MATHEMATICA Policy Research

REPORT

FINAL REPORT

Measuring Teacher Value Added in DC, 2013–2014 School Year

August 28, 2014

Eric Isenberg

Elias Walsh

Submitted to:

Office of the State Superintendent of Education
810 1st Street, NE
Washington, DC 20002

Project Officer: Jessica Enos

District of Columbia Public Schools
1200 1st Street, NE
Washington, DC 20002

Project Officer: Alden Wells

Submitted by:

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

Project Director: Alexandra Resch

Reference Number: 40379.503

CONTENTS

TABLES

ACKNOWLEDGMENTS

We would like to thank the Office of the State Superintendent of Education of the District of Columbia (OSSE) and the District of Columbia Public Schools (DCPS) for funding the work. Several people at these agencies played key roles in building and implementing the value-added model described in this report, including Jeffrey Noel and Jessica Enos at OSSE, and Jason Kamras and Alden Wells at DCPS.

At Mathematica Policy Research, Emma Kopa and Mary Grider oversaw a team of programmers, including Matt Jacobus, Juha Sohlberg, Dejene Ayele, Jeremy Page, Dylan Ellis, Becca Wang, Alma Moedano, and Molly Crofton, who processed the data and provided expert programming. Duncan Chaplin provided valuable comments. Betty Teller edited the report, and Kimberly Ruffin and Lisa Walls provided word processing and production support.

I. OVERVIEW

In this report, we document our approach to estimating a value-added model of teacher effectiveness in the District of Columbia Public Schools (DCPS) and eligible DC charter schools participating in Race to the Top (RTT) during the 2013–2014 school year. A value-added model that includes school, teacher, and student data from both DCPS and DC charter schools was used to generate estimates for all DC teachers. The DC Office of the State Superintendent of Education (OSSE) and DCPS agreed to make no substantive changes in how value added was calculated for teachers in the 2013–2014 school year compared to the method of calculation in the previous year. Therefore, there were only minor changes to the value-added model between the 2012–2013 and 2013–2014 school years.

The 2013–2014 school year was the fifth year of IMPACT, a teacher evaluation system for DCPS that relies in part on value-added estimates of teacher and school effectiveness. Under IMPACT, teachers who earn a highly effective rating receive performance pay, and those who earn low ratings are dismissed. Thirty-five percent of the total evaluation score for eligible teachers is composed of an individual teacher value-added score.

The past school year was also the third year in which 26 DC charter school local education agencies (LEA) used a teacher-level value-added model in their teacher evaluation systems as part of RTT. Each charter LEA designed its own evaluation system for its personnel. As required by OSSE, value-added estimates accounted for 30 to 50 percent of eligible charter school teachers' overall evaluation scores in participating LEAs. DC charter schools have used teacher evaluation scores to inform personnel decisions, which may include offering teachers performance pay for outstanding performance or dismissing teachers for poor performance.

In the remainder of this report, we describe the data used to estimate teacher value added in the 2013–2014 school year (Chapter II) and provide the technical details of the statistical methods used to estimate value added in 2013–2014 (Chapter III). We include tables of diagnostic information that summarize the population of students and teachers on which the value-added estimates were based, as well as the results from the statistical model used to produce those estimates. For a broader discussion of value added as a measure of effective teaching, the use of value added within teacher evaluation systems in DC schools, and a nontechnical description of the steps used to estimate value added in DC, please refer to the technical report from the 2011–2012 school year, the first year of RTT (Isenberg and Hock 2012).

II. DATA

In this chapter, we review the data used to generate value-added estimates of teacher effectiveness in math and reading. We discuss the standardized assessment used in DC schools, the data on student background characteristics, and how we calculated the amount of time that students spent in more than one school or with more than one teacher. We also provide an overview of the roster confirmation process that allows teachers to confirm whether and for what portion of the year they taught math or reading to students. We used data on students and teachers in the 2013–2014 and 2012–2013 school years. We evaluated teacher effectiveness during the 2013–2014 school year. We used data from the 2012–2013 school year as an auxiliary year to better estimate the relationships between student and classroom characteristics and student achievement. These data did not otherwise affect the value-added estimates of teachers in the 2013–2014 school year.

A. DC CAS test scores

The outcomes we analyzed were the scores from the 2014 DC Comprehensive Assessment System (CAS) test in reading in grades 4 to 10 and in math in grades 4 to 8.[1] Starting with 4th- to 10th-grade students with a DC CAS test from 2014 (the post-test), we excluded students if they lacked a DC CAS test from the previous grade in the same subject in 2013 (the pre-test).[2] We then excluded students from the analysis file if student background data were missing.[3] We also excluded students who repeated or skipped a grade, as they lacked pre- and post-test scores in consecutive grades and years. Finally, we excluded students not linked to a teacher eligible to receive a value-added estimate for the student's grade level either because the students (1) were included in the roster file (described below) but not claimed by a teacher or (2) were claimed only by a teacher with fewer than seven students in his or her grade (as we do not estimate a value-added measure for teachers unless they are linked to at least seven students). We applied analogous rules for inclusion in the model in the auxiliary year to students with a DC CAS post-test from 2013. After applying these rules, we reported estimates only for teachers who taught 15 or more students over the course of the 2013–2014 school year in at least one subject across all grades. For example, we would report an estimate in reading for a teacher who claimed eight students in reading in grade 4 and seven students in grade 5. For a teacher who claimed nine students in reading in grade 4 and six students in grade 5, however, the grade 5 students would not be linked to the teacher, as they would not meet the seven-student minimum in that grade level. Because such a teacher would be linked to only nine students across all grades, we would not report a value-added estimate in reading for this teacher.

Table II.1 shows the total number of students who could have been included in the analyses, the reasons they were excluded, and the total included in the models. The first two columns show

[1] OSSE chose not to estimate value-added models for high school math teachers because the topics covered on the grade 10 math test cover general math concepts rather than the content of particular courses for which teachers are responsible.

[2] We considered some students with scores on the DC CAS post-test to be missing test score data because the scores were flagged as incomplete by CTB/McGraw Hill.

[3] We included students who were missing some background characteristics, but excluded those for whom no data on background characteristics were available.

the totals for students in math, and the last two columns show the totals for reading. The top row of the table shows the total number of students who received test scores for the math or reading DC CAS test. The number of students was larger in reading than in math because we included students with post-tests in grades 4 to 10 for reading, but only those with post-tests in grades 4 to 8 for math. The next four rows show the reasons why students who had these post-test scores were excluded from the analysis file. As shown in the bottom row of the table, 90 percent of students with 2014 test scores were included in the value-added model for math and 88 percent for reading. The most common reason students were excluded was lack of pre-test scores.

For each subject, the DC CAS is scored so that the first digit is a grade indicator that does not reflect student achievement. For example, a 3rd-grade student could receive a scale score from 300 to 399, a 4th-grade student from 400 to 499, and so on. The range for grade 9 and grade 10 students is 900 to 999.[4] For the value-added model, we dropped the first digit and used the rest of the scores, which ranged from 0 to 99.

Table II.1. Reasons students tested in 2014 were excluded from the analysis files, by subject

	Math (grades 4 to 8)		Reading (grades 4 to 10)	
	Number	Percent	Number	Percent
Students with post-test scores	18,254	100.0	22,854	100.0
(1) Missing same-subject pre-test scores	1,164	6.4	1,619	7.1
(2) Skipped or repeated a grade	253	1.4	582	2.5
(3) Not linked to an eligible teacher	348	1.9	442	1.9
Total excluded	1,765	9.7	2,643	11.6
Total included in value-added model	16,490	90.3	20,212	88.4

Note: Students are included in this table only if they were included in roster confirmation and could be linked to background characteristics. This excludes students enrolled in charter schools not participating in Race to the Top if they were not also enrolled in a participating charter school or in a DCPS school.

Students were excluded sequentially in the order presented and therefore are not counted for more than one reason in this table.

The value-added model includes DCPS and charter school students in grades 4–8 for math and grades 4–10 for reading.

The resulting scores may be meaningfully compared only within grades and subjects; math scores, for example, generally are more dispersed than reading scores within the same grade. Therefore, before using the test scores in the value-added model, we transformed the test scores using a two-part process. First, we created subject-, year-, and grade-specific z-scores by subtracting the mean and dividing by the standard deviation within a subject-grade-year combination.[5] This step translated math and reading scores in every grade and year into a common metric. Second, we created a measure with a range resembling the original DC CAS

[4] The DC CAS test score file also indicated the grade level for each student, which allowed us to distinguish students in grades 9 and 10.

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

point metric by multiplying each z-score by the average standard deviation across all grades within each subject and year.

B. Student background data

We used data provided by OSSE and DCPS to construct variables that accounted for the following student background characteristics in the value-added model:

- Prior achievement in same subject as post-test
- Prior achievement in other subject (we control for math and reading pre-tests regardless of post-test)
- Poverty status
- Limited English proficiency status
- Existence of a specific learning disability
- Existence of other types of disabilities requiring special education
- Transfer of students across schools during the school year
- Proportion of days the student attended school during the previous year

We also accounted for two classroom characteristics:

- Average classroom pre-test scores
- Standard deviation of classroom pre-test scores

Attendance is a measure of student motivation. We used 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 would. The proportion of the days a student attended school is a continuous variable that could range from zero to one. Aside from pre-test variables and attendance, the other variables are binary, taking a value of either zero or one.

To account for poverty status, we used data on free or reduced-price lunch (FRL) eligibility for the current year and the three prior school years. For each of these years, we created up to four indicators to measure student poverty status: (1) eligible for a free lunch, (2) eligible for a reduced-price lunch, (3) ineligible for a free or reduced-price lunch, and (4) attended a community-eligible (CE) school/eligibility unknown.[6] When using data from the 2009–2010 to the 2011–2012 school years, students were placed in the last category either because they attended a CE school or because they were missing information on FRL status that year for another reason. For data from the 2012–2013 and 2013–2014 school years, we used this category only for students attending CE schools. We imputed FRL status for students missing FRL information for other reasons, as described below. Except for students with an imputed FRL status, each student belonged to one of the four categories for each school year.

[6] For students in the 2013–2014 school year, we accounted for student poverty status by using data on student FRL dating back to the 2010–2011 school year. For students in the auxiliary year (the 2012–2013 school year), we used data on student FRL from the 2009–2010 to 2012–2013 school years.

Individual student data on FRL eligibility is lacking for students attending a CE school because these schools do not collect annual information about individual student poverty status. Schools are eligible to become CE if at least 40 percent of their student population has an identified need for free lunch based on direct certification, where students qualify based on their families' participation in state welfare or food stamp programs. These schools provide free breakfasts and lunches to all enrolled students and save on administrative costs by forgoing the collection of individual student FRL applications. Between the 2011–2012 and 2013–2014 school years, the number of CE schools in DC grew from 66 DCPS schools and no charter schools to 89 DCPS schools and 69 charter schools (OSSE 2013). Some individual students attending a CE school were indicated as being eligible for free lunch via a direct certification process and included as known to be eligible for free lunch in the analysis file; a small number of other students had an individual status indicating a paid-lunch status and were included as such.

We calculated the two classroom achievement measures (average pre-test achievement and standard deviation of pre-test achievement) using the same-subject pre-test scores. The classroom characteristic calculations were weighted by the teacher dosage associated with the teacher-student combination. The classroom characteristics were measures of a student's peers in the classroom, excluding that student.

We imputed data for students who were included in the analysis file but had missing values for one or more student characteristics. Our imputation approach used the values of nonmissing student characteristics to predict the value of the missing characteristic. We did not generate imputed values for poverty status in the 2010–2011 or 2011–2012 school years; instead, we included these students in the fourth category of CE/unknown eligibility. Few non-CE students were missing poverty information in the 2012–2013 or 2013–2014 school years, however; for these students, we imputed values for the first three poverty status categories for that year using the same imputation method used for students missing other student background data. For students who did not attend a DC school for part of the previous year, we used a Bayesian method to impute missing attendance data, based on other student characteristics in addition to attendance during the portion of the year spent in DC.[7] Finally, we did not generate imputed values for the same-subject pre-test; rather, we dropped from the analysis file any students with missing same-subject pre-test scores.[8]

Table II.2 shows the characteristics of students included in the reading value-added model. The characteristics of students in the math value-added model differed from those in the reading model by no more than one percentage point for grades 4–8, the grades common to the value-added model for both subjects.

[7] We generated a predicted value by using the values of nonmissing student characteristics, and combined this information with the actual attendance data for the part of the year spent in DC. With this method, the more time a student spends in a DC school, the more his or her imputed attendance measure relies on actual attendance data from the part of the year spent in DC. Conversely, the less time spent in DC, the more the imputed attendance measure relies on the predicted value. We implemented this approach by using a beta distribution with beta/binomial updating (Lee 1997).

[8] Less than 1 percent of students in the value-added analysis file had missing opposite-subject pre-test scores in any grade-subject combination, with the exception of grade 10 reading. The opposite-subject pre-test score for grade 10 reading is from the grade 8 math test and was missing for 14 percent of grade 10 students.

Table II.2. Characteristics of students from the 2013–2014 school year in the reading value-added model, by grade span

	Grades 4 and 5		Grades 6 to 8		Grades 9 and 10	
	Number	Percent	Number	Percent	Number	Percent
Included in value-added model	7,182	100.0	9,000	100.0	4,030	100.0
Known to be eligible for free lunch	4,354	60.6	5,789	64.3	2,571	63.8
Known to be eligible for reduced-price lunch	178	2.5	331	3.7	152	3.8
Known to be ineligible for free or reduced-price lunch	1,564	21.8	1,891	21.0	913	22.7
Unknown free or reduced-price lunch eligibility	1,087	15.1	989	11.0	394	9.8
Limited English proficiency	383	5.3	505	5.6	186	4.6
Specific learning disability	494	6.9	667	7.4	334	8.3
Other learning disability	478	6.7	533	5.9	192	4.8
Transferred schools during the school year	96	1.3	145	1.6	91	2.3

Notes: The total of the counts across grade spans in the top row corresponds to the total for reading in the final row of Table II.1.

All percentages are based on the counts in the top row.

Student characteristics were calculated as a weighted average for students enrolled in both DCPS and charter schools. The counts and percentages were not weighted in any other way.

Participation in the reading value-added model for grades 9 and 10 was optional for charter school LEAs.

The poverty status variables indicate students' poverty status in the 2013–2014 school year.

Students counted in the "status unknown" row include those who attended community-eligible schools without another data source to certify their free-lunch status and those whose status was unknown for other reasons.

For all student characteristics in this table, less than 1 percent of students have missing data.

C. Dosage

The value-added model accounted for the proportion of time each student was enrolled with a teacher, also known as the "dosage." Therefore, we collected information that allowed us to calculate the dosage for each teacher-student pair. We used school enrollment data to define the proportion of the year students spent at each school, and used teacher-student link data from administrative rosters that had been confirmed by teachers to determine which students were taught by a teacher, and for how long. The dosage combines the proportion of the year spent in a teacher's school and the time spent with a given teacher within the school.

For charter school teachers and students, the roster confirmation data provided by OSSE (described in detail below) defined the eligible teachers and students included in the analysis file. Similarly, DCPS provided roster confirmation data that defined its eligible teachers and students. Individual LEAs participating in RTT were responsible for contributing lists of teachers of math and reading in grades 4 to 8; in addition, some LEAs, including DCPS, chose to include

reading/English language arts (ELA) teachers in grades 9 and 10.[9] Each LEA determined the list of eligible teachers, using guidance from OSSE that teachers with primary responsibility for providing instruction in math and reading/ELA in the relevant grades should be included. In general, only regular education teachers were eligible to receive value-added estimates; special education teachers were not, although resource teachers in charter LEAs and Read 180 teachers were eligible. Resource teachers provided additional instruction to students and tended to work with a large number of students throughout the school year. Read 180 teachers provided supplemental instruction to students.

1. School dosage

Based on administrative data from OSSE and DCPS, which contained dates of school withdrawal and admission, we assigned every student a dosage for each school attended. School dosage equals the fraction of the first three quarters of the school year that a student was officially enrolled in that school. In determining dosage, we used school calendars from each participating LEA. We used only the first three quarters of the year because students in most LEAs start taking the DC CAS shortly after the end of the third quarter.[10] We assume that learning accumulated at a constant rate and therefore treat days spent at one school as interchangeable with days spent at another. For example, if a student split time equally between two schools, we set the dosage of each school to 50 percent, regardless of which school the student first attended.

2. Teacher dosage

To determine which students received math and reading instruction from eligible teachers during the 2013–2014 school year, DC schools conducted a roster confirmation among teachers of math in grades 4 through 8 and teachers of reading/ELA in grades 4 through 10. In most cases, teachers received lists of students who appeared on their course rosters, and had the option of adding students to their rosters. In other cases, teachers did not receive a list and so were responsible for populating the rosters with their students themselves. For each of the first three quarters, teachers indicated whether they taught each subject to each student and, if so, the proportion of time they taught the student. For example, if a student spent two days a week in an eligible teacher's classroom learning math and three days per week in another classroom with a special education teacher while the student's classmates learned math with the eligible teacher, the student was recorded as having spent 40 percent of instructional time with the eligible teacher. In recording the proportion of time they spent with each student in a given class and subject, teachers rounded to the nearest 20 percent, such that the possible responses were 0, 20, 40, 60, 80, and 100 percent. If a teacher claimed a student for less than 100 percent in any quarter, the teacher indicated the reason for the reduction, but was not responsible for naming other teachers who taught the student. OSSE ensured that all eligible charter school teachers completed the roster confirmation. Within each charter LEA, a central office administrator,

[9] As an additional reference to ensure that ineligible teachers were not mistakenly included in the analysis file, DCPS provided an official comprehensive list of teachers of math and reading in grades 4 to 10 who were eligible to receive individual value-added scores.

[10] There are two exceptions. One LEA uses trimesters rather than quarters, and in 23 schools in 9 other LEAs, the third quarter ends after the DC CAS tests are administered. In both of these cases, we used enrollment data from the first two terms and from the third term before the beginning of the testing window for the DC CAS.

principal, or designee verified the confirmed rosters. Likewise, in DCPS, principals or assistant principals verified eligible teacher-confirmed rosters. Central office staff at DCPS also followed up with DCPS teachers as necessary.

To create teacher dosage, we multiplied the school dosage for a teacher-student pair for each term by the percentage from the roster confirmation. For example, if a student spent half of the year in a teacher's school, and the teacher claimed the student for 60 percent dosage during the time spent in the school, the teacher dosage would be 0.50*0.60 = 0.30.

When two or more teachers claimed the same students at 100 percent during the same term, we assigned each teacher full credit for the shared students. This reflects a decision by OSSE that solo-taught and co-taught students should contribute equally to teachers' value-added estimates. We thus did not subdivide dosage for co-taught students. When the same teacher claimed the same student in multiple classrooms, we assigned the teacher credit for the student in each classroom based on the percentages the teacher claimed for the student in the roster confirmation. We used the same procedures to construct teacher-student links for the 2012–2013 school year. Table II.3 shows how many teachers shared students with another teacher in the value-added model and what percentage of their students was shared. As shown in the table, 14 percent of math teachers and 18 percent of reading teachers shared all of their students with other teachers, whereas 74 percent of math teachers and 59 percent of reading teachers did not share any students with another teacher. On average, math teachers shared 18 percent of their students and reading teachers shared 26 percent of their students with other teachers.

Table II.3. Teachers receiving value-added estimates in the 2013–2014 school year, by subject and extent of co-teaching

Percentage of students shared with another teacher in value-added model	Math		Reading	
	Number	Percent	Number	Percent
None	319	73.7	318	58.6
1–10 percent	22	5.1	41	7.6
11–50 percent	17	3.9	49	9.0
51–99 percent	16	3.7	40	7.4
All students	59	13.6	95	17.5
Total	433	100.0	543	100.0
Average percentage of students shared		18.3		26.1

Notes: Teachers received estimates if they were linked to at least 15 eligible students.

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

III. ESTIMATING VALUE ADDED

A. Regression estimates

We developed a linear regression model to estimate effectiveness measures for teachers. After assembling the analysis file, we estimated the regression model separately by subject (math or reading) and grade span. We used two grade spans in math (grades 4 and 5 and grades 6 to 8) and three in reading (grades 4 and 5, 6 to 8, and 9 and 10). For most characteristics, we allowed the relationship between the characteristic and achievement to take on one value for each grade span (elementary and middle school for math, and elementary, middle, and high school for reading). The grade-span approach represents an intermediate option between pooling relationships across all grades and estimating the relationships at each grade level. It balances the trade-off between (1) obtaining more-precise estimates of the relationships between student characteristics and achievement and (2) accurately reflecting differences across grades. Since prior achievement is usually a much stronger predictor of end-of-year achievement than other characteristics, however, we allowed for a different relationship to be estimated for each grade level, without losing much precision. Thus, relationships between pre-test scores and achievement took on five values in math (for grades 4 to 8) and seven values in reading (for grades 4 to 10).

In the regression equation, the post-test score depends on prior achievement, student background characteristics, classroom characteristics, the student's teacher, and unmeasured factors. For a given teacher *t* and student *i* in classroom *c*, school year *j*, and grade *g*, the regression equation may be expressed formally as:

$$(1)\quad Y_{ticjg} = \lambda_{jg} S_{i(j-1)} + \omega_{jg} O_{i(j-1)} + \boldsymbol{\beta}'\mathbf{X}_{ij} + \boldsymbol{\pi}'\mathbf{C}_{ticj} + \boldsymbol{\delta}'\mathbf{T}_{tijg} + \boldsymbol{\theta}'\mathbf{T}_{2tijg} + \varepsilon_{ticjg},$$

where Y_{ticjg} is the post-test score for student *i* and $S_{i(j-1)}$ is the same-subject pre-test for student *i* during the previous year. The regressions are run separately by post-test subject, so we drop a subject subscript for ease of notation. The variable $O_{i(j-1)}$ denotes the pre-test in the opposite subject. Thus, when estimating teacher effectiveness in math, *S* represents math tests and *O* represents reading tests, and vice versa. For 10th-grade reading students, for whom no pre-test score in math is available from the year before, we use a lagged pre-test score from two years earlier in grade 8. The pre-test scores help capture the impacts of prior inputs on student achievement; the associated coefficients λ_{jg} and ω_{jg} vary by grade and year. The vector $\mathbf{X}_{ij}$ denotes the control variables for individual student background characteristics. The coefficients on these characteristics, $\boldsymbol{\beta}$, vary across grade spans but are constrained to be the same across both years and the set of grades included in a grade span. The vector $\mathbf{C}_{ticj}$ represents the characteristics of classroom *c*; the coefficients on these classroom characteristics, $\boldsymbol{\pi}$, are estimated using student, teacher, and classroom information from two years (see Section C below).

To gain precision in the estimation of the relationships between student characteristics and achievement, we used two student cohorts to estimate the model. One cohort had post-tests from 2014 and pre-tests from 2013 (or 2012 for grade 8 math pre-tests used in the value-added model to estimate grade 10 reading results). The other cohort had post-tests from 2013 and pre-tests from 2012 (or 2011 for grade 8 math pre-tests). We pooled the data to estimate the parameters of the model in all but three cases. We estimated coefficients separately by year on (1) FRL

characteristics, because different categories of data were available for the two cohorts in different years; (2) pre-tests, to allow relationships between pre-tests and achievement to differ between years; and (3) teacher indicators, so that value added would measure performance solely in the 2013–2014 school year. In other words, each teacher's value added was based directly on students taught in the 2013–2014 school year who had post-tests from 2014 and pre-tests from 2013.

The vector $\mathbf{T}_{tijg}$ included a binary variable for each teacher-grade-year combination. For example, a teacher who taught math in grades 4 and 5 during both the 2012–2013 and 2013–2014 school years had four variables in $\mathbf{T}_{tijg}$. Under the Full Roster-Plus Method (FRM+), described in detail in Section B below, we ensured that each student contributes the same total dosage to the calculation of the parameters. We did this by duplicating each teacher-student-classroom-year link in the analysis file. The teacher links for the duplicate student-classroom-year observations—known as "shadow teachers" —composed a distinct set of indicators $\mathbf{T}_{2tijg}$ in the regression. Each teacher-student-classroom-year observation has one nonzero element in $\mathbf{T}_{tijg}$ or $\mathbf{T}_{2tijg}$. Measures of teacher effectiveness for the 2013–2014 school year were contained in the coefficient vector $\boldsymbol{\delta}$ for teacher-grade combinations from that school year. To measure teacher effectiveness in the 2013–2014 school year, we did not directly use estimates of teacher effectiveness in the 2012–2013 school year or the coefficients of the shadow teacher variables. Rather than dropping one element of $\mathbf{T}_{tijg}$ or $\mathbf{T}_{2tijg}$ from the regression, we estimated the model without a constant term. We also mean-centered the control variables so that each element of $\boldsymbol{\delta}$ represents a teacher-, grade-, and year-specific intercept term for a student with average characteristics.

Table III.1 shows the coefficient estimates and standard errors of the control variables in the model by subject and grade span. The top panel of Table III.1 shows the average association between the pre-tests and achievement on the 2013 DC CAS (measured in points on the test), accounting for all other characteristics that are included in the value-added model. The second panel shows the association between a given student characteristic and achievement. The bottom panel shows the association between each of the two classroom characteristics and achievement. Statistical significance is not indicated in this table since the point estimates predict student post-test scores whether they are statistically significant or not.

To account for team teaching, we used the FRM+, whereby each student contributed one observation to the model for each teacher to whom he or she was linked based on the roster confirmation process. Thus, the unit of observation in the analysis file was a teacher-student-classroom-year combination. This method of accounting for team teaching is based on the assumption that teachers who contribute equally to student achievement within each team receive equal credit (Hock and Isenberg 2012). To allow for the inclusion of classroom characteristics, teacher-student observations were included for each classroom shared by the teacher-student pair during a year. The model included teacher-student-classroom links from the 2012–2013 and 2013–2014 school years.

Because some students contributed multiple observations, we estimated the coefficients by using weighted least squares (WLS) rather than ordinary least squares (OLS). The teacher-grade-year variables in $\mathbf{T}_{tijg}$ are binary, and each teacher-student combination is weighted by the teacher dosage associated with that combination. In this model, the error terms are correlated, because individual students have multiple records, and heteroskedastic, due to differences across

students in how well the model can predict post-test scores based on the background characteristics included in the regressions. Therefore, we used a cluster-robust sandwich variance estimator (Liang and Zeger 1986; Arellano 1987) to produce consistent standard errors in the presence of heteroskedasticity and correlation in the regression error term.

In practice, to account for classroom composition and measurement error in the pre-tests, we estimated equation (1) using a multistep method described in Sections C and D below. The method includes three regression steps:

1. **Account for classroom characteristics** (average and standard deviation of pre-test achievement). We accounted for the relationship between the characteristics of students' peers in the same classroom and individual student achievement using data from multiple classrooms for each teacher. To obtain estimates of the contribution of classroom composition, we constrained the coefficients on the teacher variables to be the same across classrooms, including classrooms taught in different years. This constraint allowed us to leverage variation across classrooms to identify the contribution of classroom composition to student achievement. We then subtracted the contributions of classroom characteristics from the post-test to create an adjusted post-test measure.
2. **Calculate measures of teacher effectiveness for the 2013–2014 school year**. Because the first-stage regression pools teacher variables across grade and years, a second regression step was necessary to obtain estimates of teacher effectiveness based on student achievement only from the 2013–2014 school year. In the second-stage regression, we used the adjusted post-test from the first stage as the outcome variable and included distinct teacher variables for each grade and year. Because we used the adjusted post-test, this regression excludes the classroom characteristics but includes individual student background characteristics and pre-tests. We then calculated a second adjusted post-test score that nets out the contribution of pre-test scores.
3. **Calculate the precision of the estimates.** Both the first- and second-stage regressions applied a method to address measurement error in the pre-tests. However, because of computational limitations with this method, we could not obtain measures of the precision of value-added estimates from the second-stage regression. In this final regression step, we used newly adjusted post-tests to produce standard errors for the estimates of teacher effectiveness that accounted for multiple observations for each student in the regression. The regression in this step was identical to that from the second step except that we used the newly adjusted post-tests instead of controlling for pre-test.

The final teacher regression yields separate value-added coefficients for each grade-year combination in which a teacher was linked to students. We estimated a grade- and year-specific coefficient for a teacher only if the teacher had at least seven students in that grade.[11] We then aggregated teacher estimates across grades to form a single estimate for each teacher (see Section E below).

[11] Although teachers must teach at least 15 students for DCPS to evaluate them on the basis of individual value added, we included in the regression those teachers with 7 to 14 students for two reasons. First, we expected that maintaining more teacher-student links would lead to more accurate coefficient estimates. Second, we expected that value-added estimates for these teachers would provide useful data to include in the standardization and shrinkage procedures described below.

How to Interpret Table III.1

Table III.1 describes the relationships between the characteristics of DC students and achievement on the post-test by displaying the regression coefficients from the value-added model. The coefficients give the amount of the increase (or decrease, if the coefficient is negative) in the predicted score when a characteristic increases by one unit. For example, the coefficient of 0.72 in the first row of the first column of the table indicates that an increase by one DC CAS point on a student's pre-test score is associated with a 0.72 point increase in the student's predicted score on the 4th- or 5th-grade math post-test. Similarly, the coefficient on the fraction of the prior year a student attended school indicates that a student who attended 100 percent of the prior year is predicted to score 6.28 points higher than the prediction if the student instead attended for none of the prior year. More than 99 percent of students attended 75 percent of the prior year or more, so the typical contribution of prior attendance to the predicted score is much smaller than this change of 6.28 points might suggest; the change in predicted scores associated with a change in attendance from 75 to 100 percent is 1.57 DC CAS points.

For characteristics that are yes/no indicators, the coefficient gives the increase in the predicted score for a student who has that characteristic relative to a student who does not. For example, students in grades 4 and 5 in math who transferred between schools during the year are predicted to score 2.91 points lower than students who did not transfer. For the four indicators of student poverty status, the coefficients measure the difference in the predicted score of a student with that status (for example, students known to be ineligible for FRL) relative to a student who is eligible for free lunch.

Each regression coefficient describes a relationship after accounting for the other characteristics. Accordingly, the coefficient on a characteristic gives the change in predicted achievement when the characteristic is changed from no to yes or increased by one point, assuming that the students' other characteristics remain the same. Consequently, coefficients may not reflect the relationship we would observe had the other characteristics not been accounted for in the value-added model. This feature of multiple regression coefficients can produce counterintuitive relationships between characteristics and achievement if the contributions of one characteristic are accounted for largely by another characteristic in the model. For example, coefficients on limited English proficiency status would likely be consistently negative and greater in magnitude if the model did not also account for students' pre-test scores, because students with limited English proficiency tend to have lower pre-test scores.

The magnitude of the coefficients can be compared to the typical range of student achievement on the DC CAS. The standard deviation of student achievement on the grade 4 math post-test was 17.4 DC CAS points, indicating that about two-thirds of students scored within 17.4 points above or below the average score on the assessment. The standard deviations for other grades ranged from 15.6 to 17.3 points in math and from 11.8 to 15.3 points in reading.

The number in parentheses below each coefficient is the *standard error of the coefficient*—a measure of precision. A more precise coefficient indicates with more certainty that a coefficient reflects the actual relationship between the characteristic and achievement. Coefficients with smaller standard errors are more precise. The coefficients on the pre-tests are more precise than those on individual background characteristics. Roughly, a coefficient that is at least twice as large as its standard error is said to be statistically significant, meaning that it is likely that the direction of the relationship—whether positive or negative—reflects the actual relationship between the characteristic and achievement, and is not produced by chance.

Table III.1. Coefficients on covariates in the value-added models, by subject and grade span

Variable	Math		Reading		
	Grades 4 and 5	Grades 6 to 8	Grades 4 and 5	Grades 6 to 8	Grades 9 and 10
Pre-test scores (average coefficients)					
Same subject, all grades in span	0.72	0.60	0.66	0.68	0.65
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Opposite subject, all grades in span	0.15	0.23	0.15	0.13	0.05
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
Individual student background characteristics					
Known to be eligible for reduced-price lunch	-0.44	0.40	-0.84	-0.65	-0.32
	(0.81)	(0.52)	(0.62)	(0.42)	(0.61)
Known to be ineligible for free or reduced-price lunch	-0.38	0.69	-0.04	0.02	1.10
	(0.53)	(0.38)	(0.46)	(0.34)	(0.44)
Unknown free or reduced-price lunch eligibility	0.19	0.36	-0.06	0.79	0.93
	(0.49)	(0.47)	(0.38)	(0.38)	(0.60)
Limited English proficiency	0.79	0.23	-1.25	-0.70	0.38
	(0.37)	(0.33)	(0.35)	(0.27)	(0.47)
Specific learning disability	-0.90	-1.15	-2.62	-2.04	-2.66
	(0.40)	(0.33)	(0.35)	(0.28)	(0.42)
Other learning disability	-1.43	-2.49	-2.89	-2.37	-2.41
	(0.40)	(0.42)	(0.35)	(0.33)	(0.59)
Transferred schools during the school year	-2.91	-0.98	-1.14	-0.68	-2.15
	(0.84)	(0.71)	(0.60)	(0.54)	(0.99)
Fraction of the prior year student attended school	6.28	9.15	-2.73	1.22	0.45
	(1.77)	(1.77)	(1.45)	(1.41)	(2.12)
Classroom characteristics					
Average classroom pre-test score	0.04	0.08	0.13	0.05	0.12
	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)
Standard deviation of classroom pre-test scores	-0.18	0.02	-0.05	-0.04	-0.03
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)

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, using 2014 test scores as post-tests and 2013 test scores as pre-tests. The associated standard errors similarly represent averages across grades. These numbers are presented for descriptive purposes only and should not be used to conduct rigorous statistical tests. The table excludes coefficients on pre-test variables estimated separately for students from the 2012–2013 school year.

For students in grades 4–9, pre-test scores are from the prior grade in the 2012–2013 school year. In reading for 10th-grade students, same-subject pre-test scores are from 9th-grade reading test scores in the 2012–2013 school year, and opposite-subject pre-test scores are from 8th-grade math test scores in the 2011–2012 school year.

The table excludes coefficients on variables that indicate poverty status between the 2009–2010 and 2012–2013 school years. The poverty status variables reported in the table indicate students' poverty status in the 2013–2014 school year. All coefficient estimates of variables that indicate students' poverty status in previous years are no larger than 2.0 DC CAS points in absolute value.

Coefficients on the poverty status variables are relative to students who are known to be eligible for free lunch—the excluded category.

B. Full Roster-Plus Method

The Full Roster-Plus Method (FRM+) gives equal credit to teachers of co-taught students by replicating student observations to link the student to each of his or her teachers, as described in Section A, above. Because these students would otherwise carry extra weight in the estimation of the coefficients on student background characteristics, the FRM+ method equalizes the contribution of students taught by multiple teachers to the estimation of these coefficients by including an additional teacher indicator for each teacher, known as a shadow teacher (Isenberg and Walsh 2013). Under FRM+, students count toward the estimation of student background characteristics equally, regardless of how many eligible teachers claim them and the amount of time they spend with eligible teachers. To do this, we replicated observations in the data set and assigned dosage to the replicated observations so that all students have the same amount of total dosage in the analysis file. We linked the new records to artificial teacher indicators so that each teacher in the data set received a shadow teacher who absorbed the extra dosage for each student required to assign each student the same total dosage. The shadow teacher links were recorded in $\mathbf{T}_{2tijg}$, distinct from $\mathbf{T}_{tijg}$, the teacher links in the original observations. We did not change dosage for the original observations in this process; dosage measures the proportion of the year students spend with an eligible teacher. Each student thereby contributed equally to the estimates of student characteristics without affecting the proportional contributions of co-taught students to measures of teachers' effectiveness.[12]

C. Accounting for classroom characteristics

We accounted for the characteristics of students' peers in the same classroom in addition to individual student characteristics in the first-stage regression. The vector $\mathbf{C}_{ticj}$ in equation (1) represents two classroom characteristics included in the model: mean classroom pre-test score, and the standard deviation of classroom pre-test scores.[13] If these classroom characteristics influence student achievement, it is possible that omitting them would produce biased measures of teacher effectiveness if the mix of students in each classroom directly affects individual student achievement, a phenomenon sometimes called peer effects (Hoxby and Weingarth 2006; Sacerdote 2011).

Estimation of a model accounting for classroom characteristics required a multistep strategy, because we aimed to leverage multiple years of teacher performance to estimate the relationship between classroom characteristics and individual student achievement but also intended to recover a measure of teacher effectiveness based on only the 2013–2014 school year. Consequently, in the first stage, we constrained the teacher effects to be the same across years when estimating classroom characteristics. Pooling teacher variables across years in a first-stage regression allowed us to leverage variation across classrooms to estimate $\boldsymbol{\pi}$. So when estimating the contribution of classroom characteristics, we included only a single variable for each teacher across all of that teacher's classrooms, including classrooms taught in different years and students in different grades. In a later step, we calculated a single-year effect for teachers so that

[12] Standard errors are clustered at the student level for the FRM+. Thus, adding additional observations for shadow teachers and altering the maximum dosage does not artificially increase the precision of the teacher estimates.

[13] The classroom composition measures are calculated for student *i* based on all other students in the classroom excluding this student.

their performance in 2012–2013 did not directly affect a measure of their performance in 2013–2014.[14]

The first-stage value-added model is described by the equation

$$(2)\quad Y_{ticjg} = \lambda_{1jg} S_{i(j-1)} + \omega_{1jg} O_{i(j-1)} + \boldsymbol{\beta}_1' \mathbf{X}_{ij} + \boldsymbol{\pi}_1' \mathbf{C}_{ticj} + \boldsymbol{\delta}_1' \mathbf{T}_{ti} + \boldsymbol{\theta}_1' \mathbf{T}_{2ti} + \kappa d_j + \boldsymbol{\rho}' \mathbf{G}_g + \varepsilon_{1ticjg}\,.$$

The subscript *1* distinguishes the first-stage coefficients from those in equation (1) and in subsequent steps. The vectors $\mathbf{T}_{ti}$ and $\mathbf{T}_{2ti}$ include variables for each teacher, pooled across classrooms from all grades and years. To avoid potential bias that might arise from the sorting of teachers and students across schools, when estimating the contribution of classroom composition, we did not pool classrooms across schools for teachers who changed schools from one year to the next. Instead, for purposes of estimating equation (2), we treated these teachers as a separate teacher for each school in which he or she taught. We included the variable d_j, a binary indicator for the 2012–2013 school year, and the vector $\mathbf{G}_g$, a binary variables for each grade in a grade span, to measure differences across grades and years. We included these variables in equation (2) but not in the subsequent regression steps, because the teacher variables in (2) are not specific to a particular grade or year. We corrected for measurement error in the pre-test scores when estimating equation (2). In Section D, below, we describe how we corrected for measurement error in this first stage as well as in subsequent steps.

Based on the results of estimating equation (2), we calculated an *adjusted* post-test for each grade and subject that nets out the contribution of the measures of classroom composition:

$$(3)\quad A_{1ticjg} \equiv Y_{ticjg} - \hat{\boldsymbol{\pi}}_1' \mathbf{C}_{ticj}\,.$$

The vector A_{1ticjg} represents the student post-test outcome, net of the estimated contribution of classroom composition. To calculate equation (3) for students in most classrooms, we used the observed values of $\mathbf{C}_{ticj}$ from equation (1). For students in small classrooms (with fewer than 10 students) and for classrooms taught by resource teachers, we imputed the classroom characteristics in $\mathbf{C}_{ticj}$, using information about other classrooms in the same school and the values of individual student characteristics to predict the values of each classroom characteristic.

We used the adjusted post-test in place of the actual post-test to estimate single-year measures of teacher effectiveness for the 2013–2014 school year. In Section D, below, we describe how we used the adjusted post-test to produce single-year estimates of teacher effectiveness.

[14] We excluded records from classrooms with fewer than 10 students to estimate $\boldsymbol{\pi}$ because classroom characteristics based on classrooms with few students may be more likely to be mismeasured and exercise undue influence on the contribution of classroom characteristics to student achievement. We included these records in subsequent steps. We also excluded groupings of students taught by resource teachers, who provided supplementary instruction instead of regular classroom instruction. Six percent of records were excluded from the first-stage elementary and middle school grade-span regressions, and 8 percent of records were excluded from the first-stage high school grade-span regression.

D. Measurement error in the pre-tests

We corrected for measurement error in the pre-tests by using grade-specific reliability data available from the test publisher (CTB/McGraw Hill 2011, 2012, 2013). As a measure of true student ability, standardized tests contain measurement error, causing standard regression techniques to produce biased estimates of teacher effectiveness. To address this issue, we implemented a measurement error correction based on the reliability of the DC CAS 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 required two additional steps because of computational limitations with the measurement-error correction method related to producing measures of precision. Having estimated the first-stage regression given by equation (2), we used the classroom characteristic-adjusted post-tests from equation (3) to estimate a second regression step. In both of the first two regression steps, we applied the errors-in-variables correction. The second regression step included distinct teacher variables for each teacher-grade-year combination. A third and final regression step was necessary to calculate standard errors on teachers' estimates because of computational limitations with the measurement error correction method.

We used a dosage-weighted errors-in-variables regression to obtain unbiased estimates of the pre-test coefficients for each grade and year. For students in grades 4 to 9, we used the published reliabilities associated with the 2013 DC CAS for records from the 2013–2014 school year and the 2012 DC CAS for records from the 2012–2013 school year. For grade 10 students in the 2013–2014 school year, we used the reliabilities associated with the 2013 DC CAS for reading and the 2012 DC CAS for math, because the math pre-test is from grade 8. For grade 10 students in the 2012–2013 school year, we used the reliabilities associated with the 2012 DC CAS for reading and the 2011 DC CAS for math.

We estimated the second-stage regression in equation (4) to obtain pre-test relationships adjusted for measurement error based on a specification that included distinct teacher variables for each teacher-grade-year combination. Instead of the post-test, the dependent variable in equation (4) has been replaced with the adjusted post-test from equation (3). The subscript *2* distinguishes the second-stage coefficients from those in other steps.

$$(4)\quad A_{1ticjg} = \lambda_{2\,jg} S_{i(j-1)} + \omega_{2\,jg} O_{i(j-1)} + \boldsymbol{\beta}_2' \mathbf{X}_{ij} + \boldsymbol{\delta}_2' \mathbf{T}_{tijg} + \boldsymbol{\theta}_2' \mathbf{T}_{2tijg} + \varepsilon_{2ticjg}\,.$$

We then used the measurement-error-corrected values of the pre-test coefficients to calculate a second adjusted post-test that, in addition to the contribution of classroom characteristics, also nets out the contribution of the pre-tests:

$$(5)\quad A_{2ticjg} = A_{1ticjg} - \lambda_{2\,jg} S_{i(j-1)} - \omega_{2\,jg} O_{i(j-1)}\,.$$

The vector A_{2ticjg} represents the student post-test outcome, net of the estimated contribution attributable to the student's pre-test and classroom characteristics.

We estimated a third and final regression step to obtain standard errors that are consistent in the presence of both heteroskedasticity and clustering at the student level, because the regression includes multiple observations for the same student. This third-stage regression is necessary

because it is not practically feasible to simultaneously account for correlation in the error term ε_{2ticjg} across multiple observations and apply the numerical formula for the errors-in-variables correction. Thus, we obtained the new adjusted post-test in equation (5) and then estimated the final regression in (6):

(6) $A_{2ticjg} = \boldsymbol{\beta}_2' \mathbf{X}_{ij} + \boldsymbol{\delta}_2' \mathbf{T}_{tijg} + \boldsymbol{\theta}_2' \mathbf{T}_{2tijg} + \varepsilon_{2ticjg}$.

As in equation (4), the regression in equation (6) includes distinct teacher variables for each teacher-grade-year combination and includes data from small classrooms. The same subscript *2* appears on the coefficients in equation (6) as on those in equation (4) because the two regressions produce identical coefficient estimates; equation (6) applies a correction only to the standard errors.

This multistep method likely underestimates the standard error of the estimated $\boldsymbol{\delta}$ because the adjusted gain in equation (5) relies on the estimated values of λ, ω, and $\boldsymbol{\pi}$. This implies that the error term in equation (6) is clustered within grade-year combinations and within classrooms. This form of clustering typically results in estimated standard errors that are too small, because the subsequent regression steps do not account for variability in post-test scores related to pre-test scores or classroom characteristics. In view of the small number of grade-year combinations, standard techniques of correcting for clustering will not correct the standard errors effectively (Bertrand et al. 2004). Correcting for clustering at the classroom level is also problematic, given small numbers of classrooms per teacher, especially for homeroom teachers. Nonetheless, with the large within-grade and within-year sample sizes, the pre-test coefficients (λ and ω) were precisely estimated, likely leading to a negligible difference between the robust and clustering-corrected standard errors. However, the relationships between classroom composition and the post-test ($\boldsymbol{\pi}$) were less precisely estimated than the pre-test coefficients. Hence, not accounting for that source of variation could lead to more substantial underestimation of the standard errors.

Underestimated standard errors could result in insufficient shrinkage of some teachers' value-added estimates, discussed in Section F, below. When using value-added point estimates for teacher evaluations, the key concern is not whether the standard errors of the estimates are universally underestimated, but whether the standard errors for some teachers are disproportionately underestimated, which can lead to some teacher estimates shrinking too little relative to other teacher estimates in the final step. Thus, there is a trade-off in the design of the model between insufficient shrinkage for some teachers and accounting for classroom characteristics. This approach emphasizes accuracy and face validity of teachers' value-added estimates over any consequences of underestimated standard errors for the shrinkage procedure.[15]

[15] For example, accounting for classroom characteristics may address potential bias from tracking of students into classrooms (Protik et al. 2013).

E. Generalizing estimates to be comparable across grades

1. Transforming estimates into generalized DC CAS points

Both the average and variability of value-added estimates may differ across grade levels, leading to a potential problem when comparing teachers assigned to different grades. 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 sought to compare all teachers 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 did not want to penalize or reward teachers simply for teaching in a grade with atypical test properties, we translated teachers' grade-level estimates from the 2013–2014 school year so that each set of estimates is expressed in a common metric of "generalized" DC CAS points. Aside from putting value-added estimates for teachers onto a common scale, this approach leads to distributions of teacher estimates that are more equal across grades. It does not reflect a priori knowledge that the true distribution of teacher effectiveness is similar across grades. Rather, without a way to distinguish cross-grade differences in teacher effectiveness from cross-grade differences in testing conditions, the test instrument itself, or student cohorts, this approach reflects an implicit assumption that the distribution of true teacher effectiveness is the same across grades.

We standardized the estimated regression coefficients so that the mean and standard deviation of the distribution of teacher estimates is the same across grades. First, we subtracted from each unadjusted estimate the average of all estimates within the same grade. We then divided the result by an estimate of the standard deviation within the same grade. To reduce the influence of imprecise estimates obtained from teacher-grade combinations with few students, we calculated the average using weights based on the total dosage of students taught by each teacher. Our method of calculating the standard deviation of teacher effects also downweights imprecise individual estimates. Finally, we multiplied by the square root of the teacher-weighted average of the grade-specific variances, obtaining a common measure of effectiveness on the generalized DC CAS-point scale.

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

$$(7)\quad \hat{\eta}_{tg} = \left(\frac{\left(\hat{\delta}_{tg} - \overline{\hat{\delta}_g} \right)}{\hat{\sigma}_g} \right) \left(\sqrt{\left(\frac{1}{K} \sum\nolimits_h K_h \hat{\sigma}_h^2 \right)} \right),$$

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

[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. Thus, it is not possible to control directly for grade in the value-added regressions.

The teacher-weighted average of variances is across seven grades for reading and five for math. The calculation in equation (7) is based only on teacher estimates from the 2013–2014 school year; we discarded estimates based on the 2012–2013 school year and all shadow teacher estimates obtained from the regression in equation (6).

In equation (7), we used an adjusted standard deviation that removes estimation error to reflect the 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. Using the unadjusted standard deviations to scale estimates for combining across grades could lead to over- or underweighting one or more grades when the extent of estimation error differs across grades. This is because doing so would result in estimates with the same amount of total dispersion—the true variability of teacher effectiveness and the estimation error combined—in each grade, but the amount of true variability in each grade would not be equal. Instead, we scaled the estimates using the adjusted standard deviation, spreading out the distribution of effectiveness in grades with relatively imprecise estimates so that estimates of teacher effectiveness in each grade have the same true standard deviation.[17]

We calculated the error-adjusted variance of teacher value-added estimates 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 chose 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 estimation error.

Table III.2 shows the adjusted standard deviation of the initial estimates of teacher effectiveness derived from the value-added regression, as well as the weighted average across all grades produced by equation (7). A higher standard deviation for a grade-year combination indicates more dispersion in underlying teacher effectiveness before the transformation into generalized DC CAS points. The standard deviation of value-added estimates ranged from 2.1 to 4.3 DC CAS points in math and from 1.2 to 2.1 DC CAS points in reading. By comparison, the range of the standard deviations of student-level achievement across grades was 15.6 to 17.4 DC CAS points in math and 11.8 to 15.3 points in reading.

2. Combining estimates for teachers of multiple grades

To combine effects across grades into a single effect, denoted as $\hat{\eta}_t$, for a teacher with students in multiple grades, we used a weighted average of the grade-specific estimates (expressed in generalized DC CAS points). We set the weight for grade *g* equal to the proportion of students of teacher *t* in grade *g*. Because combining teacher effects across grades may cause the overall average to be nonzero, we re-centered the estimates on zero before proceeding to the next step.

We computed the variance of each teacher's combined effect as a weighted average of the grade-specific squared standard errors of the teacher's estimates. We set the weight for grade *g*

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

equal to the squared proportion of students of teacher *t* in grade *g*. For simplicity, we assumed that the covariance across grades is zero. In addition, we did not account for uncertainty arising because $\overline{\hat{\delta}_g}$ and $\hat{\sigma}_g$ in equation (7) are estimates of underlying parameters rather than known constants. Both decisions imply that the standard errors will be underestimated slightly.

Table III.2. Adjusted standard deviations of value-added estimates, by subject and grade

	Grade							
Model	4	5	6	7	8	9	10	Average
Math	3.7	3.8	4.3	2.1	3.4	n.a.	n.a.	3.6
Reading	1.9	1.7	1.8	2.1	1.2	1.7	1.3	1.8

Notes: Teachers are included in the calculation of the standard deviation for each grade that they teach.
The average standard deviation is weighted by the number of teachers in each grade.
n.a. = not applicable

F. Shrinkage procedure

To reduce the risk that teachers, particularly those with relatively few students in their grade, will receive a very high or very low effectiveness measure by chance, we applied the empirical Bayes (EB) shrinkage procedure (Herrmann et al. 2013). Using the EB procedure outlined in Morris (1983), we computed a weighted average of an estimate for the average teacher with an estimate from the 2013–2014 school year 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 arrived at the following:

$$(8) \quad \hat{\eta}_t^{EB} \approx \left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\sigma}_t^2} \right) \hat{\eta}_t ,$$

where $\hat{\eta}_t^{EB}$ is the EB estimate for teacher *t*, $\hat{\eta}_t$ 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

[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.

estimate of teacher *t*, and $\hat{\sigma}$ is an estimate of the standard deviation of teacher effects (purged of sampling error), which is constant for all teachers. The term $[\hat{\sigma}^2/(\hat{\sigma}^2+\hat{\sigma}_t^2)]$ 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 smaller $\hat{\sigma}_t^2$ is—the closer $[\hat{\sigma}^2/(\hat{\sigma}^2+\hat{\sigma}_t^2)]$ is to one and the smaller the shrinkage in $\hat{\eta}_t$. Conversely, the larger the variance of the initial estimate, the greater the shrinkage in $\hat{\eta}_t$. 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 calculated the standard error for each $\hat{\eta}_t^{EB}$ using the formulas provided by Morris (1983). As a final step, we removed any teachers with fewer than 15 students and re-centered the EB estimates on zero.

G. Translating value-added results to scores for evaluation systems

We provided OSSE with the original scores in generalized DC CAS points, the percentile rankings for each teacher compared with all DC teachers, and scores on a scale from 1.0 to 4.0. OSSE determined the method for converting the scores in consultation with the technical support committee, a group of representatives from six DC LEAs. The conversion method can be approximately described as follows. First, the value-added estimate stated in terms of generalized DC CAS points (following EB shrinkage, as described in Section F) is rounded to the nearest decimal place. It is then multiplied by a scale factor, and 3.0 is added to each estimate. Next, each estimate is rounded to one decimal place. As a final step, the minimum and maximum scores are imposed, producing a final score between 1.0 and 4.0. By definition, the average DC teacher (including DCPS and charter school teachers) has a score of zero after completing the steps described in Section F, so the average DC teacher receives a score of 3.0 after the conversion. The value-added component constitutes 30 to 50 percent of the total evaluation score for eligible charter school teachers. Each charter LEA determines how it will incorporate this information into its evaluation system.

We provided DCPS with value-added results for DCPS teachers only. Because the other components of IMPACT (the evaluation system for DCPS school-based personnel) are based on DCPS norms, DCPS determined that value-added scores for its teachers should exclude comparisons to charter school teachers. For this reason, before giving the scores to DCPS, we re-centered them using only DCPS teachers; consequently, a DCPS teacher with a score of zero generalized DC CAS points is an average teacher relative to other DCPS teachers. We also provided DCPS with percentile rankings for each teacher compared with DCPS teachers, along with converted scores, known as Individual Value-Added (IVA) scores, from 1.0 to 4.0. The method we used to convert the scores was determined by DCPS and is similar to OSSE's conversion process, but DCPS uses different scale factors for teachers who are above and below average. The average DCPS teacher on this scale receives a score of 3.0.[19] These scores are incorporated into IMPACT.

[19] In reading, DCPS teachers with a value-added estimate of -1.7 or below (in terms of generalized DC CAS points) mapped to an IVA score of 1.0 on the 1.0 to 4.0 scale. For other teachers who were at or below average, the rounded value-added estimates were -1.6 to 0.0, a range that included 17 possible values. These estimates were converted to IVA scores from 1.1 to 3.0, a range that includes 20 possible values. Consequently, there were three IVA scores that did not correspond to any value-added estimate. In particular, there were no final IVA scores in reading of 1.3, 1.9, or 2.4.

Because the generalized DC CAS-based scores provided to DCPS were shifted to be relative to the average DCPS teacher, the scores are not comparable to the scores of charter school teachers. Likewise, the scores on a scale from 1.0 to 4.0 are not comparable between DCPS and charter school teachers because OSSE and DCPS use different comparison groups and different methods of converting scores.

REFERENCES

Arellano, Manuel. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics*, vol. 49, no. 4, November 1987, pp. 431–34.

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.

CTB/McGraw-Hill. *Technical Report for Spring 2011 Operational Test Administration of DC CAS*. Monterey, CA: CTB/McGraw-Hill, 2011.

CTB/McGraw-Hill. *Technical Report for Spring 2012 Operational Test Administration of DC CAS*. Monterey, CA: CTB/McGraw-Hill, 2012.

CTB/McGraw-Hill. *Technical Report for Spring 2013 Test Administration of DC CAS*. Monterey, CA: CTB/McGraw-Hill, 2013.

Herrmann, Mariesa, Elias Walsh, Eric Isenberg, and Alexandra Resch. "Shrinkage of Value-Added Estimates and Characteristics of Students with Hard-to-Predict Achievement Levels." Washington, DC: Mathematica Policy Research, April 2013.

Hock, Heinrich, and Eric Isenberg. "Methods for Accounting for Co-Teaching in Value-Added Models." Washington, DC: Mathematica Policy Research, June 2012.

Hoxby, Caroline, and Gretchen Weingarth. "Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects." Working paper. Cambridge, MA: Harvard University, 2006.

Isenberg, Eric, and Heinrich Hock. "Measuring School and Teacher Value Added in DC, 2011–2012 School Year." Washington, DC: Mathematica Policy Research, 2012.

Isenberg, Eric, and Elias Walsh. "Accounting for Co-Teaching: A Guide for Policymakers and Developers of Value-Added Models." Washington, DC: Mathematica Policy Research, 2013.

Lee, P. *Bayesian Statistics: An Introduction.* Second Edition. New York: John Wiley and Sons, 1997.

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

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

Office of the State Superintendent of Education of the District of Columbia. "Office of the State Superintendent of Education Announces Eligible Schools in the Community Eligibility Option 2013–2014." Washington, DC: Office of the State Superintendent of Education of the District of Columbia, 2013. Available at http://osse.dc.gov/sites/default/files/dc/sites/osse/publication/attachments/SY%2013-14%20DC%20Schools%20Eligible%20for%20CEO.pdf. Accessed June 9, 2014.

Protik, Ali, Elias Walsh, Alexandra Resch, Eric Isenberg, and Emma Kopa. "Does Tracking of Students Bias Value-Added Estimates for Teachers?" Washington, DC: Mathematica Policy Research, March 2013.

Sacerdote, Bruce. "Peer Effects in Education: How Might They Work, How Big Are They, and How Much Do We Know Thus Far?" in *Handbook of the Economics of Education*, vol. 3, edited by Eric Hanushek, Stephen Machin, and Ludger Woessmann. Oxford, UK: Elsevier, 2011.

www.mathematica-mpr.com

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

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

MATHEMATICA
Policy Research