Measuring School and Teacher Value Added for IMPACT and TEAM in DC Public Schools

Final Report

August 20, 2010

Eric Isenberg
Heinrich Hock

MATHEMATICA
Policy Research, Inc.

MPR Reference Numbers:
06742.150
06325.415

Submitted to:
DC Public Education Fund
1534 14th Street, NW
Washington, DC 20005
Project Officer: Cate Swinburn

District of Columbia Public Schools
1200 First Street, NE
Washington, DC 20002
Project Officer: Hella Bel Hadj Amor

New Leaders for New Schools
30 West 26th Street
New York, NY 10010
Project Officer: Dianne Houghton

Submitted by:
Mathematica Policy Research
600 Maryland Avenue, SW
Suite 550
Washington, DC 20024-2512
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Project Director: Eric Isenberg (06742)
Project Director: Duncan Chaplin (06325)

Measuring School and Teacher Value Added for IMPACT and TEAM in DC Public Schools

Final Report

August 20, 2010

Eric Isenberg
Heinrich Hock

ACKNOWLEDGMENTS

We are grateful to the many people who contributed to this report. Special thanks go to Chris Mathews and Dianne Houghton at New Leaders for New Schools and Cate Swinburn of the District of Columbia Public Education Fund for their support of our work. We thank Hella Bel Hadj Amor, Jason Kamras, and Erin McGoldrick at the District of Columbia Public Schools for working together to build a value-added model that meets the needs of the District of Columbia Public Schools. We also thank Eric Hanushek and Tim Sass, the independent reviewers who made valuable suggestions for improvement.

At Mathematica Policy Research, Mary Grider, assisted by Emma Ernst, Francesca Palik, and Jeremy Page, processed the data and provided expert programming. Duncan Chaplin and Steven Glazerman provided valuable comments. Amanda Bernhardt and Betty Teller edited the report, and Lisa Walls and Jackie McGee provided word processing and production support.

CONTENTS

I. OVERVIEW

The District of Columbia Public Schools (DCPS) has incorporated measures of school and teacher effectiveness, based on student test score growth, into a new teacher assessment system known as IMPACT. At the same time, New Leaders for New Schools (New Leaders) has been working with DCPS to offer financial awards to effective educators in the district. To support these efforts, both organizations asked Mathematica Policy Research to design a value-added model to measure school and teacher performance in the district. Mathematica developed these measures by adapting and tailoring methods used in our earlier work for New Leaders for other schools and districts (Booker and Isenberg 2008; Booker et al. 2008; Isenberg 2008; Potamites et al. 2009a; Potamites et al. 2009b).[1]

Implemented for the first time during the 2009–2010 school year, IMPACT is an assessment system with significant consequences. Prior to the start of the school year, DCPS categorized all staff into one of 20 groups. Everyone received an IMPACT score based on a point-based formula that was tailored to the job responsibilities of their group and availability of data. Individual value-added scores constituted half of the IMPACT score of "Group 1" teachers, who were regular education teachers in grades and subjects with sufficient student test score data.[2] Most of the remaining points for these teachers depended on a series of structured classroom observations. In addition, for almost all groups (including Group 1), school value-added scores counted for five percent of the IMPACT score. Based on their IMPACT score, teachers were placed into one of four performance categories. Teachers in the lowest category were subject to separation; those in the highest category were eligible for additional compensation.

Results from the school value-added model will also be used by DCPS and New Leaders as part of the TEAM (Together Everyone Achieves More) program. TEAM is designed to encourage and identify effective leadership and teaching practices by providing financial awards to all staff—principals, teachers, and others—in schools that produce the largest gains in student achievement as measured by their value added.

DCPS and New Leaders sought an objective, fair, and transparent value-added model to assess school and teacher effectiveness. Mathematica developed such a model in accord with these principles. It was reviewed by two independent value-added experts, Eric Hanushek of the Hoover Institution at Stanford University and Tim Sass of Florida State University. In the rest of this chapter, we describe the main features of the method in nontechnical terms. Chapter II presents the data used, and Chapter III focuses on the technical details of the statistical methods.

[1] This project has been funded by the DC Public Education Fund, DCPS, and New Leaders.

[2] DCPS plans to expand testing in future years so that more teachers will be covered by Group 1. For more details on IMPACT, see http://dcps.dc.gov/DCPS/In+the+Classroom/Ensuring+Teacher+Success/IMPACT+(Performance+Assessment).

A. Using Value-Added to Measure Performance

Many commonly used measures of school and teacher effectiveness provide an incomplete picture. In many districts, teachers are evaluated based on observations by the principal that provide a snapshot of performance but do not necessarily indicate how much students learn as a result of the teacher's talent and efforts. Schools are often ranked by their students' average test score or the percentage of students who meet state proficiency standards, measures that do not account for prior learning or other student characteristics. Although schools certainly affect students' current test scores and proficiency levels, so too do the students' prior education and nonschool factors like the influence of parents. An alternate measure of effectiveness would isolate how much a school or teacher contributes to student test score improvements apart from confounding factors outside the school's or teacher's control.

To measure the performance of schools and teachers in DCPS, we used test scores and other data in a statistical model designed to capture the students' test score growth attributable to a school or teacher compared to the progress the students would have made at the average school or with the average teacher. Known as a "value-added model" because it isolates the contribution of the school or teacher from other factors, this method has been used by a number of prominent researchers (Meyer 1997; Sanders 2000; McCaffrey et al. 2004; Raudenbush 2004; Hanushek et al. 2007) and is employed in measuring the performance of schools and/or teachers in many school districts, including Chicago, Dallas, Milwaukee, Minneapolis, and New York City.

A value-added model measures teachers' contributions to students' achievement growth and typically accounts for the effect of student background characteristics on that growth. For example, suppose that a sixth-grade reading teacher has a class of students whose average score on the fifth-grade reading test, or "pretest," was 3 points above the school district average. Further suppose that students with similar background characteristics (like poverty status or disability status) typically grow 2 points more than the district average. So, given their starting point, these students would ordinarily end the year 5 points above the school district average on the sixth-grade reading test, or "posttest." The value-added model derives a relative measure of the teacher's effectiveness by comparing the average student posttest score to this standard. In this example, if the class posttest average is exactly 5 points above average, the value-added model will identify the *teacher* as an average performer. If the class posttest average exceeds this standard, the value-added model will identify the teacher as above average, and if the average is less than the standard, the value-added model will identify the teacher as below average. Because a value-added model focuses on growth and accounts for students' initial performance, it allows any schools or teachers to be identified as high performers, regardless of whether students were high-performing or low-performing at baseline.

Value-added models provide a better measure of school or teacher effectiveness than alternate measures, such as those that rely on gains in the proportion of students achieving proficiency. Those gains measure growth only for students who cross the proficiency cut-point, while value-added models incorporate achievement gains for all students, regardless of their baseline achievement levels. In addition, unlike schoolwide proficiency rates, which are affected by changes in the composition of the student population, value-added models track individual students over time. Potamites and Chaplin (2008) used DCPS data from 2005 to 2007 to show that measures of school effectiveness based on proficiency gains were not highly correlated with measures based on value-added estimates. The low correlation was primarily due to changes in the composition of students from one year to the next.

B. A Value-Added Model for DCPS

We estimated the performance of DCPS schools and teachers using a value-added model based on District of Columbia Comprehensive Assessment System (DC CAS) tests in math and reading. We measured school and teacher effectiveness in these two subjects separately. Based in part on information gathered during focus groups it conducted with teachers, DCPS sought to limit the accountability for student performance to the time period after the creation of IMPACT. We therefore based value-added measures for schools and teachers on one year of test score growth, from the 2008-2009 school year to the 2009-2010 school year.

School performance was based on as many grades as possible. Math and reading are tested in grades 3–8 and 10. Since third-grade students do not have a pretest, we estimated school performance only for grades 4–8 and 10. Schools that cover any of those grades were eligible to be included in the model. To avoid basing a school score on few students, which could lead to an imprecise measure of school performance, DCPS requested school value-added scores be reported only if there were at least 25 eligible students in the tested grades and subjects. Elementary and middle school students were included in the model if they had a posttest from 2010 and a pretest from the same subject in the previous grade in 2009. We excluded grade repeaters in these grades so that achievement growth for all students in a grade was based on the same posttest and pretest, allowing for meaningful comparisons between schools. However, because of this, not all students were included in the value-added model. The DC CAS test was not administered in grade 9, so we took the grade 10 pretest from grade 8. Most grade 10 students took the grade 8 DC CAS tests in spring 2008, but a sizable minority—16 percent—took them in spring 2007. We therefore used tests with either a two- or three-year lag between pretest and posttest for grade 10 students.

We calculated value-added estimates of teacher effectiveness separately from estimates of school effectiveness. We included regular education teachers who taught reading and/or math in grades 4–8—subjects and grades with a posttest at the end of the year and a pretest the year before. Based on concerns about the precision of value-added estimates for teachers with few students, DCPS asked that we report estimates only for teachers with a minimum of 15 students during the 2009–2010 school year.

To avoid penalizing or rewarding schools or teachers for factors that were outside their control, we designed the value-added model to account for a set of student characteristics that could affect posttest scores. These included a student's pretest scores in math and reading, poverty status, limited English proficiency, special education status, and gender. In the teacher model, we also accounted for student attendance during the prior year. In the school model, we accounted for whether grade 10 students took the grade 8 test in 2007 or 2008 in case there might have been a systematic difference in test score growth between these two groups of students. Although a student's race/ethnicity may be correlated with factors that both affect test scores and are beyond a teacher's control, DCPS chose not to account for this characteristic because preliminary results showed a high correlation in value-added measures regardless of whether race/ethnicity was considered.

C. Challenges and Solutions

Although the basic concepts associated with using a value-added model to measure school or teacher performance are straightforward, complexities arise when applying the model to data. We discuss five challenges to estimating school or teacher effectiveness fairly, and outline our solutions.

(1) Student Mobility Across Schools. When students change schools mid-year, multiple schools are responsible for their academic growth. To credit a single school with complete responsibility for a student who changes schools, or to ignore that student entirely, would distort our measure of a school's effectiveness. In DCPS in the 2009-2010 school year, four percent of the students were educated for part of the year at multiple schools. To account for this, we allocated proportional credit based on the fraction of time the student spent at each school, which can be thought of as the "dosage." The analysis included students who moved between DCPS schools in a single year as well as those who spent part of the year outside DCPS, as long as they took the DC CAS during the prior year and current year. For the school value-added model, we measured the dosage for grade 10 students over two years since most students took the pretest two years earlier.

(2) Co-Teaching. If two teachers co-taught students, it was not generally possible to distinguish the separate effects of each teacher on these students through statistical methods. In the 2009-2010 school year, 20 percent of teachers taught students who were also educated in the same subject by another Group 1 teacher. Ten percent of teachers shared all their students; five percent shared between 10 and 99 percent of their students, and six percent shared more than zero and less than 10 percent of their students.[3] In some cases, two or more teachers were jointly responsible for a classroom of students at the same time. In other cases, groups of students were taught by one teacher for part of the year and another teacher for the remainder of the year. In these circumstances, we estimated the combined effectiveness of these teachers if they had seven or more students in common. Each teacher received the team score as their individual value-added score. For teachers who taught some students in teams and other students individually, their overall value-added score was the weighted average of individual and team measures, where the weights were proportional to the number of student-equivalents taught in each situation.[4]

(3) Small Samples of Students. Performance estimates for schools and teachers could be misleading if they were based on too few students. Some students may score high on a test due to good luck rather than good preparation, and others may score low due to bad luck. For schools or teachers with many students, good and bad luck that affects test performance will tend to cancel out. A school or teacher with few students, however, can receive a very high or very low effectiveness measure based primarily on luck (Kane and Staiger 2002). We reduced the possibility of such spurious results by (1) not reporting estimates for schools with fewer than 25 students or for teachers with fewer than 15 students and (2) using a statistical technique that combines the measure of teacher performance obtained from the data with a default assumption of average performance that we made in the absence of data (Morris 1983). For an individual teacher estimate, we relied more heavily on the default assumption of average effectiveness when we had the least amount of data—typically teachers with fewer students or students whose achievement growth was most difficult to predict with a statistical model.

---

[3] Percentages are given for math but are similar for reading. The totals for the three subgroups do not sum exactly to 20 percent due to rounding.

[4] The number of student-equivalents per teacher is based on the self-reported contact time a teacher spent with the students in his or her classes. A student who was enrolled in a school from the first day of class until the test date and was assigned to a teacher's classroom for the full amount of classroom time devoted to a particular subject (math or reading) was counted as one student-equivalent. Fractional student-equivalents were possible if (1) a student changed schools during the year; (2) a student changed teachers during the year; or (3) a student was not assigned to a teacher for the full amount of classroom time devoted to the subject. For example, the student may have participated in a pullout program two days a week.

(4) Measurement Error in the Test. Because a student's performance on a single test is an imperfect measure of ability, schools or teachers may unfairly receive credit or blame for the initial performance of their students, rather than being assessed on the gains they have produced in student learning. For example, teachers of students with very high pretest scores may receive unfair measures of their performance if these test scores are attributable in part to luck; the average pretest score might have been lower if the students had been retested the next week. In such a case, the average ability level measured will be higher than their true ability level when the students enter the teacher's classroom, so part of the learning growth that occurs that year would not be credited to this teacher. To compensate for this sort of measurement error in pretest scores, we employed a statistical technique (Kmenta 1997) that makes use of published information on the test/retest reliability of a given DC CAS test.

(5) Comparing Value-Added Estimates Across Grades. The DC CAS is not specifically designed for users to compare gains across grades. Comparing value-added measures stated in terms of raw DC CAS points cannot meaningfully establish which teacher performed better if the teachers taught different grades. To compare teachers of different grades, we translated each teacher's value-added estimate into a metric of "generalized" DC CAS points using a two-step procedure. First, we adjusted teachers' value-added scores so that the average teacher in each grade received the same value-added score. Second, we multiplied each teacher's score by a grade-specific conversion factor to ensure that the dispersion of teacher value-added scores by grade was similar. To compare schools with different grade configurations, we applied a similar strategy. We transformed each grade-level measure within a school into a measure stated in generalized DC CAS points and then averaged across grades to arrive at a composite value-added measure for the school.

D. Caveats

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

(1) Estimation Error. The value-added measures are estimates of a school or teacher's performance based on the available data and the value-added model used. As with any statistical model, there is uncertainty in the estimates produced, which implies that two teachers with similar value-added estimates are "statistically indistinguishable" from one another. We quantified the precision with which the measures were estimated by reporting the upper and lower bounds of a confidence interval of performance for each teacher. Similar to Schochet and Chiang (2010), this approach also allowed us to quantify the misclassification rate for teachers under various policy scenarios. For example, for teachers with the lowest possible IMPACT score in math—the bottom 3.6 percent of DCPS teachers—one can say with at least 99.9 percent confidence that these teachers were below average in 2010. Similarly, a DCPS teacher with the lowest possible IMPACT score in reading—in the bottom 3.8 percent—was below average with at least 99.9 percent confidence.

(2) Classroom Effects. Value-added estimates measure not only the effectiveness of the teacher but also the combined effect of all factors that affect student achievement in the classroom. This includes inputs from the school, including direct effects, like the impact the school's physical plant has on test score growth, and indirect effects that work through teachers, such as the leadership abilities of the principal. Although a value-added model uses statistical techniques to account or "control" for differences in student performance based on documented sources of information about students, such as their prior-year test score or free lunch eligibility, the model cannot control for differences in student performance that arise from sources that are not explicitly

measured. Thus some caution should be applied when comparing teachers across schools (Aaronson, Barrow, and Sander 2007).

(3) Unmeasured Differences Between Students. The implicit assumption of a value-added model is that if two classrooms contain students with identical documented characteristics, the students will not differ systematically in ways that affect test score growth but are not easily measured. For example, the students' level of motivation to succeed would be presumed to be the same in these two classrooms. If students were randomly assigned to teachers, they should not differ systematically on any characteristics. On the other hand, if the assignment of students to teachers was based on unobservable factors—for example, pairing difficult-to-teach students with teachers who have succeeded with similar students in the past—a value-added model might unfairly penalize these teachers because it cannot statistically account for factors that cannot be measured.

There is debate among value-added researchers about how important this caveat is in practice (Kane and Staiger 2008; Rothstein 2009; Koedel and Betts 2009). Using data from the Los Angeles Unified School District, Kane and Staiger (2008) offer some evidence suggesting that unobservable student characteristics based on student assignment do not play a large role in determining value-added scores. They compared (a) the difference in value-added measures between pairs of teachers based on a typical situation in which principals assign students to teachers, and (b) the difference in student achievement between the teachers the following year, in which they taught classrooms that were formed by principals but then randomly assigned to the teachers. Kane and Staiger found that the differences between teachers' value-added scores before random assignment were a statistically significant predictor of achievement differences when classrooms were assigned randomly. These results were gathered in schools in which the principal was willing to allow random assignment of classrooms to teachers; it is not clear if they generalize to other contexts.

Given these caveats, DCPS has chosen not to use value-added measures as the sole determinant of a teacher's IMPACT score.

II. DATA

In this chapter, we review the data used to generate the value-added measures. We discuss the standardized assessment used in DC and the data on student background characteristics. We then discuss how we calculated the amount of time that students with multiple schools or teachers spent with each school or teacher. This discussion includes an overview of the roster confirmation process that allowed teachers to confirm whether and for how long they taught students math and/or reading and a description of how we identified team-teaching situations.

A. DC CAS Test Scores

When estimating the effectiveness of schools, we included elementary and middle school students if they had a DC CAS test from 2010 (the posttest) and a DC CAS test from the previous grade in the same subject in 2009 (the pretest). Students in grade 10 were included if they had a pretest from grade 8 in the same subject in either 2007 or 2008.[5] Beginning with 16,124 students for whom we had posttest scores, we excluded students from the analysis file if there were missing or conflicting data.[6] Of this group, 8 students had conflicting duplicate 2010 test score records, 41 students lacked corresponding information in the student background data, and 5 students had test scores that were outside the valid range of scores for the grade in which they were enrolled in 2010. The most common reason we excluded students was for lack of a pretest score, which could occur if they were not enrolled in a DC school during the testing period in April 2009 or if they missed the testing date. A total of 1,890 students, or 11.7 percent, were excluded for this reason. Finally, elementary and middle school students who repeated or skipped a grade were excluded so that achievement growth for all students in a grade was based on the same posttest and pretest.[7] This led to the exclusion of 197 students, or 1.6 percent of the remaining sample; at the school level, the percentage of students excluded for this reason ranged from zero to 8.1 percent. The resulting analysis file for math contained 13,983 students at 111 schools, an average of 126.0 students per school.

To obtain the most accurate and precise estimates of teacher effectiveness, we estimated the value-added model for teachers using all students in grades 4–8 who were in the analysis file for school-level measures. This included some students who were not linked to a Group 1 teacher. We did not include grade 10 students in the teacher-level analysis because they lacked pretest data from the prior year. Of the remaining 12,121 students, 1,090, or 9.0 percent, were not linked to a Group 1 DCPS teacher because they (1) were not linked to a DCPS school for at least 10 days, (2) were included in the roster file but not claimed by a teacher, or (3) were claimed only by a teacher with fewer than 7 students (we did not estimate a value-added measure for teachers with so few students). We reported estimates for teachers who taught 15 or more students in at least one subject. This included 480 teachers; of this group, 113 taught math only, 121 taught reading only, and 246 taught

[5] DCPS provided us with DC CAS test scores in math and reading from 2007 to 2009 and OSSE (Office of the State Superintendent of Education, which oversees DCPS and DC charter schools) provided data for 2010.

[6] Unless noted otherwise, sample sizes in this section are given for math. Sample sizes for reading were very similar.

[7] Students in grade 10 who either skipped a grade or repeated two grades between taking the grade 8 and grade 10 test were counted among the total who lacked a pretest.

both subjects, for a total of 359 teachers of math and 367 teachers of reading. In both subjects, teachers averaged 31.8 students.

For each subject, the DC CAS is scored so that each student receives a scale score from 300 to 399 for third-grade students, 400 to 499 for fourth-grade students, and so on. The range for 10th-grade students is 900 to 999. The first digit is a grade indicator only; it does not reflect the student's ability. The rest of the score, which ranges from 0 to 99, can only be meaningfully compared within grades and within subjects; math scores, for example, are generally more dispersed than reading scores within the same grade. To address this issue, 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.[8] 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.

B. Student Background Data

We used data provided by DCPS to construct variables that were used as controls in the value-added models for student background characteristics. In both the school and teacher value-added models, we controlled for the following:

- Pretest in same subject as posttest
- Pretest in other subject (so we controlled for math and reading pretests regardless of posttest)
- Gender
- Free lunch eligibility
- Reduced-price lunch eligibility
- Limited English proficiency status
- Having a specific learning disability
- Having other types of disabilities requiring special education

In the school model, we also controlled for:

- Taking the grade 8 DC CAS test in 2007 rather than in 2008 (for some grade 10 students)

In the teacher model, we also controlled for:

- Proportion of days that the student attended school during the prior year

[8] Subtracting the mean score for each subject and grade creates a score that has a mean zero of zero in all subject-grade combinations, effectively removing the uninformative first digit.

The last variable measures student motivation. We used prior- rather than current-year attendance to avoid confounding student attendance with current-year teacher quality, as a good teacher might be expected to motivate students to attend more regularly than a weaker teacher. Attendance is a continuous variable that could range from zero to one. Aside from pretest variables, the other variables are indicator variables taking the value zero or one.

The selection of these variables was based on data availability and careful judgment. For example, there were multiple categories of special education available in the administrative data, including information on students who received special test accommodations in one year but not in another. The choice of two categories for special education reflected a trade-off between a detailed specification, which allows for differences among different types of special education students, and a parsimonious specification, which avoids the problem of generating estimates that may be sensitive to outliers in the data.

C. School and Teacher Dosage

Because some students moved between schools or were taught by a combination of teachers, we apportioned their achievement growth among multiple schools or teachers. We refer to the fraction of time the student was enrolled with each teacher and at each school as the dosage.

1. School Dosage

Based on DCPS administrative data, which contain dates of school withdrawal and admission, we assigned every student a dosage for each school the student attended. School dosage equals the fraction of the school year that the student was officially enrolled at that school. Since students do not take tests on the last day of each school year, this measure covered the first three terms (that is, the fall semester and the first half of the spring semester); the third term ended eight school days before the beginning of testing. To fully account for 100 percent of each student's time during the first three terms, we also recorded the portion of the school year the student was enrolled in schools outside DCPS.

Because a school is unlikely to have an appreciable educational impact on a short-term student, we set dosage equal to zero for students who spent less than two weeks at a school. Conversely, we set it to 100 percent for those who spent all but two weeks at a school. Apart from this, in the school model we assumed that learning accumulated at a constant rate and treated days spent at one school as interchangeable with days spent at another. For example, if a student split time equally between two schools, each school was assigned a dosage of 50 percent for this student, regardless of which school the student attended first. Since the grade 8 DC CAS test served as the pretest for students in grade 10, we based dosage variables for grade 10 students on the schools they attended during the 2008-2009 and 2009-2010 school years (regardless of whether they had taken the grade 8 DC CAS two or three years earlier).

2. Teacher Dosage

To determine which students were taught math and reading by a given teacher during the 2009–2010 school year, DCPS conducted roster confirmation in March 2010, covering teachers of math and reading in grades 4–8. Teachers were provided with a list of students who appeared on their course rosters at some point during the year. For each of the first three terms, teachers indicated whether they taught each subject to each student, and if so, the proportion of time they taught the student relative to the full amount of time the teacher spent on that subject for the typical

student. For example, if a student spent two and a half days per week in a Group 1 teacher's classroom learning math and two and a half days per week in another classroom with a special education teacher while other students were learning math with the Group 1 teacher, then this student spent 0.5 of the instructional time with the Group 1 teacher. In recording the proportion of time spent with a student, teachers rounded to the nearest quarter, so 0 percent, 25 percent, 50 percent, 75 percent, and 100 percent were the possible responses. For students who spent less than 100 percent of the time with a teacher, teachers did not indicate the name of the other teacher. Staff in the DCPS central office followed up with teachers who had many unclaimed students on their roster and in other anomalous cases.

We used the confirmed class rosters to construct teacher-student links. If the roster confirmation data indicated that a student had one math or reading teacher at a school, the teacher-student weight equaled the school dosage. If a student changed teachers from one term to another, we used the school calendar to determine the number of days the student spent with each teacher, and we subdivided the school dosage among teachers accordingly. When teachers claimed the same students during the same term, as in a team-teaching situation, DCPS decided to assign each teacher full credit for the shared students, reflecting DCPS' preference to weight students equally, whether they were taught individually or by co-teachers. We therefore did not subdivide dosage for students of team teachers. Finally, similar to tracking time spent at all schools outside DCPS, we tracked the time a student spent with any teachers who were not recorded in the confirmed class rosters, which we called the "non-Group 1 teacher(s)."

3. Teacher Teams

We created variables for teacher teams to model two situations: (1) "co-teaching," in which students received instruction from more than one Group 1 teacher for the same subject during the same term or (2) "sequential teaching," in which a group of students switched from one teacher during one term to another teacher for another term. We formed teams only within schools and within grades. A teacher who taught more than one grade could therefore have multiple individual and/or team estimates.

To prevent the estimates of teachers who shared students with unidentified teachers from becoming "contaminated" with the estimate of the catchall non-Group 1 teacher(s), we formed teams between a Group 1 teacher and the "non-Group 1 teacher(s)." Otherwise, because the estimate for the non-Group 1 teacher(s) was typically negative, the statistical model might compensate by attributing especially effective teaching to a Group 1 teacher who shared students to balance the negative estimate attributed to the non-Group 1 teacher(s).

Estimating value-added measures for individuals or teams based on too few students unacceptably increases the risk of introducing imprecise or biased estimates. Therefore, we required that an individual teacher or teacher team have at least seven students. We assigned students of individual teachers with fewer than seven total students to the catchall category of non-Group 1 teacher(s). Students in teams with fewer than seven students were assigned to the individual teachers.

As an example of how our method worked, consider a teacher who taught a classroom of 24 fourth-grade students, where 8 of the students participated in a pullout program with a special education teacher for half of the regular reading time. This teacher would receive two estimates for reading. For 16 students, there would be an estimate of the teacher's individual effect; for the other 8 students, there would be an estimate of the joint effect of the Group 1 teacher and the special

education teacher. Because the teacher would receive 0.5 dosage for each of these half-time students, the total number of student-equivalents from this group for the teacher would be $0.5 \times 8 = 4$. The teacher's overall value-added estimate would be a student-equivalent weighted average of the two estimates, with the weights equal to $4/(4 + 16) = 1/5$ for the team and $16/(4 + 16) = 4/5$ for the individual effect.

III. THE VALUE-ADDED MODEL

A. Estimation Equation

We developed a value-added model that we used to measure four outcomes separately: school effectiveness in math, school effectiveness in reading, teacher effectiveness in math, and teacher effectiveness in reading. After assembling the analysis file, the first step for each outcome was to estimate a linear regression that combined students of all grade levels in the data. For school outcomes, we estimated an equation in which the posttest score depends on prior achievement, student background characteristics, schools attended, and a set of unmeasured factors. The equation can be expressed formally as:

$$(1) \quad Y_{ig} = \lambda_{1g} Y_{i(g-1)} + \omega_{1g} Z_{i(g-1)} + \boldsymbol{\alpha}_1' \mathbf{X}_i + \boldsymbol{\beta}' \mathbf{S}_{ig} + \varepsilon_{1ig},$$

where Y_{ig} is the standardized posttest score for student i in grade g and $Y_{i(g\text{-}1)}$ is the standardized same-subject pretest for student i in grade g-1 during the prior year. The variable $Z_{i(g\text{-}1)}$ denotes the pretest in the opposite subject. Thus, when estimating school effectiveness in math, Y represents math tests with Z representing reading tests, and vice-versa. The pretest scores captured prior inputs into student achievement, and the associated coefficients, λ_{1g} and ω_{1g}, varied by grade. The vector $\mathbf{X}_i$ denotes the control variables for the individual student background characteristics listed in Chapter II. The coefficients on these characteristics, $\boldsymbol{\alpha}_1$, were constrained to be the same across all grades.[9]

$\mathbf{S}_{ig}$ is a vector of school dosage variables containing one variable for each school-grade combination. The measures of school effectiveness are school-grade effects contained in $\boldsymbol{\beta}$, the coefficients of the dosage variables represented by $\mathbf{S}_{ig}$. For each grade, we included a measure of the combined effectiveness of all schools students attended that were outside of DCPS. The dosage for a given element of $\mathbf{S}_{ig}$ was set equal to the percentage of the year student i was enrolled in grade g at that school. The value of any element of $\mathbf{S}_{ig}$ was zero if student i was not taught in grade g in that school during the school year. Because $\mathbf{S}_{ig}$ accounted for student attendance throughout the school year, its elements always summed to one. Rather than dropping one of the school dosage variables from the regression, we estimated the model without a constant term. We also mean centered the control variables so that each element of $\boldsymbol{\beta}$ represented a school- and grade-specific intercept term for the average student.[10] We assumed that the error term, ε_{ig}, is heteroskedastic.

[9] We estimated a common, grade-invariant set of coefficients of student background characteristics because a preliminary investigation revealed substantial differences in sign and magnitude of grade-specific coefficients on these covariates. These cross-grade differences appeared to reflect small within-grade samples of individuals with certain characteristics rather than true differences in the association between student characteristics and achievement growth. Estimating a common set of coefficients across grades allowed us to estimate the association between achievement and student characteristics using information from all grades, which smoothed out the implausibly large between-grade differences in these coefficients.

[10] Mean centering the student characteristics and pretest scores tends to reduce the estimated standard errors of the school effects (Wooldridge 2008).

The regression produced separate value-added coefficients for each grade within a school. To reduce the likelihood of obtaining statistically imprecise estimates, we did not include dosage variables for school-grade combinations with fewer than five student equivalents.[11] The estimated coefficients were aggregated into a single measure for each school, as explained in Section D below.

The teacher model is directly analogous to the school model. The value-added regression equation can be expressed as:

$$(2) \quad Y_{ig} = \lambda_{2g} Y_{i(g-1)} + \omega_{2g} Z_{i(g-1)} + \boldsymbol{\alpha}_2' \mathbf{X}_i + \boldsymbol{\eta}' \mathbf{T}_{ig} + \varepsilon_{2ig},$$

where the notation largely parallels that for the school model described by equation (1). The main difference is that the vector $\mathbf{T}_{ig}$ includes dosage variables for teachers and teams, rather than schools. Consequently, the measures of teacher effectiveness are contained in $\boldsymbol{\eta}$, the coefficients of the dosage variables represented by $\mathbf{T}_{ig}$. Table III.1 shows the coefficients and standard errors of the control variables in the school and teacher models.

We used the regression to estimate value-added coefficients for each teacher and team separately in each grade for the 2009–2010 school year. We did not impose a student-equivalent threshold for including dosage variables since we had already applied a seven-student count threshold when determining whether to include teacher and team variables. For teachers with multiple value-added coefficients, such as teachers who were members of multiple teams, or who had both individually taught and team-taught students, we first combined these coefficients to obtain a total grade-specific estimate, as described in Section C below. We then aggregated teacher estimates across grades to form a single estimate for each teacher, as explained in Section D.

B. Measurement Error in the Pretests

We corrected for measurement error in the pretests using grade-specific reliability data available from the test publisher (CTB/McGraw Hill 2009). As a measure of true student ability, standardized tests contain measurement error, causing an ordinary least-squares regression to produce biased estimates of teacher or school effectiveness. To address this issue, we implemented a measurement error correction that makes use of 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.

Correcting for measurement error required a two-step procedure. Our statistical model included distinct pretest coefficients for each grade but common coefficients on student characteristics. As such, it was not computationally possible to apply the numerical formula for the errors-in-variables correction to all grades simultaneously. Therefore we estimated the errors-in-variables correction in the first step on a grade-by-grade basis and then estimated a second step regression with common (rather than grade-specific) coefficients on the student characteristics. We describe the procedure in the context of school measures; the procedure for the teacher measures is analogous.

[11] In practice, this occurred only in auxiliary models in which we restricted the sample to students who belonged to specific categories, such as special education. For school-grade combinations that did not meet the five student-equivalent threshold, we reassigned the dosage for these students to "schools outside DCPS."

Table III.1 Coefficients on Covariates in the School and Teacher Value-Added Models, by Subject

	School Model		Teacher Model	
Variable	Math	Reading	Math	Reading
Male	0.912	-0.451	0.646	-0.605
	(0.179)	(0.159)	(0.190)	(0.174)
Eligible for Free Lunch	-1.453	-1.452	-1.472	-1.598
	(0.237)	(0.213)	(0.256)	(0.236)
Eligible for Reduced-Price Lunch	-0.912	-1.281	-1.385	-1.642
	(0.360)	(0.321)	(0.385)	(0.354)
Limited English Proficiency	-0.552	-1.414	-0.891	-1.447
	(0.404)	(0.380)	(0.428)	0.406
Specific Learning Disability	-4.114	-7.052	-3.106	-6.682
	(0.395)	(0.379)	(0.438)	(0.442)
Other Learning Disability	-4.306	-7.336	-3.480	-7.329
	(0.538)	(0.526)	(0.546)	(0.553)
Took Additional Time Between Grades 8 and 10	-8.416	-7.002		
	(0.912)	(0.867)		
Fraction of the Prior Year Student Attended School			7.424	-0.042
			(2.187)	(2.093)
Pretest Scores				
Same Subject, All Grades, Standard Time Between Grades	0.627	0.546	0.599	0.518
	(0.059)	(0.066)	(0.046)	(0.052)
Opposite Subject, All Grades, Standard Time Between Grades	0.106	0.163	0.081	0.171
	(0.071)	(0.059)	(0.061)	(0.054)
Same Subject, Grade 10, Additional Time Between Grades	0.351	0.391		
	(0.063)	(0.066)		
Opposite Subject, Grade 10, Additional Time Between Grades	0.029	0.005		
	(0.075)	(0.059)		

Notes: Standard errors are in parentheses.

The reported coefficient estimates of pretest scores for "All Grades" represent averages of the coefficients estimated separately for grades 4-8 in the teacher model and grades 4-8 and 10 in the school model. The associated standard errors account for both the average estimation error and variability of the estimates across grades.

In the first step, we used errors-in-variables regression to obtain unbiased estimates of the pretest coefficients for each grade. For grades 4–8, we used the published reliabilities associated with the 2009 DC CAS. For grade 10 students, we estimated separate pretest coefficients for students who took the grade 8 test in 2008 and for those who took the test in 2007, using DC CAS

reliabilities from the appropriate year for each type of student.[12] We then used the measurement-error corrected values of the pretest coefficients to calculate the adjusted gain for each student in each grade. The adjusted gain, expressed as:

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

represents the student posttest outcome, net of the estimated contribution attributable to the student's starting position at pretest.

The second step pooled the data from all grades and used the adjusted gain as the dependent variable in a single equation:

$$(4) \quad \hat{G}_{ig} = \boldsymbol{\alpha}_1' \mathbf{X}_i + \boldsymbol{\beta}' \mathbf{S}_{ig} + \varepsilon_{ig},$$

The grade-specific estimates of school effectiveness, $\hat{\boldsymbol{\beta}}$, were obtained by applying ordinary least squares (OLS) regression estimation to equation (4).

This two-step method will tend to underestimate the standard errors of $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\beta}}$ for two reasons. First, OLS regression does not account for heteroskedasticity, which arises if there are systematic differences in the variability of test outcomes across different types of students or at different schools. Second, because the adjusted gain in equation (3) relies on the estimated value of λ_g, the error term in equation (4) is clustered within grades. Both heteroskedasticity and clustering will typically result in estimated standard errors that are too small, since all sources of variability are not accounted for. We use the Huber-White estimator to correct the standard errors for heteroskedasticity (Huber 1967; White 1980). However, the standard nonparametric cluster-robust sandwich estimators, as described by Froot (1989) and implemented in Stata, would not effectively correct the standard errors due to the small number of grades (Bertrand et al. 2004). Nonetheless, because of the large within-grade sample sizes, the pretest coefficients were estimated precisely, which should lead to a negligible difference between the robust and clustering-corrected standard errors.

C. Creating Total Grade-Specific Teacher Estimates

When teachers taught some students individually and others as part of a team, or when teachers were members of multiple teams, our model produced more than one estimate of effectiveness for the same teacher. In these cases, it was necessary to combine these measures to understand the total effectiveness of the teacher at a given grade level. We combined multiple grade-specific teacher estimates into a total estimate by computing a weighted average of the value-added coefficients associated with that teacher. The weights were based on the number of student-equivalents the teacher taught. For team teaching, we tracked the number of student-equivalents taught by the team that were attributable to each teacher. For teams formed when students changed teacher mid-year,

[12] To account for any systematic difference in test score growth between grade 10 students taking the test three years instead of two years prior, the vector of student characteristics in the school regressions included a dummy variable indicating whether a grade 10 student took the DC CAS in 2007.

the number of student-equivalents attributable to a given teacher depended on the proportion of the year they spent with that teacher.

We calculated the variance of the total estimate $\hat{\beta}_{jg}$ for teacher *j* in grade *g* based on:

$$(5) \quad Var\left[\hat{\beta}_{jg}\right] = \sum_{m} p_{jgm}^{\ 2} Var\left[\hat{\eta}_{gm}\right] \times \frac{N_{gm}}{n_{jgm}},$$

where p_{jgm} equaled the proportion of teacher *j*'s student-equivalents in grade *g* that were taught by team *m*, $Var\left[\hat{\eta}_{gm}\right]$ is the variance of the coefficient for team *m* obtained from the regression, N_{gm} is the total number of student-equivalents in grade *g* taught by team *m*, and n_{jgm} is the number of student-equivalents on team *m* for which teacher *j* was responsible. In Equation (5), we assumed that the covariance terms are zero because different students contributed to each estimate.

We corrected the variance to account for the fact that some individual teachers were not responsible for students in a team for the entirety of the first three terms. When estimating the variance of a teacher team effect, we included all students taught by that team of teachers. For individual teachers within the team, a naïve estimate would have treated the estimate of the teacher variance as if the teacher taught all students for the first three terms. To correct the variance for a given teacher, we therefore inflated the variance of a team estimate for an individual teacher when that teacher was responsible for teaching fewer student-equivalents than the full number upon which the team estimate was based. We did this by multiplying each of the variance terms by N_{gm} / n_{jgm}. This inflates the variance component associated with team *m* for teacher *j* so that it is proportional to the ratio of the total number of students taught by the team divided by the number of students for which teacher *j* was actually responsible. When students were taught by two teachers simultaneously, both teachers bore full responsibility for the performance of the students. In this case, $N_{gm} / n_{jgm} = 1$, and there was no inflation of the variance. When teams arose because students switched classrooms mid-year, $N_{gm} / n_{igm} > 1$. In this case the estimated variance for teacher *j* was inflated to account for additional uncertainty about the individual contribution of teacher *j* to the effectiveness of team *m*, over and above the imprecision in the team estimate.

D. Combining Estimates Across Grades

Both the average and variability of value-added scores differed across grade levels, leading to a potential problem when comparing teachers who taught different grades or schools with different grade configurations. For example, some differences across grades may be an artifact of the tests, which were not designed to fit on a single scale, so we did not want to penalize or reward teachers simply for teaching in a grade with unusual test properties. Therefore, after obtaining grade-level estimates for schools or teachers, we translated these estimates so that they could be expressed using a common metric of "generalized" DC CAS points. This translation was necessary because all schools or teachers were compared to all others in the value-added regression, regardless of any

grade-specific factors that might have led to differences in the distributions of classroom gains.[13] Below, we describe the procedure in the context of teacher measures; the procedure for the school measures is analogous.

We adjusted the teacher effectiveness estimates so that they would be on a common scale in each grade. First, we subtracted from each unadjusted estimate the weighted average of all the estimates within the same grade. We then divided the result by the weighted standard deviation within the same grade. To reduce the influence of imprecise estimates obtained from teacher-grade combinations with few students, we based the weights on the number of students each teacher taught. Second, we multiplied by the teacher-weighted average of the grade-specific standard deviations, obtaining a common measure of effectiveness on the generalized DC CAS point scale.

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

$$(6) \quad \hat{\theta}_{jg} = \frac{\hat{\beta}_{jg} - \overline{\hat{\beta}_g}}{\hat{\sigma}_g} \times \left(\frac{1}{J} \sum\nolimits_h J_h \hat{\sigma}_h \right),$$

where $\hat{\beta}_{jg}$ is the grade g estimate for teacher j, $\overline{\hat{\beta}_g}$ is the weighted average estimate for all teachers in grade g, $\hat{\sigma}_g$ is the weighted standard deviation of teacher estimates in grade g, J_h is the number of teachers with students in grade h, and J is the total number of teachers. We excluded the estimates associated with the "non-Group 1 teacher(s)" (and with the "schools outside DCPS" estimates in the school model). The values of the standard deviations calculated for each grade are shown in Table III.2.

Table III.2. Student-Weighted Standard Deviations of Scores by Grade

Model	Grade						Weighted Average
	4	5	6	7	8	10	
School							
Math	5.3	4.9	4.6	4.2	3.9	3.4	4.7
Reading	4.5	3.5	3.1	2.5	2.6	2.4	3.4
Teacher							
Math	6.1	5.2	4.7	4.2	4.2		5.2
Reading	5.0	4.0	3.6	2.6	2.6		4.0

Aside from putting value-added estimates for teachers onto a common scale, this approach equalizes the distribution of teacher estimates across grades. This 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

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

conditions, the test instrument, or student cohorts, we have assumed that the distribution of true teacher effectiveness is the same across grades.

To combine effects across grades into a single effect ($\hat{\theta}_i$) for a given teacher, we used a weighted average of the grade-specific estimates (expressed in generalized DC CAS points). Similar to the process of combining across teams, we set the weight for grade *g* equal to the proportion of students of teacher *j* who are in grade *g*, denoted as p_{jg}. We computed the estimated variance of the teacher estimate using

$$(7) \quad Var\left[\hat{\theta}_j\right] = \sum_g p_{jg}^{\ 2} Var\left[\hat{\theta}_{jg}\right],$$

where p_{jg} is defined above and $Var\left[\hat{\theta}_{jg}\right]$ is the variance of the grade-*g* estimate for teacher *j*. For simplicity, we have assumed that the covariance across grades is zero. Additionally, we have not accounted for uncertainty arising because $\overline{\hat{\beta}_g}$ and $\hat{\sigma}_g$ in equation (6) are estimates of underlying parameters rather than known constants. Both decisions imply that the standard errors obtained from equation (7) will be slightly underestimated.

E. Shrinkage Procedure

To reduce the instability of the value-added estimates that can occur due to small sample size, we applied the empirical Bayes (EB) shrinkage procedure outlined in Morris (1983) to the sets of effectiveness estimates for schools and teachers separately. We are framing our discussion of shrinkage in terms of teachers, but the same logic applies to schools. This statistical technique averages the effectiveness estimates obtained from the data with the effect of the average teacher. There was greater weight placed on the estimate for the average teacher when the effectiveness estimates were less precisely estimated—typically for teachers with fewer students.

The EB estimate for a teacher is approximately equal to a precision-weighted average of the teacher's initial value-added estimate and the overall mean of all teacher estimates.[14] Following the standardization procedure described in Section D, the overall mean is approximately zero, with better-than-average teachers having positive scores and worse-than-average teachers having negative scores.[15] We therefore have:

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

[15] The overall mean is not exactly zero because we used a student-weighted average of value-added estimates rather than an unweighted average.

$$(8)\quad \hat{\theta}_j^{EB} \approx \left(\frac{\frac{1}{\hat{\sigma}_j^2}}{\frac{1}{\hat{\sigma}_j^2} + \frac{1}{\hat{\sigma}^2}} \right) \hat{\theta}_j + \left(\frac{\frac{1}{\hat{\sigma}^2}}{\frac{1}{\hat{\sigma}_j^2} + \frac{1}{\hat{\sigma}^2}} \right) \bar{\theta} \approx \left(\frac{\frac{1}{\hat{\sigma}_j^2}}{\frac{1}{\hat{\sigma}_j^2} + \frac{1}{\hat{\sigma}^2}} \right) \hat{\theta}_j,$$

where $\hat{\theta}_j^{EB}$ is the EB estimate for teacher *j*, $\hat{\theta}_j$ is the effectiveness estimate from the data for teacher *j*, and $\hat{\sigma}_j^2$ is the standard error of the estimate for teacher *j*. The overall mean of all teacher estimates, $\bar{\theta}$, is zero, and $\hat{\sigma}^2$, the standard deviation of all teacher estimates, is constant for all teachers. Mathematically, the term $[(1/\hat{\sigma}_j^2)/(1/\hat{\sigma}_j^2 + 1/\hat{\sigma}^2)]$ must be less than one; hence, the EB estimate is always less in absolute value than the initial estimate—that is, the EB estimate "shrinks" from the initial estimate. The greater the precision of the initial estimate—that is, the larger $(1/\hat{\sigma}_j^2)$ is—the closer $[(1/\hat{\sigma}_j^2)/(1/\hat{\sigma}_j^2 + 1/\hat{\sigma}^2)]$ is to one, and the smaller the shrinkage in $\hat{\theta}_j$. Conversely, the smaller the precision of the initial estimate, the greater the shrinkage in $\hat{\theta}_j$. By applying a greater degree of shrinkage to less precisely estimated teacher measures, this 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 $\hat{\theta}_j^{EB}$ using the formulas provided by Morris (1983). As a final step, we removed any schools with fewer than 25 students from the school model and any teachers with fewer than 15 students from the teacher model (the "reporting thresholds" set by DCPS) and re-centered the EB estimates on zero.

F. Calculating School Scores That Combine Math and Reading Estimates

To rank schools for making TEAM awards, DCPS requested that we produce a single score for each school that averages the math and reading value-added estimates using a fixed set of weights. We thereby created a weighted average such that (1) the weights summed to one, and (2) each subject-specific score contributed equally to the combined score, regardless of any difference between subjects in the overall dispersion of scores:

$$(9)\quad \hat{\theta}_{j,combined} = \left(\frac{s_{reading}}{s_{math} + s_{reading}} \right) \hat{\theta}_{j,math}^{EB} + \left(\frac{s_{math}}{s_{math} + s_{reading}} \right) \hat{\theta}_{j,reading}^{EB},$$

where s_{math} is the standard deviation of post-shrinkage math scores and $s_{reading}$ is the standard deviation of post-shrinkage reading scores across all schools. The weights in equation (9) equalize extent to which differences among schools in math and reading scores translate into differences in their combined scores, scaling each component of $\hat{\theta}_{j,combined}$ by a factor that ensures that the combined score is bounded by the two subject scores.

Because the two subject estimates were calculated from separate regressions, it was not possible to directly estimate the covariance on a school-by-school basis. Consequently, we relied on the following approximation:

$$(10)\quad Cov_j(math, reading) \cong SE^{EB}_{j,math} \times SE^{EB}_{j,reading} \times Corr(math, reading),$$

where $SE^{EB}_{j,math}$ and $SE^{EB}_{j,reading}$ are the empirical Bayes estimates of the standard errors of the math and reading scores for school j and $Corr(math, reading)$ is the overall correlation between math and reading scores, which is used as a best prediction of the correlation between residuals across subjects within each school. We then substituted this estimate of the covariance into the standard variance formula to arrive at an estimate of the variance:

$$(11)\quad Var_{j,combined} = \left(\frac{s_{reading}}{s_{math} + s_{reading}} \times SE^{EB}_{j,math} \right)^2 + \left(\frac{s_{math}}{s_{math} + s_{reading}} \times SE^{EB}_{j,reading} \right)^2 + 2 \frac{s_{reading} s_{math}}{s_{math} + s_{reading}^2} \times Cov_j .$$

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics,* vol. 25, no. 1, 2007, pp. 95–135.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, vol. 119, no. 1, 2004, pp. 248–275.

Booker, Kevin, and Eric Isenberg. "Measuring School Effectiveness in Memphis." Washington, DC: Mathematica Policy Research, 2008.

Booker, Kevin, Duncan Chaplin, and Eric Isenberg. "Measuring School Effectiveness in Charter Schools Across States." Washington, DC: Mathematica Policy Research, 2008.

CTB/McGraw-Hill. *Technical Report for the Washington, D.C., Comprehensive Assessment System (DC CAS), Spring 2009*. Monterey, CA: CTB/McGraw-Hill, 2009.

Froot, Kenneth A. "Consistent Covariance Matrix Estimation with Cross-Sectional Dependence and Heteroskedasticity in Financial Data." *Journal of Financial and Quantitative Analysis*, vol. 24, no. 3, 1989, pp. 333–355.

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

Huber, Peter J. "The Behavior of Maximum Likelihood Estimation Under Nonstandard Conditions." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1*, edited by L.M. LeCam and J. Neyman. Berkeley, CA: University of California Press, 1967.

Isenberg, Eric. "Measuring Teacher Effectiveness in Memphis." Washington, DC: Mathematica Policy Research, 2008.

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

Kane, Thomas J., and Douglas O. Staiger. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working paper #14607. Cambridge, MA: National Bureau of Economic Research, 2008.

Kmenta, Jan. *Elements of Econometrics*. Second Edition. Ann Arbor, MI: University of Michigan Press, 1997.

Koedel, Cory, and Julian R. Betts. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." Working Paper 09-02. Columbia, MO: University of Missouri, 2009.

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

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

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

Potamites, Elizabeth, and Duncan Chaplin. "Ranking DC's Public Schools Based on Their Improvement in Math from 2006-2007." Washington, DC: Mathematica Policy Research, 2008.

Potamites, Elizabeth, Kevin Booker, Duncan Chaplin, and Eric Isenberg. "Measuring School and Teacher Effectiveness in the EPIC Charter School Consortium—Year 2." Washington, DC: Mathematica Policy Research, 2009a.

Potamites, Elizabeth, Duncan Chaplin, Eric Isenberg, and Kevin Booker. "Measuring School and Teacher Effectiveness in Memphis—Year 2." Washington, DC: Mathematica Policy Research, 2009b.

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

Rothstein, Jesse. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." Working paper #14442. Cambridge, MA: National Bureau of Economic Research, 2009.

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

Schochet, Peter Z., and Hanley S. Chiang. "Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains." NCEE 2010-4004. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2010.

White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica,* vol. 48, 1980, pp. 817–830.

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