MATHEMATICA
Policy Research

REPORT

FINAL REPORT

Measuring School and Teacher Value Added in Charleston County School District, 2014–2015 School Year

November 30, 2015

Alexandra Resch
Jonah Deutsch

Submitted to:
Charleston County School District
75 Calhoun Street
Charleston, SC 29401

Project Officer: Anita Huggins
Contract Number: P1314

Submitted by:
Mathematica Policy Research
1100 1st Street, NE, 12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Project Director: Alexandra Resch
Reference Number: 40241.811

This page has been left blank for double-sided copying.

ACKNOWLEDGMENTS

We are grateful to the many people who contributed to this report. First, we would like to thank the Charleston County School District (CCSD) for funding the work. We acknowledge our project officer in CCSD, BRIDGE director Anita Huggins, and the members of CCSD's BRIDGE work groups for working together to build a value-added model that meets the needs of Charleston schools.

At Mathematica Policy Research, Nora Paxton, assisted by Randy Rosso, Liz Potamites, Juha Sohlberg, Kai Filipczak, Clare Wolfendale, and Molly Crofton, processed the data and provided expert programming, and Eric Isenberg provided valuable comments. Betty Teller edited the report, and Kimberly Ruffin provided word-processing and production support.

This page has been left blank for double-sided copying.

CONTENTS

TABLES

I. OVERVIEW

Beginning with the 2013–2014 school year, the Charleston County School District (CCSD) piloted the BRIDGE evaluation, a new teacher and principal evaluation framework. This initiative is funded by a $23.7 million Teacher Incentive Fund (TIF) grant the U.S. Department of Education awarded to CCSD in 2012. For the 2013–2014 and 2014–2015 school years—the pilot phase of the project—the only teachers and principals eligible to receive value-added measures of teacher or school effectiveness are those at BRIDGE pilot schools. This includes teachers of mathematics, English/language arts (ELA), science, and social studies in grades 4 through 8 in 10 district schools and the principals of these schools, as well as teachers of algebra, English, and biology in three district schools.[1] The pilot evaluation framework for these teachers includes three components: (1) individual value added (IVA) (35 percent of a teacher's overall evaluation rating); (2) the state's Assisting, Developing, and Evaluating Professional Teaching (ADEPT) evaluation (30 percent); and (3) the district-developed classroom observation tool (COT), designed to provide consistent, constructive feedback to all teachers and to align with the ADEPT performance standards and with key elements related to instruction and classroom environment (35 percent). The pilot model for principals includes four components: (1) the Program for Assisting, Developing, and Evaluating Principal Performance (PADEPP) rating (30 percent of a principal's overall evaluation rating); (2) stakeholder surveys (20 percent); (3) school-wide proficiency (25 percent); and (4) school-wide value added (25 percent).[2] Starting with the 2014–2015 school year, CCSD is piloting student learning objectives (SLO) as an additional measure of student growth.

CCSD contracted with Mathematica Policy Research to develop and implement the value-added models that are used to estimate teacher and school effectiveness. The district also convened several work groups to provide input into the implementation of various components of the grant. The work groups included teachers, principals, and district staff with a stake in the various grant components in both pilot and non-pilot schools. In this report, we describe the value-added models that are used as part of the BRIDGE teacher and principal evaluations. We estimate (1) teacher effectiveness for the 2014–2015 school year in the 10 BRIDGE pilot schools that have grades 4 through 8, (2) school effectiveness for these same schools, and (3) teacher effectiveness among teachers of algebra, English, and biology in 3 BRIDGE pilot schools.[3]

This report is an updated version of the technical report for the 2013–2014 school year (Resch and Deutsch 2014). The remainder of Chapter I provides an overview of value-added methods in nontechnical terms, including the key changes to the models for 2014–2015. Chapter

[1] There are 13 BRIDGE pilot schools, but only 11 schools have teachers who taught subject-grade combinations where students took tests covered by the value-added models. Of these 11 schools, one is a high school with students who took end-of-course (EOC) exams, 8 contain at least one grade level between grades 4 and 8, and 2 are schools with both students who took EOC exams and at least one grade level between grades 4 and 8.

[2] School-wide value added is calculated by averaging the value added of the teachers within a school.

[3] The BRIDGE pilot schools are Baptist Hill High School, North Charleston High School, Military Magnet, Edmund Burns, Jane Edwards, Minnie Hughes, E.B. Ellington, Morningside Academy, North Charleston Elementary, Malcolm Hursey, Pinehurst, Matilda Dunston, and Midland Park Primary School. Midland Park and Matilda Dunston do not have students and teachers in eligible grade levels.

II describes the data used to estimate teacher and school value added. Chapter III provides details of the statistical methods used to estimate teacher and school value added. The appendix contains information regarding the characteristics of teachers and students included in the analysis, and diagnostic information from value-added models.

A. Using value added to measure performance

Value added is a method of measuring teacher effectiveness that seeks to isolate how much a teacher contributes to student achievement apart from any confounding factors outside the teacher's control. We also average teacher value added within a school to measure school-wide value added. To measure the performance of CCSD teachers, we use test scores and other data in a statistical model designed to capture the achievement of students attributable to a given teacher, compared to the progress the students would have made with the average teacher. Known as a "value-added model" because it seeks to isolate the teacher's contribution from other factors, this method has been developed and employed by a number of prominent researchers (Meyer 1997; Sanders 2000; McCaffrey et al. 2004; Raudenbush 2004; Hanushek et al. 2007) and has been used as one of several measures to evaluate the performance of schools and/or teachers in many districts, including Chicago, Houston, Los Angeles, New York City, and Washington, DC. Spurred in some cases by the federal government's Race to the Top initiative, whole states have adopted value-added models to measure teacher performance, including Florida, New York, Oklahoma, Pennsylvania, and Tennessee.

Value-added modeling is motivated by the idea that we can use test scores to measure student learning, and then make inferences about teacher effectiveness based on how much a teacher's students learned. Rather than simply using the proportion of students who meet a certain test score threshold, as was done under the original versions of the No Child Left Behind Act, value-added models take advantage of multivariate regression. This technique accounts for prior achievement and student characteristics, isolating the effects of teachers from the components of student achievement that are beyond a teacher's control. The basic approach of this value-added model is to predict the test scores that each student would have obtained with the average CCSD teacher and then compare the average actual scores of a given teacher's students to the average predicted scores. The difference between these two scores—how the students actually performed with a teacher versus how they would have performed with the average CCSD teacher—represents the teacher's "value added" to student achievement. For example, suppose that a 6th-grade math teacher has a class of students who, given their background characteristics, such as poverty status, disability status, and test scores on the 5th-grade math and reading tests (or "pre-tests"), typically end the year five points above the district-wide average on the 6th-grade math test (or "post-test"). The value-added model derives a relative measure of the teacher's effectiveness by comparing the average student post-test score to the average predicted score. In this example, if the class post-test average is exactly five points above average, the value-added model will identify the teacher as an average performer. If the post-test average exceeds this standard, the teacher will be identified as above average, and if the average is less than the standard, the teacher will be considered below average. Because a value-added model accounts for students' initial performance and other background characteristics, it allows teachers to be identified as high performers, regardless of whether their students were low or high performing at baseline.

B. A value-added model for CCSD

Although conceptually straightforward, producing value-added estimates for CCSD requires (1) the assembly of an analysis file of data from multiple sources and (2) the design of a value-added model that addresses several layers of complexity within CCSD's educational context to measure teachers' performance accurately and fairly. We briefly describe the key elements of the analysis file below (described more fully in Chapter II) and then provide an overview of the steps used to estimate value added (with details in Chapter III).

We estimate the performance of teachers in CCSD using a value-added model based on the ACT Aspire test in ELA and math, the Palmetto Assessment of State Standards (PASS) tests in science and social studies, and end-of-course (EOC) exams in algebra, English, and biology. We measure teacher effectiveness in each subject separately. Our value-added models differ by subject to account for different patterns of prior test score availability. For the math, ELA, algebra, and English subjects, all students have prior-year scores from the PASS test in math and ELA. However, because students in grades 3, 5, 6, and 8 in the 2013–2014 school year took either the social studies or science exam, but not both, the value-added models for social studies, science, and biology are somewhat different. In addition, 10th-grade students in the biology subject do not have prior-year PASS tests.

Value-added estimates will only be reported for eligible teachers in BRIDGE pilot schools. However, in order to compare these teachers to the average CCSD teacher, it is necessary to include all eligible teachers in CCSD in the analysis. To enable us to match teachers to students accurately, eligible teachers in BRIDGE pilot schools have participated in a process known as roster verification, in which they indicated whether and for how long they taught the students listed on their administrative rosters. We use roster-verified lists to identify whether a teacher taught a given student and, if so, for what proportion of the school year. Because teachers in other schools have not participated in roster verification this year, we derive information on which students were taught by those teachers from administrative course data.

After constructing the analysis file, we estimate the value-added model using four steps, each of which addresses a different conceptual challenge. The following steps are explained in detail later in the report:

1. **Use multiple regression.** We use multiple regression, a statistical technique that allows us simultaneously to account for a group of background factors, to avoid holding teachers accountable for factors outside their control. We account for a set of student characteristics that could be related to performance on the PASS test in the 2014–2015 school year (the post-test). These characteristics include a student's PASS and EOC test scores from the 2013–2014 school year (the pre-tests), poverty status, limited English proficiency, learning-disability status, whether the student transferred schools during the year, and attendance.[4]

[4] A student's race/ethnicity or gender may be correlated with factors that both affect test scores and are beyond a teacher's control. Preliminary results showed a high correlation between value-added measures estimated with and without race/ethnicity or gender. This suggests that either (1) other characteristics included in the value-added models capture most of the factors affecting test scores that are correlated with race/ethnicity and gender or (2) students of different races, ethnicities, or genders are more or less randomly assigned to teachers. As a result, Charleston decided not to account for race/ethnicity or gender in the value-added model.

Accounting for these characteristics, we obtain an estimate of each teacher's effectiveness. The estimate is approximately equal to the difference between the average actual post-test score of a teacher's students and the average predicted score of those students based on their characteristics. For students and teachers in BRIDGE pilot schools, we weight each student's contribution to a teacher's score by the proportion of time the student was taught by that teacher, according to the roster verification process. For students taught by multiple teachers during the year, each teacher receives credit for the students' achievement based on the amount of instructional time that teachers reported through the roster verification process.

2. **Account for imperfect measurement of pre-tests.** A student's performance on a single test is an imperfect measure of ability. If we do not account for this, it will reduce the predictive power of the pre-tests on end-of-year achievement, which in turn can infect teachers' value-added estimates. To avoid this problem, we compensate for imperfect measurement in pre-test scores by employing a statistical technique that uses published information on the reliability of the PASS and EOC exams.

3. **Compare teachers across grades.** Value-added estimates from different grade levels are not necessarily comparable to each other. We therefore take steps to place teachers of different grades on a common scale. First, we adjust grade-level estimates so that the average teacher in each grade receives the same score. Second, we multiply the estimates by a grade-specific conversion factor to equalize the dispersion of teacher value-added estimates by grade.[5] This procedure ensures that teachers of all grade levels have a similar opportunity to be identified as high performers.[6]

4. **Account for imprecisely estimated measures based on few students.** Value-added estimates for teachers with fewer students will be less likely to pinpoint the teachers' true effectiveness. By virtue of guessing some answers correctly or incorrectly, some students may over-perform on exams, whereas others will underperform. With enough students for a given teacher, this balances out; however, teachers with very few students are more likely to receive a very high or very low effectiveness measure by chance than teachers with many students (Kane and Staiger 2002). We reduce the possibility of such spurious results by (1) not reporting estimates for teachers with fewer than 10 students and (2) using a statistical technique that combines the effectiveness measure of a particular teacher (from step 3) with the overall average to produce a final value-added estimate (Morris 1983). We rely more heavily on a default assumption of average effectiveness for teachers with few students or with students whose achievement is most difficult to predict with a statistical model.

[5] For teachers with students in more than one grade, we take a student-weighted average of their grade-specific value-added estimates

[6] To compare schools with different grade configurations, we apply a similar strategy. We transform each grade-level estimate within a school into generalized PASS/ACT Aspire points and then average the grade-level estimates across grades to arrive at a composite value-added estimate for the school.

C. Caveats

It is important to recognize the limitations of any performance measures, including those generated by a value-added model. Below, we discuss three caveats that are especially important for interpreting and using the results of a value-added model like the one created for CCSD:

1. **Estimation error.** The value-added measures are estimates of a teacher's performance based on the available data and the value-added model used. As with any performance measure based on limited information, there is uncertainty in the estimates produced. Therefore, it would not be appropriate to make fine-grained distinctions between two teachers with similar value-added estimates. We quantify the precision with which the measures are estimated by reporting the upper and lower bounds of a confidence interval of performance for each teacher—the range within which differences between this teacher and his or her colleagues are most likely to be due to chance differences rather than actual underlying differences in performance.

2. **Unmeasured differences between students.** A value-added model uses statistical techniques to account (or "control") for differences in student performance based on documented sources of information about students, such as their prior-year test score or free lunch eligibility. However, the model cannot account for differences in student performance that arise from sources that are not explicitly measured. For example, we cannot account for disruptive events that occur at home on the day before the test, as this information is not reported and catalogued. Similarly, we lack direct measures of intrinsic motivation on the part of students or their families. For this reason, policymakers may have concerns about how to interpret value-added estimates. For example, one concern might be that teachers at certain schools would be unfairly rewarded if especially motivated parents choose schools for their children in ways that are not accounted for by the student characteristics in the value-added model. Similarly, if the assignment of students to teachers within schools is based on factors for which we lack data—for example, pairing difficult-to-teach students with teachers who have succeeded with similar students in the past—a value-added model might unfairly penalize these teachers because it cannot statistically account for such factors. A related concern is that teacher-level value added might reflect the efficacy of school inputs, such as the leadership of the principal or a consistent, school-wide student behavior code.

   Partly for these theoretical reasons, two research organizations have cautioned against using value-added models (ASA 2014; AERA 2015); however, researchers have shown that there is little empirical support for many of these concerns (Chetty et al. 2014). Empirical work in experimental settings (Kane and Staiger 2008; Kane et al. 2013) and quasi-experimental settings (Chetty et al. 2014a) suggests these factors do not play a large role in determining teacher value added. Using data from six large school districts, Kane et al. (2013) compared (1) the difference in value-added measures between pairs of teachers based on a typical situation in which principals assign students to teachers and (2) the difference in student achievement between the teachers in the following year, when they taught classrooms that were formed by principals but then randomly assigned to the teachers. The authors found that the differences between teachers' value-added estimates before random assignment were an exceptionally strong predictor of achievement differences when classrooms were assigned randomly. Chetty et al. (2014a) complemented these findings, using longitudinal data from a large urban district. They showed that the value added of teachers who change schools persists in their new settings. This suggests that value added reflects a teacher's performance in the classroom, not school factors or some unmeasured characteristic of the teacher's students.

3. **Single versus multiple measures.** Value-added estimates measure a teacher's contribution to student achievement based on standardized test scores. Additional measures of teacher effectiveness may improve the predictive power of teacher evaluation systems (Kane et al. 2013) or the future effectiveness of teachers (Taylor and Tyler 2011). CCSD uses multiple measures of teacher effectiveness. In addition to value added, these include the COT, a measure designed to capture effective lesson planning and instructional delivery, and ADEPT, the state teacher evaluation measure. Starting in 2014–2015, CCSD will also include student learning objectives as another measure of teachers' impact on student learning.

D. Key changes to the 2014–2015 value-added model

1. **Changes to post-tests for the math and ELA subjects for students in grades 4 through 8.** The 2015 exam for the math and ELA subjects switched from PASS to ACT Aspire. In addition, we use two separate ACT Aspire tests for the ELA subject (the English and reading subtests).[7]

2. **Estimating value added for students and teachers in algebra, English, and biology.** At the request of CCSD, we generate value-added estimates of teachers in courses in which students take EOC exams.[8] Students take these courses across a range of grade levels and take the same exam at the end of the course regardless of grade level. We do not calculate school value added for these subjects, because students and teachers included in these models represent only a small fraction of those in their respective schools.

3. **Change in method for converting value-added estimates into IVA scores.** To incorporate the algebra, English, and biology scores into the system for converting value-added estimates into IVA scores, we combine value-added estimates from math and algebra into one score, reading and English into one score, and science and biology into one score.

4. **Using attendance data from the current year, rather than the prior year, as a student characteristic.** At the request of CCSD, we account for student attendance in the current year, rather than student attendance in the prior year, in our value-added model.[9]

[7] The ACT Aspire exam includes three subtests in the ELA subject: English, reading, and writing. It also includes one composite scale score that incorporates all three subtests, which ACT Aspire refers to as "ELA." CCSD requested that we not include the writing subtest. As a result, we used only the English and reading subtests, since the ELA composite score includes the writing subtest.

[8] CCSD also gives a U.S. history EOC exam. However, most students who take this exam are in 11th grade; as a result, most students do not have any relevant pre-tests. Since accounting for students' prior achievement is key in estimating teacher effectiveness, we do not estimate a model for U.S. history.

[9] Using the 2013–2014 value-added models, we tested whether using current-year versus prior-year attendance made a significant difference in teacher estimates. The results showed that any differences were very small. Based on these results, CCSD chose to use current-year attendance in the 2014–2015 models.

In this chapter, we review the data used to generate the value-added measures. We discuss the standardized assessments used in CCSD schools, the data on student background characteristics, and how we calculate the amount of time that students spent with more than one teacher. We also provide an overview of the roster verification process that allows eligible teachers in BRIDGE pilot schools to verify whether and for what portion of the year they taught students.

A. Teacher, school, and student lists

CCSD provided an official, comprehensive list of schools with eligible teachers. Teacher and student lists from BRIDGE pilot schools came from the roster verification system, whereas those in other schools came directly from CCSD administrative course data. In general, only regular education teachers were eligible to receive value-added estimates. For the EOC model, CCSD provided a list of courses, by subject, that were eligible to take the EOC exams. Only students and teachers in these courses are included in the EOC model.

B. PASS, ACT Aspire, and EOC test scores

When estimating the effectiveness of CCSD teachers, we include elementary and middle school students in the model for a given subject if they have a test score in that subject from 2015 (the post-test) and a test score from the previous grade in math and ELA in 2014 (the pre-test). In most grades, students will have a pre-test score in math, ELA, and either social studies or science, so we do not require that students have social studies or science scores. That is, for the models that estimate social studies, science, and biology value-added measures, we do not require that students have a same-subject pre-test score. We found, empirically, that using math and ELA pre-tests for all students, plus same-subject pre-test scores for the students who had them, was sufficient to produce reasonably precise value-added estimates for teachers.[10] We exclude students who repeated or skipped a grade because they lack pre-test and post-test scores in consecutive grades and years. In addition, CCSD has decided to only include students who were enrolled on both the 45th and 135th days of the school year. We report estimates only for teachers who taught 10 or more students over the course of the year in at least one subject in a BRIDGE pilot school.

To obtain estimates of school effectiveness that include data on all students who attend a school and have taken the appropriate post-tests and pre-tests, we estimate the value-added model using all students in grades 4 through 8, even those not linked to an eligible teacher in the roster file or administrative course data.

In the subjects pertaining to students and teachers in grades 4 through 8, test scores may be meaningfully compared only within grades and within subjects. Therefore, before using the test scores in the value-added model, we create subject- and grade-specific standardized scores by subtracting the mean and dividing by the standard deviation within a subject-grade combination.

[10] The alternative would have been to exclude students from the social studies and science models unless they had same-subject pre-tests. This would have resulted in very few students being included in the calculations, and unreliable value-added estimates.

After calculating the value-added estimates, we then convert the estimates back to the PASS or ACT Aspire scale by multiplying them by the subject- and grade-specific standard deviation.

C. Student background data

We use data provided by CCSD to construct variables used in the value-added models to account for the following student background characteristics:

- Pre-test in math and ELA
- Pre-test in the same subject for science and social studies models (when available)
- Free lunch eligibility
- Reduced-price lunch eligibility
- English as a second language status
- Existence of a specific learning disability
- Existence of other types of disabilities requiring special education
- Whether the student transferred schools during the school year
- Proportion of days the student attended school during the current year
- Whether the student took the same-subject EOC exam in the prior year (EOC models only)
- Grade level (EOC models only)

The biology EOC model contains students in grades 9 and 10. All 9th-grade students have math and ELA pre-test scores and about half also have science test scores. For those in 10th grade, we used only students who had an English EOC pre-test score. Some of these students also had an algebra EOC pre-test score, which we used when available.

In the EOC models, we accounted for whether or not a student took that same EOC exam in the prior year, snce that may impact their achievement on the post-test. However, we did not account for this pre-test score in these models because too few students had these pre-tests.

Attendance is a measure of student motivation. We use current-year—rather than prior-year—attendance, at the request of CCSD. Attendance is a continuous variable that could range from zero to one. Aside from pre-test variables, the other variables are binary ones taking the value zero or one. We impute data for students who are included in the analysis file but who have missing values for one or more student characteristics. Our imputation approach uses the values of non-missing student characteristics to predict the value of the missing characteristic. We do not generate imputed values for pre-tests; we drop from the analysis file any students with missing pre-test scores in math or ELA. In addition, for the biology model, we drop 10th-grade students who do not have an English EOC pre-test score. For the EOC models, we do not include students in grade levels that had fewer than 30 students with the data needed for inclusion in the model for a particular subject.

D. Teacher dosage

For students who are taught by a combination of teachers, we apportion their achievement among multiple teachers. We refer to the fraction of time the student was enrolled with each teacher as the "dosage," which ranges from 0 to 100 percent.

1. Teacher dosage at BRIDGE pilot schools

BRIDGE pilot schools underwent a process called roster verification in spring of the 2014–2015 school year. Eligible teachers in pilot schools received lists of students who appeared on their course rosters. Teachers indicated whether they taught each subject to each student and, if so, the proportion of time they taught the student during each month prior to the testing window. For example, if a student spent half of the instructional time each week in an eligible teacher's classroom learning math and the other half in another classroom with a special education teacher while other students learned math with the eligible teacher, the student was recorded as having spent 50 percent of instructional time with the eligible classroom teacher. In recording the proportion of time spent with a student in a given class and subject, teachers chose from qualitative responses—None, Some, Shared, Most, All—that were translated to numeric responses of 0, 25, 50, 75, and 100 percent, respectively. If a teacher claimed a student for less than 100 percent of the time in any month, the teacher was not responsible for naming other teachers who taught that student. Teachers could also add students to their rosters. Principals were responsible for verifying the accuracy of the rosters that teachers submitted.

We use the verified class rosters to construct teacher-student links. If the roster verification data indicate that a student had one math or reading teacher for the entire year, we set the teacher dosage equal to one. If a student changed teachers from one term to another, we determine the amount of time the student spent with each teacher, subdividing the dosage between teachers accordingly. If two or more teachers claimed the same students at 100 percent during the same time, we assign each teacher full credit for the shared students. This reflects an assumption that solo-taught and co-taught students contribute equally to teachers' value-added estimates; we therefore do not subdivide dosage for co-taught students. Finally, we track and report on the time a student spent with any teachers not recorded in the verified class rosters.

2. Teacher dosage at other schools

At non-pilot schools, we rely on administrative course data provided by CCSD. These data do not capture instances of students switching from one teacher to another. Instead, they may show that a student had two different math teachers in one year. Without being able to determine whether the student had these teachers simultaneously or consecutively, we treat these as cases of simultaneous co-teaching of the student for the entire year. Following the same procedure as in the pilot schools, we do not subdivide dosage for these students.

This page has been left blank for double-sided copying.

A. Regression estimates

We have developed linear regression models to estimate effectiveness measures for both schools and teachers. Within each subject, we use the same regression model for both analyses because our school value-added estimates are simply aggregates of the teacher estimates within each school.

After assembling the analysis file, we estimate regressions separately for each subject. For math, ELA, science, and social studies, we estimate separate regression for students in elementary grades (4 and 5) and middle school grades (6 to 8) to allow for the possibility that the associations between student characteristics and post-test scores are different across these grade spans. For example, the relationship between achievement and English language learner status in 4th grade could be different than in 8th grade. For algebra, English, and biology, we estimate only one regression per subject, since each subject contains only two grade levels.

There are a number of differences in test-taking patterns across subjects; these require regression models that differ somewhat from subject to subject. Section A describes the regression models used for each subject and Sections B through D describe procedures performed on value-added estimates separately within each subject. Section E describes how we translate value-added estimates into IVA scores and combine across subjects.

1. The regression model for math

In the following equation, the post-test score depends on prior achievement, student background characteristics, teacher-student links, and unmeasured factors.

$$(1)\quad Y_{itg} = \lambda_g^m M_{i(g-1)} + \omega_g^m E_{i(g-1)} + \boldsymbol{\alpha^{m\prime} X_i} + \boldsymbol{\eta^{m\prime} T_{itg}} + \varepsilon_{itg}^m,$$

where Y_{itg} is the post-test score for student i taught by teacher t in grade g, $M_{i(g-1)}$ is the math pre-test score for student i in grade $g-1$ during the previous year, and $E_{i(g-1)}$ denotes the pre-test score in ELA. Y_{itg} represents math post-test scores. The pre-test scores capture prior inputs into student achievement, and the associated coefficients, λ_g^m and ω_g^m, vary by grade.[11] The vector $\boldsymbol{X_i}$ denotes the control variables for individual student background characteristics. The coefficients on these characteristics, $\boldsymbol{\alpha^m}$, are constrained to be the same across all grades within a grade span.[12]

[11] The m superscript serves to denote that these are the coefficients for the math model. Other models contain different superscripts to illustrate that they may differ by subject.

[12] Constraining the coefficients to be the same across grades within a grade span allows them to borrow strength from one another, so that small numbers of students who have a given characteristic within one grade will not lead to unstable estimates of the coefficient associated with that characteristic.

The vector $\boldsymbol{T_{itg}}$ includes a grade-specific variable for each teacher and includes a variable for a catchall teacher in each grade and school to account for student dosage that cannot be attributed to a particular eligible teacher.[13] A student contributes one observation to the model for each teacher to whom he or she is linked, based on the roster verification process for students in BRIDGE pilot schools and on administrative data for students in other schools. Each teacher-student observation has one nonzero element in $\boldsymbol{T_{itg}}$. Value-added estimates for each teacher in each grade that he or she taught are contained in the coefficient vector $\boldsymbol{\eta^m}$.

To account for multiple observations on the same student, we estimate the coefficients by using weighted least squares rather than ordinary least squares. In this method, the teacher-grade variables in $\boldsymbol{T_{itg}}$ are binary and we weight each teacher-student combination by the teacher dosage associated with that combination. We address the correlation in the error term, ε_{itg}, across multiple observations of the same student by using a cluster-robust sandwich variance estimator (Liang and Zeger 1986; Arellano 1987) to obtain standard errors that are consistent in the presence of both heteroskedasticity and clustering at the student level. This method—using the student-teacher link as the unit of observation and including the teacher dosage as a weight—is known as the full roster method (Hock and Isenberg 2012).

The regression will produce separate value-added coefficients for each teacher-grade combination. We aggregate the estimated coefficients into a single measure for each teacher (see Section C below).

2. The regression model for ELA

In ELA, there are two post-tests: the reading and English subtests from the ACT Aspire exam. To incorporate both of these tests, we estimate a "stacked" model, where each student-teacher combination within the ELA subject has two observations, rather than one. The two observations are identical, except that one uses the English score for the post-test, whereas the other used the reading score. We estimate one set of value-added estimates, where every estimate is based on both scores. To ensure that different scaling of these tests does not impact our value-added estimates, we standardize each score to have a mean of 0 and standard deviation of 1. We use value-added estimates from this scale to produce the IVA scores described in Section E. To produce value-added estimates that are in the same scale as the post-tests for reporting purposes, we multiply each teacher-grade specific estimate by the average standard deviation of the two post-tests for the relevant grade, and then followed the procedures described in Sections C and D below.

The regression model for ELA is very similar to the one for math, except that for every pre-test variable and coefficient in math, there are two for ELA: one for the reading post-test and one for the English post-test. Since these two subtests may have slightly different properties, and cover different skills and/or content, the relationship between each of them and the pre-test may be somewhat different. To account for this, we allow the relationship between the post-test and

[13] Although CCSD will not use value-added estimates for teachers in BRIDGE schools with fewer than 10 students, we include teachers with 5 to 9 students in the regression because maintaining more teacher-student links may more accurately estimate the coefficients on the covariates. If a teacher has fewer than five students in a grade, we reallocate those students to a grade-specific catchall teacher.

the pre-test to vary by the post-test type (English or reading), by estimating two sets of regression coefficients. We also include an indicator for whether the post-test is in English or reading. The regression model is the following:

$$(2)\quad Y_{itgs} = \lambda_{eg}^{r} M_{ei(g-1)} + \omega_{eg}^{r} E_{ei(g-1)} + \lambda_{rg}^{r} M_{ri(g-1)} + \omega_{rg}^{r} E_{ri(g-1)} + \boldsymbol{\alpha}^{r\prime}\boldsymbol{X_i} + \boldsymbol{\eta}^{r\prime}\boldsymbol{T_{itg}} + \kappa P_s + \varepsilon_{itgs}^{r},$$

where the outcome, Y_{itgs}, now differs by subtest s. $\lambda_{eg}^{r} M_{ei(g-1)}$ corresponds to the math coefficient and pre-test score for the set of observations with the English subtest as the outcome (denoted by the e subscript), and $\omega_{eg}^{r} E_{ei(g-1)}$ corresponds to the ELA coefficient and pre-test score for this set of observations. $\lambda_{rg}^{r} M_{ri(g-1)}$ corresponds to the math coefficients and pre-test score for the set of observations with the reading subtest as the outcome, and $\omega_{rg}^{r} E_{ri(g-1)}$ corresponds to the ELA coefficient and pre-test score for these observations. P_s is an indicator for having the English subtest as the outcome, where κ is the associated coefficient.

3. Models for algebra and English

The models for algebra and English are very similar to the model for math, with only two differences. The first is that in the algebra and English subjects, as well as in biology, students take the same test regardless of grade level. As a result, we do not estimate separate effects for each teacher-grade combination, but rather one effect for each teacher. That is, we use $\boldsymbol{T_{it}}$, which contains one indicator per teacher, rather than $\boldsymbol{T_{itg}}$. The second is that we include an indicator for each grade level. Other than these differences, accounting for grade level is unnecessary in the models for math, ELA, science, and social studies, because those models include separate variables for each teacher-grade combination. This implicitly accounts for each student's grade level. However, in the algebra, English, and biology models, we must account for grade level, because we do not have separate variables for each teacher-grade combination. Indeed, accounting for grade level is especially important in the EOC models because students in different grades, with different levels of preparation for the course, take the same post-test.

4. Models for social studies and science

In the 2013–2014 school year, all students took the social studies and science exams in 4th and 7th grade, but in the other grades from 3rd to 8th, students were randomly chosen to take either the social studies or science exam. This means that the value-added model must account for the fact that only half of the students will have a pre-test score in the same subject in three of the five grade levels. To address this, the models for social studies and science differ slightly from the model for math.

A model for students with same-subject pre-tests would have three pre-test variables and their corresponding coefficients: math, ELA, and the same-subject pre-test. A model for students without same-subject pre-tests would have only two: math and ELA. We need to estimate just one model to accurately measure teacher effectiveness. A key difference between these two hypothetical models is that the coefficients associated with the math and ELA pre-tests in the model for students who have same-subject pre-tests would be significantly lower than the coefficients associated with these pre-tests in the model for students who do not have same-subject pre-tests. This is because, after accounting for the same-subject pre-test, there is less variation left in the post-test to be explained by the math and ELA pre-test coefficients. Thus, whether or not the same-subject pre-test is used alters the coefficients on the math and ELA pre-tests. The result is that we need to estimate two sets of coefficients for the math and ELA pre-tests, one for students who have the same-subject pre-test and one for those who do not.[14]

The model for science and social studies is:

$$(3)\quad Y_{itg} = \lambda_{1g}^{s} M_{1i(g-1)} + \omega_{1g}^{s} E_{1i(g-1)} + \lambda_{2g}^{s} M_{2i(g-1)} + \omega_{2g}^{s} E_{2i(g-1)} + \varphi_{g}^{s} S_{i(g-1)} + \boldsymbol{\alpha^{s\prime} X_i} + \boldsymbol{\eta^{s\prime} T_{itg}} + \varepsilon_{itg}^{s}.$$

Y_{itg}, $\boldsymbol{\alpha^{s\prime} X_i}$, $\boldsymbol{\eta^{s\prime} T_{itg}}$, and ε_{itg}^{s} are parallel to equation (1). $\lambda_{1g}^{s} M_{1i(g-1)}$ represents the math test score and associated coefficient for students who *do* have a pre-test in the same subject as the post-test, and $\omega_{1g}^{s} E_{1i(g-1)}$ represents the ELA test score and associated coefficient for these students. $\lambda_{2g}^{s} M_{2i(g-1)}$ represents the math test score and associated coefficient for students who *do not* have a same-subject pre-test, and $\omega_{2g}^{s} E_{2i(g-1)}$ represents the ELA test score and associated coefficient for these students. $S_{i(g-1)}$ represents the same-subject pre-test score for students who have a same-subject pre-test, and is set to 0 for students who do not have it. In 5th and 8th grades, $\lambda_{2g}^{s} M_{2i(g-1)}$ and $\omega_{2g}^{s} E_{2i(g-1)}$ will both be zero, since all students in these grades will have a same-subject pre-test.

5. Model for biology

The model for biology, like the models for algebra and English, also uses only one variable for each teacher, rather than for each teacher-grade combination (that is, $\boldsymbol{T_{it}}$ rather than $\boldsymbol{T_{itg}}$). Like those models, it also accounts for student grade level. However, similar to the models for science and social studies, it does not require that every student have each pre-test score. Although all 9th-grade students have math and ELA pre-tests, only some 9th-grade students have a science pre-test. Likewise, all 10th-grade students have an English pre-test, but only some have an algebra pre-test. Following the same reasoning that motivates the social studies and science models, we estimate two sets of coefficients for the math and ELA pre-tests, one set for students with the science pre-test and one set for students without it. Similarly, we estimate two English pre-test coefficients, one for students with the algebra pre-test and one for those without it. The model for biology is:

$$(4)\quad Y_{itg} = \lambda_{1}^{b} M_{1i(g-1)} + \omega_{1}^{b} E_{1i(g-1)} + \lambda_{2}^{b} M_{2i(g-1)} + \omega_{2}^{b} E_{2i(g-1)} + \varphi^{b} S_{i(g-1)} + \gamma_{1}^{b} En_{1i(g-1)} + \gamma_{2}^{b} En_{2i(g-1)} + \vartheta^{b} A_{i(g-1)} + \boldsymbol{\alpha^{b\prime} X_i} + \boldsymbol{\eta^{b\prime} T_{it}} + \varepsilon_{itg}^{b}.$$

[14] When we tested the model using data from the 2012–2013 school year, using a single coefficient for math and a single coefficient for ELA led to biased results. There were significant differences in the estimated coefficients associated with the math and ELA pre-test scores when we estimated the model separately across the two groups of students. For students with a same-subject pre-test score, this test explained much of the variation in post-test scores, so the math and ELA pre-tests consequently explained less of the variation than when students lacked a same-subject pre-test score. Combining the estimates into one coefficient led them to be estimated as an average of two different relationships, which resulted in inaccurate results for both groups of students.

The first five terms are parallel to the science and social studies models. The first two are the math and ELA pre-test scores and coefficients for students who have the science pre-test, whereas the next two terms represent pre-test scores and coefficients for students who do not. The fifth, $\varphi_g S_{i(g-1)}$ is the science pre-test and corresponding coefficient, where the pre-test is set to 0 for students who do not have it. $\gamma_1^b En_{1i(g-1)}$ and $\gamma_2^b En_{2i(g-1)}$ represent the English EOC pre-test scores and coefficients for students who do and do not have an algebra EOC pre-test, respectively. $\vartheta^b A_{i(g-1)}$ represents the algebra pre-test score and coefficient, where the pre-test score is 0 for students who do not have it. None of the coefficients in this model requires estimation separately by grade, because the math, ELA, and science pre-tests exist only for 9th graders, and the algebra and English pre-tests exist only for 10th graders.

B. Measurement error in the pre-tests

We correct for measurement error in the pre-tests by using grade-specific reliability data provided by CCSD. As a measure of true student ability, standardized tests contain measurement error, causing standard regression techniques to produce biased estimates of teacher or school effectiveness. To address this issue, we implement a measurement error correction based on the reliability of the PASS tests. By netting out the known amount of measurement error, the errors-in-variables correction eliminates this source of bias (Buonaccorsi 2010).

Correcting for measurement error requires a two-step procedure. In the first step, we use a dosage-weighted errors-in-variables regression using equations (1) and (2) to obtain unbiased estimates of the pre-test coefficients for each grade. We use the reliabilities associated with the 2014 PASS exams and, for 10th graders in the biology model, the 2014 EOC exams. We then use the measurement-error-corrected values of the pre-test coefficients to calculate an adjusted post-test score for each student. Finally, we regress these adjusted scores on the student characteristics other than pre-tests and the teacher dosage variables in equations (1) and (2). This second-stage regression is necessary because it is not computationally possible to simultaneously account for correlation in the error term across multiple observations and apply the numerical formula for the errors-in-variables correction. The expression for the other models contains different pre-test variables and coefficients than the expression for the math and ELA models, but the procedures are analogous. For simplicity, we show the procedure for the math model. The adjusted post-test score is expressed as:

(3) $$\hat{G}_{itg} = Y_{itg} - \hat{\lambda}_g M_{i(g-1)} - \widehat{\omega}_g E_{i(g-1)},$$

and represents the student post-test outcome, net of the estimated contribution attributable to the student's starting position at pre-test.

In the second step, we use the adjusted post-test as the dependent variable in a single equation expressed as:

(4) $$\hat{G}_{itg} = \boldsymbol{\alpha}_1' \boldsymbol{X}_i + \boldsymbol{\eta}_1' \boldsymbol{T}_{itg} + \varepsilon_{itg}$$

We obtain the grade-specific estimates of teacher effectiveness, $\hat{\boldsymbol{\eta}}_{\mathbf{1}}$, by applying the weighted least squares regression technique to equation (3). In model (4) we use cluster-robust standard errors to account for both heteroskedasticity and correlation in the error term between multiple observations for the same student.

This two-step method will likely underestimate the standard error of $\hat{\boldsymbol{\eta}}_{\mathbf{1}}$ because the adjusted post-test in equation (3) relies on the estimated value of λ_g, which implies that the error term in equation (4) is clustered within grades. This form of clustering typically results in estimated standard errors that are too small, because the second-step regression does not account for a common source of variability affecting all students in a grade. In view of the small number of grades, standard techniques of correcting for clustering will not effectively correct the standard errors (Bertrand et al. 2004). Nonetheless, with the large within-grade sample sizes, the pre-test coefficients are likely to be estimated precisely, leading to a negligible difference between the robust and clustering-corrected standard errors.

Teacher and school estimates. After estimating model (4), we have one estimate for each teacher in each grade level he or she taught. To calculate school value added (for math, ELA, social studies, and science only), we first aggregate each teacher-grade estimate within each school-grade. We do this by taking a weighted average of teacher-grade estimates, where the weights are the teacher dosage associated with a given teacher-grade estimate divided by the total teacher dosage across the teacher estimates associated with that school-grade. This will give us a value-added estimate of the effectiveness of that school-grade combination.[15]

Once we have an estimate for each school-grade, we proceed with the steps described below to combine these estimates across grades and apply shrinkage, resulting in one estimate for each school. For simplicity, the steps below refer to the procedure for teachers, but the procedure for schools is analogous.

C. Combining estimates across grades for math, ELA, science, and social studies

Both the average and the variability of value-added estimates may differ across grade levels, leading to a potential problem when comparing teachers assigned to different grades or comparing schools with different grade configurations. The main concern is that factors beyond teachers' control may drive cross-grade discrepancies in the distribution of value-added estimates. For example, the standard deviation of adjusted post-test scores might vary across grades as a consequence of differences in the alignment of tests or the retention of knowledge between years. However, we seek to compare each teacher to all others in the regression regardless of any grade-specific factors that might affect the distribution of gains in student performance between years.[16] Because we do not want to penalize or reward teachers simply for

[15] We calculate the standard error of this estimate using a method that incorporates the covariance between teacher estimates within the same school. This covariance could be important because a student may have two teachers within the same grade and school, causing those teacher estimates to be correlated.

[16] Because each student's entire dosage with eligible teachers was accounted for by teachers in a given grade, the information contained in grade indicators would be redundant to the information contained in the teacher variables. Therefore, it is not also possible to control directly for grade in the value-added regressions.

teaching in a grade with unusual test properties, we standardize grade-level estimates for schools and teachers so that each set of estimates is expressed in a common metric of "generalized" test score points. Below, we describe the procedure in the context of teacher measures; the procedure for school measures is analogous. We only use this procedure for the math, ELA, science, and social studies models, because the EOC models estimate one coefficient per teacher, rather than per teacher-grade combination. For the EOC models, we simply center the teacher estimates around 0 before continuing to the shrinkage step (Section D).

We standardize the effectiveness estimates so that the mean of the estimates is the same across grades. First, we subtract from each unadjusted estimate the average of all estimates within the same grade. To reduce the influence of imprecise estimates obtained from teacher-grade combinations with few students, we calculate the average using weights based on the number of students taught by each teacher.

We then divide the result by the adjusted standard deviation of the estimates within the same grade. The adjusted standard deviation removes estimation error to reflect the true dispersion of underlying teacher effectiveness. The unadjusted standard deviation of the value-added estimates will tend to overstate the true variability of teacher effectiveness; because the scores are regression estimates, rather than known quantities, the standard deviation will partly reflect estimation error. The extent of estimation error may differ across grades, and the resulting fluctuations in the unadjusted standard deviation of teacher scores could lead to over- or underweighting one or more grades when combining scores across grades. Scaling the estimates using the adjusted standard deviation will ensure that estimates of teacher effectiveness in each grade have the same true standard deviation by spreading out the distribution of effectiveness in grades with relatively imprecise estimates.[17] Our method of calculating the standard deviation of teacher effects also downweights imprecise individual estimates. Finally, we multiply by the square root of the teacher-weighted average of the grade-specific adjusted variances, obtaining a common measure of effectiveness on the generalized scale.

Formally, the value-added estimate expressed in generalized test score points is the following:

$$(5) \quad \hat{\delta}_{tg} = \frac{\left(\hat{\eta}_{tg} - \bar{\hat{\eta}}_g\right)}{\hat{\sigma}_g} \times \sqrt{\left(\frac{1}{K}\sum_g K_g \hat{\sigma}_g^2\right)},$$

where $\hat{\eta}_{tg}$ is the grade-g estimate for teacher t, $\bar{\hat{\eta}}_g$ is the weighted average estimate for all teachers in grade g, $\hat{\sigma}_g$ is the estimate of the adjusted standard deviation of teacher effectiveness in grade g, K_g is the number of teachers with students in grade g, and K is the total number of teachers.

[17] For teachers in grades with imprecise estimates, the shrinkage procedure, described in Section D, counteracts the tendency for these teachers to receive final estimates that are in the extremes of the distribution.

We calculate the error-adjusted variance of teacher value-added scores separately for each grade as the difference between the weighted variance of the grade-g teacher estimates and the weighted average of the squared standard errors of the estimates. The error-adjusted standard deviation, $\hat{\sigma}_g$, is the square root of this difference. We choose the weights based on the empirical Bayes (EB) approach outlined by Morris (1983). In this approach, the observed variability of the teacher value-added scores is adjusted downward according to the extent of the estimation error.

To combine effects across grades into a single effect $\hat{\delta}_t$ for a given teacher, we use a weighted average of the grade-specific estimates (expressed in generalized test score points). We set the weight for grade *g* equal to the proportion of students of teacher *t* in grade *g*, denoted as p_{tg}. We then compute the variance of each teacher's estimated effect by using:

$$(6)\quad Var(\hat{\delta}_t) = \sum_g (p_{tg})^2 Var(\hat{\delta}_{tg}),$$

where $Var(\hat{\delta}_{tg})$ is the variance of the grade-g estimate for teacher *t*. For simplicity, we assume that the covariance across grades is zero. In addition, we do not account for uncertainty arising because $\overline{\hat{\eta}_g}$ and $\hat{\sigma}_g$ in equation (6) are estimates of underlying parameters rather than known constants. Both decisions imply that the standard errors obtained from equation (6) will be slightly underestimated. Because combining teacher effects across grades may cause the overall average to be nonzero, we re-center the estimates on zero before proceeding to the next step.

D. Shrinkage procedure

To reduce the risk that teachers or schools, particularly those with relatively few students in their grade, will receive a very high or very low effectiveness measure by chance, we apply the EB shrinkage procedure, as outlined in Morris (1983), separately to the sets of effectiveness estimates for teachers and schools. Again, we frame our discussion of shrinkage in terms of teachers, but the same logic applies to schools. Using the EB procedure, we compute a weighted average of an estimate for the average teacher (based on all students in the model) and the initial estimate based on each teacher's own students. For teachers with relatively imprecise initial estimates based on their own students, the EB method effectively produces an estimate based more on the average teacher. For teachers with more precise initial estimates based on their own students, the EB method puts less weight on the value for the average teacher and more weight on the value obtained from the teacher's own students.

The EB estimate for a teacher is approximately equal to a precision-weighted average of the teacher's initial estimated effect and the overall mean of all estimated teacher effects.[18] Following the standardization procedure, the overall mean is zero, with better-than-average teachers having positive scores and worse-than-average teachers having negative scores. We therefore arrive at the following:

[18] In Morris (1983), the EB estimate does not exactly equal the precision-weighted average of the two values, due to a correction for bias. This adjustment decreases the weight on the estimated effect by a factor of (K – 3)/(K – 1), where K is the number of teachers. For ease of exposition, we have omitted this correction from the description given here.

$$(7)\quad \hat{\delta}_t^{EB} = \left(\frac{\frac{1}{\hat{\sigma}_t^2}}{\frac{1}{\hat{\sigma}_t^2}+\frac{1}{\hat{\sigma}^2}}\right)\hat{\delta}_t,$$

where $\hat{\delta}_t^{EB}$ is the EB estimate for teacher *t*, $\hat{\delta}_t$ is the initial estimate of effectiveness for teacher *t* based on the regression model (after combining across grades), $\hat{\sigma}_t$ is the standard error of the estimate of teacher *t*, and $\hat{\sigma}$ is an estimate of the adjusted standard deviation of teacher effects (purged of sampling error), which is constant for all teachers. The term $[(1/\hat{\sigma}_t^2)/(1/\hat{\sigma}_t^2 + 1/\hat{\sigma}^2)]$ must be less than one. Thus, the EB estimate always has a smaller absolute value than the initial estimate—that is, the EB estimate "shrinks" from the initial estimate. The greater the precision of the initial estimate—that is, the larger $(1/\hat{\sigma}_t^2)$ is—the closer $[(1/\hat{\sigma}_t^2)/(1/\hat{\sigma}_t^2+1/\hat{\sigma}^2)]$ is to one and the smaller the shrinkage in $\hat{\delta}_t$. Conversely, the smaller the precision of the initial estimate, the greater the shrinkage in $\hat{\delta}_t$. By applying a greater degree of shrinkage to less-precisely estimated teacher measures, the procedure reduces the likelihood that the estimate of effectiveness for a teacher falls at either extreme of the distribution by chance. We calculate the standard error for each $\hat{\delta}_t^{EB}$ using the formulas provided by Morris (1983). As a final step, we remove any teachers with fewer than 10 students and re-center the EB estimates on zero.

E. Translating value-added results to scores for BRIDGE

We convert value-added estimates to an individual value-added (IVA) score on a scale from 1.0 to 4.0, according to a method that CCSD determined in consultation with the senior leadership team. CCSD determined that the aggregate IVA score will be a weighted average of the subject-specific IVA scores, where the weights will be proportional to the number of students a teacher taught in each subject.[19] We provide CCSD with the original value-added estimates in each subject for teachers in BRIDGE pilot schools, percentile rankings for individual teachers compared to all CCSD teachers, IVA scores converted to a scale from 1.0 to 4.0, and aggregate IVA scores across all the subjects for each teacher. We report original school-wide value-added results for BRIDGE pilot schools separately by subject, and also convert these measures to a scale that fits with the principal evaluation framework developed for CCSD. The aggregate IVA score constitutes 35 percent of the BRIDGE composite measure of teacher effectiveness for eligible teachers in BRIDGE pilot schools.

The value-added estimates we report for teachers in each of the seven subjects are described above; however, we use a slightly different routine to generate the value-added estimates that are converted IVA scores. We calculate subject-specific IVA scores that combine across aligned subjects, where we align (1) math and algebra, (2) ELA and English, and (3) science and biology. In order to accommodate the fact that aligned subjects have very different scales, we use

[19] In the 2013–2014 school year, students took either the social studies or science exams in grades 5, 6, and 8. Since value-added estimates require a post-test score, the number of students that a teacher taught was expected to be about twice the number of students included in estimating a teacher's value-added score in these subjects and grade levels. CCSD decided to weight by the number of students taught, not the number of students included in the calculation, so that a teachers' value-added scores in social studies and science are not down-weighted due to the testing regime. However, in the 2014–2015 school year, students took the social studies and science exams in all grade levels, so this correction was not necessary. Consequently, CCSD decided to weight by the number of students included in the calculation for the 2014–2015 school year.

estimates on the scale of student standardized scores to combine across aligned subjects. Specifically, we calculate IVA scores using the following steps:[20]

1. We use the grade-specific value-added estimates from the math, reading, and science models that are on the scale of student standardized test scores.
2. We use the value-added estimates from the algebra, English, and biology models that are on the scale of student standardized test scores, and treat them as if they come from an additional grade level. That is, a teacher with grade-specific math estimates in 7th and 8th grade would be treated the same as a teacher with a math estimate in 8th grade and an algebra estimate.
3. We then follow all of the steps described in Sections C and D above to combine value-added estimates across grade levels and shrink the final combined estimates. We are then left with scores for the three “aligned” subjects: math-algebra, ELA-English, and science-biology, plus our scores from social studies.
4. We then convert these estimates to IVA scores using the procedure used in 2013–2014. That is, we scale the value-added estimates to have an average of 3 and a standard deviation of 0.75, and set the minimum value to 1 and the maximum to 4.
5. Finally, we average IVA scores across subjects for teachers with scores in more than one subject.

[20] Because the social studies subject does not have an aligned EOC subject, the following steps do not apply to social studies scores.

APPENDIX A

This page has been left blank for double-sided copying.

A. STUDENTS AND TEACHERS INCLUDED IN THE VALUE-ADDED MODELS

Not all students with a post-test in a given subject can be included in that subject's value-added model. Of the students in the models pertaining to grades 4 through 8 (math, ELA, social studies, and science), 82 to 88 percent of them are included in the model. Students can be excluded because (1) they do not have a math or ELA pre-test, (2) they skipped or repeated a grade, or (3) they are not linked to a teacher. Tables A.1 and A.2 show the total number of students who could have been included in the analyses, the reasons why we exclude students, and the total included in the models. The top row of these tables shows the total number of CCSD students who received the relevant test scores on the PASS, ACT Aspire, or EOC tests; students must have a 2015 test score to be included in the value-added models. The next three rows show the reasons we exclude students who have 2015 test scores from the analysis file. In general, we include a slightly lower percentage of students in the EOC models, as shown in the last row of Table A.2.

Table A.1. Reasons that students tested in 2015 were excluded from the analysis files (subjects pertaining to grades 4 through 8)

	Math		ELA		Social studies		Science	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Students with post-test scores	16,649	100.0	16,602	100.0	16,660	100.0	16,665	100.0
(1) Missing math/ELA pre-test scores	1,608	9.7	1,574	9.5	1,634	9.8	1,636	9.8
(2) Skipped or repeated a grade	151	0.9	151	0.9	127	0.8	126	0.8
(3) Student is not linked to a teacher	1,180	7.1	486	2.9	318	1.9	314	1.9
Total excluded	2,939	17.7	2,211	13.3	2,079	12.5	2,076	12.5
Total included in value-added model	13,710	82.3	14,391	86.7	14,581	87.5	14,589	87.5

Notes: Students are included in this table only if they could be linked to background characteristics. Students are excluded sequentially in the order presented and so do not count for more than one reason in this table.

Table A.2. Reasons that students tested in 2015 were excluded from the analysis files (EOC subjects)

	Algebra		English		Biology	
	Number	Percent	Number	Percent	Number	Percent
Students with post-test scores	2,196	100.0	2,709	100.0	1,854	100.0
(1) Missing pre-test scores	326	14.8	433	16.0	396	21.4
(2) Skipped or repeated a grade	2	0.1	6	0.2	13	0.7
(3) Student is not linked to a teacher	141	6.4	21	0.8	38	2.0
Total excluded	469	21.4	460	17.0	447	24.1
Total included in value-added model	1,727	78.6	2,249	83.0	1,407	75.9

Notes: Students are excluded sequentially in the order presented and so do not count for more than one reason in this table. In these subjects, students with EOC post-tests that are not enrolled in EOC-eligible courses are included in category 3.

The value-added models account for a variety of student background characteristics. Table A.3 illustrates the distribution of these characteristics for students included in the ELA value-added model. For all characteristics other than the percentage of days absent in the current year, the value represents the proportion of students with that characteristic. We show the table that corresponds to students included in the ELA value-added model because the values that correspond to the other subjects differ from those below by no more than three percentage points. All students in grades 4 through 8 take the PASS and ACT Aspire tests. Since the samples of students across the math, ELA, social studies, and science models overlap to a great degree, it is not surprising that the characteristics of students included in these value-added models are very similar.

Table A.3. Characteristics of students from the ELA value-added model

	Grades 4 and 5	Grades 6 to 8
	Average/ percent	Average/ percent
Eligible for free lunch	47.6	44.6
Eligible for reduced-price lunch	4.1	5.2
Ineligible for free or reduced-price lunch	48.4	50.2
English as a second language	7.1	5.8
Specific learning disability	4.7	5.4
Speech/language impairment	1.0	0.7
Other learning disability	3.6	3.3
Percent absent in current year	3.5	5.2
Overage for grade	8.5	12.8
Transferred schools during school year	3.3	2.7

Notes: The table shows percentages for students in the ELA value-added model. The characteristics of students in the math, social studies, and science value-added model differed from those in the ELA model by no more than three percentage points.

Table A.4 shows the characteristics of the students in the EOC value-added models, separately by subject. Since each subject contains a mix of students from varying grade levels, the samples of students across the EOC subjects do not overlap. In particular, the biology models contains mostly 10th-grade students, who have a different distribution of characteristics than the 8th- and 9th-grade students included in the algebra and English models.

Table A.4. Characteristics of students from the EOC value-added models

	Algebra	English	Biuology
	Average/ percent	Average/ percent	Average/ percent
Eligible for free lunch	36.5	38.4	30.8
Eligible for reduced-price lunch	6.9	5.4	3.9
Ineligible for free or reduced-price lunch	56.6	56.2	65.3
English as a second language	3.0	3.7	2.4
Specific learning disability	2.4	3.3	2.7
Speech/language impairment	0.2	0.1	0.2
Other learning disability	1.9	2.5	2.3
Percent absent in prior year	4.9	4.9	5.1
Overage for grade	10.3	11.8	12.3
Transferred schools during school year	1.6	2.8	2.1
Repeat EOC exam	7.9	1.6	0.2
Percentage of students in each grade			
Grade 8	47.7	6.4	0.0
Grade 9	52.3	93.6	29.9
Grade 10	0.0	0.0	70.1

Notes: The table shows percentages for students in the EOC value-added models.

One of the reasons to conduct roster verification (described in Chapter II, Section D) is to accurately capture teacher dosage in cases in which students have more than one teacher in the same subject. The majority of the teachers in BRIDGE pilot schools who participated in roster verification shared at least some of their students with other teachers, particularly in math and ELA. Table A.5 shows the extent of co-teaching among these teachers. In this table, “co-teaching” refers to teachers who taught the same student during the 2014–2015 school year, either during the same term or in separate terms. We do not show the analogous table for the EOC subjects, as there were no more than seven teachers in any of these subjects.

Table A.5. Teachers who received value-added estimates, by subject and extent of co-teaching (subjects pertaining to grades 4 through 8)

Percentage of students shared with another teacher in value-added model	Math		ELA		Social Studies		Science	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
None	9	16.4	2	2.8	15	31.3	14	30.4
1–10 percent	14	25.5	5	7.0	17	35.4	18	39.1
11–50 percent	16	29.1	28	39.4	6	12.5	6	13.0
51–99 percent	4	7.3	14	19.7	10	20.8	7	15.2
All students	12	21.8	22	31.0	0	0.0	1	2.2
Total	55	100.0	71	100.0	48	100.0	46	100.0

Notes: Teachers received estimates if they were linked to at least 10 eligible students.

A co-teacher is any teacher who shares at least one student with another teacher in the value-added model for the listed subject.

The table includes only teachers in BRIDGE pilot schools who received value-added estimates for 2014–2015.

B. STATISTICAL RESULTS FROM THE VALUE-ADDED MODELS

As explained in Chapter I, the value-added regressions account for student characteristics that could be related to performance on the 2015 tests. Accurate estimation of these student characteristics is essential for measuring teacher effectiveness fairly. Table A.6 shows the estimated association between student characteristics and the 2015 ACT Aspire tests (the post-test) for math and ELA in the teacher value-added regression models. The top panel of the table shows the average association between the pre-tests and the post-test. The bottom panel shows the association between a given student characteristic and the change in achievement on the 2015 exam (measured in points on the test). Although the estimates for the student background characteristics were common across grades within grade spans, we estimated a separate pre-test coefficient for each grade. Table A.7 displays all of the estimated associations described above for the value-added models for social studies and science, and Table A.8 does the same for the EOC subjects.

Although we measure value-added estimates in terms of points on the relevant 2015 test, we estimated the regressions themselves after standardizing the post-test and pre-test scores so that the mean and standard deviations for each subject-grade combination were 0 and 1, respectively. This means that the coefficients listed in the tables can be interpreted as student standard deviations of the post-test. For example, the coefficient on the first background characteristic in the first column of Table A.6 indicates that, holding all else equal, students ineligible for free or reduced-price lunch (that is, “paid” lunch students) score 0.08 standard deviations higher, on average, than students eligible for free lunch (the omitted category). Similarly, the first coefficient listed under the pre-test panel indicates that an increase of one standard deviation in the same-subject pre-test is associated with a 0.64 standard deviation increase in the post-test.

This page has been left blank for double-sided copying.

Table A.6. Coefficients on covariates in the math and ELA value-added models, by grade span

Variable	Math: Grads 4 and 5	Math: Grades 6 to 8	ELA: Grades 4 and 5	ELA: Grades 6 to 8
Pre-test scores (average coefficients)				
Same subject, all grades in span	0.64 (0.03)	0.68 (0.03)		
Opposite subject, all grades in span	0.16 (0.03)	0.08 (0.03)		
Same subject (English post-test), all grades in span			0.59 (0.02)	0.54 (0.03)
Opposite subject (English post-test), all grades in span			0.20 (0.02)	0.24 (0.02)
Same subject (reading post-test), all grades in span			0.68 (0.02)	0.62 (0.03)
Opposite subject (reading post-test), all grades in span			0.14 (0.02)	0.17 (0.02)
English post-test			0.01 (0.01)	-0.01 (0.01)
Individual student background characteristics				
Ineligible for free or reduced-price lunch	0.08 (0.03)	0.08 (0.02)	0.09 (0.02)	0.03 (0.02)
Eligible for reduced-price lunch	-0.01 (0.05)	0.04 (0.03)	0.04 (0.03)	0.05 (0.02)
English as a second language	0.10 (0.03)	0.16 (0.03)	0.00 (0.03)	0.02 (0.02)
Specific learning disability	0.20 (0.04)	0.06 (0.03)	-0.05 (0.03)	-0.20 (0.03)
Speech/language impairment	0.06 (0.08)	0.09 (0.07)	0.01 (0.06)	0.02 (0.08)
Other learning disability	0.03 (0.05)	0.03 (0.04)	-0.06 (0.04)	-0.12 (0.03)
Percent absent in current year	-0.34 (0.27)	-0.32 (0.16)	-0.33 (0.22)	-0.33 (0.15)
Missing percent absent value	-0.63 (0.10)	0.00 (0.06)	0.21 (0.08)	0.02 (0.05)
Overage	-0.04 (0.03)	-0.10 (0.02)	-0.05 (0.02)	-0.10 (0.02)
Transferred schools during the school year	0.00 (0.05)	-0.04 (0.04)	-0.02 (0.04)	-0.02 (0.04)

Notes: Standard errors are in parentheses.

Table A.6. (*continued*)

The reported coefficient estimates of pre-test scores represent averages of the coefficients estimated separately for the grades included in the grade span for the column. The associated standard errors similarly represent averages across grades. The standard errors therefore do not account for the variability of the estimates across grades. These numbers are presented for descriptive purposes only and should not be used to conduct statistical inference.

For the ELA model, we estimate separate pre-test coefficients for the English post-test versus the reading post-test.

The coefficient on English post-test is interpreted relative to the predicted post-test score for the reading post-test.

The coefficients on free/reduced-price lunch eligibility variables are interpreted relative to the predicted post-test score for students who are eligible for free lunch, the excluded category.

Table A.7. Coefficients on covariates in the social studies and science value-added models, by grade span

Variable	Social studies: Grads 4 and 5	Social studies: Grades 6 to 8	Science: Grades 4 and 5	Science: Grades 6 to 8
Pre-test scores (average coefficients)				
Same subject, all grades in span	0.62 (0.03)	0.58 (0.03)	0.60 (0.05)	0.54 (0.04)
Missing same subject pre-test	0.03 (0.02)	0.02 (0.01)	0.01 (0.02)	-0.02 (0.01)
ELA, all grades in span, missing same subject pre-test	0.60 (0.03)	0.52 (0.04)	0.48 (0.03)	0.40 (0.03)
Math, all grades in span, missing same subject pre-test	0.18 (0.03)	0.26 (0.03)	0.35 (0.03)	0.44 (0.03)
ELA, all grades in span, not missing same subject pre-test	0.19 (0.04)	0.15 (0.04)	0.18 (0.04)	0.14 (0.04)
Math, all grades in span, not missing same subject pre-test	0.04 (0.03)	0.11 (0.03)	0.13 (0.03)	0.19 (0.03)
Individual student background characteristics				
Ineligible for free or reduced-price lunch	0.09 (0.02)	0.06 (0.02)	0.08 (0.02)	0.11 (0.02)
Eligible for reduced-price lunch	0.08 (0.04)	0.00 (0.03)	0.06 (0.03)	0.02 (0.03)
English as a second language	0.11 (0.03)	0.09 (0.02)	0.04 (0.03)	0.04 (0.02)
Specific learning disability	0.04 (0.03)	-0.04 (0.02)	0.07 (0.03)	-0.02 (0.02)
Speech/language impairment	0.08 (0.08)	0.05 (0.07)	0.10 (0.08)	0.03 (0.08)
Other learning disability	-0.08 (0.04)	0.00 (0.03)	-0.03 (0.04)	-0.10 (0.03)
Percent absent in current year	-0.72 (0.25)	-0.58 (0.13)	-0.17 (0.24)	-0.41 (0.14)
Missing percent absent value	-1.01 (0.10)	0.08 (0.06)	-0.61 (0.10)	0.03 (0.06)
Overage	-0.04 (0.03)	0.01 (0.02)	-0.03 (0.03)	-0.03 (0.02)
Transferred schools during the school year	-0.05 (0.04)	0.02 (0.04)	-0.04 (0.04)	-0.03 (0.03)

Notes: Standard errors are in parentheses.

The reported coefficient estimates of pre-test scores represent averages of the coefficients estimated separately for the grades included in the grade span for the row. The associated standard errors similarly represent averages across grades. The standard errors therefore do not account for the variability of the estimates across grades. These numbers are presented for descriptive purposes only and should not be used to conduct statistical inference.

Table A.7. (*continued*)

The two opposite subject pre-test variables were required for each of ELA and math in the social studies and science models. One represents the coefficient for those students missing their same-subject pre-test, whereas the other represents the coefficient for those students who took the same-subject pre-test.

The coefficients on missing the same-subject pre-test are interpreted relative to the predicted post-test scores for students who are not missing the same-subject pre-test.

The coefficients on free/reduced-price lunch eligibility variables are interpreted relative to the predicted post-test score for students who are eligible for free lunch, the excluded category.

Table A.8. Coefficients on covariates in the algebra, English, and biology value-added models

Variable	Algebra	English	Biology
Pre-test scores			
Same subject, grades 8 and 9	0.44 (0.03)	0.47 (0.04)	0.35 (0.11)
Opposite subject, grades 8 and 9	0.13 (0.03)	0.11 (0.04)	
Missing science pre-test, grade 9			-0.01 (0.05)
Math, not missing same-subject pre-test, grade 9			0.09 (0.12)
Math, missing same-subject pre-test, grade 9			0.36 (0.07)
ELA, not missing same-subject pre-test, grade 9			0.29 (0.08)
ELA, missing same-subject pre-test, grade 9			0.30 (0.06)
Algebra pre-test, grade 10			0.22 (0.03)
Missing algebra pre-test, grade 10			-0.05 (0.04)
English, not missing algebra pre-test, grade 10			0.54 (0.04)
English, missing algebra pre-test, grade 10			0.82 (0.04)
Individual student background characteristics			
Ineligible for free or reduced-price lunch	0.06 (0.04)	0.16 (0.04)	0.10 (0.05)
Eligible for reduced-price lunch	0.06 (0.06)	0.04 (0.05)	-0.02 (0.09)
English as a second language	0.13 (0.07)	-0.12 (0.07)	-0.06 (0.14)
Specific learning disability	0.00 (0.11)	-0.19 (0.08)	-0.23 (0.12)
Speech/language impairment	-0.07 (0.17)	-0.05 (0.12)	-0.36 (0.22)
Other learning disability	-0.35 (0.11)	-0.26 (0.09)	-0.20 (0.13)
Percent absent in current year	-1.51 (0.36)	-0.34 (0.28)	-1.59 (0.40)
Missing percent absent value	0.02 (0.10)	0.03 (0.11)	0.09 (0.14)
Overage	-0.08 (0.05)	-0.12 (0.04)	-0.03 (0.05)

Table A.8. (*continued*)

Variable	Algebra	English	Biology
Transferred schools during the school year	-0.03 (0.10)	0.07 (0.08)	-0.11 (0.13)
Repeat EOC	0.30 (0.07)	0.16 (0.10)	-0.22 (0.13)
Grade 9	-1.03 (0.10)	-0.39 (0.12)	
Grade 10			-0.49 (0.05)

Notes: Standard errors are in parentheses.

The reported coefficient estimates of pre-test scores in the first two rows represent averages of the coefficients estimated separately for the grades, where applicable. The associated standard errors similarly represent averages across grades. The standard errors therefore do not account for the variability of the estimates across grades. These numbers are presented for descriptive purposes only and should not be used to conduct statistical inference.

The two sets of pre-test variables for math, ELA, and English were required in the biology model. One represents the coefficient for those students missing their same-subject pre-test, whereas the other represents the coefficient for those students who took the same-subject pre-test.

The coefficients on missing science pre-test and missing algebra pre-test are interpreted relative to the predicted post-test score for students who are not missing those pre-tests.

The coefficients on free/reduced-price lunch eligibility variables are interpreted relative to the predicted post-test score for students who are eligible for free lunch, the excluded category.

American Educational Research Association (AERA). “AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs.” November 2015.

American Statistical Association (ASA). “ASA Statement on Using Value-Added Models for Educational Assessment.” April 2014.

Arellano, Manuel. “Computing Robust Standard Errors for Within-Groups Estimators.” Oxford Bulletin of Economics and Statistics, vol. 49, no. 4, November 1987, pp. 431–434.

Bertrand, M., E. Duflo, and S. Mullainathan. “How Much Should We Trust Differences-in-Differences Estimates?” Quarterly Journal of Economics, vol. 119, no. 1, 2004, pp. 248–275.

Buonaccorsi, John P. Measurement Error: Models, Methods, and Applications. Boca Raton, FL: Chapman & Hall/CRC, 2010.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates.” American Economic Review, vol. 104, no. 9, 2014, pp. 2593–2632.

Chetty, Raj, John Friedman, and Jonah Rockoff. “Discussion of the American Statistical Association’s Statement (2014) on Using Value-Added Models for Educational Assessment,” May 2014a. Available at http://obs.rc.fas.harvard.edu/chetty/ASA_discussion.pdf

Hanushek, Eric A., John F. Kain, Steven G. Rivkin, and Gregory F. Branch. “Charter School Quality and Parental Decision Making with School Choice.” Journal of Public Economics, vol. 91, nos. 5–6, 2007, pp. 823–848.

Hock, Heinrich, and Eric Isenberg. “Methods for Accounting for Co-Teaching in Value-Added Models.” Working paper #6. Washington, DC: Mathematica Policy Research, June 2012.

Kane, Thomas J., and Douglas O. Staiger. “Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation.” No. w14607. National Bureau of Economic Research, 2008.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. “Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment.” Research paper, MET project. Bill & Melinda Gates Foundation, 2013.

Liang, Kung-Yee, and Scott L. Zeger. “Longitudinal Data Analysis Using Generalized Linear Models.” Biometrika, vol. 73, no. 1, April 1986, pp. 13–22.

McCaffrey, Daniel F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton. "Models for Value-Added Modeling of Teacher Effects." Journal of Educational and Behavioral Statistics, vol. 29, no. 1, 2004, pp. 67–102.

Meyer, Robert H. "Value-Added Indicators of School Performance: A Primer." Economics of Education Review, vol. 16, no. 3, 1997, pp. 283–301.

Morris, Carl N. "Parametric Empirical Bayes Inference: Theory and Applications." Journal of American Statistical Association, vol. 78, no. 381, 1983, pp. 47–55.

Raudenbush, Stephen W. "What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice?" Journal of Educational and Behavioral Statistics, vol. 29, no. 1, 2004, pp. 121–129.

Resch, Alexandra, and Jonah Deutsch. "Measuring School and Teacher Value Added for Charleston County School District, 2013–2014 School Year." Washington, DC: Mathematica Policy Research, November 2014.

Sanders, William L. "Value-Added Assessment from Student Achievement Data—Opportunities and Hurdles." Journal of Personnel Evaluation in Education, vol. 14, no. 4, 2000, pp. 329–339.

Taylor, Eric S., and John H. Tyler. "The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-Career Teachers." Working paper 16877. Cambridge, MA: National Bureau of Economic Research, 2011.