WORKING PAPER 23

MATHEMATICA Policy Research

Working PAPER

BY ERIC ISENBERG, BING-RU TEH, AND ELIAS WALSH

Elementary School Data Issues: Implications for Research Using Value-Added Models

October 2013

Abstract

Researchers conducting research using administrative data often presume that data from grades 4 and 5 are better than data from grades 6 to 8 for conducting research on teacher effectiveness that uses value-added models because (1) elementary school teachers teach all subjects to their students in individual, self-contained classrooms and (2) elementary school classrooms are more homogenous, with little academic tracking used when assigning students to teachers, unlike in middle school. We examined both assumptions. First, we used data on teacher–student links from DC Public Schools that have undergone a roster confirmation process whereby teachers verify which subjects and students a particular teacher taught. We compared the teacher–student links that resulted from this process to data that approximate the quality of teacher–student links in unconfirmed administrative data. Second, to examine the extent to which tracking of students by achievement segregates students at the middle school level compared to the upper elementary school level, we compared the variation in baseline student achievement at upper elementary and middle school grades within classes at the same school, between classes at the same school, and between schools. Results show that departmentalization of teaching instruction across math and reading/ELA is actually quite common in grades 4–5, at least in DCPS; in the unconfirmed administrative data, about one in six teachers of these subjects is linked to a subject that he or she does not teach. In addition, we found more within-school variation in pre-test scores in middle school grades but an offsetting amount of between-school variation in upper elementary grades. As an example of how using unconfirmed administrative data can affect results, we examined how calculations of the year-to-year and cross-subject stability of value-added estimates depended on the quality of the data used.

I. INTRODUCTION

The No Child Left Behind Act of 2002 required, among other things, that all children attending public schools be tested annually in math and reading in grades 3 to 8, and once during high school. A side benefit of this testing regime is that it has provided some basic data for estimating value-added models of teacher effectiveness in a wide array of school districts. Given that value-added models depend on having at least one year of pre-test scores from the prior school year for a teacher's students, merging test score data with student demographic data and teacher–student links has made it possible to estimate value added for teachers of math and reading/ English language arts (ELA) in grades 4 to 8. Through the Teacher Incentive Fund, Race to the Top, and other initiatives, the federal government has encouraged the use of student growth models (of which value added is one example) in evaluations of teachers in these grades and subjects. Researchers have also made use of these data to examine topics ranging from the effectiveness of different types of schools, teacher mobility patterns, and the distribution of teacher effectiveness, to diagnosing the utility of value-added models themselves.

In conducting research that estimates teacher value added using administrative data sets, it is often presumed that upper elementary school grades 4 and 5 provide better data than middle school grades 6 to 8. Two reasons are principally cited for this: (1) elementary school teachers teach all subjects to their students in individual, self-contained classrooms and (2) there is little tracking of students to teachers in elementary school compared to middle school, resulting in more homogeneous classrooms at the elementary school level (Harris and Anderson 2013).

The first reason to favor elementary school grades depends on having accurate student–teacher links, a critical data element for estimating value-added models of teacher effectiveness. However, misclassification errors in roster data, such as linking teachers to subjects or students they did not teach, could lead to bias in value-added estimates. This bias arises because teachers are attributed value-added estimates that do not reflect their actual contributions to student achievement.

Researchers at the Value-Added Research Center at the University of Wisconsin have identified a number of reasons why administrative rosters may not capture the realities of teacher–student links, including "student attendance and mobility, teacher attendance and mobility, cross-subject course content, instructional supports and curriculum specialists, team teaching, special education and English language learner accommodations, and other divergences from the standard isolated classroom model" (Kluender et al. 2011). According to Battelle for Kids (BFK), a nonprofit organization that has assisted many states and districts in linking teachers and students, administrative data for districts typically do not link students and teachers accurately enough for use with high-stakes evaluation systems. BFK (2009) has cataloged a number of sources of errors in district administrative data: data systems do not capture the regrouping or switching of students between teachers during the school year, moment-in-time data systems do not accurately describe student mobility, data systems are unable to account for co-teaching or identify the amount of time students spend with each teacher, the process of aligning courses to tested subjects may not be straightforward, charter schools within the district may use data systems that are difficult to incorporate into the district's system, and many elementary schools do not capture accurate teacher–student links in their course scheduling systems. In particular, if instruction is departmentalized in upper elementary grades, this may not be reflected in district administrative data.

Ordinarily, researchers assume that students linked to elementary school teachers who teach a "self-contained" class receive instruction in both math and reading/ELA from these teachers. There are many examples of research that have made this assumption when using rich data from state data systems (Rothstein 2010; Sass et al. 2012). BFK has reported, however, that during the 2012–2013 school year, across a range of districts and states in which they worked, "nearly one in four teachers had incorrect or incomplete content area association" and "more than one in three rosters were inaccurate and required changes during roster verification" (BFK 2013).

Using data from District of Columbia Public Schools (DCPS), we examined the assumptions on self-contained classrooms and tracking. In brief, we found that departmentalization of teaching instruction across math and reading/ELA is actually quite common in grades 4 and 5 in DCPS, where about one in six teachers of these subjects is linked to a subject in the administrative data that the teacher does not teach. Put another way, about one in four DCPS upper elementary school students were taught by a departmentalized teacher. To answer the second research question on how tracking could affect the differences in classroom composition across teachers, we computed the percentage of variance in pre-test scores explained by between-school, within-school, and within-classroom variation. We found that there is more within-school variation in pre-test scores in middle school grades, but an offsetting amount of greater between-school variation in upper elementary grades. These findings could be consistent with either a greater degree of tracking in middle schools or more homogenous attendance areas for elementary schools, or both. As an example of how using unconfirmed administrative data can affect results, we examined how calculations of the year-to-year and cross-subject stability of value-added estimates depend on the quality of the data used.

We conclude that because of missing data on departmentalization of elementary school teaching, one should be cautious about assuming that data from grades 4 and 5 are necessarily superior to data from grades 6 to 8 for conducting research that involves value-added models of teacher effectiveness. We discuss implications on drawing conclusions from research that uses value-added estimates that are based on unconfirmed roster data. From a policy perspective, this finding also highlights the importance of the roster confirmation process in minimizing the probability of attributing value-added estimates to teachers that do not reflect their actual contributions to student achievement.

II. THE ROSTER CONFIRMATION PROCESS IN DCPS

IMPACT, which has been used at DCPS since the 2009–2010 school year, is an example of a high-stakes evaluation system that includes value added as an important component (Isenberg and Hock 2012). Under IMPACT, DCPS has rewarded teachers who earn a highly effective rating with performance pay, and dismissed teachers who earned an ineffective rating that year or a minimally effective rating for two consecutive years. Mathematica Policy Research has assisted DCPS since the start of IMPACT by designing and implementing a value-added model of teacher effectiveness based on specifications determined by DCPS and later by the Office of the State Superintendent of Education (OSSE). Depending on the school year, value added has counted for either 35 percent or 50 percent of a teacher's overall evaluation. To ensure students were correctly linked to the math and reading/ELA teachers from whom they received instruction, DCPS has conducted an annual roster confirmation among teachers of math and reading/ELA in grades 4 to 8.

This roster confirmation process in DCPS allows a teacher to confirm exactly which students he or she taught during the year, for which subject(s), and for what portion of the year, to improve upon the original administrative data.[1] DCPS first determines which teachers are eligible to receive value-added estimates by creating a list of teachers who are general education instructors of math and/or reading/ELA in grades 4 to 8. These teachers are asked to confirm their administrative rosters. (Teachers who are not on this list are evaluated under IMPACT according to a different set of components.) In most cases, eligible teachers receive lists of students who appear on their course rosters. 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. For example, suppose students in a classroom typically receive math instruction from a particular teacher five days a week. If a student spent two days a week in the teacher's classroom learning math and spent the remaining time allocated for math instruction on the other three days in another classroom with a special education teacher, the student would have spent 40 percent of math instructional time with the teacher. In recording the proportion of time spent with a student in a given class and subject, teachers rounded to the nearest 25 percent in the 2010–2011 school year and to the nearest 20 percent in the 2011–2012 school year. If a teacher claimed to have taught a student for less than 100 percent in any quarter, the teacher was not responsible for naming other teachers who taught the student. Teachers could also add students to their rosters. In a few cases, a prefilled roster was unavailable and teachers added all of their students. Central office staff at DCPS followed up with DCPS teachers as necessary to resolve apparent anomalies. In 2011–2012, an additional step was added in which school principals verified confirmed rosters. The confirmed roster data are not necessarily a perfect reflection of the students that teachers taught—no survey guarantees perfection—however, because these data are used for a high-stakes evaluation system, errors in the subjects that teachers taught are unlikely, and there are considerably fewer errors in the links to individual students than in the administrative data.

To approximate the typical quality of administrative rosters a researcher would receive from a district that does not implement a roster confirmation process, we used administrative course scheduling data from DCPS from the 2010–2011 and 2011–2012 school years to create unconfirmed rosters. The DCPS course scheduling data have two components: (1) a list of students and their homeroom teachers and (2) a list of students showing courses in which they were enrolled in October and teachers associated with those courses. Teachers on the first list were assumed to have taught both math and reading/ELA. Teachers on the second list were classified as math and/or reading/ELA teachers based on the courses with which they were associated. Teachers who appeared on both lists were removed from the first list, as checks against the confirmed rosters suggested that they taught only the subject to which they were linked on the second list. These unconfirmed rosters included all math and reading/ELA teachers who had students in grades 4 through 8.

[1] This paragraph describes the roster confirmation process in the 2010-2011 and 2011-2012 school years, the years of the data used in Sections III and IV.

III. RESULTS

A. Teacher and Student Misclassification

We would expect unconfirmed and confirmed rosters to disagree to some extent: the difference between the two is the value of roster confirmation. If elementary school teachers primarily teach in self-contained classrooms and middle school teachers primarily teach single subjects, there would be no reason to think that discrepancies would be more or less likely to arise across grade levels. However, if there is a large amount of departmentalization of instruction in grades 4 and 5 that is undocumented in administrative rosters, there might be greater discrepancies between the unconfirmed and confirmed rosters at these grade levels.

As a first look at how administrative elementary and middle school data compare to roster-confirmed data and to each other, we examined teachers and students who appeared either on unconfirmed rosters, confirmed rosters, or both rosters. The top panel of Table III.1 shows results separately for teachers and students in the upper elementary grades; the bottom panel shows results for teachers and students in the middle school grades. "All" teachers refers to any math and/or reading/ELA teacher who appeared in the rosters. We used data from two school years, 2010–2011 and 2011–2012.

Most teachers and students appeared in both the confirmed and the unconfirmed rosters, although some showed up in only one of the rosters. In middle school grades, 83.1 percent of teachers appeared in both rosters in at least one subject; in elementary school grades, 83.0 percent of teachers were listed in at least one subject. Of course, this means that about one in six teachers appeared in only one roster. Of these, more teachers appeared in the confirmed roster than in the unconfirmed roster. In some cases, team teachers who were excluded from unconfirmed rosters appeared on confirmed rosters.[2] Another type of misclassification occurred at the teacher–student level: students either were linked to teachers who do not teach them or were not linked to teachers who did.

Looking by subject, however, the results show a discrepancy in the percentage of misclassified teachers by grade span, with more upper elementary school teachers than middle school teachers in the unconfirmed rosters for each subject. For math, 18.7 percent of upper elementary teachers appear only in the unconfirmed data, compared to 6.7 percent of middle school teachers. For reading/ELA, the comparable percentages are 16.9 percent compared to 7.1 percent. Percentages are more similar to one another for teachers who appear as either a math or a reading/ELA teacher in the unconfirmed data—6.0 percent for upper elementary grades compared to 4.7 percent for middle school grades. This suggests that unconfirmed rosters may in fact be failing to pick up departmentalization of instruction in upper elementary grades.

[2] A small number of differences may be explained by teacher mobility, since the unconfirmed rosters were compiled in the fall and the confirmed rosters in the spring. Some teachers may have moved in and out of DCPS during the school year. A few mismatches may also have occurred because the unconfirmed rosters included only teacher names (but not teacher IDs), so we may have inadvertently failed to match a few teachers who were present in both rosters, such as married women who changed their last names.

Table III.1. Percentage of Teachers and Students in the Confirmed and Unconfirmed Rosters

	In Confirmed Roster Only	In Unconfirmed Roster Only	In Both Rosters	Sample Size (In Both Rosters)
Grades 4 to 5				
Teachers				
Math	8.8%	18.7%	72.5%	626
Reading/ELA	8.8%	16.9%	74.4%	628
All	10.8%	6.0%	83.1%	646
Students				
Math	4.9%	5.5%	89.6%	11,072
Reading/ELA	4.6%	4.7%	90.7%	10,954
All	5.0%	3.4%	91.6%	11,104
Grades 6 to 8				
Teachers				
Math	10.7%	6.7%	82.7%	225
Reading/ELA	13.3%	7.1%	79.6%	240
All	12.3%	4.7%	83.0%	424
Students				
Math	6.2%	3.7%	90.1%	11,661
Reading/ELA	11.8%	3.4%	84.8%	11,516
All	4.7%	1.2%	94.1%	12,043

Source: District of Columbia Public Schools and Office of the State Superintendent of Education administrative data.

Notes: Statistics are based on two years of data, from the 2010–2011 and 2011–2012 school years. The sample size is the number of teacher-year or student-year observations.

Math teachers refer to teachers who taught math only or math and reading/ELA. Reading/ELA teachers refer to teachers who taught reading/ELA only or math and reading/ELA. "All teachers" refers to teachers who taught math only, reading/ELA only, or both subjects.

Math students are students who were linked to math courses only or to both math and reading/ELA. Reading/ELA students are students who were linked to reading/ELA courses only or to both reading/ELA and math. "All students" refers to students who were linked to math courses only, reading/ELA courses only, or both.

As a second of way of analyzing the data, we compared the two rosters once more, limiting the analysis to teachers and students who were present in both rosters, and treating the confirmed rosters as the correct information against which the unconfirmed rosters were compared. To calculate the percentage of teachers who were linked by the unconfirmed administrative data to subjects they did not teach, we created a list of teachers who taught only reading/ELA or math according to the confirmed rosters and looked for instances in the unconfirmed rosters where they were linked to both subjects. To calculate the percentage of teacher–student links that were incorrect, we divided the number of teacher–student links that appeared in only the unconfirmed rosters (but not in the confirmed rosters) by the total number of unique teacher–student links in the unconfirmed rosters. We combined results from the 2010–2011 and 2011–2012 school years. Table III.2 summarizes the prevalence of both types of misclassification.

Table III.2. Prevalence of Misclassification in Unconfirmed Roster Data

	Percentage of Teachers Linked to Subjects They Did Not Teach	Percentage of Teacher–Student Links That Were Incorrect	Number of Teachers	Number of Teacher–Student Links
Math				
Grade 4	12.8	21.2	234	5,072
Grade 5	16.4	18.2	226	4,845
Grades 4 and 5	**14.8**	**19.7**	**453**	**9,917**
Grade 6	5.4	5.5	93	3,524
Grade 7	0.0	4.2	83	3,618
Grade 8	0.0	4.1	77	3,369
Grades 6 to 8	**2.7**	**4.6**	**188**	**10,511**
Reading/ELA				
Grade 4	16.7	18.1	245	5,069
Grade 5	18.1	19.3	232	4,867
Grades 4 and 5	**17.1**	**18.7**	**467**	**9,936**
Grade 6	6.1	6.7	99	3,314
Grade 7	0.0	7.7	84	3,245
Grade 8	0.0	7.4	79	3,212
Grades 6 to 8	**3.1**	**7.3**	**191**	**9,771**

Source: District of Columbia Public Schools and Office of the State Superintendent of Education administrative data.

Notes: Statistics are based on two years of data, from the 2010–2011 and 2011–2012 school years. The sample size is the number of teacher-year or student-year observations.

In this table, we only considered teachers and students whom we were able to match and identify across the confirmed and unconfirmed rosters.

The proportion of teachers linked to subjects they did not teach was larger among those who taught elementary school grades than among those who taught middle school grades. In math, 14.8 percent of upper elementary school teachers were misclassified, compared to 2.7 percent of middle school teachers. In reading/ELA, 17.1 percent of upper elementary teachers were misclassified, compared to 3.1 percent of middle school teachers. Consequently, there were also higher percentages of incorrect teacher–student links in elementary school grades. For grades 7 and 8, no teachers were misclassified by subject for either year. All misclassifications for middle school grades were in grade 6; all misclassified grade 6 teachers were in schools that spanned grades K–6 or K–8, where grade 6 homeroom teachers were assumed to be teaching both subjects. Sixth-grade teachers were never misclassified by subject in traditional middle schools housing grades 6–8.

Administrative rosters are used for course scheduling in middle schools, but not in elementary schools. Consequently, it is easier to track the identities of a student's math and reading/ELA teachers in a middle school. In elementary schools, most classrooms are self-contained, so administrative rosters are not often used for course scheduling. As a result, the departmentalization of elementary classrooms is easy to miss when it does happen. Although it is not evident from the unconfirmed data, the confirmed data show that in DCPS, 27 percent of upper elementary school students are taught by a departmentalized teacher in math and 28 percent in reading/ELA, with greater departmentalization in grade 5 than in grade 4.

B. Tracking and Heterogeneity of Classrooms

A second reason why one might prefer to restrict studies that measure teacher effectiveness using value-added models to elementary school grades is that middle students may be more likely to be grouped by achievement levels ("tracked") into different academic tracks—that is, some students will be placed in more advanced courses than others. Having unobservable differences in students by their academic track, few teachers who teach students on both tracks, or differing degrees of alignment between the post-tests and advanced versus basic courses can all pose problems for estimating teacher value added (Jackson 2012; Protik et al. 2013; Harris and Anderson 2013). Loveless (2009) writes that "middle school is where tracking begins, providing a bridge between the heterogeneously grouped classes of elementary school and the tracked classes of high school."

To examine the degree of tracking by grade level, we measured the percentage of variation in student pre-test scores arising from differences between students at different schools, between students in different classrooms within a school, and between students within the same class. If there is more tracking at the middle school level, one would expect to see a greater amount of variation in test scores accounted for within schools. We used a three-level unconditional hierarchical linear model, with students nested within classrooms, and classrooms nested within schools, to decompose the variance of pre-test scores into these three components.[3] Data are from the 2011–2012 school year because data from 2010–2011 do not contain indicators distinguishing separate classrooms within a teacher. We used only confirmed roster data for these calculations. The course titles used in DCPS suggest that, at the middle school level, math courses are targeted to students at different levels of prior achievement, but it is not clear whether this is the case in reading/ELA (Protik et al. 2013).

As expected, there is a greater share of the total variation explained by within-school variation in the middle school grades compared to the upper elementary school grades, suggesting more tracking in middle school (Table III.3). For math, 3 percent of the variation in student pre-test scores is explained by within-school differences in upper elementary grades, compared to 22 percent in middle school grades. For reading/ELA, the difference is 4 percent for upper elementary grades compared to 15 percent for middle school grades. For both subjects, the percentage of the variance explained by variation between classrooms within a school grows by grade level. For example, in math, it grows from 1 percent for grade 4 to 27 percent for grade 8. This is evidence that more schools adopt tracking as grade levels progress.

[3] We excluded grades in schools with a single classroom since it was not possible for a school to create separate tracks in this case. Instead, because there can be no variation between classrooms in these schools, we included them in the results by assigning a value of 0 for between teacher/class variation and weighting the HLM estimates by one minus the share of students who were excluded. This was a substantial share of schools in DCPS: 19 percent of school–grade combinations had one classroom in math and 20 percent had one classroom in reading. If we weight each school–grade combination by the number of students, 9 percent of school–grade combinations had one classroom.

Table III.3. Evidence of Tracking by Grade Span: Decomposition of Variance of Pre-test Scores

	Percentage of Variance Between Schools	Percentage of Variance Between Classrooms Within Schools	Percentage of Variance Between Students Within Classrooms	Number of Students
Math				
Grade 4	0.30	0.01	0.69	3,364
Grade 5	0.29	0.05	0.66	3,208
Grades 4 and 5	**0.28**	**0.03**	**0.69**	**6,572**
Grade 6	0.14	0.15	0.71	2,213
Grade 7	0.10	0.21	0.69	1,902
Grade 8	0.14	0.27	0.59	2,019
Grades 6 to 8	**0.12**	**0.22**	**0.66**	**6,134**
Reading/ELA				
Grade 4	0.24	0.01	0.75	3,424
Grade 5	0.26	0.06	0.68	3,395
Grades 4 and 5	**0.23**	**0.04**	**0.73**	**6,819**
Grade 6	0.15	0.09	0.76	2,398
Grade 7	0.10	0.13	0.77	2,354
Grade 8	0.16	0.20	0.64	2,394
Grades 6 to 8	**0.12**	**0.15**	**0.73**	**7,146**

Source: District of Columbia Public Schools and Office of the State Superintendent of Education administrative data.

Notes: The decomposition of the variance of pre-test scores at each level was carried out using a three-level, unconditional hierarchical linear model.

Data are from students enrolled in DCPS in the 2011–2012 school year.

The greater variation between classrooms within a school at the middle school level is offset by greater variation between schools at the upper elementary school level, however, as would result from more homogenous attendance areas for elementary schools compared to larger attendance areas for middle schools. Consequently, the percent of variation explained by differences between students within a classroom has a smaller range across grade levels than the percentage explained by between-school or between-classroom differences. For math, despite a difference of 19 percentage points between the upper elementary and middle school grade levels in the percentage of variance explained by differences between classrooms within a school, the percentage of variance explained by variation between schools is greater at the upper elementary school level by 16 percentage points—28 percent versus 12 percent. For reading/ELA, a difference of 11 percentage points between classrooms within schools is exactly offset by greater between-school differences at the upper elementary school level. One caveat in generalizing these results to other districts is that there were many small schools in DCPS in the 2010–2011 school year; if smaller attendance areas contributed to more homogenous groups of students at each school in DCPS, then the amount of between-school variation may be higher than would be expected in a district with larger schools.

For value-added models, if one estimates differences for teachers within a school by including school fixed effects (for example, Jackson and Bruegmann 2009), there could be advantages to focusing on upper elementary school grades. For models that seek to compare all teachers in a district, the potential advantage of using elementary school grades disappears, as greater within-school differences in middle school grades are offset by greater between-school differences in students in upper elementary school grades. Variations in test scores are not by themselves problematic, since value-added models account for pre-test scores as a key control

variable. However, greater between-school heterogeneity might result in difficulties in estimating value added at the upper elementary school level that parallel those present at the middle school level. For example, one might be concerned that there are unobservable differences in parents across schools at the upper elementary school level, few teachers who teach students with overlapping distributions of pre-test achievement, or better alignment between post-tests and the curriculum as implemented in some schools compared to others.

IV. APPLICATION: CORRELATION OF VALUE-ADDED ESTIMATES ACROSS YEARS AND GRADES

As an application of what would happen if we were to base judgments about value-added models only on unconfirmed data, we measured the correlation in value-added estimates across years and grades using the unconfirmed data and again using the roster-confirmed data. Interest in year-to-year correlation in value has increased as school districts have adopted value-added estimates as a component of teacher evaluation systems. These systems implicitly assume that a teacher's evaluation this year is a good predictor of his or her effectiveness the next year. Cross-subject correlations are important in an evaluation context if teachers are evaluated on tested subjects but not on untested subjects. Correlations between mismeasured estimates and other estimates for the same teachers in another subject or another year, which may or may not also be mismeasured, could appear more or less stable than correlations based on roster-confirmed data without misclassification errors. This is because basing value added on better data will better reflect actual fluctuations in teacher effectiveness between subjects and years.

Recent studies suggest that year-to-year correlations of value-added estimates range from 0.2 to 0.7 and that cross-subject correlations range from 0.3 to 0.6 (Loeb and Candelaria 2012). Goldhaber and Hansen (2010) estimated year-to-year correlations of 0.3 in reading and 0.6 in math, based on single-cohort value-added models for grade 5. Looking at five large Florida school districts, McCaffrey et al. (2009) found year-to-year correlations of teacher value added in math from 0.2 to 0.5 in elementary school and from 0.3 to 0.7 in middle schools. Aaronson et al. (2007) and Koedel and Betts (2007) observed that teachers at the top or bottom of the value-added distribution in one year are more likely to end up at the same end of the distribution the following year. Examples of cross-subject correlations include Koedel and Betts (2007), who found a correlation between math and reading value added of 0.4, and Loeb et al. (2012) and Goldhaber et al. (2012), who found correlations between 0.5 and 0.6.[4]

We estimated value-added models for math and reading/ELA teachers in the 2010–2011 and 2011–2012 school years separately for the analysis files created using confirmed rosters and

[4] Although value-added estimates that vary from year to year or between subjects may reflect real fluctuations in a teacher's effectiveness, they might also indicate some imprecision and bias in the process for estimating value-added estimates. Correlations that adjust for the amount of imprecision in the estimates—and therefore reflect year-to-year correlations in underlying teacher effectiveness if the estimates are unbiased—are larger. For example, the year-to-year correlations in Goldhaber and Hansen (2010) increased to 0.6 in reading and 0.7 in math after adjusting for imprecision. Also, the cross-subject correlation in Koedel and Betts (2007) increased to 0.6 after this adjustment. However, if value-added estimates are biased, these correlations may not reflect true stability. Furthermore, the direction of the effect on stability of removing bias will depend on the source and form of the bias.

those created using unconfirmed rosters. We applied the same methods and exclusion rules to both sets of files. After having assembled two data sets based on the confirmed and unconfirmed class rosters, we proceeded to estimate value added by year (2010–2011 and 2011–2012) and subject (math and reading) for all teachers with at least 15 students in the subject–year combination. These value-added estimates were made based on one year of student growth data. We included all teachers who appeared on each of the rosters and did not restrict the sample of teachers analyzed as we did in Table III.2. However, only teachers who taught 15 or more students over the course of the school year in at least one subject have value-added estimates. Full details of the value-added model are given in Appendix A, and details for the other data used to estimate value added are given in Appendix B.

A. Year-to-Year Correlations

For teachers who have same-subject value-added estimates for both the 2010–2011 and 2011–2012 academic years, we examined the stability of these value-added estimates from one year to the next. These correlations are presented by grade and grade span in Table IV.1.

Table IV.1. Year-to-Year Correlations of Estimated Teacher Value Added (2010–2011 and 2011–2012)

	Unconfirmed Roster (1)	Number of Teachers (Unconfirmed Rosters) (2)	Confirmed Roster (3)	Number of Teachers (Confirmed Roster) (4)
Math Teachers				
Grade 4	0.30	47	0.33	50
Grade 5	0.28	57	0.20	57
Grades 4 and 5	**0.25**	**118**	**0.22**	**112**
Grades 6 to 8	**0.26**	**44**	**0.23**	**47**
Grades 4 to 8	**0.26**	**169**	**0.21**	**164**
Reading/ELA Teachers				
Grade 4	0.32	46	0.53	48
Grade 5	0.48	56	0.33	62
Grades 4 and 5	**0.34**	**115**	**0.40**	**119**
Grades 6 to 8	**0.16**	**44**	**0.08**	**56**
Grades 4 to 8	**0.23**	**167**	**0.29**	**179**

Source: District of Columbia Public Schools and Office of the State Superintendent of Education administrative data.

Notes: The numbers of teachers at individual grade levels do not necessarily add up to the total number of teachers within grade spans. This is because some teachers teach multiple grades and individual grade-level correlations only consider teachers who taught the same grade in both 2010–2011 and 2011–2012.

The correlations in Table IV.1 suggest no systematic gain (or loss) in year-to-year stability of value-added estimates transitioning from using unconfirmed roster data to confirmed roster data. The correlation decreased from 0.26 to 0.21 for math teachers overall, and increased from 0.23 to 0.29 for reading/ELA teachers overall. However, all of the increase in the correlation for reading/ELA is due to grade 4 teachers. The stability of year-to-year value-added estimates of grade 4 teachers appears to have increased in both math and reading/ELA when confirmed roster data was used in place of unconfirmed roster data, though we did not find an increase in any

other grade or grade span.[5] We chose not to adjust the correlations in Table IV.1 to account for imprecision, so the results reflect year-to-year fluctuations in value added due to both underlying teacher effectiveness and imprecision.

In elementary grades, value-added estimates that use unconfirmed roster data need not lead to lower year-to-year correlations than those that use confirmed roster data, as long as classrooms of students rotate as a group from one teacher to another. For example, suppose teacher A in classroom A and teacher B in classroom B were erroneously classified as teaching both math and reading/ELA, when teacher A actually teaches math to both classes and teacher B teaches reading/ELA to both classes. The unconfirmed roster correlations would include value-added estimates for both teachers in both subjects, whereas the confirmed data would include estimates for the teachers only in the subject they actually teach. Teacher A's reading value added in the unconfirmed data would actually measure teacher B's contributions in classroom A. In this example, including the misclassified value-added information in the correlation calculation would not obviously lead to lower stability, because teacher B's contributions are linked year to year, even though they are attributed to teacher A and based on just one classroom.

B. Cross-Subject Correlations

In addition to examining the correlation of teachers' value-added estimates from one year to the next, we also looked at whether teachers who taught both math and reading/ELA within the same year and had a high value-added estimate for one subject also tended to have a high value-added estimate for the other subject. We excluded 7th- and 8th-grade teachers from this analysis because these grades are fully departmentalized in our data. We did, however, examine the small number of grade 6 teachers who taught both math and reading/ELA in self-contained classroom settings, although correlations using grade 6 data are aggregated with those for grades 4 and 5. Again, we chose not to adjust these correlations to account for imprecision, so the results reflect fluctuations in value added across subjects due to both underlying teacher effectiveness and imprecision.

Comparing the correlations in columns (1) and (3), it appears that the cross-subject correlations in elementary grades are higher when we use confirmed roster data to estimate value added than when we used unconfirmed roster data. As shown in Table IV.2, for grades 4 to 6 combined, the cross-subject correlation in the unconfirmed rosters is 0.60 but rises to 0.65 in the confirmed rosters. One explanation is to consider again the case in which teacher A was erroneously classified as teaching both subjects in classroom A and teacher B was erroneously classified as teaching both subjects in classroom B, although teacher A actually teaches math to both classes and teacher B teaches reading/ELA to both classes. The confirmed roster cross-subject correlations will not include teachers A and B, because they each teach only one subject. However, both teachers' cross-subject correlations will be included in the case of the unconfirmed roster. If teacher A's effectiveness is very different from teacher B's, the erroneous cross-subject correlations for both teachers will be low.

[5] We did not compare correlations for individual middle school grades because of small within-grade sample sizes.

Table IV.2. Cross-Subject Correlations of Estimated Teacher Value Added

	Unconfirmed Roster (1)	N (2)	Confirmed Roster (3)	N (4)
Grade 4	0.67	237	0.69	182
Grade 5	0.54	235	0.64	187
Grades 4 to 6	**0.60**	**488**	**0.65**	**366**

Source: District of Columbia Public Schools and Office of the State Superintendent of Education administrative data.

Notes: Statistics are based on two years of data, from the 2010–2011 and 2011–2012 school years. The sample size is the number of teacher-year observations.

Grade-level correlations for grades 4 and 5 compare value-added estimates of teachers who taught both subjects for that grade. Because some teachers taught multiple grades, teachers could be counted separately within a year for grades 4 and 5, but would only be counted once for that year when combining results across grades 4 to 6.

As a test of this potential explanation, we calculated correlations based on a restricted sample of 250 teacher-year observations over two years. These teachers were linked to both subjects in both the confirmed and unconfirmed rosters. Consistent with the explanation of subject assignment misclassifications, the cross-subject correlations were 0.61 in both data sets. These results suggest that cross-subject correlations based on administrative data may be lower than correlations based on roster-confirmed data as a result of subject assignment misclassifications.

V. CONCLUSIONS

It is often assumed that elementary school data are better suited than middle school data to research that measures teacher effectiveness using value-added models because (1) elementary school teachers have a single, self-contained classroom of students to whom they teach all subjects and (2) researchers can avoid potential difficulties associated with tracking by achievement level in middle schools. We have shown, however, that, at least with DCPS data, it would often be incorrect to assume no departmentalization of instruction in grades 4 and 5. Attributing subject assignments to teachers based on the homeroom assignments of students in grades 4 and 5 would lead to about one in six teachers being misclassified in each subject. In contrast, subject assignments for middle school teachers are rarely incorrect. As for concerns that classrooms of the same grade and subject are stratified by achievement in middle school, we confirmed that there is greater variation in student achievement across classrooms in grades 6 to 8, but this is offset by more stratification by achievement across schools at the upper elementary school level.

The important question for using administrative data for research on teachers is when and how much these misclassification errors matter. In the example we examined, misclassification errors mattered much more for measuring cross-subject correlations in value-added estimates than in measuring cross-year correlations. In general, if the same students stay together when they are taught by different teachers, misclassification will not matter as long as the identity of the teacher is unimportant to the application being studied. This phenomenon appears to contribute to our finding little difference in the measure of year-to-year correlations when using roster-confirmed versus unconfirmed data. If knowing the identity of the teacher is important, however—as in the example of cross-subject correlations or, for example, if value-added data are

to be linked to teacher personnel data—then misclassification errors that arise using unconfirmed data may more seriously affect the results.

In the long run, as more districts adopt roster confirmation as a way of incorporating value added into teacher evaluation systems, one side benefit will be better data on teacher–student links for research. Researchers may then no longer need to be concerned about problems of using unconfirmed data. In the meantime, because of missing data on departmentalization of elementary school teaching, one should be cautious about assuming that data from upper elementary school grades are necessarily superior to data from middle school grades for conducting research that involves value-added estimates. The extent of problems that arise with using unconfirmed data depend on the way in which teacher value-added estimates are merged or otherwise linked to other data on teachers.

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.

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

Battelle for Kids. "The Importance of Accurately Linking Instruction to Students to Determine Teacher Effectiveness." Columbus, OH: Battelle for Kids, October 2009.

Battelle for Kids. "Roster Verification." Columbus, OH: Battelle for Kids, 2013.

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

Buonaccorsi, John P. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: Chapman & Hall/CRC, 2010.

CTB/McGraw-Hill. *Technical Report for Spring 2010 Operational Test Administration of DC CAS.* Monterey, CA: CTB/McGraw-Hill, 2010.

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

Goldhaber, Dan, James Cowan, and Joe Walch. "Is a Good Elementary Teacher Always Good? Assessing Teacher Performance Estimates Across Subjects." Seattle, WA: Center for Education Data and Research, 2012.

Goldhaber, Dan, and Michael Hansen. "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance." Seattle, WA: Center for Education Data and Research, 2010.

Harris, Douglas, and Andrew Anderson. "Does Value-Added Work Better in Elementary Than in Secondary Grades?" Washington, DC: Carnegie Knowledge Network, May 2013.

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

Isenberg, Eric, and Heinrich Hock. "Design of Value Added Models for IMPACT and TEAM in DC Public Schools, 2010–2011 School Year." Washington, DC: Mathematica Policy Research, May 2011.

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

Jackson, Kirabo. "Teacher Quality at the High-School Level: The Importance of Accounting for Tracks." Working Paper #17722. Cambridge, MA: National Bureau of Economic Research, 2012.

Jackson, C. Kirabo, and Elias Bruegmann. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers." *American Economic Journal: Applied Economics*, vol. 1, no. 4, October 2009, pp. 1–27.

Kluender, Ray, Chris Thorn, and Jeff Watson. "Why Are Student-Teacher Linkages Important? An Introduction to Data Quality Concerns and Solutions in the Context of Classroom-Level Performance Measures." Madison, WI: Center for Educator Compensation Reform, 2011.

Koedel, Cory, and Julian Betts. "Re-Examining the Role of Teacher Quality in the Educational Production Function." Working paper, University of Missouri, 2007.

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

Loeb, Susanna, Tara Beteille, and Demetra Kalogrides. "Effective Schools: Teacher Hiring, Assignment, Development and Retention." *Education Finance and Policy*, vol. 7, no. 3, 2012, pp. 269–304.

Loeb, Susanna, and Christopher Candelaria. "How Stable Are Value-Added Estimates Across Years, Subjects, and Student Groups?" Carnegie Knowledge Network, October 2012.

Loveless, Tom. "Tracking and Detracking: High Achievers in Massachusetts Middle Schools." Washington, DC: Fordham Institute, 2009.

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy,* vol. 4, no. 4, fall 2009, pp. 572–606.

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

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

Rothstein, Jesse. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics*, vol. 125, no. 1, February 2010, pp. 175–214.

Sass, Tim, Jane Hannaway, Zeyu Xu, David Figlio, and Li Feng. "Value Added of Teachers in High-Poverty Schools and Lower-Poverty Schools." *Journal of Urban Economics*, vol. 72, no. 2, 2012, pp. 104–122.

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

APPENDIX A

Estimating Teacher Effectiveness

1. Value-Added Model

After assembling the analysis file, we estimated a regression separately for math and reading using students at upper elementary (grades 4 and 5) and middle school (grades 6, 7, and 8) levels in the data. In each regression equation, the post-test score depends on prior achievement, student background characteristics, variables linking students to schools or teachers, and unmeasured factors.

The model accounts for team teaching that occurs at the teacher level and the unit of observation is a teacher–student combination. We assume that the combined effort of team teachers constitutes a single input into student achievement (Hock and Isenberg 2012). For a given teacher *t* and student *i*, the regression equation may be expressed as

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

where Y_{ig} is the post-test score for student *i* in grade *g* and $Y_{i(g-1)}$ is the same-subject pre-test for student *i* in grade $g-1$ during the previous year. The variable $Z_{i(g-1)}$ denotes the pre-test in the opposite subject. Thus, when estimating teacher value added in math, *Y* represents math tests and *Z* represents reading tests, and vice versa. The pre-test scores capture prior inputs into student achievement, and the associated coefficients, λ_g and ω_g, vary by grade. The vector $\mathbf{X}_i$ denotes the control variables for individual student background characteristics. The coefficients on these characteristics, α, are constrained to be the same across all grades within the relevant grade span.[6] The vector $\mathbf{T}_{tig}$ includes a grade-specific variable for each teacher and includes a variable for a catchall ineligible teacher in each grade to account for student dosage that cannot be attributed to a particular teacher who is eligible to receive a value-added estimate. A student contributes one observation to the model for each teacher the student is linked to, 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 $\boldsymbol{\eta}$ and we mean centered the control variables so that each element of $\boldsymbol{\eta}$ represents a teacher- and grade-specific intercept term for a student with average characteristics.[7]

[6] We estimated 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 allowed us to base the association between achievement and student characteristics on information from all grades, which should smooth out the between-grade differences in these coefficients.

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

To account for multiple observations on the same student, we estimated the coefficients by using weighted least squares rather than ordinary least squares. In this method, the teacher–grade variables in $\mathbf{T}_{tig}$ are binary, and we weighted each teacher–student combination by the teacher dosage associated with that combination. We addressed the correlation in the error term, ε_{2tig}, across multiple observations by using a cluster-robust sandwich variance estimator (Liang and Zeger 1986; Arellano 1987) to obtain standard errors that are consistent in the presence of both heteroskedasticity and clustering at the student level.

This 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 estimated a grade-specific coefficient for a teacher only if he or she teaches at least seven students in that grade.[8] We then aggregated teacher estimates across grades to form a single estimate for each teacher (see Section 3 below).

2. Measurement Error in the Pre-Tests

We corrected for measurement error in the pre-tests by using grade-specific reliability data available from the test publisher (CTB/McGraw Hill 2010; 2011). 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 implemented a measurement error correction based on the test/retest reliability of the DC CAS tests. By netting out the known amount of measurement error, the errors-in-variables correction eliminates this source of bias (Buonaccorsi 2010).

Correcting for measurement error requires a two-step procedure. In the first step, we used a dosage-weighted errors-in-variables regression based on Equation (1) to obtain unbiased estimates of the pre-test coefficients for each grade. We used the published reliabilities associated with the 2010 and 2011 DC CAS. We then used the measurement-error corrected values of the pre-test coefficients to calculate the adjusted gain for each student in each grade. The adjusted gain is expressed as

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

and represents the student post-test outcome, net of the estimated contribution attributable to the student's starting position at pre-test.

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

In the second step, we pooled the data from all grades and used the adjusted gain as the dependent variable in a single equation expressed as

$$(3) \quad \hat{G}_{tig} = \boldsymbol{\alpha}_2' \mathbf{X}_{2i} + \boldsymbol{\eta}' \mathbf{T}_{tig} + \varepsilon_{tig} .$$

We obtained the grade-specific estimates of teacher effectiveness, $\hat{\boldsymbol{\eta}}$, by applying the weighted least squares regression technique to Equation (3).

This two-step method will likely underestimate the standard error of $\hat{\boldsymbol{\eta}}$ because the adjusted gain in Equation (2) relies on the estimated value of λ_g, which implies that the error term in Equation (3) is clustered within grades. This form of clustering typically results in estimated standard errors that are too small because the second-step regression does not account for a common source of variability affecting all students in a grade. In view of the small number of grades, standard techniques of correcting for clustering will not effectively correct the standard errors (Bertrand et al. 2004). Nonetheless, with the large within-grade sample sizes, the pre-test coefficients are likely to be estimated precisely, leading to a negligible difference between the robust and clustering-corrected standard errors.

3. Combining Estimates Across Grades

Both the average and the variability of value-added estimates may differ across grade levels, leading to a potential problem when comparing teachers assigned to different grades. 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 estimates. For example, the standard deviation of adjusted gains might vary across grades as a consequence of differences in the alignment of tests or the retention of knowledge between years. However, we sought to compare all teachers to all others in the regression regardless of any grade-specific factors that might affect the distribution of gains in student performance between years.[9] Because we did not want to penalize or reward teachers simply for teaching in a grade with unusual test properties, we translated grade-level estimates for teachers so that each set of estimates is expressed in a common metric.

We standardized the estimated regression coefficients so that the mean and standard deviation of the distribution of teacher estimates are the same across grades. First, we subtracted from each unadjusted estimate the weighted average of all 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 taught by each teacher.

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

[9] Because each student's entire dosage was 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.

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 assumed that the distribution of true teacher effectiveness is the same across grades.

To combine effects across grades into a single effect for a given teacher, we used a weighted average of the grade-specific estimates. We set the weight for grade *g* equal to the proportion of students of teacher *t* in grade *g*. Because combining teacher effects across grades may cause the overall average to be nonzero, we re-centered the estimates on zero before proceeding to the next step.

4. Shrinkage Procedure

To reduce the risk that teachers, particularly those with relatively few students in their grade, will receive a very high or very low effectiveness measure by chance, we applied the empirical Bayes (EB) shrinkage procedure, as outlined in Morris (1983), separately to the sets of effectiveness estimates for teachers. Using the EB procedure, we computed a weighted average of an estimate for the average teacher (based on all students in the model) and the initial estimate based on each teacher's own students. For teachers with relatively imprecise initial estimates based on their own students, the EB method effectively produces an estimate based more on the average teacher. For teachers with more-precise initial estimates based on their own students, the EB method puts less weight on the value for the average teacher and more weight on the value obtained from the teacher's own students.

The EB estimate for a teacher is approximately equal to a precision-weighted average of the teacher's initial estimated effect and the overall mean of all estimated teacher effects.[10] We calculated the standard error for each shrinkage estimate using the formulas provided by Morris (1983). As a final step, we removed any teachers with fewer than 15 students from the model and re-centered the EB estimates on zero.

[10] Following 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.

APPENDIX B

Data

When estimating the effectiveness of DCPS teachers, we included students in grades 4-8 if they had a post-test score from 2011 or 2012 and a pre-test from the previous grade in the same subject in the previous year. We excluded students from the analysis file in the case of missing or conflicting school enrollment data, test-score data, or student background data.[11,12] We also excluded students who repeated or skipped a grade because they lacked pre-test and post-test scores in consecutive grades and years. Between 86 and 89 percent of students with post-test scores were included in each of the subject- and year-specific value-added models.

In addition to the confirmed and unconfirmed rosters discussed in the preceding section, we used other administrative data provided by DCPS and the Office of the State Superintendent of Education (OSSE) to estimate value-added models. In particular, we used data on student test scores, student background characteristics, and school enrollment.

The DC Comprehensive Assessment System (DC CAS), administered in April, is the set of standardized tests used for accountability purposes in DCPS. Test scores on the DC CAS are not vertically scaled and may therefore be meaningfully compared only within grades and within subjects. Math scores, for example, are generally more dispersed than reading scores within the same grade. Therefore, before using the test scores in the value-added model, we created subject- and grade-specific z-scores by subtracting the mean and dividing by the standard deviation within a subject-grade combination. This step allowed us to translate math and reading scores in every grade and subject into a common metric.

We used data provided by OSSE and DCPS to construct variables used in the value-added models as controls for student background characteristics. All value-added models account for the following: pre-test in same and opposite subjects, gender, race/ethnicity, 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, and the proportion of days that the student attended school during the previous year. We included attendance because it could reflect some aspects of student motivation. We used previous—rather than current-year—attendance to avoid confounding student attendance with current-year teacher effectiveness; that is, a good teacher versus a weaker teacher might be expected to motivate students to attend school more regularly. Attendance is a continuous variable that could range from zero to one. Aside from pre-test variables, the other variables are binary variables taking the value zero or one. Table B.1 shows the characteristics of students from the confirmed rosters included in the value-added models. The composition of students from the unconfirmed rosters included in the value-added models is similar.

[11] We considered a student who answered fewer than five questions on the DC CAS post-test to be missing test score data.

[12] We included students who were missing individual student background characteristics but excluded those for whom no data on background characteristics were available.

Table B.1. Mean Characteristics of Students from the Confirmed Rosters Included in the Value-Added Models

	Math		Reading	
	2010–2011	2011–2012	2010–2011	2011–2012
Male	0.49	0.48	0.49	0.50
White	0.11	0.11	0.13	0.12
Black	0.72	0.72	0.71	0.73
Hispanic	0.17	0.17	0.17	0.15
Eligible for Free Lunch	0.64	0.64	0.68	0.69
Eligible for Reduced-price Lunch	0.06	0.06	0.05	0.05
Limited English Proficiency	0.08	0.07	0.07	0.06
Proportion of the Prior Year Student Attended School	0.95	0.95	0.94	0.94
Specific Learning Disability	0.06	0.07	0.08	0.09
Other Learning Disability	0.04	0.04	0.05	0.05

Source: District of Columbia Public Schools and Office of the State Superintendent of Education administrative data.

Notes: Free and reduced-price lunch status was imputed using data from prior years for approximately 12 percent of students in the teacher model. For all other student characteristics, less than 1 percent of students had missing data.

We imputed data for students who were included in the analysis file but who had missing values for one or more student characteristics. Our imputation approach used the values of nonmissing student characteristics to predict the value of the missing characteristic.[13] We did not generate imputed values for the same-subject pre-test; we dropped from the analysis file any students with missing same-subject pre-test scores.

Given that some students moved between schools or were taught by a combination of teachers, we apportioned their achievement among more than one school or teacher. We refer to the fraction of time the student was enrolled at each school and with each teacher as the "dosage." We created "school dosage" for each school–student combination based on school enrollment data. If the roster data indicated that a student had one math or reading/ELA 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 determined the number of days the student spent with each teacher, subdividing the school dosage among teachers accordingly. When two or more teachers claimed the same students during the same term, we assigned each teacher full credit for the shared students. We therefore did not subdivide dosage for co-taught students.

[13] For missing data on free or reduced-price lunch status, we used an alternative imputation procedure because these data are missing for DCPS students attending Provision 2 schools, which do not collect information on free and reduced-price lunch status every year. We also used an alternative imputation method to impute missing attendance data for students who did not attend a DC school for part of the previous year. These methods are described in Isenberg and Hock (2012).

Authors' Note

We thank the Office of the State Superintendent of Education of the District of Columbia (OSSE) and the District of Columbia Public Schools (DCPS) for providing the data for this study. We are grateful to Duncan Chaplin and Steve Glazerman for their helpful comments. Emma Kopa, assisted by Maureen Higgins, provided excellent programming support. The paper was edited by Betty Teller and produced by Jackie McGee. The text reflects the views and analyses of the authors alone and does not necessarily reflect views of Mathematica Policy Research, OSSE, or DCPS. All errors are the responsibility of the authors.

About the Series

Policymakers require timely, accurate, evidence-based research as soon as it's available. Further, statistical agencies need information about statistical techniques and survey practices that yield valid and reliable data. To meet these needs, Mathematica's working paper series offers policymakers and researchers access to our most current work. For more information about this paper, contact Eric Isenberg, senior researcher, at ejisenberg@mathematica-mpr.com, or Elias Walsh, researcher, at ewalsh@mathematica-mpr.com.

www.mathematica-mpr.com

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

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

MATHEMATICA
Policy Research