MATHEMATICA
Policy Research

REPORT

FINAL REPORT

Measuring School and Teacher Value Added in Charleston County School District, 2013–2014 School Year

August 13, 2014

Alexandra Resch
Jonah Deutsch

Submitted to:
Charleston County School District
75 Calhoun Street
Charleston, SC 29401
Project Officer: Anita Huggins and Audrey Lane
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: 40241.300

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 officers in CCSD, BRIDGE co-directors Audrey Lane and 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, Mary Grider, assisted by Randy Rosso, Liz Potamites, Eric Lundquist, Serge Lukashanets, and Eli Kaplan processed the data and provided expert programming. Eric Isenberg provided valuable comments. Cindy George edited the report, and Sharon Clark provided word-processing and production support.

CONTENTS

I. OVERVIEW

During 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 from the U.S. Department of Education that CCSD was awarded in 2012. For the 2013–2014 school year—the pilot phase of the project—the only teachers and principals eligible to receive value-added measures of teacher or school effectiveness were those at BRIDGE pilot schools. This includes teachers of mathematics, English/language arts (ELA), science, and social studies in grades 4 through 8 in 11 district schools and the principals of these 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 will pilot 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 will be 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 include teachers, principals, and district staff with a stake in the various grant components in both pilot and non-pilot schools. Additionally, the district formed a steering committee charged with weighing work group recommendations and making recommendations to the superintendent, Dr. Nancy McGinley, and her senior leadership team. In this report, we describe the value-added models that will be used as part of the BRIDGE teacher and principal evaluations. We will estimate (1) teacher effectiveness for the 2013–2014 school year in the 11 BRIDGE pilot schools that have grades 4 through 8 and (2) school effectiveness for these same schools.[3] The remainder of Chapter I provides an overview of value-added methods in nontechnical terms. Chapter II describes the data used to estimate teacher and school value added. Chapter III provides the details of the statistical methods used to estimate teacher and school value added. In fall 2014, we will produce a statistical addendum that provides empirical information about the students and teachers included in the value-added model and the statistical models used to produce the results.

[1] There are 14 BRIDGE pilot schools, but only the 11 schools that contain grades 4 through 8 are eligible to receive value-added estimates in 2013–2014.

[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, Military Magnet, C.C. Blaney, Edmund Burns, Jane Edwards, Minnie Hughes, E.B. Ellington, EXCEL at Morningside, North Charleston Elementary, Malcolm Hursey, and Pinehurst.

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 from any confounding factors outside the teacher's control. We will also aggregate teacher value added to estimate school-wide value added. To measure the performance of CCSD teachers, we will 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 is used as one of several measures used 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 test score averages, as was done under the original incarnations of the No Child Left Behind Act, value-added models take advantage of statistical techniques, such as multivariate regression, to account for prior achievement and student characteristics, isolating the effects of teachers from the components of student achievement that are outside of 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 5 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 5 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 will 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 will estimate the performance of teachers in CCSD using a value-added model based on the Palmetto Assessment of State Standards (PASS) tests in math, ELA, science, and social studies. We will measure teacher effectiveness in these subjects separately, and will calculate value added for teachers of these subjects in grades 4 to 8 during the 2013–2014 school year. Students take a PASS test in math and ELA in each grade from 3 to 8, but, because students take either the social studies or science exam, but not both, in grades 3, 5, 6, and 8, the value-added models for social studies and science will be somewhat different, and will include different sets of students than the models for math and ELA.

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 will use roster-verified lists to create a dosage for each teacher-student pair, indicating whether a teacher taught a given student, and if so, for how long. However, because teachers in other schools have not participated in roster verification this year, we will not report their value-added estimates. Instead, we will derive information on which students were taught by these teachers from administrative course data.

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

1. **Multiple regression.** We will 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 will account for a set of student characteristics that could be related to performance on the PASS test in the 2013–2014 school year (the post-test). These characteristics include a student's PASS test scores from the 2012–2013 school year (the pre-tests), poverty status, limited English proficiency, learning-disability status, whether they transferred schools during the year, and attendance during the 2012–2013 school year.[4] Accounting for these characteristics, we will 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 will 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

[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.

teachers during the year, each teacher will receive credit for the students' achievement based on the amount of instructional time that teachers reported through the roster verification process.

2. **Accounting for imperfect measurement of pre-tests.** A student's performance on a single test is an imperfect measure of ability. If we did not take this into account, teachers may be unfairly held accountable for the initial performance of their students. Imperfect measurement of students' ability on the pre-test can dampen the observed relationship between pre- and post-test scores, compared to the true relationship between student achievement at the beginning and end of the year. So, if we were to use the observed relationships without any adjustments, teachers of students with low pre-test scores might be held partly accountable for the performance of their students before they entered their classrooms. To avoid this problem, we will compensate for imperfect measurement in pre-test scores by employing a statistical technique that uses published information on the reliability of the PASS exams.

3. **Comparing teachers across grades.** The PASS is not designed to allow the comparison of scores across grades. We will therefore place teachers on a common scale by translating each teacher's value-added estimate into a metric of "generalized" PASS points. This translation will be based on a three-stage procedure. First, we will translate student scores into a common metric in which each student test score is measured relative to other test scores within the same year, grade, and subject. We will then use these scores to produce initial teacher value-added estimates. Second, we will adjust these estimates so that the average teacher in each grade receives the same score. Third, we will multiply the resulting estimates by a grade-specific conversion factor to ensure that the dispersion of teacher value-added estimates is similar by grade. This ensures that features of the PASS scale scores that vary by grade level do not influence teachers' value-added estimates. For teachers with students in more than one grade, we will take a student-weighted average of their grade-specific value-added estimates.[5]

4. **Accounting for imprecisely estimated measures based on few students.** Value-added estimates for teachers with fewer students will be less precise—that is, less likely to pinpoint the teachers' true effectiveness. By virtue of guessing some answers correctly or incorrectly, individual students may overperform on exams while 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 will 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 will 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] To compare schools with different grade configurations, we will apply a similar strategy. We will transform each grade-level estimate within a school into generalized PASS 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 will 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 true, 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 reasons, some statisticians have cautioned against using value-added models (ASA 2014); 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. 2011) 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. (2011) 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 pilot a process of developing SLOs as another measure of teachers' impact on student learning.

II. DATA

In this chapter, we review the data used to generate the value-added measures. We discuss the standardized assessment used in CCSD schools, the data on student background characteristics, and how we will 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. The schools on this list included at least one of the grades from 4 to 8. 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. CCSD also provided the data from which to construct a student list that indicates students' official grade level.

B. PASS test scores

When estimating the effectiveness of CCSD teachers, we will include elementary and middle school students in the model for a given subject if they have a test score in that subject from 2014 (the post-test) and a test score from the previous grade in math and ELA in 2013 (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 and science value-added measures, we do not require that students have a same-subject pre-test score. While same-subject pre-test scores generally play a central role in value-added models, our empirical investigation found that math and ELA pre-tests, and the same-subject pre-test scores for the students who had them, were sufficient to produce reasonably precise estimates.[6] We will 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 day of the school year (October 23, 2013) and on the first day of PASS testing in the spring of 2014. We will 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 will estimate the value-added model using all students in grades 4 through 8, even those not linked to an eligible teacher. Some students who have the requisite post-test and pre-test scores are not linked to an eligible teacher because they (1) were included in the roster file but not claimed by a teacher (in BRIDGE pilot schools); (2) were not linked to a CCSD teacher in the administrative course data (in other

[6] 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 value-added estimates that would be unreliable.

schools); or (3) were claimed only by a teacher with fewer than five students in his or her grade (since we do not calculate a value-added estimate for teachers with so few students).[7]

Test scores may be meaningfully compared only within grades and within subjects. Therefore, before using the test scores in the value-added model, we will create subject- and grade-specific standardized scores by subtracting the mean and dividing by the standard deviation within a subject-grade combination.[8] After calculating the value-added estimates, we will then convert the estimates back to the PASS scale by multiplying them by the subject- and grade-specific standard deviation.

C. Student background data

We will 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 student transferred schools during the school year
- Proportion of days that the student attended school during the previous year

Attendance is a measure of student motivation. We will use previous—rather than current-year—attendance to avoid confounding student attendance with current-year teacher effectiveness; that is, a good teacher might be expected to motivate students to attend school more regularly than a weaker teacher. Attendance is a continuous variable that could range from zero to one. Aside from pre-test variables, the other variables are binary variables taking the value zero or one.

We will 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 will use the values of non-missing student characteristics to predict the value of the missing characteristic. We will not generate imputed values for the math and ELA pre-tests; we will drop from the analysis file any students with missing pre-test scores in math or ELA.

[7] We require a minimum of five students within a grade to estimate value added for a teacher for a given subject and grade. We will report value-added estimates for teachers linked to at least 10 students in the model for a given subject (across all grades).

[8] Subtracting the mean score for each subject and grade creates a score with a mean of zero in all subject-grade combinations.

D. Teacher dosage

Given that some students are taught by a combination of teachers, we will apportion their achievement among multiple teachers. We refer to the fraction of time the student was enrolled with each teacher as the "dosage."

1. Teacher dosage at BRIDGE pilot schools

BRIDGE pilot schools underwent a process called roster verification in the spring of the 2013–2014 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 (that is, None, Some, Shared, Most, All) that were translated to numeric responses of 0, 25, 50, 75, and 100 percent. If a teacher claimed a student for less than 100 percent of time in any month, the teacher was not responsible for naming other teachers who taught the student. Teachers could also add students to their rosters. Principals verified, or assigned staff to verify, the accuracy of the rosters that teachers submitted.

We will 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 will set the teacher dosage equal to one. If a student changed teachers from one term to another, we will determine the amount of time the student spent with each teacher, subdividing the dosage among teachers accordingly. If two or more teachers claimed the same students at 100 percent during the same time, we will 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 will not subdivide dosage for co-taught students. Finally, we will 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 will 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 will 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 will not subdivide dosage for these students.

III. ESTIMATING VALUE ADDED

A. Regression estimates

We have developed one linear regression model to estimate effectiveness measures for both schools and teachers. We will 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 will estimate regressions separately for math, ELA, science, and social studies, and separately for students in elementary grades (4 and 5) and middle school grades (6 to 8), for a total of eight regressions. We will separate students by these grade spans 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 very different than in 8th grade.

1. The regression model for math and ELA

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

$$(1)\quad Y_{itg} = \lambda_g M_{i(g-1)} + \omega_g E_{i(g-1)} + \boldsymbol{\alpha}_1' \boldsymbol{X}_i + \boldsymbol{\eta}_1' \boldsymbol{T}_{itg} + \varepsilon_{itg},$$

where Y_{itg} is the post-test score for student i taught by teacher t in grade g, $M_{i(g\text{-}1)}$ is the math pre-test score for student i in grade $g\text{-}1$ during the previous year, and $E_{i(g\text{-}1)}$ denotes the pre-test score in ELA. Y_{itg} represents math post-test scores when evaluating math teachers and ELA post-test scores when evaluating ELA teachers. The pre-test scores capture prior inputs into student achievement, and the associated coefficients, λ_g and ω_g, vary by grade. The vector $\boldsymbol{X}_i$ denotes the control variables for individual student background characteristics. The coefficients on these characteristics, $\boldsymbol{\alpha}_1$, are constrained to be the same across all grades within a grade span.[9]

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.[10] The catchall teacher may represent, for example, one or more teachers with very small classes. A student contributes one observation to the model for each teacher to whom the student 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 $\boldsymbol{\eta}_1$.

[9] 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.

[10] Although CCSD will not use value-added estimates for teachers in BRIDGE schools with fewer than 10 students, we will 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 will reallocate those students to a grade-specific catchall teacher.

To account for multiple observations on the same student, we will 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 will exclude teacher-grade combinations with a total dosage of fewer than five students. We will aggregate the estimated coefficients into a single measure for each teacher (see Section C below).

2. Models for social studies and science

In CCSD, all students take the social studies and science exams in 4th and 7th grade, but in the other grades from 3rd to 8th, students are randomly chosen to take either the social studies or science exam. This has two implications for our analysis: (1) we will exclude from the analysis students who do not take a given exam in a given grade as their post-test, and (2) 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. The alternative would be to exclude students without a same-subject pre-test score from the model, which would have resulted in too few students to calculate reliable estimates for the majority of teachers. To address the second implication, the model for social studies and science must differ slightly from the model for math and ELA. The model for science and social studies is

$$(2)\quad Y_{itg} = \lambda_{1g}M_{1i(g-1)} + \omega_{1g}E_{1i(g-1)} + \lambda_{2g}M_{2i(g-1)} + \omega_{2g}E_{2i(g-1)} + \varphi_g S_{i(g-1)} + \boldsymbol{\alpha_2'X_i} + \boldsymbol{\eta_2'T_{itg}} + \varepsilon_{itg}.$$

Y_{itg}, $\boldsymbol{\alpha'_2X_i}$, $\boldsymbol{\eta'_2T_{itg}}$, and ε_{itg} are parallel to equation (1). $\lambda_{1g}M_{1i(g-1)}$ 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 $\omega_{1g}E_{1i(g-1)}$ represents the ELA test score and associated coefficient for these students. $\lambda_{2g}M_{2i(g-1)}$ represents the math test score and associated coefficient for students who *do not* have a same-subject pre-test, and $\omega_{2g}E_{2i(g-1)}$ 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. In 5th and 8th grades, $\lambda_{2g}M_{2i(g-1)}$ and $\omega_{2g}E_{2i(g-1)}$ will both be zero, since all students in these grades will have a same-subject pre-test. We include two pre-test variables and associated coefficients for both math and ELA because, in testing of the model using the prior year's data, 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.

B. Measurement error in the pre-tests

We will 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 will 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 will 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 will use the reliabilities associated with the 2013 PASS exams from grades 3 to 7. We will then use the measurement-error corrected values of the pre-test coefficients to calculate an adjusted post-test score for each student. Finally, we will regress these adjusted scores on the student characteristics, other than pre-tests, and 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 social studies and science models contains more pre-test variables and coefficients than the expression for the math and ELA models, but the procedures are analogous. For simplicity, we will focus on the procedure for the math and ELA models. The adjusted post-test score is expressed as:

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

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 will use the adjusted post-test as the dependent variable in a single equation expressed as:

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

We obtain the grade-specific estimates of teacher effectiveness, $\hat{\boldsymbol{\eta}}_1$, 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 $\hat{\boldsymbol{\eta}}_1$ 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 will have one estimate for each teacher in each grade level he or she taught. To calculate school value added, we will first aggregate each teacher-grade estimate within a 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.[11]

Once we have an estimate for each school-grade, we will 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

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 gains 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.[12] Because we do not want to penalize or reward teachers simply for teaching in a grade with unusual test properties, we will standardize grade-level estimates for schools and teachers so that each set of estimates is expressed in a common metric of "generalized" PASS points. Below, we describe the procedure in the context of teacher measures; the procedure for school measures is analogous.

We will standardize the effectiveness estimates so that the mean of the estimates is the same across grades. First, we will 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 will calculate the average using weights based on the number of students taught by each teacher.

We will 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-

[11] We will 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 these teacher estimates to be correlated.

[12] 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.

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.[13] Our method of calculating the standard deviation of teacher effects also downweights imprecise individual estimates. Finally, we will 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 PASS-point scale.

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

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

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.

We will 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 will choose the weights based on the empirical Bayes 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 will use a weighted average of the grade-specific estimates (expressed in generalized PASS points). We will set the weight for grade *g* equal to the proportion of students of teacher *t* in grade *g*, denoted as p_{tg}. We will then compute the variance of each teacher's estimated effect by using:

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

where $Var\left(\hat{\delta}_{tg}\right)$ 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

[13] 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.

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 will 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 will apply the empirical Bayes (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 will 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.[14] 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:

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

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 will calculate the standard error

[14] 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.

for each $\hat{\delta}_t^{EB}$ using the formulas provided by Morris (1983). As a final step, we will 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 will 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.[15] The aggregate IVA score constitutes 35 percent of the BRIDGE composite measure of teacher effectiveness for eligible teachers in BRIDGE pilot schools. We will 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 will 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. Because some pilot schools may have only one or two eligible teachers for each subject, to protect confidentiality we will report school-wide value-added results for these schools in a manner determined by CCSD.

[15] Because students take either the social studies or science exams in grades 5, 6, and 8, and because value-added estimates require a post-test score, the number of students that a teacher taught is 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.

REFERENCES

American Statistical Association. “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. “The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood.” Cambridge, MA: National Bureau of Economic Research, 2011.

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 2014. 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.” Washington, DC: Mathematica Policy Research, June 2012.

Kane, Thomas J., and Douglas O. Staiger. “The Promise and Pitfalls of Using Imprecise School Accountability Measures.” *Journal of Economic Perspectives*, vol. 16, no. 4, 2002, pp. 91–114.

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.

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.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ WASHINGTON, DC

MATHEMATICA
Policy Research