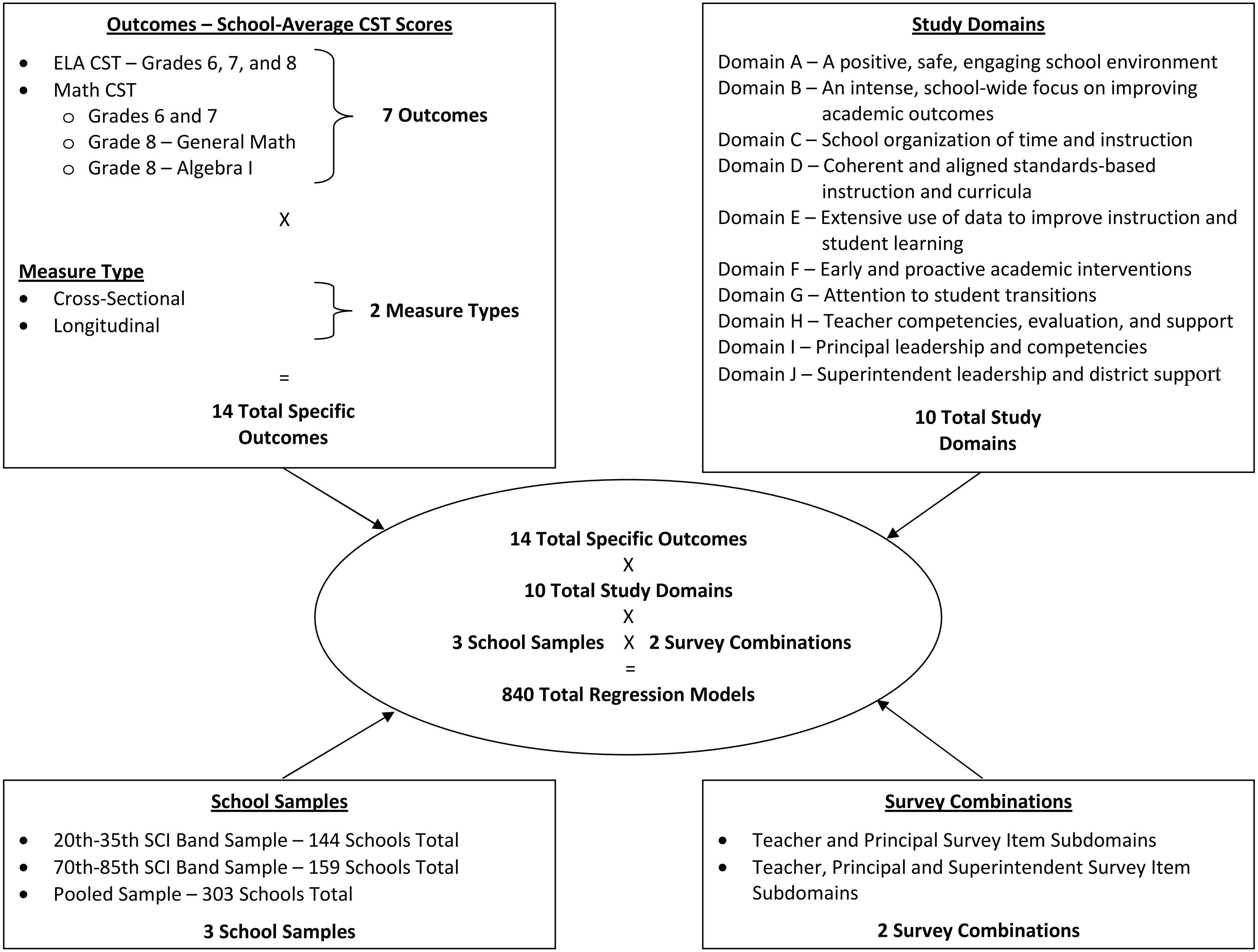
11. Using the baseline variables, the retained set of subdomains from the final principal plus teacher regression models in Step 7, and superintendent subdomains for the pooled sample of schools perform a forward stepwise regression procedure locking in all baseline, principal, and teacher subdomains (and corresponding high-SCI band interactions), but allowing the newly introduced superintendent subdomains designated for possible inclusion to enter equations. The required significance level for the superintendent subdomain variables to be retained in models is set to p-value less than or equal to 0.10. Construct lists of all subdomains that were retained in these regressions.

12. Make lists of the superintendent subdomains identified as significant (retained) in Steps 9, 10, and 11 above, and create interaction variables of these with the high-SCI band indicator. Using the pooled sample of schools, enter all baseline variables as well as all retained subdomain variables from Steps 9, 10, and 11 and corresponding high-SCI band interactions into regression models. Perform check of VIFs and drop any subdomains with a VIF greater than 10 from further consideration.

13. Using the lists of principal, teacher, and superintendent subdomain variables (and corresponding interactions) retained in Step 12, run final principal plus teacher plus superintendent models by performing a forward stepwise regression procedure that locks in all baseline variables, principal plus teacher subdomains and corresponding high-SCI band interactions, and the superintendent subdomains, but allows the superintendent high-SCI band interactions designated for possible inclusion to enter equations. The required significance level for the superintendent subdomain/high-SCI band interactions to be retained in models is set to p-value less than or equal to 0.10.

For each of the final principal plus teacher as well as the principal plus teacher plus superintendent models, wherever a high-SCI band interaction term was included, Stata was used to calculate the p-value for a significance test of whether the sum of the regression coefficients for the interaction term and its corresponding subdomain was different from zero. Note that where a high-SCI band interaction was included, the subdomain coefficient is interpreted as the effect of the variable in 20th-35th SCI band schools only, and the interaction term is interpreted as the contrast between the effects for the 20th-35th and the 70th-85th SCI band schools. It follows that the sum of these two coefficients represents the estimated effect for the 70th-85th band schools only. Thus, calculation of these sums and the corresponding significance tests just described facilitated interpretation of regression findings.

### Exhibit 1: Total Number of Regression Models Estimated for Study



**Outcomes – School-Average CST Scores**

- ELA CST – Grades 6, 7, and 8
- Math CST
  - Grades 6 and 7
  - Grade 8 – General Math
  - Grade 8 – Algebra I

**7 Outcomes**

X

**Measure Type**

- Cross-Sectional
- Longitudinal

**2 Measure Types**

=

**14 Total Specific Outcomes**

**Study Domains**

Domain A – A positive, safe, engaging school environment
Domain B – An intense, school-wide focus on improving academic outcomes
Domain C – School organization of time and instruction
Domain D – Coherent and aligned standards-based instruction and curricula
Domain E – Extensive use of data to improve instruction and student learning
Domain F – Early and proactive academic interventions
Domain G – Attention to student transitions
Domain H – Teacher competencies, evaluation, and support
Domain I – Principal leadership and competencies
Domain J – Superintendent leadership and district support

**10 Total Study Domains**

**14 Total Specific Outcomes**
X
**10 Total Study Domains**
X
**3 School Samples** X **2 Survey Combinations**
=
**840 Total Regression Models**

**School Samples**

- 20th-35th SCI Band Sample – 144 Schools Total
- 70th-85th SCI Band Sample – 159 Schools Total
- Pooled Sample – 303 Schools Total

**3 School Samples**

**Survey Combinations**

- Teacher and Principal Survey Item Subdomains
- Teacher, Principal and Superintendent Survey Item Subdomains

**2 Survey Combinations**

### *Statistical Comparisons Across Study Domains*

The analyses described above produced a wealth of findings, but there were obvious challenges in synthesizing and interpreting so much information. The first major step taken was to compare the explanatory power of the ten domains within each of the seven sets of cross-sectional principal plus teacher models, the seven sets of cross-sectional principal plus teacher plus superintendent models, the seven sets of longitudinal principal plus teacher models, and the seven sets of longitudinal principal plus teacher plus superintendent models. For ease of exposition, consider a single outcome and models using a single combination of survey data, say the principal plus teacher models for the cross-sectional 6th grade ELA CST (ELA6) outcome. Then, there are ten models to be compared, one each for Domains A through J. The steps in this comparison were, first, to quantify the explanatory power of each of the ten domain-specific models, and then to compare these across domains.

The various policies and practices defined by subdomains within a domain are not independent, but neither are they perfectly correlated. In order to determine how much variance in ELA6 can be accounted for by the subdomains in a given domain, the regression model for that domain was used to predict the (conditional) ELA6 score for a set of hypothetical schools with identical demographics, differing only in their domain-specific practices.[5] The standard deviation of the predicted ELA6 means for these hypothetical schools is a measure of the outcome variation accounted for by variation in the domain-specific practices. As shown in Figure A-1, 220 schools were included in the principal plus teacher regression models for ELA6. Thus, 220 hypothetical schools were considered, all with identical demographics, but each matching one of the 220 actual schools with respect to domain-specific practices. The standard deviations of the predicted scores for the ten domains calculated in this manner were as shown in the second column of Figure A-4.

**Figure A-4: Illustration of Predicted Power of the Ten Domains in Explaining Variation in Cross-Sectional 6th Grade ELA CST Outcomes Using Final Principal Plus Teacher Model**

| Domain | Predicted Standard Deviations | Standardized Predicted Standard Deviations |
|---|---|---|
| A | 2.14 | 0.111 |
| B | 3.21 | 0.166 |
| C | 3.68 | 0.191 |
| D | 4.73 | 0.245 |
| E | 2.88 | 0.149 |
| F | 4.85 | 0.251 |
| G | 2.94 | 0.152 |
| H | 3.87 | 0.201 |
| I | 4.14 | 0.214 |
| J | 2.91 | 0.151 |

[5] Each of the demographic variables was set to its mean value calculated over the full pooled sample of schools.

As shown in Figure A-4, schools that differed only with respect to their Domain A practices—for example, matching the observed Domain A variation among the schools actually sampled—would have a distribution of predicted mean ELA6 scores with a standard deviation of just over 2 scale score points. To facilitate comparison across outcomes as well as further pooling of findings, these results were next standardized. Each standard deviation was divided by the standard deviation of observed (unconditional) school means for the outcome variable involved, yielding the values in the third column of Figure A-4.

For instance, the standard deviation of observed ELA6 school means for the 220 schools in the example above was 19.30 points, so this represents an effect size in school-level standard deviation units of 0.111 (equal to the predicted standard deviation of 2.11 divided by the observed standard deviation of 19.30). (Note that the effect size would be smaller if expressed in standard deviation units for individual student scores.) It is also evident that the ratio of these standard deviations between the highest and lowest domains is more than 2:1. Clearly, the domains differ substantially in their explanatory power.

A total of 28 tables like the ones in Appendix C (Figures C1-C4) could be generated using the pooled sample. As stated, these would pertain to the seven outcomes for each of the two survey combination models (principal and teacher subdomains with and without superintendent subdomains) for both the cross-sectional and longitudinal outcomes. Another 28 tables were generated using the 20th-35th SCI Band schools only, and an additional 28 tables were generated using the 70th-85th SCI band. Tables with domain-specific averages were also generated by averaging across the three ELA outcomes (ELA6, ELA7, and ELA8), averaging across the four math outcomes (Math6, Math7, Math8Gen, and Math8Alg), and averaging across all seven outcomes. This brought the total number of tables to 120:

2 Survey Combination Models (principal/teacher versus principal/teacher/superintendent)
x
3 School Samples (pooled versus 20th-35th SCI band versus 70th-85th SCI band)
x
2 Outcome Types (cross-sectional versus longitudinal)
x
10 Outcomes (7 individual ELA and math tests versus 3 ELA/math test averages)
=
120 Total Tables[6]

Each of these 120 comparisons yielded a ranking of the ten domains from highest to lowest in explanatory power, but these ranks were quite unstable in cases where two or more domains had very similar standardized standard deviations (SSDs). To facilitate interpretation of these comparisons, a test was devised for the statistical significance of differences in SSDs within each set. The significance test treated the regression model as fixed, but evaluated the standard error of the contrast between two SSDs across hypothetical resamplings of schools. Because the SSD is monotonically

[6] Averaging was performed by first squaring each standardized SD, then taking the square root of the mean of the squared values. In other words, standardized variances, not standardized standard deviations, were averaged.

related to the standardized variance—which is simply the SSD squared—a significance test of the difference between two standardized variances (SVs) could be used. For two standardized variances, let $SV_1$ = Var($X_1$) and $SV_2$ = Var($X_2$), where $X_1$ is the variable constructed by dividing the predicted values for domain 1 by the standard deviation of school means, and similarly for $X_2$. The test statistic is then ($SV_1$ – $SV_2$)/[standard error of ($SV_1$ – $SV_2$)]. The denominator of this expression (the standard error of the difference between $SV_1$ and $SV_2$) is Sqrt(Var($SV_1$) + Var($SV_2$) – 2Cov($SV_1$, $SV_2$)). The sampling variance of a variance is twice the square of the population variance divided by the sample size. Substituting sample values and adjusting the number of observations (N) accordingly, Var($SV_1$) is estimated by $2(SV_1)^2/(N-1)$, and likewise for Var($SV_2$). Similarly, the sampling variance of a covariance between two sample variances is twice the square of the population covariance divided by the sample size. Substituting sample values and adjusting N accordingly, Var(Cov($SV_1$, $SV_2$)) is estimated by $2\text{Cov}(X_1 , X_2)^2/(N-1)$. Using these formulas, all pairwise tests could be carried out for individual CST outcomes.

Tests for means across outcomes (all ELA, all math, or all seven outcomes) were more complicated because all of the variances and covariances among the variables involved were required. Also, these variances were based on different, partially overlapping sets of schools. (Different schools might contribute to SSDs for Math8Gen versus Math8Alg, for example.) Simulations were carried out to investigate appropriate estimators of the covariance between two variances based on overlapping samples. A formula that performed very well in simulations was as follows: Let $N_1$ be the number of observations for $X_1$, $N_2$ be the number of observations for $X_2$, and $N_{12}$ be the number of observations common to both $X_1$ and $X_2$. Then Cov($SV_1$, $SV_2$) = $2(N_{12}-1)\ \text{Cov}(X_1 , X_2)^2/((N_1-1)(N_2-1))$. Note that this formula reduces to the standard formula if $N_1 = N_2 = N_{12}$.

EdSource is an independent, impartial, not-for-profit organization whose sole mission is to clarify complex education issues and to promote thoughtful decisions about public school improvement. Founded in California in 1977, EdSource is a respected source of information for K–14 education policy, data, and research.

EdSource®
Clarifying Complex Education Issues®

520 San Antonio Rd, Suite 200, Mountain View, CA 94040-1217 | 650/917-9481
Fax: 650/917-9482 | edsource@edsource.org
www.edsource.org | www.ed-data.org