Measuring Teacher Value Added in DC, 2012–2013 School Year

Final Report

January 17, 2014

Eric Isenberg
Elias Walsh

MATHEMATICA
Policy Research

Mathematica Reference Number:
06838.502

Submitted to:
Office of the State Superintendent of Education
810 1st Street, NE
Washington, DC 20002
Project Officer: Jeffrey Noel

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 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Eric Isenberg

Measuring Teacher Value Added in DC, 2012–2013 School Year

Final Report

January 17, 2014

Eric Isenberg
Elias Walsh

MATHEMATICA
Policy Research

ACKNOWLEDGMENTS

We are grateful to the many people who contributed to this report. First, 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 building the value-added model described in this report during research we conducted in the 2011–2012 school year or helping us to implement it in the 2012–2013 school year. This includes individuals currently or formerly at OSSE or DCPS: Robin Chait and Jeffrey Noel at OSSE, and Jason Kamras, Cate Swinburn, Sam Pearcy, and Olivia Meeks at DCPS. We also thank the members of OSSE's 2011–2012 Technical Support Committee, composed of representatives of six DC local education agencies (Joshua Boots, KIPP PCS; Anna Gregory, DCPS; Zachary Morford, Friendship PCS; Alexandra Pardo, Thurgood Marshall PCS; Richard Pohlman, E.L. Haynes PCS; and Megan Reamer, Capital City PCS) for working together to build a value-added model that meets the needs of the DC schools. The members of the Technical Advisory Board—Steve Cantrell, Gates Foundation; Laura Hamilton, RAND Corporation; Eric Hanushek, Hoover Institution at Stanford University; Kati Haycock, Education Trust; David Heistad, Bloomington (MN) Public Schools; Jonah Rockoff, Columbia University; Tim Sass, Georgia State University; and Jim Wyckoff, University of Virginia—provided valuable suggestions.

At Mathematica Policy Research, Emma Kopa and Mary Grider oversaw a team of programmers, including Juha Sohlberg, Dejene Ayele, Jeremy Page, Dylan Ellis, and Becca Wang, who processed the data and provided expert programming. Duncan Chaplin provided valuable comments. Molly and Jim Cameron edited the report, and Jackie McGee provided word processing and production support.

CONTENTS

TABLES

I. OVERVIEW

A. Background

In this report, we update our approach to estimating value-added models 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 2012–2013 school year. To better meet the needs of DC schools, the Office of the State Superintendent of Education (OSSE) approved nine changes to how value added was calculated for the 2012–2013 school year. OSSE authorized Mathematica Policy Research to conduct research during the 2011–2012 school year to explore how changes to the value-added model might affect teachers' value-added estimates if the changes were implemented during the 2012–2013 school year. At the conclusion of that research, OSSE sought our recommendations, gathered advice from an eight-member Technical Advisory Board, and consulted with the Technical Support Committee, composed of six representatives from DCPS and DC charter schools. The changes are (1) expanding the value-added model to include 9th- and 10th-grade reading/English Language Arts teachers; (2) accounting for the influence of classroom characteristics on student achievement; (3) allowing the relationships between student characteristics and student achievement to vary by elementary, middle, and high school grade spans; (4) estimating relationships between student characteristics and student achievement using two student cohorts; (5) accounting for the relationship between student achievement and transferring between schools during the school year; (6) accounting for poverty status using multiple years of data on student free and reduced-price lunch (FRL) status; (7) equalizing student weights in estimating the association between student characteristics and student achievement; (8) explicitly accounting for estimation error when calculating the standard deviation of value-added scores within each grade; and (9) removing students not taught by an eligible teacher from the analysis file.

We provide an overview of the changes to the value-added model in nontechnical terms (remainder of Chapter I); update last year's technical report by describing the data used to estimate teacher value added in the 2012–2013 school year (Chapter II); and provide the details of the statistical methods used to estimate teacher and school value added in 2012–2013, including technical details of the changes to the value-added model (Chapter III). In Chapters II and III, we include tables of diagnostic information that summarize the population of students and teachers on which the value-added estimates are 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 non-technical description of the steps used to estimate value added in DC, please refer to last year's technical report (Isenberg and Hock 2012).

B. Updates to the DC Value-Added Model

1. Include Teachers of 9th- and 10th-Grade Students in Reading/English Language Arts

For the first time, we estimated the value added of reading/English Language Arts (ELA) teachers of 9th- and 10th-grade students, in addition to calculating value added for teachers of 4th- through 8th-grade students. We used 9th-grade reading tests that were given to DCPS students from 2011 to 2013 and to charter school students in 2012 and 2013. We also used the 10th-grade reading tests given to all DC students in 2012 and 2013. OSSE has chosen not to extend value added to high school math teachers because math courses tend to be content specific (for example, algebra or geometry), so a single citywide math test would be unlikely to capture the material for any teacher's particular course. Charter local education agencies (LEAs) chose individually whether to participate in providing data for value-added estimates for high school reading/ELA teachers. Therefore, value-added estimates for high school teachers should be interpreted as how a teacher compares to teachers at other participating schools, not to those at all RTT schools.

As with students in upper elementary and middle school grades, we accounted for prior test scores. For 9th-grade students, we accounted for math and reading pre-test scores from 8th grade. Because there is not a statewide standardized test in math given in 9th grade, for 10th-grade students we used a reading pre-test from 9th grade and a math pre-test from 8th grade. As in the past, the relationships between pre-test and post-test scores were allowed to vary by grade level—we estimated distinct relationships for each pre-test grade level from grades 3 through 7 in math and grades 3 through 9 in reading.

2. Account for Classroom Characteristics

In addition to individual student characteristics, we accounted for a set of classroom characteristics that allow for the possibility that students may perform differently in different classroom environments. For example, students may perform better when they have higher-performing or more advantaged peers. Another possibility is that it may be more difficult for a teacher to target instruction in a class with a greater diversity of skill levels. Thus, at all grade levels, we accounted for average classroom achievement and the variability of classroom achievement. For grade 6 and above, we also accounted for the proportion of students eligible for FRL.

Our methodology allows us to distinguish the effect of classroom composition on student achievement from the correlation between classroom composition and teacher effectiveness that can arise from the way in which teachers are matched to schools and classrooms. For example, assume that more effective teachers are incentivized to teach at schools with many disadvantaged students. In this example, even if students did not affect one another's performance in class, teacher-level student characteristics and effective teaching would be correlated. However, this association would be caused by the way in which teachers are matched to students, rather than by how students affect each other's achievement.

Therefore, we estimated the effects of classroom composition on the achievement of a given student by comparing students in multiple classrooms for a given teacher. By relying on differences in classroom composition for multiple classrooms of the same teacher, we isolated the relationship between classroom composition and student achievement without confounding the effect of classroom composition with the way in which teachers are matched to students. For homeroom teachers, we used two years of data to obtain more than one classroom per teacher. To obtain multiple classrooms for departmentalized math and reading teachers (more common at the middle school level), we used data on multiple classrooms within each year as well as between years. (Because definitions of student FRL changed from year to year, we relied on variation across classrooms within the 2012-2013 school year to estimate the relationship between the proportion of FRL students and individual student achievement for middle school students. We excluded the proportion FRL from the value-added model for grades 4 and 5 since we relied mainly on year-to-year variation in the composition of a teacher's classroom in these grades.)

3. Allow Relationships Between Characteristics and Achievement to Vary by Grade Span

We allowed the relationships between all student characteristics and student achievement to vary for each of three grade spans: grades 4 and 5, grades 6 through 8, and grades 9 and 10. By allowing for separate relationships by grade span, we targeted relationships between student characteristics and achievement more accurately. In the past, only relationships between pre-test scores and achievement were allowed to vary for each individual grade level; for other student characteristics, estimates were pooled so that the relationships did not vary across grade levels. The inclusion of high school reading/ELA teachers in the value-added model increased concern about pooling estimates across grades: the relationships between student characteristics and achievement in 4th grade may differ from those in 10th grade. In addition, the relationships between classroom composition and student achievement may differ at elementary, middle, and high school grade levels, further reinforcing the desirability of allowing for more variation by switching from a pooled approach to a grade-span approach.

In prior years, by pooling all grades together and "borrowing strength" from other grade spans, we were able to estimate relationships more precisely than by doing so within a grade span. Compared to value-added models from the 2010–2011 school year and before, however, the inclusion of charter school students adds precision by increasing the number of students on which value-added models are based. To further increase precision, we added an additional cohort of data, as explained in the next section. The grade-span approach represents an intermediate option between the pooled approach and an individual grade-level approach. 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.

4. Estimate Relationships Between Characteristics and Achievement Using Two Student Cohorts

We used two student cohorts to estimate the model, although teachers' value added was calculated based on their students in the 2012–2013 school year with post-tests from 2013 and pre-tests from 2012. We included the student cohort from the 2011–2012 school year (with post-tests from 2012 and pre-tests from 2011) to (1) provide more precise estimates of the relationships between student characteristics and achievement and (2) provide multiple classrooms to estimate the association between classroom composition and student achievement within a teacher. For homeroom teachers who teach a single classroom each year, the second student cohort allowed us to estimate classroom composition relationships. For teachers in middle and high school grades, who generally had multiple classrooms within a year, adding a second cohort improved the precision of these estimates by adding more classrooms per teacher. Including the extra cohort of students also improved the estimates of the relationships between student characteristics and achievement, but did not affect the students or students' achievement data that directly inform a teacher's value-added estimate. Teachers' value-added estimates measure their contributions to student achievement during the 2012–2013 school year, not the contributions made during the 2011–2012 school year.

5. Account for Relationship Between Student Achievement and Transferring Schools During the School Year

In addition to accounting for past test scores, poverty status, and other characteristics, we accounted for whether a student transferred into or out of a teacher's school in the middle of the year. Students who transferred schools during the school year may have experienced a disruptive environment outside of school that led to lower performance compared to similar students who did not transfer schools. By accounting for mid-year transfers, teachers of a transfer students were not held accountable for circumstances outside their control that were related to these disruptions.

6. Account for Poverty Status Using Multiple Years of Data

We used information on the poverty status of individual students for up to four school years, beginning with the 2009–2010 school year. In the past, we included indicators to account for differences in student achievement by paid, free, and reduced-price lunch eligibility in the current year and imputed missing data for these variables for students enrolled in community-eligible (CE) schools, which do not collect annual information about individual student poverty status. The number of CE schools grew considerably in the 2012–2013 school year, however, decreasing the viability of imputing individual-level data missing for this reason. Thus, for the 2012–2013 value-added model, we included indicators tracking a student's status in each year since the 2009–2010 school year. For each of those four years, we categorized students as belonging to one of these groups: ineligible for FRL, eligible for free lunch, eligible for reduced-price lunch, attending a CE school, or lacking information on FRL eligibility that year. Although many CE students did not have an individual status in the 2012–2013 school year, some of these students did have an individual status in a previous year when this information was collected. Including students' prior poverty status information can help address the limitations of information from the current school year, so that teachers are not held accountable for factors beyond their control. Because not all poverty status categories were universally available for all years, the specific categories included vary by year.

7. Equalize Student Weights (Full Roster-Plus Method)

We equalized the weight given to each student when estimating the association between student characteristics and student achievement, while still apportioning teachers' responsibility for the student based on the fraction of the year they spend with the teacher, regardless of how many teachers claimed them. To achieve this, we designed a method of accounting for team-teaching called the Full Roster-Plus Method (FRM+), an improvement of the Full Roster Method (FRM) that we had developed previously. Under the FRM, we accounted for team teaching using multiple teacher-student links for each student and weights based on the amount of time each teacher taught the student (Hock and Isenberg 2012). The FRM allows for value-added estimates to be calculated for any variety of team-teaching circumstances, preserving the relative weights of students within a teacher. A side effect of the FRM, however, is that it effectively double-counts students taught by two teachers, so that these students received a full weight with each of their teachers. This causes these students to receive extra weight when calculating the relationship between student characteristics and achievement.

The FRM+ has all of the features of the FRM but offers an additional benefit: the FRM+ ensures that each student contributed equally to the calculation of the relationship between student characteristics and achievement. The FRM+ accomplishes this overall equal weighting by duplicating each teacher-student link, assigning weights to the new links so that each student has records that sum to the same total weight. Each student thereby contributes equally to the estimates of student characteristics without affecting the proportional contributions of co-taught students to teachers' scores.

8. Account for Estimation Error in Calculating the Standard Deviation of Value Added

In the past, after calculating initial teacher value-added estimates using a multiple regression model, we have multiplied each teacher's estimate by a grade-specific conversion factor to ensure that the dispersion of teacher value-added scores is similar in each grade. This assumes that the variation in teachers' effectiveness is similar across grade levels. The grade-specific conversion factor was equal to the average standard deviation of value-added estimates across all grades, divided by the standard deviation of value-added estimates for the teacher's grade.

We changed the method for calculating the standard deviation of teacher effects within each grade to explicitly account for "estimation error"—that is, the imprecision with which teacher effectiveness is measured. The measured variability of value-added estimates reflects two parts—the true underlying variance of teacher effectiveness and the estimation error. By including both the true variance and estimation error, the unadjusted standard deviation of value-added estimates may overstate the true variability of teacher value added and could lead to over- or underweighting one or more grades when combining grade-specific scores for teachers of multiple grades. The new method better reflects our assumption that the variation in teachers' true effectiveness is the same for all grades.

9. Remove Students Not Linked to an Eligible Teacher from the Analysis File

We excluded from the analysis file those students not taught by a teacher eligible to receive a value-added estimate. In the past, to increase the precision of the estimates, we included these students in the model by linking them to a single catch-all ineligible teacher. However, the estimation of the relationship between student characteristics and achievement then was then based in part on cross-school differences in students. To ensure better accuracy, we dispensed with links to a catch-all ineligible teacher, including links for students who were also linked to an eligible teacher for at least part of the school year.

II. DATA

In this chapter, we review the data used to generate the value-added measures. 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.

A. DC CAS Test Scores

We included in the analysis file 4th- through 10th-grade students with a DC Comprehensive Assessment System (CAS) test from 2013 (the post-test) if they had a DC CAS test from the previous grade in the same subject in 2012 (the pre-test).[1] We excluded students from the analysis file in the case of missing or conflicting test score or student background data.[2,3] 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 not taught by a DC teacher for at least 5 percent of the school year, (2) were included in the roster file but not claimed by a teacher, or (3) 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 with so few students). We applied analogous rules for inclusion in the model to students with a DC CAS test from 2012.[4] After applying these rules, we reported estimates only for teachers who taught 15 or more students over the course of the 2012–2013 school year in at least one subject. For example, we would report an estimate for a teacher who claimed eight students in grade 4 and seven students in grade 5. For a teacher who claimed nine students in grade 4 and six students in grade 5, however, the grade 5 students would not be linked to the teacher because they would not meet the seven-student minimum in that grade level. So we would not report a value-added estimate 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 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 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, 87.3 percent of students with 2013 test scores were

---

[1] We excluded most grade 10 charter school students in the 2011–2012 school year from the analysis file because the grade 9 reading test was first given to charter school students in 2012, so we could not include a pre-test from the prior grade and year. Of these students, 13 remain in the analysis file because they were tested in DCPS in 2011.

[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 individual student background characteristics but excluded those for whom no data on background characteristics were available.

[4] DCPS provided us with DC CAS test scores for DCPS students from 2010 and 2011. OSSE provided us with DC CAS scores for charter school students from 2011 and for all DC students in 2012 and 2013.

included in the value-added model for math, and 83.9 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.[5] For the value-added model, we dropped the first digit and used the rest of the score, which ranged from 0 to 99.

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 created subject- and grade-specific z-scores by subtracting the mean and dividing by the standard deviation within a subject-grade combination.[6] This step allowed us to translate math and reading scores in every grade and subject into a common metric. To create a measure with a range resembling the original DC CAS-point metric, we then multiplied each test score by the average standard deviation across all grades within each subject and year.

Table II.1. Reasons Students Tested in 2013 Were Excluded from the Analysis Files

	Math		Reading	
	Number	Percent	Number	Percent
Students with Post-Test Scores	18,330	100.0	23,664	100.0
(1) Student has conflicting post-test scores	0	0.0	2	0.0
(2) Missing same-subject pre-test scores	1,519	8.3	2,531	10.7
(3) Skipped or repeated a grade	305	1.7	643	2.7
(4) Not linked to an eligible teacher	508	2.8	639	2.7
Total Excluded	2,332	12.7	3,815	16.1
Total Included in Value-Added Model	15,998	87.3	19,849	83.9

Notes: Students are included in this table only if they were included in roster confirmation and could be linked to background characteristics. This excludes only those enrolled in charter schools not participating in Race to the Top.

Students are excluded sequentially in the order presented and so do not count 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.

B. Student Background Data

We used data provided by OSSE and DCPS to construct variables used in the value-added model as controls for student background characteristics. The value-added model accounts for the following:

---

[5] 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.

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

- Pre-test in same subject as post-test
- Pre-test 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 will account for three classroom characteristics:

- Average classroom pre-test scores
- Standard deviation of classroom pre-test scores
- Fraction of free or reduced-price eligible students in the classroom

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, the other variables are binary, taking a value of zero or one.

To account for poverty status, we used data on FRL eligibility for multiple school years, beginning with 2009–2010. Including students' prior poverty status is particularly important for those attending a CE school, because these schools do not collect annual information about individual student poverty status. Schools are eligible to become CE if they have a student population composed of at least 40 percent with 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 forgo collecting individual student FRL applications. In the 2012–2013 school year, the number of CE schools in DC grew from 66 DCPS schools and no charter schools to 88 DCPS schools and 27 charter schools.

In addition to creating indicators for paid, free, or reduced-price lunch status, we created additional indicators for students who attended a CE school and those who had no available information that year. Some individual students attending a CE school were indicated as eligible for free lunch via a direct certification process and included as eligible for free lunch in the analysis file, and a small number of others had an individual status indicating a paid lunch status and were included as such. The CE category was used only for the 2012–2013 school year because no charter schools had adopted community-eligibility status before the 2012–2013 school year. Before 2012–2013, DCPS students attending a CE school and not identified as eligible for free lunch via direct certification were indicated as having no information. Each student belonged to one category for each school year. For predicting post-test scores of students for the 2011–2012 school year, we did not include information on their status from the 2012–2013 school year. The student poverty status indicators for each school year are summarized in Table II.2.

Table II.2. Method of Accounting for Student Poverty Status

School Year of Status	Ineligible for Free or Reduced-Price Lunch	Eligible for Free Lunch	Eligible for Reduced-Price Lunch[a]	Attended Community-Eligible School[b]	No Information[c]
2009–2010	Included	Included			Included
2010–2011	Included	Included	Included		Included
2011–2012	Included	Included	Included		Included
2012–2013	Included	Included	Included	Included	

Notes: Each student is defined as belonging to a single category for each status year.

Community-eligible schools are schools that do not annually collect poverty status for individual students.

[a]We included students eligible for reduced-price lunch with those eligible for a free lunch in the 2009–2010 school year because data are not available to distinguish free and reduced-price lunch status for some students that year.

[b]We included students who attended community-eligible schools in the 2009–2010, 2010–2011, or 2011–2012 school years with those eligible for free lunch if another data source was available to certify their free lunch status and in the "no information" category otherwise. We included those who attended community-eligible schools in the 2012–2013 school year with students eligible for free lunch if another data source was available to certify their free lunch status and in the "attended community-eligible school" category otherwise.

[c]For students in the "no information" category in the 2012–2013 school year, we imputed poverty status for the three included categories, using the same imputation method as that for students missing other student background data.

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. We calculated the percentage of students in a classroom who were eligible for FRL only for those in grades 6 to 8 in the 2012–2013 school year. We calculated this percentage as the number of students identified in the 2012–2013 school year as eligible for free lunch, reduced-price lunch, or who attended a CE school, divided by the number of students in the classroom and excluding students with no information.[7] All three 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; for each student, the classroom characteristics measured the characteristics of others 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 2009–2010 to 2011–2012 school years; instead, we included a "no information" indicator for these students. Because there were few students with missing poverty

[7] In contrast to the individual poverty status CE variable, we consider all students attending a CE school as free or reduced-price eligible for the classroom characteristic calculation, including a small number whose individual status was paid lunch. This choice ensures that the relationship between classroom poverty status and achievement is estimated based on variation between classrooms in schools that collected individual poverty status information in the 2012–2013 school year.

information in the 2012–2013 school year, however, a "no information" indicator would have been relatively imprecisely estimated. Therefore, for these students, we imputed values for the 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.[8] Finally, we did not generate imputed values for the same-subject pre-test; we dropped from the analysis file any students with missing same-subject pre-test scores.[9]

Table II.3 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 1.0 percentage point.

Table II.3. Characteristics of Students from the 2012–2013 School Year in the Reading Value-Added Model

	Grades 4 and 5		Grades 6 to 8		Grades 9 and 10	
	Number	Percent	Number	Percent	Number	Percent
Included in Value-Added Model	6,867	100.0	9,372	100.0	3,610	100.0
Eligible for free lunch	3,729	54.3	5,645	60.2	2,037	56.4
Eligible for reduced-price lunch	194	2.8	355	3.8	159	4.4
Attended community-eligible school	1,630	23.7	1,750	18.7	716	19.8
Ineligible for free or reduced-price lunch	1,314	19.1	1,622	17.3	698	19.3
Limited English proficiency	428	6.2	553	5.9	231	6.4
Specific learning disability	475	6.9	768	8.2	293	8.1
Other learning disability	462	6.7	580	6.2	189	5.2
Transferred schools during school year	106	1.5	151	1.6	67	1.9

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 a DCPS and charter school. 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 2012–2013 school year.

Students who attended community-eligible schools are included with students eligible for free lunch if another data source was available to certify their free lunch status.

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

---

[8] 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).

[9] Less than 1 percent of students in the value-added analysis file had missing opposite 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 16 percent of grade 10 students.

C. School and Teacher Dosage

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/ELA teachers in grades 9 and 10.[10] Each LEA determined the list of eligible teachers, using guidance from OSSE that teachers with primary responsibility for providing instruction in math and reading/English language arts 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, but resource and Read 180 teachers were eligible. Resource teachers provide additional instruction to students and tend to work with a large number of students throughout the school year.

Given that some students moved between schools or were taught by a combination of teachers, we apportioned their achievement among more than one school or teacher. We refer to the proportion of time the student was enrolled at each school and with each teacher as the "dosage."

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 their tests shortly after the end of the third quarter.[11]

Recognizing that a teacher is unlikely to have an appreciable educational impact on a short-term student, we set dosage as equal to zero for students who spent less than 5 percent of the year at a school. Apart from this, 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 2012–2013 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. Teachers could also add

---

[10] 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 through 10 who were eligible to receive individual value-added scores.

[11] There are two exceptions. One LEA uses trimesters rather than quarters. In a school at another LEA, 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.

students to their rosters. In other cases, teachers were responsible for adding all of the students to their rosters. 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 this instance, the dosage was 40 percent. 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, principal, or designee verified the confirmed rosters. Likewise, in DCPS, 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 the three terms for a given LEA were of equal length, a student spent the first two terms in a teacher's school, and the teacher claimed the student for 60 percent dosage during those terms, the teacher dosage would be 0.67*0.60 = 0.40. 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 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 2011–2012 school year. Table II.4 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, 14.8 percent of math teachers and 16.3 percent of reading teachers shared all of their students with other teachers, whereas 73.8 percent of math teachers and 59.1 percent of reading teachers did not share any students with another teacher.

Table II.4. Teachers Receiving Value-Added Estimates, 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	315	73.8	315	59.1
1–10 percent	25	5.9	46	8.6
11–50 percent	17	4.0	57	10.7
51–99 percent	7	1.6	28	5.3
All students	63	14.8	87	16.3
Total	427	100.0	533	100.0

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 through 8) and three in reading (grades 4 and 5, 6 through 8, and 9 and 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 variable $O_{i(j-1)}$ denotes the pre-test in the opposite subject. Thus, when estimating teacher effectiveness in math, S represents math tests with O representing reading tests, and vice versa. The pre-test scores capture prior inputs into student achievement. 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 capture prior inputs into student achievement, and 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, $\boldsymbol{\beta}$, 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).

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 2011–2012 and 2012–2013 school years had four variables in $\mathbf{T}_{tijg}$. Under the FRM+, described in detail in Section B, we duplicated each teacher-student-classroom-year link in the analysis file so that each student contributes the same total dosage to the calculation of the parameters. We included teacher links for the duplicate—or "shadow"—student-classroom-year observations as 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 2012–2013 school year were contained in the coefficient vector $\boldsymbol{\delta}$ for teacher-grade combinations from that school year (j = 2013). We did not directly use $\boldsymbol{\delta}$ for j = 2012 or the coefficient vector $\boldsymbol{\theta}$ to measure teacher effectiveness in the 2012–2013 school year. 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.[12]

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). The second panel shows the association between a given student characteristic and achievement. The bottom panel shows the association between each of the three classroom characteristics and achievement.

To account for team teaching, we used the Full Roster-Plus Method, 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 contribute equally to student achievement within each team (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 2011–2012 and 2012–2013 school years.

Because some students contributed multiple observations, we estimated the coefficients by using weighted least squares (WLS) rather than ordinary least squares (OLS). In this method, the teacher-grade-year variables in $\mathbf{T}_{tijg}$ are binary, and we weighted each teacher-student combination by the teacher dosage associated with that combination. We addressed the correlation in the error term, ε_{tijg}, across multiple observations 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.

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. The method includes three regression steps:

1. **Account for classroom characteristics (average and standard deviation of pre-test achievement and percentage eligible for FRL).** 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 2012–2013 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

[12] Mean centering the student characteristics and pre-test scores tends to reduce the estimated standard errors of the teacher effects (Wooldridge 2008).

on student achievement only from the 2012–2013 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.[13] We then aggregated teacher estimates across grades to form a single estimate for each teacher (see Section E below).

[13] 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 coefficients on the covariates that are estimated more accurately. 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 displays the regression coefficients from the value-added model. In other words, it describes the relationships between the characteristics of DC students and achievement on the post-test. 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 0.77 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 0.77 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 7.46 points higher than the prediction if the student instead has 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 7.46 points might suggest; the change in predicted scores associated with a change in attendance from 75 to 100 percent is 1.87 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 have limited English proficiency are predicted to score 0.07 points lower than students who do not have limited English proficiency. 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 all other characteristics included in the model. Put another way, 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 all of 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 counter-intuitive 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.

To put the magnitude of the coefficients into perspective, they 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 16.4 DC CAS points, indicating that about two- thirds of students scored within 16.4 points above or below the average score on the assessment. The standard deviations for other grades ranged from 14.6 to 17.3 points in math and from 12.2 to 14.8 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 Math and Reading Value-Added Models, by Grade Span

	Math		Reading		
Variable	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	0.77 (0.01)	0.68 (0.01)	0.61 (0.01)	0.62 (0.01)	0.58 (0.01)
Opposite subject, all grades in span	0.09 (0.01)	0.14 (0.02)	0.16 (0.01)	0.16 (0.01)	0.07 (0.01)
Individual Student Background Characteristics					
Limited English proficiency	-0.07 (0.32)	0.47 (0.32)	-1.56 (0.33)	-0.90 (0.30)	-0.11 (0.53)
Specific learning disability	-2.08 (0.36)	-0.97 (0.31)	-2.81 (0.35)	-1.86 (0.26)	-2.80 (0.45)
Other learning disability	-1.98 (0.38)	-1.96 (0.39)	-2.85 (0.36)	-2.73 (0.33)	-1.67 (0.68)
Transferred schools during the school year	-1.66 (0.69)	-1.66 (0.62)	-1.10 (0.57)	-0.72 (0.52)	-2.55 (1.02)
Fraction of the prior year student attended school	7.46 (1.67)	7.64 (1.65)	-0.80 (1.44)	0.96 (1.34)	-1.52 (2.27)
Ineligible for free or reduced-price lunch	0.51 (0.51)	0.68 (0.40)	0.08 (0.48)	0.62 (0.35)	0.42 (0.50)
Eligible for reduced-price lunch	-0.58 (0.61)	-0.17 (0.55)	0.06 (0.52)	-0.46 (0.46)	-0.16 (0.74)
Attended community-eligible school	-0.07 (0.38)	0.73 (0.37)	0.24 (0.29)	0.52 (0.30)	-0.37 (0.56)
Classroom Characteristics					
Average classroom pre-test score	-0.08 (0.02)	0.10 (0.01)	0.03 (0.02)	0.04 (0.01)	0.18 (0.02)
Standard deviation of classroom pre-test scores	-0.09 (0.02)	0.07 (0.02)	-0.09 (0.02)	-0.04 (0.02)	0.02 (0.03)
Fraction free or reduced-price lunch in classroom	n.a.	-1.56 (0.49)	n.a.	0.62 (0.39)	0.88 (0.99)

Notes: Standard errors are in parentheses.

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

For students in grades 4–9, pre-test scores are from the prior grade in the 2011–2012 school year. For 10th-grade students, same-subject pre-test scores are from 9th-grade test scores in the 2010–2011 school year, and opposite-subject pre-test scores are from 8th-grade test scores in the 2009–2010 school year.

The table excludes coefficients on pre-test variables estimated separately for students from the 2011–2012 school year. Additionally, the table excludes coefficients on variables that indicate poverty status between the 2009–2010 and 2011–2012 school years. The poverty status variables reported in the table indicate students' poverty status in the 2012–2013 school year. All coefficient estimates of variables that indicate students' poverty status in previous years are no larger than 1.9 DC CAS points in absolute value.

Students who attended community-eligible schools are included with students eligible for free lunch if another data source was available to certify their free lunch status.

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

n.a. = not applicable

B. Full Roster-Plus Method

The Full Roster-Plus Method (FRM+) is a modification of the FRM used in the DC value-added model estimated in 2011–2012. FRM+ handles co-teaching exactly as the FRM, but equalizes the contribution of students taught by multiple teachers to the estimation of the coefficients on student background characteristics. 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.[14,15]

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. If these classroom characteristics influence student achievement, it is possible that omitting them would produce biased measures of teacher effectiveness. For two reasons, classroom characteristics may predict student achievement. First, the mix of students in each classroom could affect individual student achievement. These direct impacts of classroom peers on a student's achievement sometimes are called peer effects (Hoxby and Weingarth 2006; Sacerdote 2011). Second, classroom characteristics may help to account for measurement error. The method used to account for measurement error in the pre-tests (described in Section D) may not account for all measurement error. Including classroom characteristics could address this additional measurement error if students' true achievement levels are related to their peers' characteristics. The effect of measurement error on a classroom characteristic coefficient estimate could be positive or negative, depending on the direction of the relationship between the measurement error and achievement. Thus, the classroom coefficient estimate could be positive or negative, depending on the magnitude of this effect.

The vector $\mathbf{C}_{ticj}$ in equation (1) represents three classroom characteristics included in the model: mean classroom pre-test score, the standard deviation of classroom pre-test scores, and the

[14] Just as in the FRM, 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.

[15] Because we did not create links to a catch-all ineligible teacher, some students had a total dosage of less than one hundred percent across all teacher-student links. Under FRM+, as under FRM, co-teachers are functionally considered a team that receives the same value-added estimate, so the value-added estimate of an eligible teacher who shared students with ineligible teacher(s) would not have directly changed whether or not we had created teacher-student links to a catch-all ineligible teacher.

proportion of FRL students in the classroom.[16,17] We used all three variables for the middle and high school grade spans, but excluded FRL from the elementary grade span. Additionally, we calculated the proportion of FRL students for records from the 2012–2013 school year but not from 2011–2012. We calculated this proportion only for a single school year to ensure that all comparisons in the proportion of FRL students were across classrooms in the same year. Comparisons across years could lead to spurious relationships with student achievement because of previously described changes in the FRL data between years (Table II.2). The exclusion of the proportion of FRL students from the elementary school grade span was a consequence of our decision to calculate this proportion for a single year. Multiple years of FRL data would have been necessary to obtain the multiple classrooms needed to account for classroom characteristics of 4th- and 5th-grade teachers, most of whom are homeroom teachers.

Estimation of a model accounting for classroom characteristics required a multistep strategy because we constrained a teacher effect 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 a 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 their performance in 2011–2012 did not directly affect a measure of their performance in 2012–2013.[18]

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, we did not pool classrooms across schools for teachers who changed schools from one year to the next when estimating the contribution of classroom composition. 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 $\mathbf{G}_g$, of 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 the subsequent regression steps, because the teacher variables in (2) are not grade and year specific. We

---

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

[17] The fraction of FRL students in the classroom was calculated only for classrooms in the 2012–2013 school year, and set to zero for all classrooms in 2011–2012. Consequently, the relationship between classroom poverty status and achievement is estimated based on poverty data from only the 2012–2013 school year.

[18] 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 classrooms taught by resource teachers, as indicated by DCPS or OSSE. Six percent of records were excluded from the first-stage elementary and middle school grade-span regressions, and 10 percent of records were excluded from the first-stage high school grade-span regression.

corrected for measurement error in the pre-test scores when estimating equation (2). In Section D, 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 (3) for students in most classrooms, we used the same values of $\mathbf{C}_{ticj}$ from equation (1). For students in small classrooms, 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 2012–2013 school year. In Section D, we describe how we used the adjusted post-test to produce single-year estimates of teacher effectiveness.

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 2010, 2011, 2012). 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 test/retest 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 through 9, we used the published reliabilities associated with the 2012 DC CAS for records from the 2012–2013 school year and the 2011 DC CAS for records from the 2011–2012 school year. 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, because the math pre-test is from grade 8. For grade 10 students in the 2011–2012 school year, we used the reliabilities associated with the 2012 DC CAS for reading (because no 2011 reliability data were available) and the 2010 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.

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

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:

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

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 computationally possible 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):

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

As in (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 it did in 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 relationships, which 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. 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.[19]

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.[20] 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 2012–2013 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 number 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.

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

[20] 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.

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

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

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. 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 2012–2013 school year; we discarded estimates based on the 2011–2012 school year and all shadow teacher estimates also 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. 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 ensures 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.[21]

We calculated 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 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 2.9 to 4.1 DC CAS points in math and from 1.3 to 2.3 DC CAS points in reading. By comparison, the range of the standard deviations of student-level achievement across grades was 14.6 to 17.3 DC CAS points in math and 12.2 to 14.8 points in reading.

[21] 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.

Table III.2. Student-Weighted Standard Deviations of Value-Added Estimates

Model	Grade 4	5	6	7	8	9	10	Weighted Average
Math	4.0	3.5	4.1	2.9	3.0	n.a.	n.a.	3.6
Reading	2.3	2.1	1.7	1.8	1.3	2.3	1.3	1.9

Notes: Teachers are included in the calculation of the standard deviation for each grade that they teach, weighted by the number of students they teach in that grade.

n.a. = not applicable

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 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.

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 2012–2013 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.[22] Following the standardization procedure, the overall mean is zero, with better-than-average teachers having

[22] 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 increases the weight on the overall mean by (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.

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 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 generalized DC CAS point score, percentile rankings for individual teachers compared to all DC teachers, and a score converted to a scale from 1.0 to 4.0. OSSE determined the method for converting the score in consultation with the Technical Support Committee, a group of representatives from six DC LEAs. In this system, the average DC teacher (including DCPS and charter school teachers) receives a score of 3.0. The value-added component constitutes 30 to 50 percent of the total evaluation score for eligible charter school teachers, but each charter LEA determines the exact way in which it will incorporate this information into its evaluation system.

We provided DCPS with value-added results only for DCPS teachers. Because the other components of a teacher's evaluation in IMPACT (the evaluation system for DCPS school-based personnel) are based on DCPS norms, DCPS determined that value-added scores for their teachers should exclude comparisons to charter school teachers. For this reason, we re-centered the scores using only DCPS teachers before we provided value-added scores to DCPS. 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 compared to DCPS teachers, with a converted score that runs from 1.0 to 4.0 based on a method determined by DCPS. The average DCPS teacher on this scale receives a score of 3.0. The score on the 1.0–4.0 scale is incorporated into IMPACT.

Given that the generalized DC CAS point scores provided to DCPS were shifted to be relative to the average DCPS teacher, DC CAS-point scores provided to DCPS teachers and DC CAS-point scores provided to charter school teachers are not comparable. 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 2010 Operational Test Administration of DC CAS*. Monterey, CA: CTB/McGraw-Hill, 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.

Herrmann, Mariesa, Elias Walsh, Eric Isenberg, and Alex 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.

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.

Protik, Ali, Elias Walsh, Alex 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.

Wooldridge, Jeffrey. *Introductory Econometrics: A Modern Approach*. Fourth Edition. Mason, OH: South-Western/Thomson, 2008.

MATHEMATICA
Policy Research

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