WORKING PAPER

Does Tracking of Students Bias Value-Added Estimates for Teachers?

March 2013

Ali Protik
Elias Walsh
Alexandra Resch
Eric Isenberg
Emma Kopa

MATHEMATICA
Policy Research

ABSTRACT

We compare two alternative methods to account for the sorting of students into academic tracks. Using data from an urban school district, we investigate whether including track indicators or accounting for classroom characteristics in the value-added model is sufficient to eliminate potential bias resulting from the sorting of students into academic tracks. We find that accounting for two classroom characteristics—mean classroom achievement and the standard deviation of classroom achievement—may reduce bias for middle school math teachers, whereas track indicators help for high school reading teachers. However, including both of these measures simultaneously reduces the precision of the value-added estimates in our context. In addition, we find that while these different specifications produce substantially different value-added estimates, they produce small changes in the tails of value-added distribution.

I. INTRODUCTION

Federal initiatives and private efforts—including Race to the Top, the Teacher Incentive Fund, and the Gates Foundation's Intensive Partnerships to Empower Effective Teachers—have provided incentives in recent years for states and school districts to base teacher personnel decisions in part on teachers' contributions to student achievement. Increasingly, school districts—including the District of Columbia, New York City, and Los Angeles—are using value-added models to measure teacher effectiveness. The increasing availability of longitudinal student test-score data in grades targeted by the No Child Left Behind law and a growing emphasis among states and school districts on policies that reward teachers and schools based on performance have led to substantial empirical work on value-added models.

Value added represents the unique contribution each teacher makes to student achievement, holding constant factors that are outside the teacher's control. As such, value-added models predict individual student achievement based on the student's characteristics, including baseline achievement, and compare the prediction to the actual achievement of a teacher's students. The prediction is derived by using data on other students in the state or district and represents what we would expect the student to achieve if he or she were taught by the average teacher. The difference between how a teacher's students actually performed and how they were predicted to perform is the estimate of the teacher's "value added" to student achievement.

A central concern of value-added models is the potential bias in teacher effects that arise from the sorting of students to classrooms based on characteristics related to achievement but not captured by the observable student characteristics included in the model. Students may be sorted nonrandomly across or within schools according to unobservable characteristics such as ability, motivation, or behavioral factors. One type of sorting is informal, whereby students are nonrandomly assigned to different teachers for the same course because of decisions made by either schools or students. Another type of sorting is more formal and involves nonrandom placement of students into different academic tracks—different levels of the same course or different courses for the same subject. Academic tracks may be regular or advanced, such as advanced tracks in math (pre-algebra or algebra in grade 7) or honors track in reading, and are likely to be more common in middle and high schools.

In elementary schools, value-added estimates of teachers with nonrandomly assigned students have been shown to predict sufficiently the same teachers' estimates in a subsequent year with random assignment (Kane and Staiger 2008; Kane et al. 2013). However, similar analyses have not been conducted in middle or high schools, where nonrandom sorting of students into academic tracks is likely to occur. Further, Jackson (2012) has shown that other courses a student takes with related content may be associated with the student's academic track. Thus, students may also be exposed to track-specific instruction, also potentially biasing the estimated teacher effects.

While some studies have estimated middle and high school teacher effects by using value-added models similar to the ones used for elementary school teachers (Aaronson et al. 2007; Clotfelter et al. 2010; Goldhaber et al. 2010; Koedel 2008), Jackson (2012) is the only study, to our knowledge, that explicitly addresses potential bias from student sorting to different academic tracks. Whereas standard approaches leverage variation in student achievement across academic tracks, Jackson compares teachers within the same academic track and school. This approach can mitigate bias in

value-added estimates arising from the sorting of teachers and students into academic tracks. However, because comparisons between teachers are limited to those within track and school, Jackson's strategy requires either several teachers teaching students within the same track and school or several teachers across years teaching students in the same track and school. This approach has two key limitations. First, it may not always be feasible. For example, many schools rely on a single teacher to teach all students in an academic track, or school districts may prefer to rely on value-added estimates based on a single year because such estimates represent the most recent evidence of a teacher's effectiveness. The second limitation is that relying on comparisons of teachers within each track and school restricts the ability to make meaningful comparisons between teachers across schools and tracks.

In this paper, we compare two alternative methods to account for student sorting into academic tracks in middle and high schools without the above limitations. Specifically, we account for academic tracks directly by using indicators for each academic track as in Jackson (2012). However, we define tracks as honors or advanced courses for each subject (math or reading) in each grade in the school district, whereas Jackson defined tracks as a combination of courses, level of courses, and school. Thus, our approach allows for comparison of teachers across schools, which is more meaningful for school districts interested in using value added to measure their teachers' effectiveness. But, if instruction within academic tracks varies systematically by school and thus contributes to bias, our approach will fail to take into account such variation. Alternatively, we account for measures of the baseline achievement of students' peers in the same classroom to account for sorting. If students are sorted into different academic tracks based on their prior achievement, then classroom characteristics could account for student sorting into tracks. In fact, variables related to a student's prior achievement can also proxy for unobserved abilities and are included in all the alternative value-added models we consider in this paper. Thus, measures of the baseline achievement of a student's peers in the same classroom are additional proxies for sorting into academic tracks.[1]

To compare the above two alternative methods to account for student sorting into academic tracks, we ask the following four research questions:

1. Does including track indicators affect teacher value-added estimates or their precision?
2. Does accounting for classroom characteristics affect teacher value-added estimates or their precision?
3. Can classroom characteristics substitute for track indicators in value-added models?
4. How does accounting for sorting by including both classroom characteristics and track indicators affect value-added estimates as compared to estimates based only on classroom characteristics?

We found that, under both value-added models, the inclusion of track indicators and classroom characteristics systematically changed teacher value-added estimates for middle school math, but

[1] Kane and Staiger (2008) found that the relationship between value-added estimates based on random and nonrandom assignment was stronger when the specification included average peer characteristics.

more so in the classroom characteristics model. Specifically, estimates for teachers of higher-ability students were lower in these models compared to the basic model that does not include these variables. Thus, both models addressed bias in value-added estimates in relation to higher-ability student sorting into advanced academic tracks. In addition, the precision of value-added estimates improved more under the classroom characteristics model. In high school reading, only the classroom characteristics model appeared to address bias, but it was also more imprecise.

II. HOW SORTING CAN LEAD TO BIAS IN VALUE-ADDED ESTIMATES

The direction of bias from the sorting of students to classrooms and teachers depends on the nature and underlying reasons for the sorting. For example, principals may want to group lower-ability students of a grade and assign them to the strongest teacher in that grade. Principals' evaluations of students' abilities might be based on unobservable student characteristics and thus may not be accounted for in the value-added models. In such cases, the estimated value added for the affected teacher would be lower than the teacher's value added that would be predicted according to what we observe and/or under random sorting of students. In other words, in this case, the strongest teachers would receive value-added estimates that are too low—a downward bias in the estimated value added. The opposite would be true for teachers assigned to the group of higher-ability students; the teachers' value added would be higher than would be predicted, resulting in an upward bias in the estimated value added of those teachers.

Sorting of students conditional on their covariates may be more common in middle and high schools, where students are placed in or choose academic tracks depending on their previous academic records, ability, motivation, and/or aspirations. For example, middle school students in grade 7 take general math, pre-algebra, or algebra depending on their prior achievement and or ability, which may be assessed by the grade 6 teacher. Similarly, advanced high school students in English Language Arts take Honors English. If teachers specialize in teaching students in specific tracks, then value-added estimates for teachers of students in the advanced academic tracks will be biased upward; the opposite would hold for teachers of students in regular academic tracks. In addition, Jackson (2012) points out that other courses a student takes with related content may be associated with the student's academic track. For example, students who take algebra may be more likely to take physics, which, in turn, may help them with algebra. In this example, the algebra teachers' value-added estimate would also reflect the contributions of the physics teacher and thus would be too high.

One way of accounting for sorting into academic tracks is to compare teachers who teach students in the same academic tracks. That way, teachers are compared to other teachers who teach students with the same prior achievement and/or ability, essentially eliminating the bias associated with students with varying abilities nonrandomly sorting into different academic tracks. This is the approach taken by Jackson (2012). He uses detailed course information from all public middle and high schools in North Carolina and defines academic tracks as a combination of schools, groups of courses, and levels of those courses (advanced, regular, and basic). Thus, he compares teachers of students enrolled in the same school, enrolled in the same courses, and taking the same levels of those courses.

An alternative way to account for sorting is to account directly for the classroom characteristics of a teacher in terms of prior achievement, which amounts to controlling for the characteristics of peer students in the value-added model. Even though students are sorted according to their unobserved

(to the researcher) ability, motivations, and/or aspirations, these unobserved characteristics are likely to be highly correlated with prior achievements. The prior achievements represented by a classroom of students in a value-added model, in addition to a given student's prior achievement, provide additional proxies for a student's unobserved ability. For two reasons, classroom characteristics may predict student achievement. First, the mix of students in each classroom could affect individual student achievement. Accounting for average classroom achievement allows for the possibility that students perform better when they learn among higher-performing peers, and accounting for the standard deviation of classroom achievement allows the distribution of baseline achievement in a classroom to affect student performance. The direct impacts of classroom peers on a student's achievement are sometimes called peer effects (Hoxby and Weingarth 2006; Sacerdote 2011).

Second, classroom characteristics may help in accounting for measurement error. For example, the basic value-added model includes a method to account for measurement error in the pre-tests, but the method may not account for all measurement error. Inclusion of the class average pre-test score could address this additional measurement error if students' true achievement levels are related to their peers' test scores. Through this channel, the direction of the change could theoretically work in either direction. For example, students with higher-achieving peers could have higher or lower predicted post-test scores, depending on the direction of the measurement error in an individual student's test score.[2]

III. EMPIRICAL APPROACH

A. VALUE-ADDED MODELS

To investigate how accounting for student sorting into academic tracks by using within-track variation and variation in classroom characteristics affected value-added estimates of teachers, we estimate four value-added models: (1) a basic model that accounts for individual student characteristics but does not account for student sorting into academic tracks; (2) the basic model augmented with track indicators for each student to allow within-track comparison; (3) the basic model augmented with classroom characteristics; and (4) the basic model augmented with both track indicators and classroom characteristics (the "full model"). We estimate all four models separately by subject (math and reading).

Our analysis is based on students in grades 7 and 8 for math and in grades 9 and 10 for reading during the 2011–2012 school year, with both pre-test and post-test scores.[3] We obtained data from the District of Columbia Public Schools (DCPS) and the Office of the State Superintendent of Education of the District of Columbia (OSSE). The data contain information on students' test scores, demographic characteristics, and course enrollment. Our analysis focuses on grades 7 through 10 because these are the grades in which we can identify students' tracks based on their

[2] The measurement error correction method in the basic model assumes that, whether a student receives an average, high, or low score on an assessment, the score would be measured with equal accuracy. However, as is typical for standardized tests, DC CAS scores are more accurate for students scoring in the middle than at the top or bottom of the scale (CTB/McGraw Hill 2011). In addition, the method assumes that the relationship between pre-test and post-test scores is linear. If not, one can construct examples in which the method over- or under-corrects for measurement error.

[3] We subsequently refer to the 2011–2012 school year as the 2012 school year.

course enrollment. Analysis in grades 9 and 10 is limited to reading achievement because students in these grades are not tested in math. Analysis in grades 7 and 8 is limited to math achievement because the course enrollment data in these grades do not allow us to distinguish tracks for reading courses. The data also allow students to be linked to their teachers and classrooms, permitting us to identify which teachers teach specific courses in math and reading and to calculate measures of students' peers within the same classroom.

In DC, students are tested in math and reading by using the District of Columbia Comprehensive Assessment System (DC CAS) tests. To enable us to compare students across grades, we standardize student test scores within subject, year, and grade to have a mean of zero and a standard deviation of one. We exclude students who are repeating the grade so that, in each grade, we compare only students who completed the same examination.

We estimate the following basic value-added model for a given teacher *t* of student *i* in grade *g*:

$$Y_{tig} = \lambda_g S_{ig} + \omega_g O_{ig} + \boldsymbol{\beta}' \mathbf{X}_i + \boldsymbol{\delta}' \mathbf{T}_{tig} + \varepsilon_{tig}, \quad (1)$$

where Y_{tig} is the post-test score for student *i* in grade *g* and S_{ig} is the same-subject pre-test score for student *i* during the previous year. The variable O_{ig} denotes the pre-test score in the opposite subject. Thus, when we estimate teacher effectiveness in math, *S* represents math tests, with *O* representing reading tests and vice versa. The pre-test scores capture prior inputs into student achievement. The vector $\mathbf{X}_i$ denotes the control variables for individual student background characteristics, specifically, indicators for free-lunch eligibility, reduced-price lunch eligibility, special education status, and race/ethnicity.

The vector $\mathbf{T}_{tig}$ includes an indicator variable for each teacher in grade *g*. A student contributes one observation to the model for each teacher to whom the student is linked. The contribution is based on a roster confirmation process that allows teachers to indicate whether and for how long they have taught the students assigned to them according to their administrative rosters and to add students to their rosters. Students are weighted in the regression according to their "dosage," which indicates the amount of time the teacher taught the student.[4] The coefficient $\boldsymbol{\delta}$ represents the teacher effect. Finally, ε_{tig} is the random error term.

The value-added model in (1) is estimated in two regression steps and two subsequent steps to adjust estimates for comparability across grades and to account for imprecise estimates:

- **Measurement error correction.** Given that measurement error in the pre-test scores attenuates the estimated relationship between the pre-test scores and the

[4] To estimate the effectiveness of teachers who share students, we use a technique called the Full Roster Plus Method that attributes equal credit to teachers of shared students. In this method, each student contributes one observation to the model for each teacher to whom he or she is linked, and students are weighted according to the dosage they contribute. Then, we create additional observations to equalize the dosage that each student contributes to the model. The proportion of dosage that each student contributes to a teacher remains unchanged because the additional observations are linked to a distinct set of "shadow" teacher indicators. Coefficient estimates for the shadow teachers are discarded.

post-test scores, we adjust for measurement error by using an errors-in-variables correction (eivreg in Stata) that relies on published information on the test-retest reliability of the DC CAS. We use errors-in-variables regression to regress the post-test score on the pre-test scores, student background characteristics, grade indicators, and teacher indicators. Because the errors-in-variables regression does not allow for standard errors to be clustered by student, we use this step to obtain adjusted gain scores that are equal to the post-test scores minus the predicted post-test scores where the predictions are based on the pre-test. We then use the adjusted gains to obtain the teacher effects in the next step.

- **Main regression.** We estimate teacher effects by regressing the adjusted gain scores from the first step on student background characteristics, grade indicators, and teacher indicators and then by clustering standard errors by student. The teacher value-added estimates are the coefficients on the teacher indicators in this regression, $\boldsymbol{\delta}$, with their variance given by the squared standard errors of the value-added estimates.

- **Combine teachers' estimates across grades.** We combine teachers' estimates into a single value-added estimate when the teacher taught students in several grades. We make teachers' estimates comparable across grades and then combine them by using a weighted average. We standardize the estimated regression coefficients so that the mean and standard deviation[5] of the distribution of teacher estimates are the same across grades. When combining the standardized estimates, we base the weights on the number of students taught by each teacher to reduce the influence of imprecise estimates obtained from teacher-grade combinations with few students.

- **Empirical Bayes procedure.** We use an Empirical Bayes (EB) procedure as outlined in Morris (1983) to generate the shrinkage estimates, which are approximately a precision-weighted average of the teacher's initial estimated effect and the overall mean of all estimated teacher effects.[6] We calculate the standard error for each shrinkage estimate by 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 re-center the shrinkage estimates to have a mean of zero.

Next, to compare teachers with students in the same academic tracks, we augment the basic model (1) as follows:

$$Y_{tiga} = \lambda_g S_{ig} + \omega_g O_{ig} + \boldsymbol{\beta}' \boldsymbol{X}_i + \boldsymbol{\delta}' \boldsymbol{T}_{tiga} + \boldsymbol{\gamma}' \boldsymbol{I}_{iga} + \varepsilon_{tiga}, \quad (2)$$

[5] The grade-specific estimates are standardized to account for sampling error. Consequently, estimates in grades with less precise estimates also receive less weight in the combined value-added score.

[6] In Morris (1983), the EB estimate does not exactly equal the precision-weighted average of the two values because of a correction for bias. The 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 the correction from the description given here.

where the subscript *a* denotes the academic track of student *i*, and $\boldsymbol{I_{iga}}$ is a vector of academic track indicators. In DC, grade 7 students in math take either general math or pre-algebra or algebra, and, in grade 8, they take either pre-algebra or algebra or geometry. For estimating the effects of math teachers, the vector $\boldsymbol{I_{iga}}$ includes indicators for pre-algebra and algebra in grade 7 and for algebra and geometry in grade 8. Advanced high school students in grades 9 and 10 take Honors English classes, and the vector $\boldsymbol{I_{iga}}$ includes indicators for honors when estimating the effects of reading teachers.

Our third model augments the basic model (1) by including classroom characteristics of teachers. For a given teacher *t* and student *i* in classroom *c*, the regression is expressed as:

$$Y_{tigc} = \lambda_g S_{ig} + \omega_g O_{ig} + \boldsymbol{\beta}' \boldsymbol{X_i} + \boldsymbol{\delta}' \boldsymbol{T_{tig}} + \boldsymbol{\pi}' \boldsymbol{C_{tic}} + \varepsilon_{tigc}, \quad (3)$$

where the vector $\mathbf{C}_{tic}$ represents the characteristics of classroom *c*. The variables included in the vector are the mean classroom pre-test scores and the classroom standard deviation of pre-test scores. We use roster-confirmed data on classroom assignment and teacher-student links to calculate classroom characteristics for middle school math teachers and high school reading teachers. The data allow us to use several sections within a year to estimate the association between classroom characteristics and student achievement.[7]

Finally, for our fourth model, we augment the basic value-added model (1) with the vectors on academic track indicators from model (2) and classroom characteristics from model (3):

$$Y_{tigac} = \lambda_g S_{ig} + \omega_g O_{ig} + \boldsymbol{\beta}' \boldsymbol{X_i} + \boldsymbol{\delta}' \boldsymbol{T_{tiga}} + \boldsymbol{\gamma}' \boldsymbol{I_{iga}} + \boldsymbol{\pi}' \boldsymbol{C_{tic}} + \varepsilon_{tigac}, \quad (4)$$

B. MODEL COMPARISON APPROACH

In our analysis, we compare estimates from models (2), (3), and (4) to estimates from the basic model (1). We also compare models (2) and (3) to assess whether they perform differently in accounting for bias associated with student sorting into different academic tracks. For each of the comparisons, we use three criteria by which to judge the models:

[7] The unit of observation in the model is a teacher-classroom-student combination. To allow for the inclusion of classroom characteristics, we included student-teacher observations for each classroom shared by the student-teacher pair during the year. We use an additional regression step to estimate the model in equation (3) to leverage variation in classroom characteristics across several sections for each teacher across grades. This step precedes the two regression steps in the basic model, includes the classroom characteristics as additional regressors, and uses teacher indicators pooled across grades rather than the within-grade indicators in the subsequent regression steps. The additional step includes all other regressors included in the first-stage regression in the basic model as well as the same errors-in-variables measurement error correction. We calculate an adjusted gain by subtracting the contributions of the classroom characteristics and use the adjusted gain as the dependent variable in the subsequent regression step.

- We measure the magnitude of the changes in value-added estimates between the two models. We calculate the magnitude both as a correlation between the estimates and as the average absolute size of differences in value-added estimates for individual teachers.

- We compare the relationship between average student achievement and the difference in value-added estimates between the two models in a comparison. As described in Section II, we expect to observe upward bias in the estimated value added of teachers of advanced-track students. Thus, the direction of change in value-added estimates in each pair of models compared to average student achievement allows us to examine if one model performs better in terms of reducing bias associated with student sorting into academic tracks. For example, if teachers of higher-achieving students are estimated to have lower value added under model (1) compared to the basic model, then model (1) performs better in reducing the bias associated with student sorting into academic tracks.[8]

- We compare the precision of estimates from each model. and calculate the change in precision as a percentage increase or decrease in the average confidence interval of teachers' value-added estimates.

IV. DATA

We use data on the same teachers and students in DCPS in all four value-added models. We calculate value added for teachers of math in grades 7 and 8 and for teachers of reading in grades 9 and 10 for the 2011–2012 school year. In Table 1, we present student background characteristics separately for the middle school math students and high school reading students. In both student samples, about half of the students are male, and about three-quarters are African American. At the middle school level, 68 percent qualify for free lunch and 5.5 percent for reduced-price lunch. The corresponding numbers at the high school level are 58.5 and about 6 percent. DCPS has a relatively low percentage of students deemed English language learners or with limited English language proficiency: about 7 percent at the middle school level and about 6 percent at the high school level. More than 12 percent at the middle school level and 11 percent at the high school level have special education status.

DCPS students go into different academic tracks for both math and reading. At the middle school level, advanced math students take pre-algebra or algebra in grade 7 and algebra or geometry in grade 8. Regular math students take general math in grade 7 and pre-algebra in grade 8. Thirty-nine percent of DCPS grade 7 math students and 43 percent of grade 8 math students are in advanced tracks (Table 1). At the high school level for reading, advanced students take Honors English classes in both grades 9 and 10; in fact, 18.5 and 19.2 percent, respectively, of students in grades 9 and 10 are Honors English (Table 1).

[8] A finding of no change in the relationship shows that one model does not reduce bias related to average student achievement, but the finding does not rule out the possibility that the model reduces another source of bias.

In Table 2a and 2b, we present a list of all middle school math and high school reading courses offered in DCPS and the percentage of students enrolled in the courses by grade. The math courses are as described above, with the tracks well defined. No students take geometry in grade 7 and general math in grade 8. The list of reading courses at the high school level is longer, and students may take more than one reading course. We define students as advance-tracked if they are enrolled in an Honors English class. DCPS high schools offer one Honors English class in each of grades 9 and 10: Honors English I for grade 9 students and Honors English II for grade 10 students. We found no cross-grade enrollment (grade 9 students taking Honors English II or grade 10 students taking Honors English I).

Some teachers in DCPS middle schools specialize in teaching different math courses. About 18 percent of math teachers teach only grade 7 general math, 2 percent teach only grade 7 pre-algebra, 11 percent teach only grade 8 pre-algebra, and 3 percent teach only grade 8 algebra. None of the math teachers teaches only grade 7 algebra or only grade 8 geometry. A large percentage of middle school math teachers, 65.6 percent, teach more than one math course (Table 3a). At the high school level, 11 to 13 percent of the grade 9 and 10 reading teachers teach both Honors and non–Honors English classes while 3 percent of grade 10 teachers teach only Honors English. None of the grade 9 teachers teaches only honors English (Table 3b).

V. RESULTS

We quantify the extent to which our two methods of accounting for student sorting into different academic tracks affects the value-added estimates of teachers. We measure differences in terms of standard deviations of value-added estimates. To facilitate comparisons across models, we scale estimates from each model to have the same standard deviation as the basic model by re-centering the distribution of final value-added estimates under each model to have a zero mean and a standard deviation of one for each subject.[9]

A. MODEL INCLUDING TRACK INDICATORS COMPARED TO BASIC MODEL

As expected and in line with Jackson (2012), we find that the inclusion of track indicators in the value-added model results in substantive changes to the value-added estimates of individual teachers from estimates based on the basic model, especially in math. The correlation between the basic model and the model with track indicators is 0.92 in math and 0.96 in reading, shown in the 1st row and columns 1 and 2 in the top panel of Table 4. However, even results correlated above 0.90 may have substantive implications for the estimates of individual teachers. On average, a teacher's value-added estimate changes by 33 percent of a standard deviation of teacher value added in math and 17 percent of a standard deviation in reading, shown in the second row of the table corresponding to the first two columns. Alternatively, the average change in math represents a teacher whose value-added estimate moves from the 50th to the 63th percentile; for reading, the average change moves a teacher from the 50th to the 57th percentile. The largest changes in teachers' estimates are equal to

[9] We do not distinguish between signal and noise variance when we standardize. Consequently, comparisons of the magnitude of estimates between models tend to show larger differences when one of the models produces less precise estimates. The role of precision in the comparisons is reduced because we use Empirical Bayes shrinkage estimates for all analyses.

one standard deviation of math value added (either above or below the original estimate) while changes in reading value added exceed 0.8 standard deviation for some teachers (not shown).

Students in the advanced tracks in math, pre-algebra, and algebra in grade 7 and algebra and geometry in grade 8 may tend to have higher-than-predicted post-test scores based on their pre-test scores and background characteristics alone, as indicated by the positive coefficients of these indicators (Table 5a, Column 1). The same holds for Honors English students in grade 10 but not in grade 9, where students in the advanced track are likely to have lower post-test scores, conditional on their pre-test scores and background characteristics (Table 5b, Column 1). However, the indicator for grade 9 is not statistically significant. When we include track indicators, we see a negative correlation between value-added estimates and average pre-test scores relative to the basic model with no track indicators. In Figures 1a and 1b, we plot the change in each teacher's value-added estimate from the basic model against the average pre-test score of the teacher's students in math and reading, respectively. Teachers with a positive difference on the vertical axis receive higher value-added estimates when track indicators are included. The linear fit through the relationship in Figure 1a is downward-sloping—teachers with higher-achieving students receive lower value-added estimates in math from the model that includes track indicators relative to the basic model. This finding is consistent with our expectation that student sorting into academic tracks results in upward bias in value-added estimates for teachers of higher-ability students; thus, when track indicators are included to account for this sorting bias, the teachers' value added declines. Changes in value-added estimates are not as strongly related to average pre-test scores for reading; the relationship for reading in Figure 1b is almost flat.

Finally, including track indicators in the value-added model increases the average precision of teachers' estimates in math but not in reading. On average, the inclusion of track indicators reduces the confidence interval—or equivalently, increases precision—for math teachers' value-added estimates by 2.7 percent. However, there is no precision gain or loss for reading teachers' value-added estimates (third row and columns 1 and 2 in the top panel of Table 4).

B. MODEL INCLUDING CLASSROOM CHARACTERISTICS COMPARED TO BASIC MODEL

Next, we examine changes in value-added estimates when we account for classroom characteristics (but not track indicators) relative to the basic model. We find that accounting for classroom characteristics also leads to substantive changes in the estimates of individual teachers from estimates based on the basic model in both math and reading. The correlation between the basic and classroom characteristics models is 0.98 for both subjects, shown in the first row and columns 3 and 4 of the top panel of Table 4. On average, a teacher's value-added estimate changes by 17 percent of a standard deviation of teacher value added in math, representing a move from the 50^{th} to 57^{th} percentile, and by 15 percent of a standard deviation in reading, representing a move from the 50^{th} to 56^{th} percentile. The largest changes in teachers' estimates exceed 0.5 standard deviation of math value added (either above or below the original estimate) while changes in reading value added never exceed 0.5 standard deviation.

Average class pre-test scores are significantly related to student achievement in the value-added model. In Tables 5a and 5b, we show that a one (student-level) standard deviation increase in class average pre-test scores positively affects the predicted post-test score for a student in both math and reading. Based on these estimates, a student with higher-achieving peers in the same class is

expected to achieve higher scores in both math and reading. However, a student in a class with more diverse pre-test scores achieves higher scores in math but lower in reading, although none of the estimated relationships with classroom dispersion is statistically significant. When accounting for classroom characteristics relative to the basic model, we observe a negative correlation between value-added estimates and average pre-test scores for both math and reading teachers. In Figures 2a and 2b, we plot the change in each teacher's value-added estimate from the basic model against the average pre-test score of the teacher's students in math and reading. The downward-sloping linear fit in both figures indicates that teachers with higher-achieving students receive lower value-added estimates and that the relationship is stronger for math as illustrated by the steeper linear fit for math teachers' changes in value added in Figure 2a. Again, the finding is consistent with our expectation that student sorting into academic tracks results in upward bias in value-added estimates for teachers of higher- ability students.

Accounting for classroom characteristics increases the average precision of math teachers', but not reading teachers' estimates. On average, accounting for classroom characteristics reduces the average confidence interval for math teachers by 9.4 percent and increases the average confidence interval for reading teachers by 9.6 percent (third row corresponding to columns 3 and 4 in the top panel of Table 4).

C. COMPARING MODELS INCLUDING CLASSROOM CHARACTERISTICS WITH MODELS INCLUDING TRACK INDICATORS

We compare the model that includes track indicators and the model that includes classroom characteristics to examine whether one can substitute for the other in addressing bias from sorting of students into academic tracks. If classroom characteristics can substitute for track indicators and vice versa, then we would expect to see only small differences between teachers' estimates based on these two models.

In fact, these two models produce different results. The correlations of value-added estimates between the two models are comparatively lower, 0.93 in math and 0.94 in reading. The average differences in teachers' value-added estimates are 30 percent of a standard deviation in math teacher value added and 23 percent of a standard deviation in reading teacher value added (columns 3 and 4 in the bottom panel of Table 4). In Figures 3a and 3b, we examine the results for teachers under the two models as a function of the pre-test scores of their students. If the track indicator model and classroom characteristic model produce similar results, then the linear fit through each of the relationships would appear as a horizontal line. Instead, the downward slope in Figures 3a and 3b indicates that the models perform differently. More specifically, teachers with higher-achieving students receive smaller value added in the classroom characteristics model for both math and reading than in the model with track indicators. However, while average precision for math teachers' estimates was better under the classroom characteristics model, it was better under the track indicator model for reading teacher's estimates. Compared to the track indicator model, confidence interval of value-added estimates reduced by 6.8 percent in math and increased by 9.6 percent in reading under the classroom characteristics model (third row corresponding to columns 3 and 4 in the bottom panel of Table 4)

D. THE FULL MODEL COMPARED TO THE CLASSROOM CHARACTERISTICS MODEL

In our final "full model", we include both track indicators and classroom characteristics to determine if including both measures to account for student sorting into academic tracks improves value-added estimates by reducing bias compared to the classroom characteristics model. The correlations of value-added estimates between the full model and the classroom characteristics model are as low as the correlations between the track indicator model and the classroom characteristics model, 0.94 in both math and reading. The average differences in value-added estimates are 0.27 standard deviation in math and 0.24 standard deviation in reading (columns 5 and 6 in the middle panel of Table 4).

Compared to the classroom characteristics model, the full model changes teachers' value-added estimates in math in a way that is not consistent with our expectation that student sorting into academic tracks results in upward bias in value-added estimates for teachers of higher-ability students. Teachers with higher-achieving students receive higher value-added estimates under this model, as evident from the upward-sloping linear fit in Figures 4a. However, it was opposite for reading—the linear fit was upward sloping (Figure 4b). The precision of this model is worse compared to the classroom characteristics model, confidence interval for value-added estimates decreases by 6.9 percent in math and 7.8 percent in reading (columns 5 and 6 in the middle panel of Table 4). Thus, incorporating track indicators in addition to classroom characteristics to account for bias from sorting of students into academic tracks leads to less precise estimates in both math and reading. This finding could indicate that the identification of the relationships between classroom characteristics and student achievement requires variation within- as well as between tracks.

VI. DISCUSSION

To account for potential bias in teacher value-added estimates from sorting of students into academic tracks, we tested two alternative approaches that, respectively, include track indicators and each student's classroom characteristics in the value-added model. In most cases, both track indicators and classroom characteristics (or the combination of the two) appeared to be able to address bias associated with sorting of students into academic tracks. However, in some cases, there was a loss of precision in the estimates, most notably when using the full model to estimate results in reading.

The two specifications we tested rely on different identification strategies that rely considerably on the context of the school districts using value-added estimates to evaluate their teachers. The identification under the classroom characteristics model is derived from variation in teachers' classrooms, making it important to have a considerable number of teachers who teach in more than one classroom that vary by average classroom characteristics. Similarly, identification in the track indicator model is derived from teachers who teach students in more than one track, making it important to have a considerable number of such teachers. Also, the relative importance of these specifications depends largely on how value-added estimates are used. The consequences for teachers evaluated by a value-added system are typically for those at the tails of the distribution. Thus a lot of fluctuations in the value-added estimates in the middle of the distribution across different specifications may not be relevant for practical purposes.

To evaluate the consequences for teachers at the tails of the distribution, we examined the top 10 and the bottom 10 teachers are the same across different models. This represents 16 percents of the

math teachers we evaluated at the middle school level and 15 percent of the reading teachers we evaluated at the high school level. Table 6 presents the number of top 10 and bottom 10 math and reading teachers under different models, who were in the top 10 or bottom 10 when we used the basic model. In math, all top 10 teachers in the basic model remains in the top 10 and 7 of the bottom 10 teachers remain in the bottom 10 no matter what model we use. Thus, the model specifications including either track indicators or classroom characteristics or both is only relevant for the 10 middle school math teachers at the bottom of the distribution and all specifications result in the same teachers remaining at the bottom.

Different model specifications make more changes to the top and bottom 10 for high school reading teachers. Only four of the top 10 teachers under the basic model remain in the top 10 when we use the classroom characteristics model and the full model (same 4 teachers), while 7 remain when the track indicators model is used. Three of the 4 teachers who remain in the top 10 under the two other models are part of the 7 who remain in the top 10 under the track indicator model. Comparatively, the change in the bottom 10 is stable across models—8 of the 10 bottom teachers under the basic model remain in bottom 10 under all three alternative models and they are the same 8 teachers across models. Thus, while the track indicator, classroom characteristics, and the full model including both produce substantially different value-added estimates, they produce small changes in the bottom tail of the value-added distribution.

VII. CONCLUSIONS

Value-added models perform differently for math and reading when sorting into academic tracks could potentially introduce bias at the middle and high school levels. We find that, in our context, a classroom characteristics model provides a stronger correction for this potential problem in middle school math and that high school reading. However, these models make small changes at both the top and the bottom of the distribution, which is possibly more relevant for school districts using value added to evaluate their teachers. Also, it is important to note that there are no clear tracks in our data for reading at the middle school level. Thus, the only feasible approach to address bias from student sorting at the middle school level would be to include classroom characteristics in the value-added models.

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.

Clotfelter, Charles, Helen Ladd, and Jacob Vigdor. "Teacher Credentials and Student Achievement in High-School: A Cross-Subject Analysis with Student Fixed Effects." *Journal of Human Resources,* vol. 45, no. 3, 2010, pp. 655-681.

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

Goldhaber, D., P. Goldschmidt, P. Sylling, and F. Tseng. "Assessing Value-Added Model Estimates of Teacher Contributions to Student Learning at the High School Level." Working Paper 2011-4. Seattle, WA: Center for Education Data & Research (CEDR), University of Washington, 2011.

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

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.

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

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, Douglas O. Staiger. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." MET Project Research Paper. Bill & Melinda Gates Foundation. 2013.

Koedel, Cory. "An Empirical Analysis of Teacher Spillover Effects in Secondary School." *Economics of Education Review*, vol. 28, no. 6, 2009, pp. 682-692.

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

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

Table 1. Summary Statistics: Student Characteristics

	Math		Reading	
	Mean	Standard Deviation	Mean	Standard Deviation
Demographic				
Male	0.488	0.500	0.472	0.499
White (Non-Hispanic), Asian, Pacific Islander	0.102	0.302	0.105	0.306
African-American	0.738	0.440	0.747	0.435
White (Hispanic)	0.160	0.366	0.148	0.355
Free lunch eligible	0.684	0.455	0.585	0.489
Reduced-price lunch eligible	0.055	0.217	0.059	0.231
Academic				
ELL/LEP	0.069	0.253	0.058	0.234
Special ED— learning disability	0.088	0.283	0.083	0.276
Special ED—other	0.037	0.188	0.029	0.168
Advanced Track				
Grade 7 pre-algebra or algebra	0.39	0.488	n.a.	n.a.
Grade 8 algebra or geometry	0.453	0.498	n.a.	n.a.
Grade 9 honors English	n.a.	n.a.	0.185	0.386
Grade 10 honors English	n.a.	n.a.	0.192	0.394

Source: Administrative data from OSSE and DCPS.

Note: Math students are in grades 7 and 8; reading students are in grades 9 and 10.

n.a. = not applicable

Table 2a. List of Math Courses for Students in Grades 7 and 8

	Grade 7		Grade 8	
Course	Honors (Yes/No)	% of Students	Honors (Yes/No)	% of Students
General math	No	60.9	n.a.	n.a.
Pre-algebra	Yes	31.3	No	54.7
Algebra	Yes	7.7	Yes	38.9
Geometry	n.a.	n.a.	Yes	6.4
Total students		100.0 (N = 1,733)		100.00 (N = 1,921)

Source: Administrative data from OSSE and DCPS.

Notes: n.a. = not applicable

Table 2b. List of English Language Arts Courses for Students in Grades 9 and 10

	Grade 9		Grade 10	
Course	Honors (Yes/No)	% of Students	Honors (Yes/No)	% of Students
English I	No	73.0	No	0.0
English II	No	0.0	No	72.4
Honors English I	Yes	18.5	Yes	0.0
Honors English II	Yes	0.0	Yes	19.2
Reading workshop	No	4.8	No	3.6
Extended literacy	No	15.5	No	16.9
ELA strategies	No	0.0	No	4.7
English and humanities	No	7.5	No	7.8
Total students		1,593		1,253

Source: Administrative data from OSSE and DCPS.

Notes: Percentages in column three and five add up to more than 100 because students can take more than one English class.

Table 3a. Percentage of Teachers Teaching Different Grade 7 and 8 Math Courses

	Grade 7	Grade 8
General math only	18.0	n.a.
Pre-algebra only	1.6	11.5
Algebra only	0.0	3.3
Geometry only	n.a.	0.0
Multiple courses	65.6	
Total number of teachers	61	

Source: Administrative data from OSSE and DCPS.

Notes: n.a. = not applicable

Table 3b. Percentage of Teachers Teaching Honors and Non-Honors Reading Courses in Grades 9 and 10

	Grade 9	Grade 10
Honors	0.0	2.9
Non-honors	35.7	41.4
Both	11.4	12.9
Total number of teachers	33	40

Source: Administrative data from OSSE and DCPS.

Table 4. Comparison Between Different Value-Added Models

	Track Indicator Model		Classroom Characteristics Model		Tracking and Classroom Characteristics Model	
	Math	Reading	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)	(5)	(6)
Comparison with Basic Model						
Correlation	0.92	0.96	0.98	0.98	0.93	0.88
Average size of differences in value-added estimates	0.33	0.17	0.17	0.15	0.31	0.37
Percent change in confidence intervals	-2.7	0.0	-9.4	9.6	-3.1	18.2
Comparison with Classroom Characteristics Model						
Correlation	0.93	0.94	n.a.	n.a.	0.94	0.94
Average size of differences in value-added estimates	0.30	0.23	n.a.	n.a.	0.27	0.24
Percent change in confidence intervals	7.3	-8.7	n.a.	n.a.	6.9	7.8
Comparison with Track Indicator Model						
Correlation	n.a.	n.a.	0.93	0.94	0.99	0.91
Average size of differences in value-added estimates	n.a.	n.a.	0.30	0.23	0.03	0.34
Percent change in confidence intervals	n.a.	n.a.	-6.8	9.6	-0.4	18.1
Number of teachers	62	65	62	65	62	65

Source: Administrative data from OSSE and DCPS.

Note: Math students are in grades 7 and 8; reading students are in grades 9 and 10.

n.a. = not applicable

Table 5a. Coefficients on Track Indicators and Classroom Characteristics from Math Value-Added Models

	Track Indicators	Classroom Characteristics	Full Model
Variable	(1)	(2)	(3)
Track Indicators			
Grade 7 pre-algebra	0.11**	n.a.	0.098*
	(0.052)		(0.052)
Grade 7 algebra	0.417***	n.a.	0.388***
	(0.076)		(0.076)
Grade 8 algebra	0.185***	n.a.	0.174***
	(0.051)		(0.051)
Grade 8 geometry	0.247***	n.a.	0.218**
	(0.085)		(0.085)
Classroom Characteristics			
Average pre-test score	n.a.	0.104***	0.015
		(0.017)	(0.025)
Standard deviation of pre-test score	n.a.	0.061	0.009
		(0.04)	(0.042)

Source: Administrative data from OSSE and DCPS.

Notes: Standard errors are in parentheses.

*** Significant at the 1% level; ** significant at the 5% level; * significant at the 10% level.

The value-added models include 3655 students in math.

n.a. = not applicable

Table 5b. Coefficients on Track Indicators and Classroom Characteristics from Reading Value-Added Models

	Track Indicators	Classroom Characteristics	Full Model
Variable	(1)	(2)	(3)
Track Indicators			
Grade 9 honors	-0.039	n.a.	-0.167***
	(0.045)		(0.045)
Grade 10 honors	0.254***	n.a.	0.177***
	(0.066)		(0.066)
Classroom Characteristics			
Average pre-test score	n.a.	0.066**	0.155***
		(0.031)	(0.047)
Standard deviation of pre-test score	n.a.	-0.011	0.042
		(0.052)	(0.053)

Source: Administrative data from OSSE and DCPS.

Notes: Standard errors are in parentheses.

*** Significant at the 1% level; ** significant at the 5% level; * significant at the 10% level.

The value-added models include 2846 students in reading.

n.a. = not applicable

Table 6. Number of Teachers at the Top and Bottom of the Distribution Under Different Value-Added Models Compared to the Basic Model

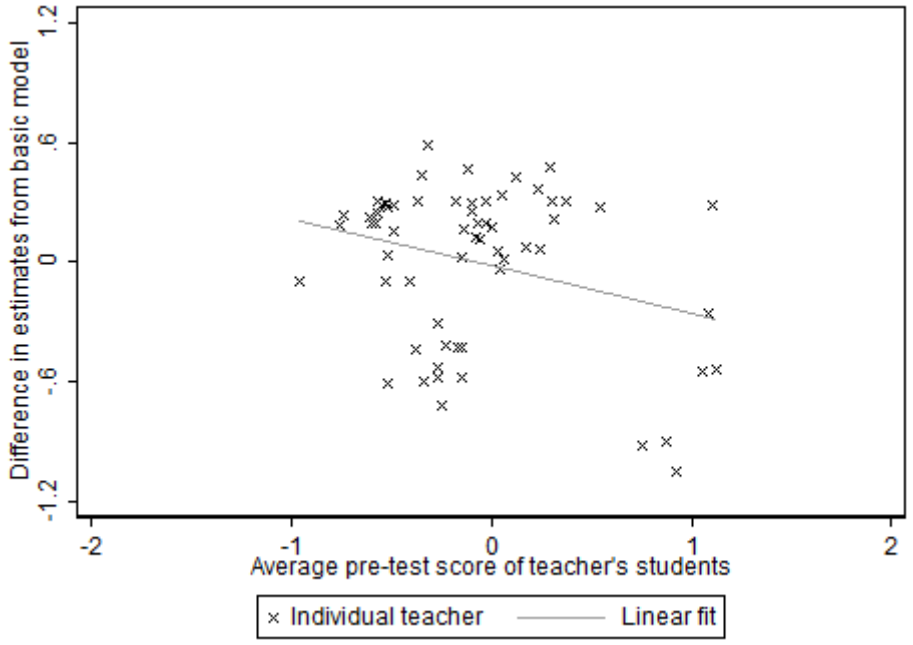
	Track Indicator Model		Classroom Characteristics Model		Tracking and Classroom Characteristics Model	
	Math	Reading	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)	(5)	(6)
Top 10	10	7	10	4	10	4
Bottom 10	7	8	7	8	7	8

Source: Administrative data from OSSE and DCPS**.**

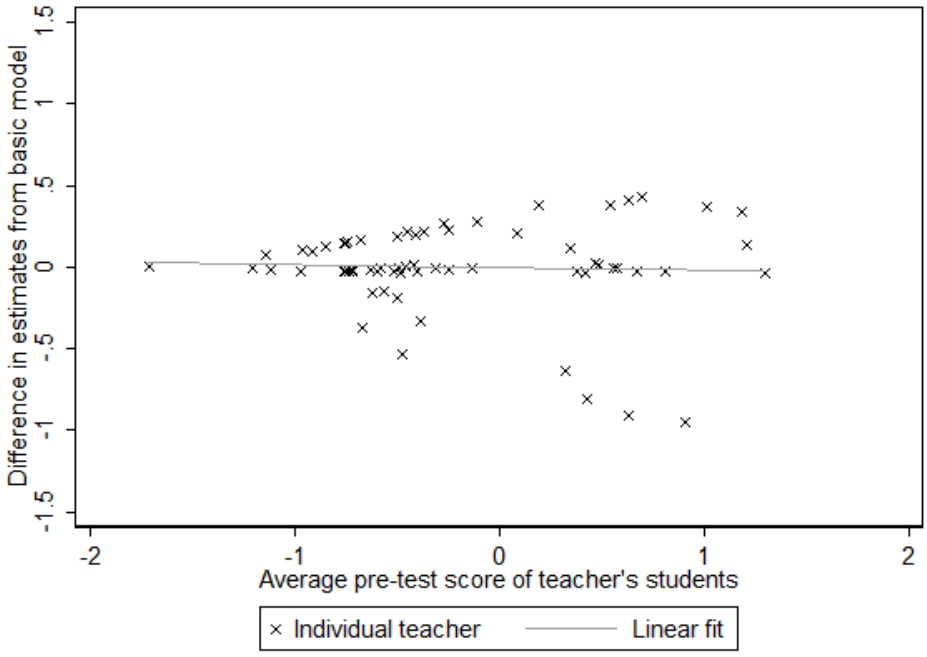
Notes: Math teachers teach students in grades 7 and 8; reading teachers teach students in grades 9 and 10.

Figure 1. Changes in the Relationship with Student Achievement When Track Indicators Are Included

(a) Math



(b) Reading

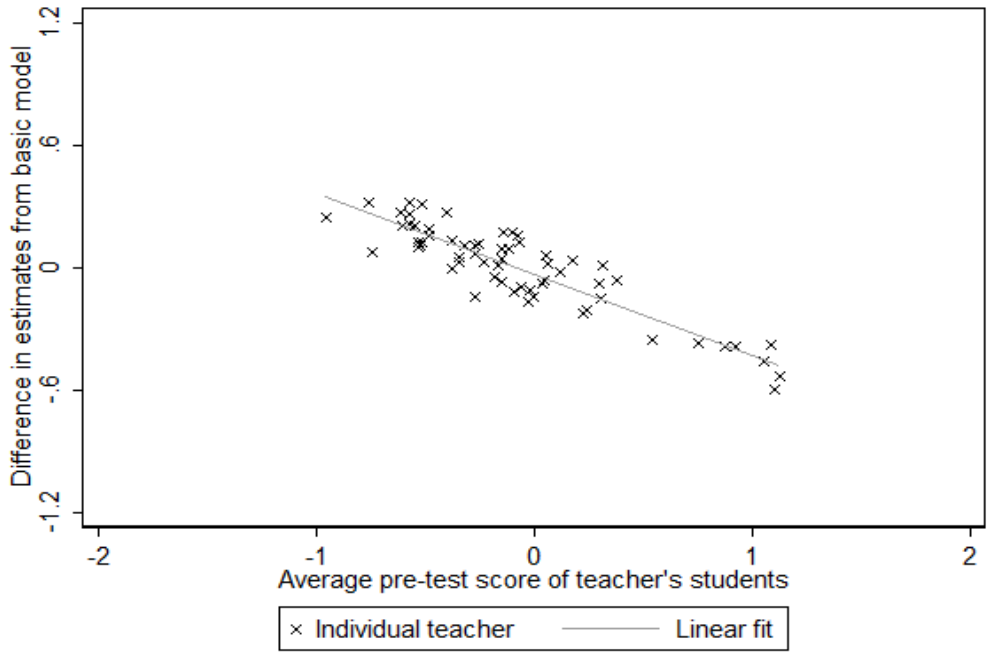


Source: Administrative data from OSSE and DCPS.

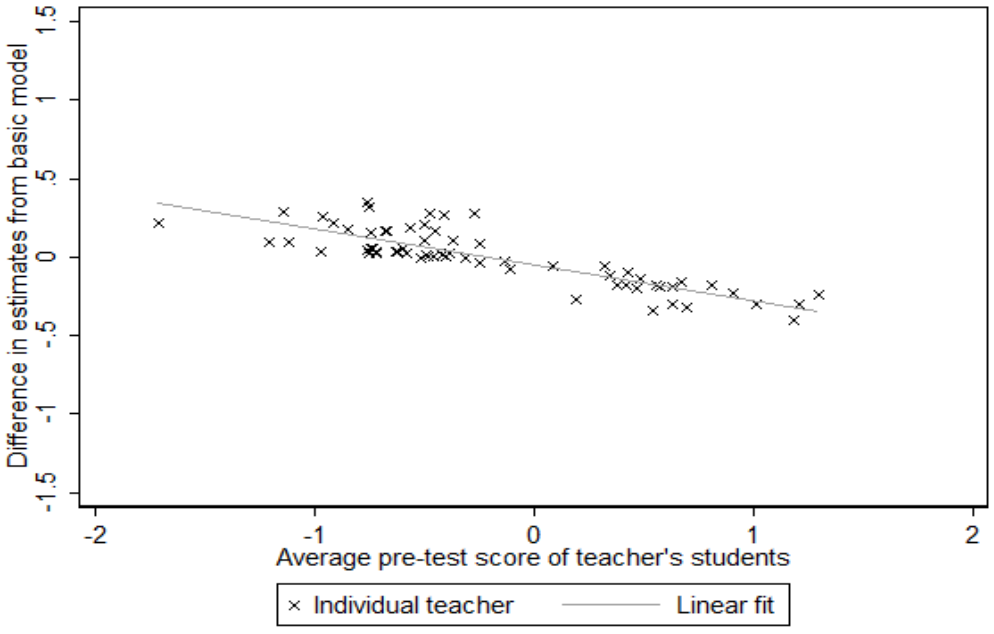
Notes: The figures include the 62 teachers with math value added and 65 teachers with reading value added. Teachers with positive differences have larger value-added estimates in the track indicator model.

Figure 2. Changes in the Relationship with Student Achievement When Accounting for Classroom Characteristics

(a) Math



(b) Reading



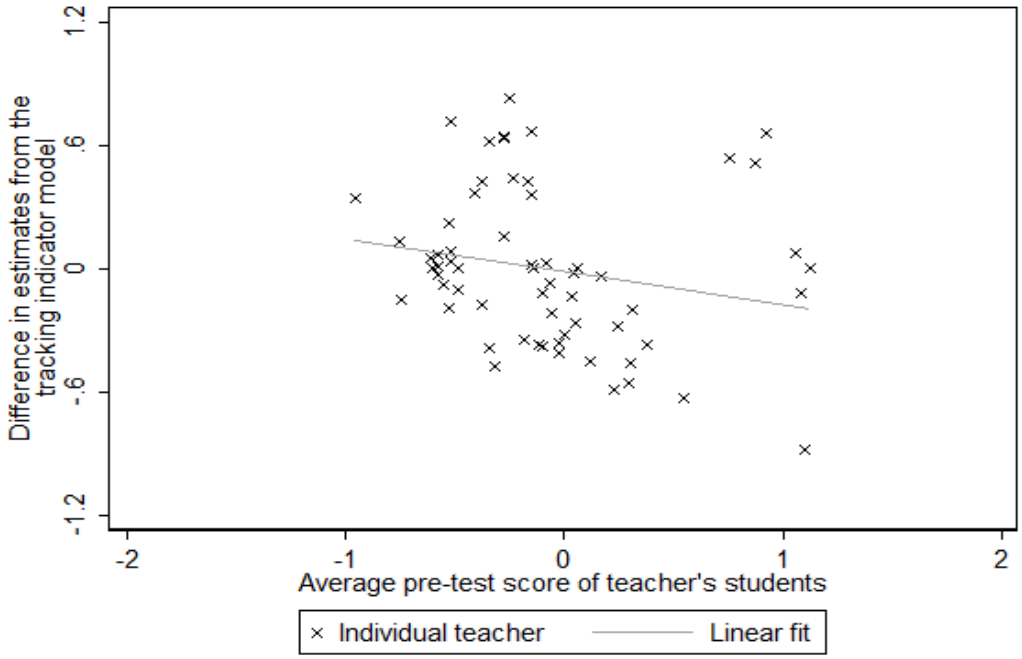
Source: Administrative data from OSSE and DCPS.

Notes: The figures include 62 teachers with math value added and 65 teachers with reading value added.

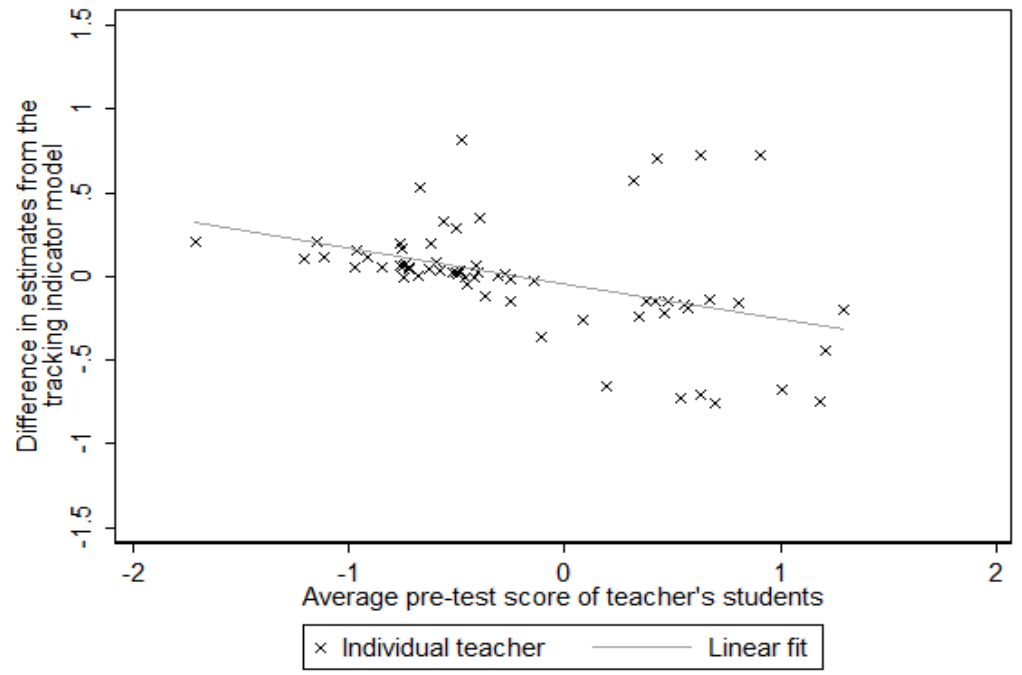
Teachers with positive differences have larger value-added estimates in the classroom characteristics model.

Figure 3. Changes in the Relationship with Student Achievement When Accounting for Classroom Characteristics Instead of Including Track Indicators

(a) Math



(b) Reading



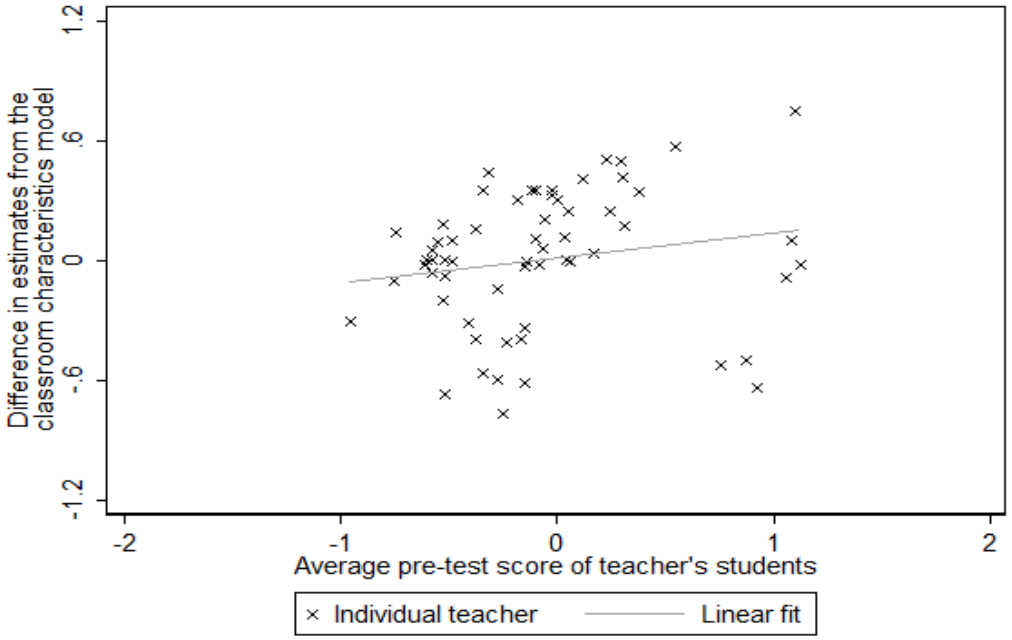
Source: Administrative data from OSSE and DCPS.

Notes: The figures include 62 teachers with math value added and 65 teachers with reading value added.

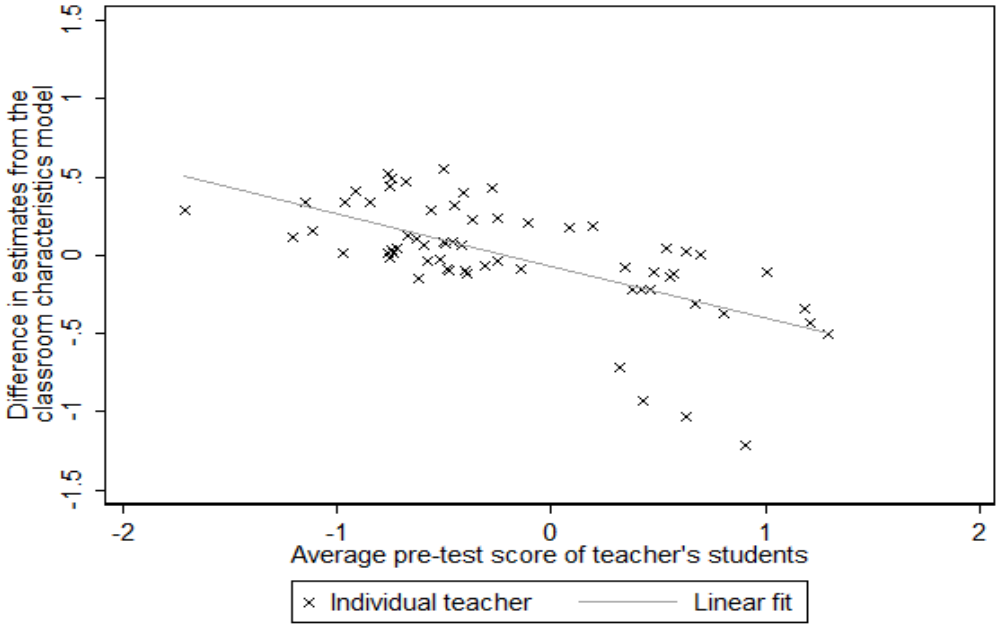
Teachers with positive differences have larger value-added estimates in the classroom characteristics model.

Figure 4. Changes in the Relationship with Student Achievement in the Full Model Compared to the Classroom Characteristics Model

(a) Math



(b) Reading



Source: Administrative data from OSSE and DCPS.

Notes: The figures include 62 teachers with math value added and 65 teachers with reading value added.

Teachers with positive differences have larger value-added estimates in the full model.

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, contact Ali Protik, researcher, at aprotik@mathematica-mpr.com; Elias Walsh, researcher, at ewalsh@mathematica-mpr.com; Alexandra Resch, researcher, at aresch@mathematica-mpr.com; Eric Isenberg, associate director of research, at eisenberg@mathematica-mpr.com; or Emma Kopa, systems analyst, at ekopa@mathematica-mpr.com.

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 Lori Taylor for their helpful comments. Juha Sohlberg, assisted by Maureen Higgins and Raúl Torres, provided excellent programming support. The paper was edited by Carol Soble and produced by Lisa Walls. 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.

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

www.mathematica-mpr.com