Design of Value-Added Models for IMPACT and TEAM in DC Public Schools, 2010–2011 School Year

Final Report

May 13, 2011

Eric Isenberg
Heinrich Hock

MATHEMATICA
Policy Research, Inc.

Mathematica Reference Number:
06860.500

Submitted to:
District of Columbia Public Schools
1200 First Street, NE
Washington, DC 20002
Project Officer: Dominique Amis

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

Design of Value-Added Models for IMPACT and TEAM in DC Public Schools, 2010–2011 School Year

Final Report

May 13, 2011

Eric Isenberg
Heinrich Hock

MATHEMATICA
Policy Research, Inc.

ACKNOWLEDGMENTS

We are grateful to the many people who contributed to this report. We thank the District of Columbia Public Schools (DCPS) for funding the work and Hella Bel Hadj Amor, Jason Kamras, and Erin McGoldrick at DCPS for working together to build a value-added model that meets the needs of the IMPACT and TEAM programs. We also thank independent reviewers Eric Hanushek and Tim Sass for their valuable suggestions.

At Mathematica Policy Research, Duncan Chaplin provided valuable comments. Carol Soble and Amanda Bernhardt edited the report, and Lisa Walls provided word processing and production support.

CONTENTS

I. OVERVIEW

This report presents the value-added models that will be used to measure school and teacher effectiveness in the District of Columbia Public Schools (DCPS) in the 2010–2011 school year. It updates our earlier technical report, "Measuring Value Added for IMPACT and TEAM in DC Public Schools." The earlier report described the methods used to estimate teacher effectiveness in DCPS during the 2009–2010 school year, the first year of the IMPACT assessment system in DCPS. Working together with officials at DCPS as well as with external peer reviewers Eric Hanushek of the Hoover Institution at Stanford University and Tim Sass of Florida State University, we modified two aspects of the value-added model in order to meet the needs of DCPS:

1. We changed the method of accounting for team teaching in the teacher model to increase the number of students accounted for in some teachers' value-added measures. The new method is also more computationally robust. Value-added estimates produced by both methods do not differ substantially.
2. At the request of DCPS, the value-added models no longer account for a student's gender when estimating teachers' and schools' effectiveness. Based on an analysis of data from previous years, we expect the change to have a negligible effect on value-added estimates for schools and teachers.

Apart from these modifications, the school and teacher value-added models that we estimated during the 2009–2010 school year are identical to those we plan to estimate for the 2010–2011 school year.

In the remainder of this chapter, we provide a detailed description of the two changes made to the value-added models for the 2010–2011 school year and then present a nontechnical overview of the resulting models used to estimate teacher and school effectiveness, showing how key features of the models help produce results designed to be as accurate and fair as possible. In the remaining chapters of this report, we update the earlier technical report by discussing the technical details of the value-added models we will estimate with data from the 2010–2011 school year, incorporating the two changes that we have made to the method. In Chapter II, we present a description of the data, and, in Chapter III, we focus on the statistical methods.

A. Updates to the Value-Added Models for the 2010–2011 School Year

1. Change to Full Roster Method for Estimation of Teacher Value Added

We will use a new approach, called the Full Roster Method, to account for teachers in the value-added model who co-teach students. As described in Isenberg and Hock (2010), we calculated estimates for the 2009–2010 school year using what we call the Teacher Team Method, in which each student was uniquely linked to either a solo teacher or a team of teachers. The multiple regression estimated value-added measures for both solo teachers and teacher teams. If a teacher instructed both solo-taught and team-taught students, two effects were estimated for that teacher, which were then combined by taking a student-weighted average. The Full Roster Method uses a different strategy by linking students to each of their teachers, thus creating unique records for each teacher-student combination. For example, we would create two records for a student team-taught by two teachers, one for each teacher-student combination. The resulting analysis file includes one measure per teacher, but some students contribute directly to the value-added estimate of multiple

teachers. In the regression analysis, each teacher-student combination is weighted according to the fraction of the year the student spent with the teacher.

Both methods are valid approaches to accounting for co-teaching, and each offers unique advantages (Hock and Isenberg 2011). The Teacher Team Method allows for a more flexible handling of co-teaching; for example, co-teaching that occurs within schools can be modeled differently from co-teaching that occurs when students move from one school to another. However, the Full Roster Method offers two offsetting, practical advantages over the Teacher Team Method. First, the Full Roster Method allows more students to contribute directly to the calculation of a value-added score for a teacher. With the Teacher Team Method, it is not practical to calculate estimates for teams with fewer than 7 students. A teacher who had both solo-taught and team-taught students could therefore be unlinked from some of those students. For example, a teacher with 15 team-taught and 5 solo-taught students would be linked in the regression only to the 15 team-taught students. By contrast, the Full Roster Method links all of a teacher's students to the teacher regardless of whether the students are also taught by other teachers. This approach is expected to improve the accuracy and precision of the value-added estimates. Second, because the Full Roster Method does not rely on identifying and accounting for potentially complicated teaming arrangements, the results can be generated more quickly and at lower cost.

A numerical comparison of the results of the Full Roster Method and the Teacher Team Method—based on data from the 2009–2010 school year—showed that the correlation between teacher value-added scores was 0.994 in math and 0.992 in reading. Thus, both methods provide similar results, but the Full Roster Method allows for more students to be incorporated directly into teachers' value-added scores and can be generated more quickly and at lower cost. For these reasons, DCPS has decided, in consultation with external reviewers Eric Hanushek and Tim Sass, to adopt the Full Roster Method for the 2010–2011 school year.

2. Gender No Longer Accounted For

Value-added models used to calculate school and teacher effectiveness will no longer account for student gender. In the 2009–2010 school year, gender was among the student characteristics that were accounted for when calculating value-added scores. At the request of DCPS, we used data from the 2009–2010 school year to compare school and teacher value-added estimates from otherwise-identical models that did and did not include student gender. Overall, we found little difference between the two sets of estimates—correlation coefficients were above 0.999 in both math and reading for both schools and teachers. Based in part on these results, DCPS has asked us to remove gender from the set of student characteristics included in the value-added model used for the 2010–2011 school year.

B. Summary of the Value-Added Model

The basic approach of a value-added model is to predict the test score performance that each student would have obtained with the average DCPS teacher and then compare the average performance of a given teacher's students to the average of the predicted scores. The difference between these two scores—how the students actually performed with a teacher and how they would

have performed with the average DCPS teacher—is attributed to the teacher as his or her value added to students' test score performance.[1]

Although conceptually straightforward, the design of a value-added model that accurately and fairly measures the performance of teachers must address several layers of complexity in the educational context. The value-added model we designed for DCPS involves four steps, each of which addresses a different conceptual challenge.

1. **Multiple Regression.** We use multiple regression, a statistical technique that allows us to avoid penalizing or rewarding teachers for factors outside their control. We account for a set of student characteristics that could be related to performance on the DC Comprehensive Assessment System (DC CAS) test in the 2010–2011 school year (the posttest). The characteristics include a student's DC CAS test scores from the 2009–2010 school year in math and reading (the pretests), poverty status, limited English proficiency, special education status, and attendance during the 2009–2010 school year.[2] By accounting for these characteristics, we obtain a net estimate of each teacher's effectiveness. The estimate is approximately equal to the difference between the average actual posttest score of a teacher's students and the average predicted score of those students based on their characteristics. We weight each student's contribution to a teacher's score by the proportion of the time the student was assigned to the teacher when the teacher was teaching that subject. A teacher's estimate is based on data from all of his or her students, including those taught by other teachers in the same subject in the 2010–2011 school year.
2. **Accounting for Measurement Error in the Pretest.** Given that a student's performance on a single test is an imperfect measure of ability, teachers may unfairly receive credit or blame for the initial performance of their students rather than being assessed on the gains they produce 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 achievement level measured on the pretest will be higher than students' true achievement level when they enter the teacher's classroom; therefore, part of the learning growth that occurs that year would not be credited to the teacher. To compensate for such measurement error in pretest scores, we employ a statistical technique that makes use of published information on the test/retest reliability of the DC CAS.
3. **Comparing Teachers across Grades.** The DC CAS is not designed to compare scores across grades. Thus, we account for attributes of the test and the tested population that could affect *changes* in student scores across grades by translating each teacher's value-added estimate into a metric of "generalized" DC CAS points. This translation is based

[1] The presentation in this section is framed in terms of teachers and teacher value added, but the discussion applies equally to schools and school value added. Subsequent footnotes highlight differences between the teacher and school models.

[2] In the school model, we do not account for attendance in the prior year but do account for whether grade 10 students took the grade 8 test in 2008 or 2009 in case there was a systematic difference in test score growth between these two groups of students.

on a three-step procedure. First, before the multiple regression step, we translate math and reading scores into a common metric in which each *student* test score is measured relative to other test scores within the same year, grade, and subject. Second, after obtaining *teachers'* value-added scores, we adjust the teacher scores so that the average teacher in each grade receives the same score. Third, we multiply each teacher's score by a grade-specific conversion factor to ensure that the dispersion of teacher value-added scores by grade is similar. For teachers with students in more than one grade, we take a student-weighted average of their grade-specific value-added estimates.[3]

4. **Accounting for Imprecisely Estimated Measures Based on Few Students.** Estimates of teacher or school effectiveness could be misleading if 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 affecting test performance tends 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 reduce the possibility of such spurious results by (1) not reporting estimates for teachers with fewer than 15 students[4] and (2) using a statistical technique that combines the effectiveness measure of a particular teacher with the overall average (Morris 1983). We rely more heavily on a default assumption of average effectiveness for teachers with fewer students or students whose achievement growth is most difficult to predict with a statistical model.

[3] To compare schools with different grade configurations, we apply a similar strategy. We transform each grade-level measure within a school into a measure stated in generalized DC CAS points and then average across grades to arrive at a composite value-added measure for the school.

[4] For schools, the threshold is 100 test scores across math and reading; for example, a school would qualify if it had 50 eligible students, all of whom took both the math and reading test.

II. DATA

In this chapter, we review the data used to generate the value-added measures. We discuss the data on student background characteristics and how we calculate the amount of time that students spent in more than one school or with more than one teacher. We provide an overview of the roster confirmation process that allows teachers to confirm whether and for how long they taught students math and/or reading. Finally, we discuss the standardized assessment used in DC schools.

A. Teacher, School, and Student Lists

DCPS provides an official comprehensive list of schools eligible for inclusion in the value-added model as well as a list of teachers of math and reading in grades 4 through 8 who are eligible to receive individual value-added scores. Teachers who are on the list and teach at least 15 students with pretest and posttest scores are termed Group 1 teachers in the IMPACT assessment system. For the purposes of the value-added model, we refer to teachers on the list as Group 1 whether or not they teach 15 students. DCPS also provides a student list that indicates students' official grade level.

B. Student Background Data

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

- Pretest in same subject as posttest
- Pretest in other subject (we control for math and reading pretests regardless of posttest)
- Free-lunch eligibility
- Reduced-price lunch eligibility
- Limited English proficiency status
- Existence of a specific learning disability
- Existence of other types of disabilities requiring special education
- In the teacher model, we also control for
- Proportion of days that the student attended school during the previous year
- In the school model, we also control for
- Taking the grade 8 DC CAS test in 2008 rather than in 2009 (for some grade 10 students)

Attendance measures student motivation. We use previous- rather than current-year attendance to avoid confounding student attendance with current-year teacher quality; that is, a good teacher versus a weaker teacher might be expected to motivate students to attend more regularly. We exclude the attendance measure from the school model because many students will have attended the same school in the previous year, which would risk confounding student attendance with school

quality. The last variable in the list is relevant only for the school model because we do not include grade 10 students in the teacher model. Attendance is a continuous variable that could range from zero to one. Aside from pretest variables, the other variables are binary variables taking the value zero or one.

We impute data for students who are present in the background data but have missing values for one or more student characteristics. Our imputation approach involves using the values of nonmissing student characteristics to predict the value of the missing characteristic. For students who did not attend a DCPS school for part of the previous year, we use 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 DCPS.[5] We do not generate imputed values for the same-subject pretest; students with missing same-subject pretest scores are dropped from the analysis file.

The selection of variables is based on data availability and careful judgment. For example, the administrative data contain several categories of special education. The choice of two categories for special education reflects 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.

The models do not include controls for race/ethnicity or gender. A student's race/ethnicity or gender may be correlated with factors that both affect test scores and are beyond a teacher's control. DCPS decided not to account for these characteristics after preliminary results showed a high correlation in value-added measures regardless of whether race/ethnicity or gender were included in the models. This suggests that the other characteristics included in the value-added models capture most of the factors affecting test scores that are correlated with race/ethnicity and gender.

C. School and Teacher Dosage

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

1. School Dosage

Based on DCPS administrative data, which contain dates of school withdrawal and admission, we assign every student a dosage for each school attended. School dosage equals the fraction of the first three quarters of the school year that the student was officially enrolled at that school. We use only the first three quarters of the year because students start taking their tests five school days after the end of the third quarter. To account fully for each student's time, we also record the portion of the school year that the student was enrolled in schools outside DCPS.

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

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 dosage to 100 percent for students who spent all but two weeks at a school. Apart from this, in the school model, 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 attended first. Given that the grade 8 DC CAS test is the pretest for students in grade 10, we base dosage variables for grade 10 students on the schools they attended during the 2009–2010 and 2010–2011 school years for students who took the grade 8 DC CAS test in 2009, and on the schools they attended during the 2008-2009 through 2010-2011 school years for students who took the grade 8 DC CAS test in 2008.

2. Teacher Dosage

To determine which students received math and reading instruction from a given teacher during the 2010–2011 school year, DCPS conducts a roster confirmation among teachers of math and reading in grades 4 through 8. Teachers receive a list of students who appeared on their course rosters at some point during the year. For each of the first three quarters, teachers indicate 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 would spend on that subject for students enrolled for the full quarter. 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 learn math with the Group 1 teacher, then the student is recorded as having spent 50 percent of instructional time with the Group 1 teacher. In recording the proportion of time spent with a student, teachers round to the nearest quarter such that the possible responses are 0, 25, 50, 75, and 100 percent. For students who spent less than 100 percent of the time with a teacher, teachers do not indicate the names of any other teachers who taught the student. Staff in the DCPS central office follow up with teachers with many unclaimed students on their roster and in other anomalous cases.

We use the confirmed class rosters to construct teacher-student links. If the roster confirmation data indicate that a student had one math or reading teacher at a school, we set the teacher-student weight equal to the school dosage. If a student changed teachers from one term to another, we use the school calendar to determine the number of days the student spent with each teacher, subdividing the school dosage among teachers accordingly. When two or more teachers claim the same students during the same term, DCPS assigns each teacher full credit for the shared students. This decision reflects the preference of DCPS that solo-taught and co-taught students contribute equally to teachers' value-added estimates. We therefore do not subdivide dosage for students who are co-taught. Finally, similar to tracking time spent at all schools outside DCPS, we track the time a student spent with any teachers not recorded in the confirmed class rosters, which we call the non–Group 1 teacher(s).

D. DC CAS Test Scores

When estimating the effectiveness of schools, we include elementary and middle school students if they have a DC CAS test from 2011 (the posttest) and a DC CAS test from the previous

grade in the same subject in 2010 (the pretest). We include students in grade 10 if they have a pretest from grade 8 in the same subject in either 2008 or 2009.[6] We exclude students from the analysis file in the case of missing or conflicting test score data or no matching student background data. We also exclude students who repeated or skipped a grade because they lack pretest and posttest scores in consecutive grades and years. Based on data from 2009–2010, we expect that the most common reason for excluding students will be that they were not enrolled in a DC school during the testing period in April 2010 or do not take the test in 2011 (Isenberg and Hock 2010).

To obtain accurate and precise estimates of teacher effectiveness, we estimate the value-added model for teachers using all students in grades 4 through 8 in the analysis file for school-level measures. We include all of these students, even those not linked to a Group 1 teacher because they (1) did not attend 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 do not estimate a value-added measure for teachers with so few students). Including these unlinked students in the analysis allows us to estimate more precisely the relationship between student characteristics and achievement for all students, including the majority of students linked to teachers. We report estimates only for teachers who taught 15 or more students in at least one subject.

For each subject, the DC CAS is scored so that each student receives a scale score from 300 to 399 for grade 3 students, 400 to 499 for grade 4 students, and so on. The range for grade 10 students is 900 to 999. The first digit is a grade indicator only and does not reflect student achievement. We drop the first digit, making use of the rest of the score, which ranges from 0 to 99.

The resulting scores may be meaningfully compared only 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 create subject- and grade-specific z-scores by subtracting the mean and dividing by the standard deviation within a subject-grade combination.[7] This step allows 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 multiply each test score by the average standard deviation across all grades within each subject and year.

[6] DCPS provided us with DC CAS test scores in math and reading from 2007 to 2009; the Office of the State Superintendent of Education (OSSE), which oversees DCPS and DC charter schools, provided data for 2010. In June 2011, DCPS will send a file with 2011 posttest scores.

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

III. ESTIMATING VALUE ADDED

A. Regression Estimates

We have developed two linear regression models to estimate effectiveness measures for schools and for teachers. After assembling the analysis file, we estimate a regression separately for math and reading using students at all grade levels in the data. In each regression equation, the posttest score depends on prior achievement, student background characteristics, variables linking students to schools or teachers, and unmeasured factors.

1. The School Regression Model

The regression equation used to estimate effectiveness measures for schools may be expressed formally as:

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

where Y_{ig} is the posttest score for student i in grade g and $Y_{i(g\text{-}1)}$ is the same-subject pretest for student i in grade g-1 during the previous 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 capture prior inputs into student achievement, and the associated coefficients, λ_{1g} and ω_{1g}, vary by grade. The vector $\mathbf{X}_{1i}$ denotes the control variables for individual student background characteristics. The coefficients on these characteristics, $\boldsymbol{\alpha}_1$, are constrained to be the same across all grades.[8]

The vector $\mathbf{S}_{ig}$ contains one dosage variable for each school-grade combination, and the associated coefficients contained in $\boldsymbol{\beta}$ measure the effectiveness of each school by grade. For students attending more than one school, each school receives partial credit based on the student's dosage. The dosage for a given element of $\mathbf{S}_{ig}$ is set to be equal to the percentage of the year student i was enrolled in grade g at that school. The vector of dosage variables ($\mathbf{S}_{ig}$) also includes a separate variable for each grade level for the fraction of the school year a student spent outside DCPS. The value of any element of $\mathbf{S}_{ig}$ is zero if student i was not taught in grade g in that school during the school year. Because $\mathbf{S}_{ig}$ accounts for student attendance throughout the school year, its elements always sum to one. Rather than dropping one of the school dosage variables from the regression, we estimate the model without a constant term. We also mean center the control variables so that each element of $\boldsymbol{\beta}$ represents a school- and grade-specific intercept term for a student with average

[8] We estimate a common, grade-invariant set of coefficients of student background characteristics because our calculations using 2009–2010 data 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 allows us to base the association between achievement and student characteristics on information from all grades, which should smooth out the implausibly large between-grade differences in these coefficients.

characteristics.[9] We assume that there are systematic differences in the variability of test outcomes across different types of students or at different schools, which implies that the error term, ε_{ig}, is heteroskedastic.

We estimate equation (1) by using ordinary least squares (OLS). Heteroskedasticity generally results in estimated errors that are too small because the regression does not account for all sources of variability. Accordingly, we calculate heteroskedasticity-robust standard errors using the Huber-White estimator (Huber 1967; White 1980).

The regression produces separate value-added coefficients for each grade within a school. To reduce the likelihood of obtaining statistically imprecise estimates, we do not include dosage variables for school-grade combinations with fewer than five student equivalents.[10] We aggregate the estimated coefficients into a single measure for each school (see Section C below).

2. The Teacher Regression Model

The teacher model differs from the school model in order to account for team teaching that occurs at the teacher level but not at the school level. The chief difference in the teacher model is that the unit of observation is a teacher-student combination rather than a student. Unlike the school model, in which schools contribute separately to the achievement of students who attend more than one school, the teacher model is based on the assumption that the combined effort of team teachers constitutes a single input into student achievement (Hock and Isenberg 2011). For a given teacher t and student i, the regression equation may be expressed as:

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

where the notation largely parallels that for the school model described by equation (1). The vector $\mathbf{T}_{tig}$ includes a grade-specific variable for each teacher and includes a variable for a non–Group 1 teacher in each grade to account for student dosage that cannot be attributed to a particular Group 1 teacher. A student contributes one observation to the model for each teacher to whom the student is linked, based on the roster confirmation process. Each teacher-student observation has one nonzero element in $\mathbf{T}_{tig}$. Measures of teacher effectiveness are contained in the coefficient vector $\boldsymbol{\eta}$.

To account for multiple observations on the same student, we estimate the coefficients by using weighted least squares (WLS) rather than OLS. In this method, the teacher-grade variables in $\mathbf{T}_{tig}$ are binary, and we weight each teacher-student combination by the teacher dosage associated with that combination. We account for the correlation in the error term ε_{2tig} across multiple observations on the same student when estimating standard errors. Specifically, we use the method described in

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

[10] In practice, this is likely to occur only in auxiliary models in which we restrict the sample to students who belong to specific categories, such as special education. For school-grade combinations that do not meet the five-student–equivalent threshold, we will reassign the dosage for these students to a variable representing time spent at all other schools in the given grade, including schools outside DCPS.

Froot (1989) to calculate heteroskedasticity-robust standard errors that are additionally clustered at the student level.

Similar to the school model, the teacher regression yields separate value-added coefficients for each grade in which a teacher is linked to students. To improve the precision of the estimates, we estimate a grade-specific coefficient for a teacher only if he or she teaches at least seven students in that grade.[11] We then aggregate teacher estimates across grades to form a single estimate for each teacher (see Section C below).

B. Measurement Error in the Pretests

We correct for measurement error in the pretests by using grade-specific reliability data available from the test publisher (CTB/McGraw Hill 2008; 2009; 2010). As a measure of true student ability, standardized tests contain measurement error, causing standard regression techniques to produce biased estimates of teacher or school effectiveness. To address this issue, we implement a measurement error correction that uses 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 requires a two-step procedure. Our statistical model includes distinct pretest coefficients for each grade but common coefficients on student characteristics. However, it is not computationally possible to apply the numerical formula for the errors-in-variables correction simultaneously to all grades. Therefore, we estimate the errors-in-variables correction in the first step on a grade-by-grade basis and then estimate a second-step regression with common (rather than grade-specific) coefficients on the student characteristics. We describe the procedure in the context of teacher measures; the procedure for the school measures is analogous.

In the first step, we use a dosage-weighted errors-in-variables regression based on equation (2) to obtain unbiased estimates of the pretest coefficients for each grade. For grades 4 through 8, we use the published reliabilities associated with the 2010 DC CAS.[12] We then use the measurement-error corrected values of the pretest coefficients to calculate the adjusted gain for each student in each grade. The adjusted gain is expressed as:

[11] Although teachers must teach at least 15 students for DCPS to evaluate them on the basis of individual value added, we include in the regression teachers with 7 to 14 students for two reasons. First, we expect that maintaining more teacher-student links will lead to coefficients on the covariates that are estimated more accurately. Second, we expect that value-added estimates for these teachers will provide useful data to include in the standardization and shrinkage procedures described below. We do not include teachers with fewer than 7 students because estimates for such teachers would be too likely to be outliers, which could skew the standardization and shrinkage procedures. If a teacher has fewer than 7 students in a grade, those students are reallocated to the grade-specific non–Group 1 teacher.

[12] In the school model, weights are not used when applying errors-in-variables regression because dosage is already accounted for in the school-grade variables. Further, we estimate separate pretest coefficients for grade 10 students taking the grade 8 test in 2009 and those taking the test in 2008, using DC CAS reliabilities from the appropriate year for each type of student. To account for any systematic difference in test score growth between grade 10 students taking the test three rather than two years previously, the vector of student characteristics in the school regressions includes a binary variable indicating whether a grade 10 student took the DC CAS in 2008.

$$(3) \quad \hat{G}_{tig} = Y_{tig} - \hat{\lambda}_{2g} Y_{i(g-1)} - \hat{\omega}_{2g} Z_{i(g-1)},$$

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

The second step pools the data from all grades and uses the adjusted gain as the dependent variable in a single equation expressed as:

$$(4) \quad \hat{G}_{tig} = \boldsymbol{\alpha}_2' \mathbf{X}_{2i} + \boldsymbol{\eta}' \mathbf{T}_{tig} + \varepsilon_{tig},$$

The grade-specific estimates of teacher effectiveness, $\hat{\boldsymbol{\eta}}$, are obtained by applying the WLS regression technique to equation (4).[13]

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

C. Combining Estimates Across Grades

Both the average and the variability of value-added scores may differ across grade levels, leading to a potential problem when comparing teachers assigned to different grades or comparing schools with different grade configurations. The main concern is that factors beyond teachers' control—rather than teacher distribution or school effectiveness—may drive cross-grade discrepancies in the distribution of value-added scores. For example, the standard deviation of adjusted gains might vary across grades as a consequence of differences in the alignment of tests or the retention of knowledge between years. However, we seek to compare all schools or 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.[14] Because we do not want to penalize or reward teachers simply for teaching in a grade with unusual test properties, we translate grade-level estimates for schools and teachers so that each set of estimates is expressed in a common metric of "generalized" DC CAS points. Below, we describe the procedure in the context of teacher measures; the procedure for school measures is analogous.

[13] In the school model, OLS is applied to a student-level expression comparable to equation (4).

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

We standardize the estimated regression coefficients so that the mean and standard deviation of the distribution of teacher estimates is the same across grades. First, we subtract from each unadjusted estimate the weighted average of all estimates within the same grade. We then divide 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 base the weights on the number of students taught by each teacher. Finally, we multiply 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:

$$(5) \quad \hat{\theta}_{tg} = \frac{\hat{\eta}_{tg} - \overline{\hat{\eta}_g}}{\hat{\sigma}_g} \times \left(\frac{1}{K} \sum\nolimits_h K_h \hat{\sigma}_h \right),$$

where $\hat{\eta}_{tg}$ is the grade-g estimate for teacher t, $\overline{\hat{\eta}_g}$ 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, K_h is the number of teachers with students in grade h, and K is the total number of teachers. We exclude the estimates associated with the non–Group 1 teachers (and with the "schools outside DCPS" estimates in the school model).

Aside from putting value-added estimates for teachers onto a common scale, this approach equalizes the distribution of teacher estimates 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, in the test instrument itself, or in student cohorts, we assume that the distribution of true teacher effectiveness is the same across grades.

To combine effects across grades into a single effect ($\hat{\theta}_t$) for a given teacher, we use 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, denoted as p_{tg}. We compute the variance of each teacher's estimated effect by using:

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

where $Var\left[\hat{\theta}_{tg}\right]$ is the variance of the grade-g estimate for teacher t. For simplicity, we assume that the covariance across grades is zero. In addition, we do not account for uncertainty arising because $\overline{\hat{\eta}_g}$ and $\hat{\sigma}_g$ in equation (6) are estimates of underlying parameters rather than known constants. Both decisions imply that the standard errors obtained from equation (6) will be slightly underestimated.

D. Shrinkage Procedure

To reduce the risk that teachers or schools, particularly those with relatively few students, will receive a very high or very low effectiveness measure by chance, we apply the empirical Bayes (EB) shrinkage procedure, as outlined in Morris (1983), separately to the sets of effectiveness estimates for teachers and schools. We frame our discussion of shrinkage in terms of teachers, but the same logic applies to schools. Using the EB procedure, we can compute a weighted average of an estimate for the average teacher (based on all students in the model) and the initial estimate that uses 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.[15] Following the standardization procedure, the overall mean is approximately zero, with better-than-average teachers having positive scores and worse-than-average teachers having negative scores.[16] We therefore arrive at the following:

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

where $\hat{\theta}_t^{EB}$ is the EB estimate for teacher *t*, $\hat{\theta}_t$ is the initial estimate of effectiveness for teacher *t* based on the regression model (after combining across grades), and $\hat{\sigma}_t^2$ is the standard error of the estimate for teacher *t*. The overall mean of all teacher estimates, $\bar{\theta}$, is approximately zero, and $\hat{\sigma}^2$, the standard deviation of all teacher estimates, is constant for all teachers. Mathematically, the term $[(1/\hat{\sigma}_t^2)/(1/\hat{\sigma}_t^2 + 1/\hat{\sigma}^2)]$ 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 is $(1/\hat{\sigma}_t^2)$—the closer $[(1/\hat{\sigma}_t^2)/(1/\hat{\sigma}_t^2 + 1/\hat{\sigma}^2)]$ is to one and the smaller the shrinkage in $\hat{\theta}_t$. Conversely, the smaller the precision of the initial estimate, the greater is the shrinkage in $\hat{\theta}_t$. 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

[15] 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.

[16] The overall mean will not be exactly zero because we use a student-weighted average of value-added estimates rather than an unweighted average.

calculate the standard error for each $\hat{\theta}_t^{EB}$ using the formulas provided by Morris (1983). As a final step, we remove any teachers with fewer than 15 students from the teacher model and any schools with fewer than 100 student tests across both subjects from the school model and then re-center the EB estimates on zero.

E. Calculating School Scores That Combine Math and Reading Estimates

To select the schools that would receive TEAM awards, DCPS asked us to produce a single score for each school using a weighted average of the math and reading value-added estimates for that school. We therefore selected fixed weights such that (1) the weights sum to one and (2) the score from each subject contributes equally to the combined score, regardless of any difference between subjects in the overall dispersion of scores. For example, if math scores were distributed across a wide range and reading scores were distributed across a narrow range, a school's rank within the distribution of overall scores generated by taking a simple average of the two scores would depend primarily on the math score. There would be comparatively little influence on the rank of a school's overall score due to the reading score. Therefore, to give equal effective weight to both subjects, for a given school j, the combined score is expressed as:

$$(8) \quad \hat{\theta}_{j,combined} = \left(\frac{s_{reading}}{s_{math} + s_{reading}} \right) \hat{\theta}_{j,math}^{EB} + \left(\frac{s_{math}}{s_{math} + s_{reading}} \right) \hat{\theta}_{j,reading}^{EB},$$

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 (8) equalize the extent to which differences between schools in math and reading scores translate into differences in their combined scores.

Given that the two subject estimates are calculated from separate regressions, it is not possible to directly estimate the covariance on a school-by-school basis. Consequently, we rely on the following approximation:

$$(9) \quad Cov_j(math, reading) \cong SE_{j,math}^{EB} \times SE_{j,reading}^{EB} \times Corr(math, reading),$$

where $SE_{j,math}^{EB}$ and $SE_{j,reading}^{EB}$ 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 substitute this estimate of the covariance into the standard variance formula to arrive at the following estimate:

$$(10) \quad Var_{j,combined} = \left(\frac{s_{reading}}{s_{math} + s_{reading}} \times SE_{j,math}^{EB} \right)^2 + \left(\frac{s_{math}}{s_{math} + s_{reading}} \times SE_{j,reading}^{EB} \right)^2 + 2 \frac{s_{reading} s_{math}}{s_{math} + s_{reading}^2} \times Cov_j,$$

which represents the estimated variance of the combined, two-subject school score.

REFERENCES

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.

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 the Washington, D.C., Comprehensive Assessment System (DC CAS), Spring 2008. Monterey, CA: CTB/McGraw-Hill, 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.

Hock, Heinrich, and Eric Isenberg. "Methods for Accounting for Co-Teaching in Value-Added Models." Washington, DC: Mathematica Policy Research, 2011.

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, and Heinrich Hock. "Measuring School and Teacher Value Added for IMPACT and TEAM in DC Public Schools." Washington, DC: Mathematica Policy Research, 2010.

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.

Lee, Peter M. Bayesian Statistics: An Introduction (Second Edition). New York: John Wiley and Sons, 1997.

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

White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, vol. 48, no. 4, 1980, pp. 817-830.

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

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