MATHEMATICA
Policy Research

Working PAPER

BY ELIAS WALSH AND ERIC ISENBERG

How Does a Value-Added Model Compare to the Colorado Growth Model?

October 2013

Abstract

We compare teacher evaluation scores from a typical value-added model to results from the Colorado Growth Model (CGM), which eight states use or plan to use as a component of their teacher performance evaluations. The CGM assigns a growth percentile to each student by comparing each student's achievement to that of other students with similar past test scores. The median growth percentile of a teacher's students provides the measure of teacher effectiveness. The CGM does not account for other student background characteristics and excludes other features included in many value-added models used by states and school districts. This may lend the CGM more transparency for educators than a value-added model, but at a possible cost of fairness to teachers. Using data from the District of Columbia Public Schools, we examine changes in evaluation scores across the two methods for all teachers and for teacher subgroups. We find that use of the CGM in place of a value-added model depresses the evaluation scores for teachers with more English language learner students and increases the evaluation scores for teachers of low-achieving students. Our analyses suggest that the results may be explained by how the CGM adjusts for prior achievement and its exclusion of other measures of student disadvantage.

I. INTRODUCTION

A. Measuring Teachers' Contributions to Student Achievement

Spurred in some cases by the federal government's Race to the Top initiative, many states and school districts have included in their performance evaluations measures of teacher effectiveness based on student achievement data. States and districts that want to measure teacher effectiveness by using test scores must choose from a menu of options that include value-added models and the Colorado Growth Model (CGM). Value added is used to measure the performance of teachers in many states and districts, including the District of Columbia, Chicago, Los Angeles, and Florida. Massachusetts, Rhode Island, New Jersey, Virginia, and other districts use the CGM.[1]

Value added provides a measure of teachers' contributions to student achievement that accounts for factors beyond the teacher's control. The basic approach of a value-added model is to predict the standardized test score performance that each student would have obtained with the average teacher and then compare the average performance of a given teacher's students to the average of the predicted scores. The difference between the two scores—how the students actually performed with a teacher and how they would have performed with the average teacher—is attributed to the teacher as his or her value added to students' test score performance.

The CGM is a student-level model that assigns percentile ranks by comparing each student's current achievement to other students with similar past test scores. Each *student growth percentile* (SGP) indicates a relative rank for the student's academic growth during the school year. The median SGP of a teacher's students provides the measure of teacher effectiveness. The CGM does not account for student background characteristics such as status as an English language learner, the existence of a learning disability, or eligibility for free or reduced-price lunch (FRL).

Some policymakers view the choice between the CGM and value added as a choice between more transparency and less bias (Hawaii Department of Education 2013). Even though the method for calculating student growth percentiles is complex, the CGM does not adjust for student background characteristics, which may lend the CGM more transparency than a value-added model. However, the CGM also might unfairly disadvantage teachers with many English language learners, special education students, or FRL students. If teachers perceive that it is more difficult to obtain a high evaluation score if they teach disadvantaged students, they might seek to avoid teaching in schools that have many disadvantaged students.

Policymakers may also prefer the CGM because student growth percentiles can be computed with publicly available software that does not require extensive customization for use by a state or district. In contrast, a value-added model typically requires numerous decisions (for example, which student characteristics to include). However, other policymakers may prefer the flexibility provided by value-added models.

[1] In Appendix Table A.1, we provide a complete list of states that use or plan to use the CGM in teacher evaluation systems.

B. Research Questions

We compared estimates of teacher effectiveness from a value-added model to those from the CGM by examining two questions:

1. How large are the changes in evaluation scores when replacing a value-added model with the CGM?
2. Are the changes related to the characteristics of teachers' students?

To answer the questions, we used data on students and teachers in the District of Columbia Public Schools (DCPS) during the 2010–2011 school year to calculate value-added estimates and CGM measures of teacher effectiveness for the same teachers.

The CGM may induce bias because it does not account for student background characteristics and because of other differences between the CGM and value-added models.[2] Although we lack a benchmark for unbiased estimates that would allow us to test directly for bias, our analysis can suggest how large the bias might be and which teachers would most likely be affected by a change from the value-added model to the CGM. However, we cannot rule out the possibility that value-added estimates are also biased because of the sorting of teachers to students on the basis of characteristics that are not accounted for in the value-added model.[3]

C. Earlier Literature Comparing Value-Added Models and the Colorado Growth Model

Most of the earlier literature comparing the CGM to value-added models focused on school-level rather than on teacher-level estimates (Castellano 2011; Ehlert et al. 2012; Catellano and Ho 2012, 2013). Some of the school-level studies found substantive differences between estimates based on these competing models. Ehlert et al. (2012) found that school-level CGM estimates were lower for high-poverty schools relative to a school-level value-added model. Goldhaber et al. (2012), examining teacher-level estimates for North Carolina teachers, found that, although estimates from the two models were highly correlated, teachers of more disadvantaged students tended to receive lower scores on the CGM compared to a value-added model that accounted for other student background characteristics in addition to prior test scores. Wright (2010) also examined teacher-level estimates and found that, compared to the EVAAS value-added model, the CGM produces lower scores for teachers with more students who are FRL-eligible. The EVAAS model is more similar to the CGM than most value-added models, including the value-added model we examine, because it does not account for student background characteristics.[4]

[2] For this discussion of bias, we have assumed that evaluation scores based on the two models are both intended to measure the same dimension of teacher effectiveness.

[3] However, growing evidence suggests that some value-added models provide measures of teacher effectiveness with small bias (Kane and Staiger 2008; Kane et al. 2013; Chetty et al. 2013).

[4] In addition, EVAAS includes random teacher effects rather than the fixed teacher effects used in the value-added model that we estimated. The random effects approach is similar to the CGM in that both rely on within- and between-teacher variation to estimate how much students' current achievement depends on prior achievement. In contrast, a fixed effect model relies only on within-teacher variation to estimate the same relationship.

We contribute to the existing literature in three ways. First, we provide new evidence of how a change from a value-added model to the CGM would affect teachers. To do so, we examine estimates in a context in which a value-added model is used with high-stakes consequences. Second, whereas earlier studies considered changes in the abstract, we document how changes would affect consequences for teachers in the DCPS IMPACT teacher evaluation system. Finally, by examining how evaluation scores would change for teachers of students with particular characteristics, we provide new evidence on the reasons for the pattern of changes.

D. Overview of Findings

The CGM estimates correlate with the value-added estimates at 0.93 in math and 0.91 in reading. Given the knifeblade nature of an evaluation system, however, even highly correlated results can cause some teachers who would be retained as a consequence of their value-added estimate to lose their jobs as a consequence of the CGM, or vice versa. Applying the rules of the multiple-measure IMPACT evaluation system to evaluation scores that substitute CGM scores for value-added scores, we found that 14 percent of teachers would change from one of the four broad performance categories to another as a result of shifting from the value-added model to the CGM. Changes in these categories have important consequences for teachers, ranging from the receipt of performance pay to dismissal.

We also found that, in general, teachers of students with low pre-test scores would fare better in the CGM than in a value-added model, but teachers with other types of disadvantaged students would fare worse. In contrast, previous work has found that teachers with many low pre-test students received lower CGM scores relative to scores from a value-added model.

We investigated the result further and found evidence that the patterns arise from a major distinction in how the CGM and value-added models are estimated. The CGM uses both within- and between-teacher comparisons to estimate how much students' current achievement depends on their prior achievement; therefore, teacher sorting could affect the estimated relationship in the case of a correlation between teacher effectiveness and student pre-test scores. To the contrary, the value-added model with teacher fixed effects uses only within-teacher comparisons to estimate this relationship; therefore, teacher sorting does not directly affect the estimated relationship. Thus, the CGM may overadjust for prior student performance and thereby help teachers with lower-achieving students. In contrast, by excluding student background characteristics, the CGM may hurt teachers with more disadvantaged students measured by these excluded characteristics.

II. IMPACT EVALUATION SYSTEM AND VALUE ADDED IN DC PUBLIC SCHOOLS

A. The IMPACT Teacher Evaluation System

In the 2009–2010 school year, DCPS launched IMPACT, a new teacher evaluation system. IMPACT, which placed teachers into one of four performance categories, carried significant consequences. Teachers who performed poorly—in the bottom category for one year or the second-lowest category for two consecutive years—were subject to separation; those in the top category were eligible for additional compensation. As part of its evaluation of teacher effectiveness, IMPACT incorporated value-added estimates.

For the 2010–2011 school year, individual value-added scores constituted 50 percent of the IMPACT score for general education DCPS teachers who taught math, reading/English language arts (ELA), or both subjects in grades 4 to 8. The rest of the IMPACT score was calculated by using a point-based formula that included scores from a series of structured classroom observations known as the Teaching and Learning Framework (TLF) (35 percent), a rating from the principal measuring the teacher's "commitment to the school community" (10 percent), and the school-level value-added score (5 percent) (District of Columbia Public Schools 2011).

Based on his or her IMPACT score, a teacher was placed into one of four performance categories that depended on strict cutoffs. Teachers in the lowest category (ineffective) were subject to separation at the end of the year. Those in the second-lowest category (minimally effective) in two consecutive years were also subject to separation. No high-stakes consequences applied to teachers in the third category (effective). Teachers in the highest category (highly effective) were eligible for additional compensation. In the 2010–2011 school year, of the teachers with value-added scores as part of their evaluation, 3 percent were ineffective, 28 percent minimally effective, 66 percent effective, and 3 percent highly effective.

To incorporate a value-added estimate into a teacher's IMPACT score, DCPS translated value-added estimates into Individual Value Added (IVA) scores based on a formula that gave each teacher a score from 1.0 to 4.0. The formula made a Fahrenheit-to-Celsius–type of translation from the original value-added estimate—measured in terms of DC Comprehensive Assessment System (CAS) test score points—to the 1.0 to 4.0 scale.[5] Scores for the other components were also on a scale from 1.0 to 4.0. All components were combined with the IVA score by using weights to form a teacher's overall IMPACT score. For teachers who taught both math and reading/ELA, the two IVA scores were averaged.

B. The DCPS Value-Added Model

We estimated teacher value added for math and reading by using data on DCPS students and teachers during the 2010–2011 school year according to the method described in Isenberg and Hock (2011), from which the description below is taken.[6] In brief, the value-added model was estimated in two regression steps and two subsequent steps to adjust estimates for comparability across grades and to account for imprecise estimates.

1. **Measurement error correction**. Measurement error in the pre-test scores will attenuate the estimated relationship between the pre- and post-test scores. We adjusted 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 used an errors-in-variables regression to regress the post-test score on the pre-test scores, student background characteristics, and grade and teacher indicators. Because the errors-in-variables regression does not allow standard errors to be clustered by student, we obtained adjusted post-test scores, which subtract the predicted effects of the

[5] The translation method used in the analysis differs slightly from the one used by DCPS in the 2010–2011 school year. It is more similar to the method used in the 2011–2012 and 2012–2013 school years.

[6] Although the value-added model used by DCPS has since incorporated several changes, our analysis is based on the value-added model used during the 2010–2011 school year.

pre-test scores from the post-test scores, the results of which we used to obtain the initial teacher effects in the next step.

2. **Main regression**. We estimated teacher effects by regressing the adjusted post-test scores from the first step on student background characteristics and teacher-grade indicators, clustering standard errors by student. In this regression, the initial teacher value-added estimates were the coefficients on the teacher indicators, with their variance given by the squared standard errors of the coefficient estimates. Appendix A provides the statistical details.

3. **Combine teachers' estimates across grades**. We combined teachers' estimates into a single value-added estimate when the teacher taught students in several grades. We made teachers' estimates comparable across grades and then combined them by using a weighted average. To do so, we standardized the estimated regression coefficients within each grade so that the means and standard deviations of their distributions were the same. When combining the standardized estimates, we based 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.

4. **Empirical Bayes procedure**. We used an Empirical Bayes (EB) procedure as outlined in Morris (1983) to account for imprecise estimates. These "shrinkage" estimates were approximately a precision-weighted average of the teacher's initial estimated effect and the overall mean of all estimated teacher effects. We calculated the standard error for each shrinkage estimate by using the formulas provided by Morris (1983). As a final step, we removed from our analysis any teachers with fewer than 15 students and recentered the shrinkage estimates to have a mean of zero.

III. COLORADO GROWTH MODEL

Damian Betebenner (2007) developed the CGM for the Colorado Department of Education (CDE). The CDE and other states and districts use the CGM to provide data on individual student academic growth as well as on school- and district-level performance (CDE 2009).[7] Betebenner et al. (2011) present a rationale for using SGP measures in teachers' evaluations that are based on multiple measures of effectiveness. Beginning in the 2013–2014 school year, the CDE will also allow local education agencies to use the CGM to measure teacher effectiveness (CDE 2013). Other states have already adopted the CGM to measure teacher effectiveness (Appendix Table A.1). For teachers' evaluations, some of these states will use student growth percentiles at the school level rather than at the teacher level.

The CGM can be implemented by using a package for the R statistical software program, which its developers freely provide online with documentation. The CGM employs a different approach than a typical value-added model. In the first step, the CGM estimates an SGP for each student (Betebenner 2007). The SGP is a student's academic achievement rank in a given year, grade level, and subject relative to the achievement of other students with the same baseline score or history of scores. Thus, a student with an SGP of 50 is said to perform on the post-test as well as or better

[7] The District of Columbia Public Charter School Board uses the CGM as a component of its school performance reports (Office of the State Superintendent of the District of Columbia Public Schools 2011).

than half of the students with the same pre-test score (or scores) while a student with an SGP of 90 is said to have exceeded the post-test performance of all but 10 percent of the students with the same pre-test score. The CGM performs the first step through a series of 100 quantile regressions conducted at each percentile of the distribution of student achievement on the post-test.

In the second step, the CGM attributes academic growth to teachers as measured by using the SGP. We obtained a measure of teacher effectiveness in DCPS based on the CGM by calculating the median SGP of students linked to the teacher.

The procedure to obtain a measure of teacher effectiveness based on the CGM differs from the procedure we used to estimate teacher value added on several dimensions, although value-added specifications can vary substantially. Unlike the value-added model, the CGM does not include information on students' opposite subject pre-tests or any student background characteristics other than same-subject pre-test scores. The CGM includes additional years of pre-test scores for students when these earlier scores are available, and allows for more flexibility in how current scores depend on prior scores. The additional information and flexibility may, in part, compensate for the CGM's exclusion of other observable characteristics. The CGM does not implement any direct correction for measurement error in the pre-test scores. For teachers linked to students in multiple grade levels, the CGM makes no distinction between SGPs of students in different grades when calculating the median SGPs, though it is possible that the quantile regression approach may make the multiple-grade issue less of a concern than it would be for a more typical value-added model.[8] Finally, no shrinkage is performed on the CGM value-added scores. Appendix B outlines the technical details of the CGM.

IV. METHODS AND DATA

We estimated value added and median growth percentiles for the 334 math teachers and 340 reading/ELA teachers in grades 4 through 8 who taught at least 15 eligible students during the 2010–2011 school year, the threshold used by DCPS to determine eligibility to receive a value-added estimate as part of the evaluation. We then compared the two sets of evaluation scores in four ways. First, we calculated correlations between value added and median growth percentiles for math and reading estimates. Second, we scaled both sets of scores to have a standard deviation of one and a mean of zero and calculated the average absolute difference in standard deviations of teacher value-added estimates for math and reading. Third, we transformed both sets of scores into percentile ranks and calculated percentiles of absolute changes in ranks between the two sets of scores for math and reading. Finally, we calculated the proportion of teachers who would change IMPACT effectiveness categories if value-added estimates were replaced with median growth percentiles. To do so, we converted both sets of evaluation scores into the IVA evaluation component and calculated two sets of IMPACT scores. We scaled the median growth percentiles to have the same standard deviation and mean as the value-added scores before converting to IVA so that the same conversion method could be applied to both sets of scores.[9]

[8] The SGP estimates do not depend on the scale of the assessment (Briggs and Betebenner 2009).

[9] We used a conversion method that differs from the method used by DCPS in the 2010–2011 school year. Our method sets cut points for each possible IVA score such that the mean value-added estimate would be mapped to 2.5, 10 percent of teachers would receive a score of 1.0, and 10 percent of teachers would receive an estimate of 4.0 if the estimates had a normal distribution.

To make comparisons of the effect of switching from value added to the CGM, we examined how teachers with "many" and "few" disadvantaged students would fare under both models. We defined a teacher as teaching many disadvantaged students by calculating the percentage of students with a given characteristic for all teachers and then finding the percentage of students that corresponded to a teacher at the 90th percentile of the distribution of teachers. Similarly, we defined a teacher as teaching few disadvantaged students at the 10th percentile. In Table 1, we present the summary statistics for the teachers in our math and reading samples. For example, math teachers with 96.1 percent of FRL-eligible students were at the 90th percentile of the distribution of teachers in terms of the proportion of such students in teachers' classrooms. Teachers at the 10th percentile of the distribution had 23.4 percent of their students FRL-eligible. For reading/ELA teachers, the percentiles were similar. Pre-test scores were measured in standard deviations of student achievement, which we adjusted to be constant across grades.

Using regression analysis, we calculated the effect of replacing value-added estimates with median growth percentiles for teachers of students with high and low levels of disadvantage. We calculated the difference between the two evaluation scores (in standard deviations of teacher value added) for each teacher and used that as the dependent variable in a teacher-level regression.[10] The explanatory variables were the proportion of each teacher's students with the characteristics in Table 1. We included these characteristics individually in separate regressions and simultaneously in one regression.[11] We then scaled the regression coefficients to reflect the effect of moving from the low level of disadvantage to the high level indicated in Table 1. We estimated separate regressions for math and reading.

At least three factors could affect which teachers achieve the largest changes between the CGM and value added. First, given that the CGM does not account for student background characteristics, an evaluation system that uses the CGM as its measure of teacher effectiveness in place of a value-added model may penalize teachers who teach many disadvantaged students.

[10] We calculated standard errors robust to heteroskedasticity. We weighted observations based on the number of students contributing to the teacher's value-added estimate.

[11] In the regression that included all characteristics simultaneously, we also included the average prior attendance and the average opposite-subject pre-test score of teachers' students because the value-added model also accounted for these characteristics.

Table 1. Comparison Levels for Teachers of Students with High and Low Levels of Disadvantage

	Math		Reading	
Characteristic	Low Level	High Level	Low Level	High Level
English Language Learner	0.0%	33.0%	0.0%	25.5%
Learning Disability	0.0%	20.9%	0.0%	20.7%
Eligible for Free or Reduced-price Lunch	23.4%	96.1%	22.6%	96.2%
Pre-test Score (same-subject)	0.73	-0.65	0.68	-0.62

Source: Administrative data from DCPS and the Office of the State Superintendent of Education of the District of Columbia (OSSE).

Notes: The high level of disadvantage is the 90th percentile of the average student characteristic at the teacher level, and the low level is the 10th percentile. The percentiles are reversed for the pre-test score.

The table is based on teacher-level averages of student characteristics for the 334 math teachers and 340 reading/ELA teachers in grades 4 through 8 with value-added estimates.

Second, whereas our value-added model accounts for measurement error in the pre-tests by using an errors-in-variables technique (Buonaccorsi 2010), the CGM does not apply a similar correction. The errors-in-variables correction produces a steeper relationship between pre-tests and achievement than that obtained from a model with no adjustment. The result is likely to be a lower level of achievement for students with relatively low pre-test scores, thereby attributing more of a student's achievement level at post-test to his or her starting point rather than to the student's current teacher. Thus, the correction may help teachers of students with lower pre-test scores. The CGM does not provide for a similar correction and therefore may reduce the evaluation scores of teachers of low pre-test students relative to the evaluation scores they would have achieved in a value-added model.[12]

However, a third factor could work in the opposite direction. Given that the value-added model includes teacher fixed effects (binary indicators for each teacher-grade combination in the analysis file), the adjustment for prior achievement is based on comparisons of students with higher and lower levels of prior achievement who were taught by the same teacher. Holding teacher quality constant, the comparisons can identify the degree to which prior achievement affects current achievement. The exclusion of such fixed effects, as in the case of the CGM, means that the adjustment for prior achievement is based in part on comparisons of students with different teachers. Thus, the CGM risks confounding the structural relationship between pre-test scores and achievement—one that is based on how much students learn during the year with an average teacher—with the way in which teachers are matched to students. For example, take the case of more effective teachers teaching at schools with higher-achieving students.[13] In this example, even if

[12] The CGM estimates a flexible relationship between pre-test scores and achievement, which could contribute to the magnitude of this difference because the relationship between pre-tests and achievement is typically flatter for high- and low-scoring students than for students in the middle of the distribution of pre-test scores. However, the effect that measurement error in pre- and post-tests has on SGP estimates is not well understood, in part because the CGM uses quantile regression. Thus, the direction and magnitude of potential bias in the CGM from measurement error is unknown.

[13] Recent work has found that, on average, disadvantaged students may be less likely to be taught by the most effective teachers, though the differences are small and depend on the districts or grade levels studied (Glazerman and Max 2011; Mansfield 2012; Sass et al. 2012).

students retained no knowledge from year to year such that there was no structural relationship between pre-test scores and achievement, pre-test scores and effective teaching would be positively correlated. Thus, the SGP for a student with a low pre-test score would reflect both the lower average effectiveness of teachers of similar students and the lower predicted achievement for the student. As a result, teachers of students with low pre-test scores would receive higher scores under the CGM than under the value-added model. It is thus an empirical question of which effect dominates.

Other differences in the production of CGM measures compared to measures produced by a value-added model could affect the distribution of changes. As recommended in Betebenner (2007), we calculate SGPs that account for as many as three prior same-subject test scores for each student, whereas the value-added model used one prior same-subject test score and one prior opposite-subject test score.[14] Unlike the value-added model, the CGM does not standardize estimates by grade level or use empirical Bayes shrinkage, both of which could also affect which teachers have larger changes.

The data contained information on students' test scores and background characteristics and, importantly for the estimation of value added, enabled students to be linked to their teachers. DCPS students in grades 3 through 8 and in grade 10 took the DC CAS math and reading tests. Our analysis was based on students who were in grades 4 through 8 during the 2010–2011 school year and who had both pre- and post-test scores. To enable us to compare teachers across grades, we standardized student test scores within subject, year, and grade to have a mean of zero and a common standard deviation. We excluded students who were repeating the grade; therefore, in each grade, we compared only students who completed the same tests. Isenberg and Hock (2011) provide complete details of the data used for the value-added model.

V. RESULTS

A. Magnitude of Differences

We found substantive differences between evaluation scores based on the GCM and the value-added model. The evaluation scores were correlated at 0.93 for math and 0.91 for reading (Table 2, row 1). The level of correlation might suggest a high degree of similarity, but even high correlations can obscure substantive differences in estimates for individual teachers. As seen in Figure 1, the changes are substantial for many teachers. A scatter plot for reading evaluation scores is similar (Appendix Figure C.1). We find that the average teachers' estimates differ by 0.29 standard deviations of teacher value added in math and 0.33 standard deviations in reading (row 2).

As an alternative approach, we compared the percentile ranks of teachers under the two models and found evidence of some substantial differences. We show the magnitude of the changes by comparing percentile ranks in Table 2, rows 3 through 5. We found that the median math and reading/ELA teacher moved 6 percentiles in the distribution of effectiveness. Five percent of teachers moved at least 22 percentiles in the distribution for math and 25 percentiles in reading. The largest change for an individual teacher was 48 percentiles.

[14] As a sensitivity test, we also compared evaluation scores based on the CGM by using only a single year of prior same-subject test scores. Results were similar to those from the CGM that used three pre-tests (Appendix Table C.1).

Replacing value-added estimates with median growth percentiles would imply changes to teacher IMPACT scores. As a result of changes to the value-added component score, 14.2 percent of teachers would change IMPACT performance categories, with a different evaluation consequence as a result (last row of Table 2).

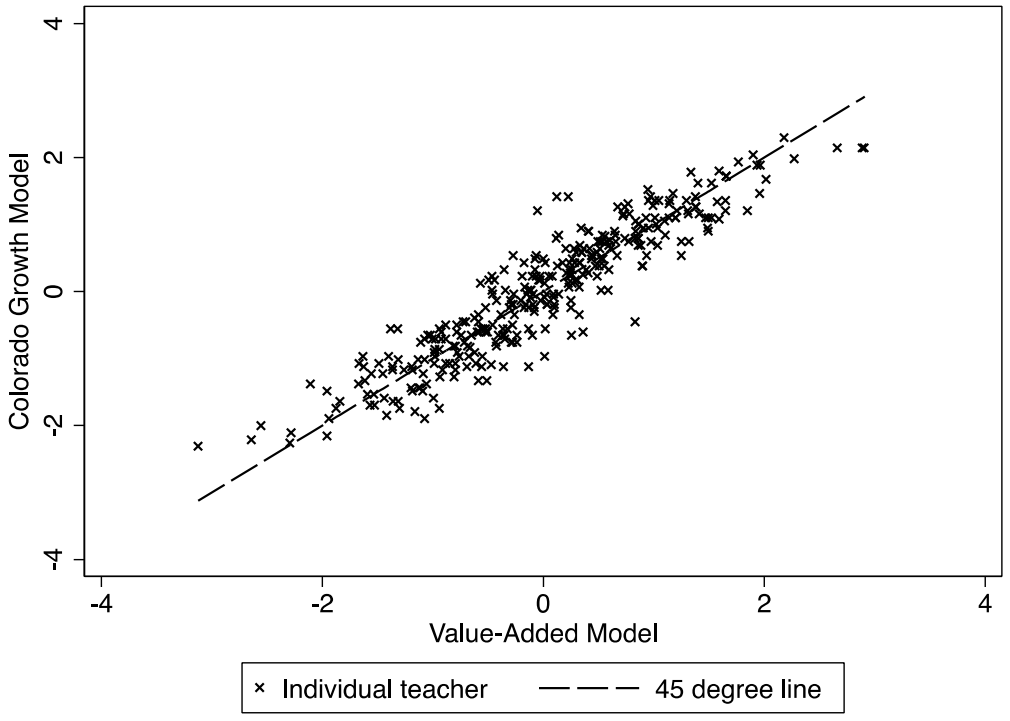
Table 2. How Much Evaluation Scores Change When Using the Colorado Growth Model in Place of a Value-Added Model

	Math	Reading	Combined
Correlation with Benchmark Value-added Model	0.93	0.91	--
Average Absolute Change in Estimates (standard deviations of teacher value added)	0.29	0.33	--
Median Absolute Change in Percentile Rank	6	7	--
95th Percentile Absolute Change in Percentile Rank	22	25	--
Largest Absolute Change in Percentile Rank	41	48	--
Percentage of Teachers Who Change IMPACT Effectiveness Categories	--	--	14.2%

Source: Administrative data from DCPS and the Office of the State Superintendent of Education of the District of Columbia (OSSE).

Notes: Correlation is calculated as a Pearson correlation.

Figure 1. Colorado Growth Model and Value-Added Evaluation Scores in Math



Source: Administrative data from DCPS and the Office of the State Superintendent of Education of the District of Columbia (OSSE).

Notes: The figure includes data for the 334 math teachers in grades 4 through 8 with value-added estimates.

The two sets of evaluation scores are scaled to have a mean of zero and a standard deviation of one.

B. Distribution of Differences

Teachers of students with low pre-test scores would earn higher evaluation scores based on the CGM relative to the value-added model, whereas teachers with more disadvantaged students in some other categories would earn lower evaluation scores. As shown in Table 3, teachers of students with low pre-test scores would gain 0.16 standard deviations of teacher value added in math and 0.17 standard deviations in reading relative to teachers of students with high pre-test scores when replacing value-added estimates with median growth percentiles. However, teachers with more English language learners would earn lower evaluation scores under the CGM than under the value-added model—by 0.16 standard deviations in reading and 0.10 standard deviations in math. Differences for teachers with many special education students and those with many FRL-eligible students are smaller in magnitude and not statistically significant.

Though standing in contrast to the findings of Goldhaber et al. (2012) and Wright (2010), our findings that the CGM provides a relative benefit to teachers with lower-achieving students (but not for other measures of student disadvantage) are consistent with an explanation related to the fixed teacher effects included in the value-added model. By basing the adjustment for student pre-test scores in part on variation between teachers, the CGM may confound the structural relationship between pre-test scores and achievement with how teachers are matched to students, perhaps leading to higher evaluation scores for teachers with low-achieving students if more effective teachers teach at schools with fewer disadvantaged students.

Table 3. How Evaluation Scores Change for Teachers of Disadvantaged Students When Using the Colorado Growth Model in Place of a Value-Added Model

	Change in Estimates at High Versus Low Level of Student Disadvantage (standard deviations of teacher value added)	
Characteristic	Math	Reading
English Language Learner	-0.16*	-0.10*
Learning Disability	-0.09	-0.04
Eligible for Free or Reduced-price Lunch	0.06	0.08
Pre-test Score (same-subject)	0.16*	0.17*

Source: Administrative data from DCPS and the Office of the State Superintendent of Education of the District of Columbia (OSSE).

Notes: The reported estimates give the average difference in evaluation scores for a teacher with more disadvantaged students relative to a teacher with fewer disadvantaged students when switching from a value-added model to the CGM.

The high level of disadvantage is the 90th percentile of the indicated characteristic; the low level is the 10th percentile. The percentiles are reversed for the pre-test score. The high and low levels of disadvantage are those in Table 1.

The regression results are not adjusted to account for other characteristics of students included in the value-added model.

A positive number indicates that a teacher with more disadvantaged students would receive higher evaluation scores from the CGM relative to the value-added model compared to a teacher with fewer disadvantaged students.

*Statistically significant at the 5 percent level.

The absence of teacher fixed effects in the CGM might have explained our results if the CGM had a greater impact on the relationship between the changes and pre-test scores than on other consequences that work in the opposite direction. Accordingly, we estimated a second version of the value-added model that does not include fixed effects but instead averages the residuals from the regression model for all students of a given teacher. Consistent with the sorting of effective teachers to schools with more disadvantaged students, we found that the CGM decreased the scores of teachers with more low-achieving students relative to the average residual value-added model (Appendix Table C.2).

Goldhaber et al. (2012) estimated a value-added model that was broadly similar to the value-added model we estimated and included fixed teacher effects; therefore, the differences in the DCPS and North Carolina samples may provide an explanation for the contrasting results. If the consequence of using fixed teacher effects explains the results in DCPS, then variation in the amount of sorting of teachers to schools between the two samples might explain the difference in results.

The results in Table 3 show how, on average, the changes are related to the listed student characteristics. They do not, however, describe the marginal change in evaluation scores associated with each characteristic, holding other characteristics constant. For example, the results do not distinguish between differences in results for two teachers with low-achieving students when only one teacher's students are also FRL-eligible. In Table 4, we report on the marginal differences.

The results that adjust for other covariates to obtain marginal changes show larger differences across the two models. In accounting for other characteristics under the CGM, teachers of students with low pre-test scores score 0.41 standard deviations better in math and 0.44 standard deviations better in reading compared to teachers of students with high pre-test scores. All else equal, reading/ELA teachers who taught more students with learning disabilities would earn lower evaluation scores under the CGM than under a value-added model by 0.16 standard deviations, and those with more FRL-eligible students would earn lower evaluation scores by 0.21 standard deviations. In math, marginal differences associated with teaching special education students or FRL-eligible students were not statistically significant. For both math and reading, differences for teachers with many English language learners were similar in the cases of both adjustments and no adjustments.

Teachers of students with low pre-test scores would generally benefit from a change to the CGM if the levels of other student characteristics in their classes were the same as those among teachers of students with high pre-test scores (Table 4). Given, however, that the students of teachers with low pre-test scores tended to have higher levels of disadvantage as measured by other characteristics, the unadjusted results for pre-test scores were far more modest (Table 3).[15] If the overlap in categories were perfect, then accounting for student background characteristics in a value-added model would be unnecessary if the model already accounted for pre-test scores. The overlap in categories of disadvantage are far from perfect, however, which is one reason for the extent of the differences we found between measures of teacher effectiveness from the CGM compared to the value-added model.

[15] The results in Tables 3 and 4 are consistent with positive correlations between the categories of student disadvantage.

Even though some of the relationships reported in Tables 3 and 4 were statistically significant, the characteristics of teachers' students cannot explain most changes between estimates generated from the value-added model and the CGM. For example, the R-squared is 0.03 for the regression of changes on average pre-test scores. The limited role of pre-test scores in explaining changes is also evident in Figure 2. For this figure, we converted teacher evaluation estimates from both the value-added model and the CGM to the number of standard deviations from the mean estimate, and then subtracted the value-added measure from the CGM measure. Figure 2 plots changes in math evaluation scores from replacing value-added estimates with median growth percentiles against the average achievement of teachers' students on the math pre-test. Positive changes on the vertical axis indicate higher scores in the CGM relative to the value-added model. The trend line is based on the relationship in Table 3, row 4. Even though the trend is statistically significant, most changes are substantially above or below the trend. (The same plot for reading is in Appendix Figure C.2.) Even with all characteristics of teachers' students included simultaneously to explain the changes, the R-squared increases only to 0.10. Thus, teachers may be more likely to change effectiveness categories if they teach students with certain characteristics, but most changes are unrelated to these characteristics.

Table 4. How Evaluation Scores Change for Teachers of Disadvantaged Students When Using the Colorado Growth Model in Place of a Value-Added Model, Adjusting for Other Student Characteristics

	Change in Estimates at High Versus Low Level of Student Characteristics (standard deviations of teacher value added)	
Characteristic	Math	Reading
English language learner	-0.15*	-0.10*
Learning disability	-0.12	-0.16*
Eligible for free or reduced-price lunch	-0.03	-0.21*
Pre-test score (same-subject)	0.41*	0.44*

Source: Administrative data from DCPS and the Office of the State Superintendent of Education of the District of Columbia (OSSE).

Notes: The reported estimates give the average difference in evaluation scores for a teacher with more disadvantaged students relative to a teacher with fewer disadvantaged students when switching from a value-added model to the CGM.

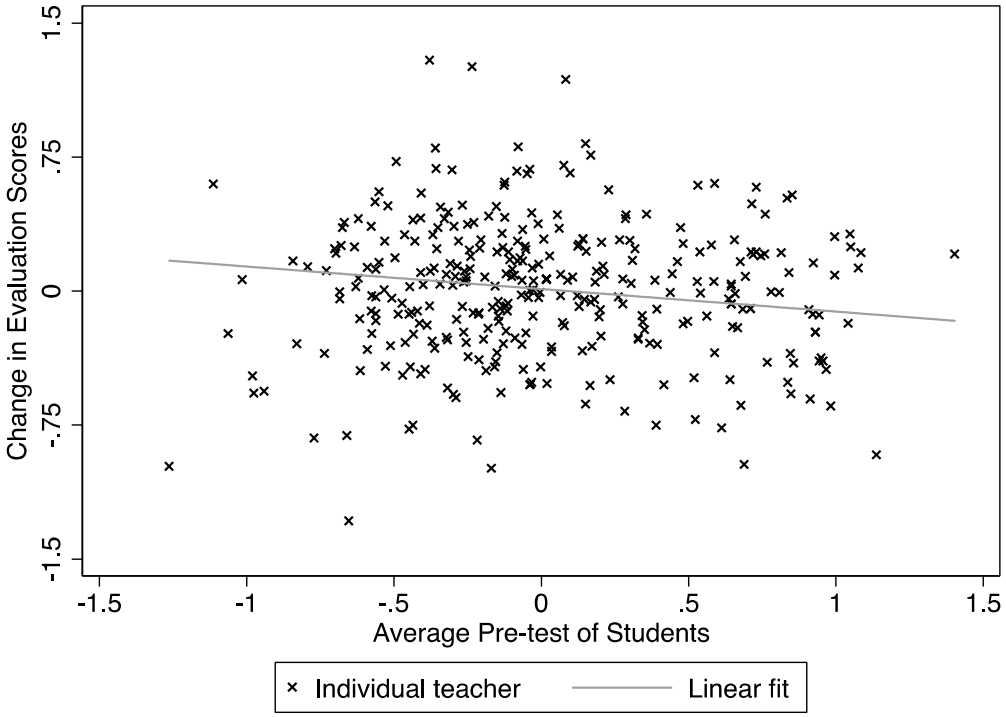
The high level of disadvantage is the 90th percentile of the indicated characteristic; the low level is the 10th percentile. The percentiles are reversed for the pre-test score. The high and low levels of disadvantage are those in Table 1.

The regression results are adjusted to account for other characteristics of students included in the value-added model.

A positive number indicates that a teacher with more disadvantaged students would receive higher evaluation scores from the CGM relative to the value-added model compared to a teacher with fewer disadvantaged students.

*Statistically significant at the 5 percent level.

Figure 2. Change in Math Evaluation Scores When Using the Colorado Growth Model in Place of a Value-Added Model, by Average Achievement of Teachers' Students



Source: Administrative data from DCPS and the Office of the State Superintendent of Education of the District of Columbia (OSSE).

Notes: The figure includes data for the 334 math teachers in grades 4 through 8 with value-added estimates.

The change is reported in standard deviations of teacher value added.

A positive change indicates that the teacher would receive higher evaluation scores from the CGM relative to the value-added model.

VI. DISCUSSION

Aside from differences in outcomes for teachers with different types of students, use of the CGM in place of a value-added model may have different consequences in a policy environment.

First, some policymakers appear to believe that the CGM is more transparent to teachers than a value-added model (Hawaii Department of Education 2013), but the CGM's transparency may be more perceived than real. The SGP as a measure of student academic growth is appealing because it combines the key feature of a value-added model—accounting for a student's level of performance at baseline—with percentiles, a familiar metric for reporting test scores. However, percentiles of test score *levels* may be familiar to many teachers, but percentiles of test score *growth* are unlikely to be familiar to many teachers. Thus, the CGM results would have to be carefully communicated to teachers. Another potential benefit of the CGM in terms of transparency is its perceived simplicity, which is a function of not accounting for student background characteristics other than same-subject pre-test scores. A third benefit is that adopting the CGM allows districts to sidestep a potentially charged discussion of which student characteristics to account for. A value-added model typically accounts for several background characteristics and prior achievement in multiple subjects.

Finally, it is not clear whether the method of quantile regression used to calculate SGPs has any transparency benefit relative to a value-added model.

Second, as teachers come to understand the metric by which they are evaluated, they will likely respond to a different set of incentives under the CGM compared to a value-added model. One difference is that a value-added model implicitly depends on the mean performance of a teacher's students, but the CGM depends on students' median performance. On the one hand, the median is robust to outliers; thus, for example, a stray unmotivated student who randomly fills in the bubbles on the test does not harm teachers. However, given that only the student at the median student growth percentile would matter for a teacher's CGM-based evaluation score, the teacher would have no incentive to attend to struggling students and the highest performing students because their student growth percentiles do not affect the median growth percentile for a teacher. Of course, an easy alternative would call for adapting the CGM by using the average growth percentile of a teacher's students rather than the median growth percentile. Using historical DCPS data, we verified that the two approaches yield highly correlated results and produce similar results (Appendix Table C.1).[16] However, the results may diverge in practice if the median were used and incentivized teachers to change their behavior accordingly.

A second incentive problem, less easily rectified, is that teachers may seek to avoid teaching at schools whose students have background characteristics associated with lower SGPs.[17] Adapting the CGM to account for student characteristics might eliminate these incentives, but it might also eliminate the CGM's perceived transparency benefits.

VII. CONCLUSION

We found evidence of substantial differences between measures of teacher effectiveness based on the CGM compared to a value-added model. We quantified the magnitude of the change by substituting CGM-based teacher evaluation scores for value-added estimates for DCPS teachers in the 2010–2011 school year. This would have resulted in 14 percent of teachers receiving different consequences under the DCPS evaluation system. The consequences ranged from receiving performance pay to dismissal. Even though some changes were related to the background characteristics of teachers' students, most teachers' evaluation scores changed for other reasons.

Our findings do not conclusively indicate bias in the CGM, but we do find that reliance on the CGM in place of a value-added model would tend to depress the evaluation scores for teachers with more English language learners and raise scores for teachers with more low-achieving students. Although the CGM may offer some advantages relative to value-added models—invariance to test score scales and more flexibility in the adjustments for pre-test scores, for example—it is also a flawed measure of teacher effectiveness because it excludes important features of value-added models that are widely thought to reduce bias. Its design also raises concerns about teacher

[16] Castellano and Ho (2012) compared mean and median SGP results for schools and found a root mean squared error of four percentiles.

[17] Similarly, teachers may seek to avoid teaching students with high pre-test scores if teaching such students is associated with lower evaluation scores, as in our results from DCPS. However, the avoidance of such assignments may require more information about the consequences of using the CGM to measure teacher effectiveness than is likely available to most teachers.

incentives associated with the avoidance of teaching assignments that involve certain types of students (such as special education students) or the tendency to devote attention to students who are more likely to influence performance outcomes based on the CGM—those students for whom a teacher expects to see growth near the median growth percentile. States or school districts considering the adoption of the CGM for teacher evaluation systems should consider whether these concerns can be resolved and whether the potential validity and incentive benefits of a value-added model can offset any perceived loss in transparency.

REFERENCES

Betebenner, D., Richard J. Wenning, and Derek C. Briggs. "Student Growth Percentiles and Shoe Leather." Dover, NH: National Center for the Improvement of Educational Assessment, 2011.

Betebenner, Damian W. "Estimation of Student Growth Percentiles for the Colorado Student Assessment Program." Dover, NH: National Center for the Improvement of Educational Assessment, 2007.

Briggs, Derek, and Damian Betebenner. "Is Growth in Student Achievement Scale Dependent?" Unpublished manuscript, 2009.

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

Castellano, Katherine E. "Unpacking Student Growth Percentiles: Statistical Properties of Regression-Based Approaches with Implications for Student and School Classifications." Doctoral dissertation. Iowa City, IA: University of Iowa, 2011.

Castellano, Katherine E., and Andrew D. Ho. "Contrasting OLS and Quantile Regression Approaches to Student 'Growth' Percentiles." *Journal of Educational and Behavioral Statistics*, vol. 38, no. 2, 2013, pp. 190-215.

Castellano, Katherine E., and Andrew D. Ho. "Simple Choices among Aggregate-Level Conditional Status Metrics: From Median Student Growth Percentiles to Value-Added Models." Unpublished manuscript, 2012.

Chetty, Raj, Jonah E. Rockoff, and John N. Friedman. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." Working paper. Cambridge, MA: National Bureau of Economic Research, 2013.

Colorado Department of Education. "Colorado's Academic Growth Model: Report of the Technical Advisory Panel for the Longitudinal Analysis of Student Assessment Convened Pursuant to Colorado HB 07-1048." Denver, CO: Colorado Department of Education, 2008.

Colorado Department of Education. "Determining a Final Educator Effectiveness Rating." Denver, CO: Colorado Department of Education, 2013.

Office of the State Superintendent of Education of the District of Columbia. "The DC Schoolwide Growth Model: Frequently Asked Questions." Washington, DC: Office of the State Superintendent of Education of the District of Columbia, 2011.

District of Columbia Public Schools. "IMPACT: The District of Columbia Public Schools Effectiveness Assessment System for School-Based Personnel, 2011–2012. Group 1: General Education Teachers with Individual Value-Added Student Achievement Data." Washington, DC: District of Columbia Public Schools, 2011.

Ehlert, Mark, Cory Koedel, Eric Parsons, and Michael Podgursky. "Selecting Growth Measures for School and Teacher Evaluations: Should Proportionality Matter?" CALDER Working Paper no. 80, National Center for Analysis of Longitudinal Data in Education Research. Washington, DC: American Institutes for Research, 2012.

Glazerman, Steven, and Jeffrey Max. "Do Low-Income Students Have Equal Access to the Highest-Performing Teachers?" NCEE Evaluation Brief. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, 2011.

Goldhaber, Dan, Joe Walch, and Brian Gabele. "Does the Model Matter? Exploring the Relationship between Different Student Achievement-Based Teacher Assessments." Seattle, WA: Center for Education Data and Research, 2012.

Hawaii Department of Education. "Hawaii Growth Model Frequently Asked Questions." Honolulu, HI: Hawaii Department of Education, 2013.

Hock, Heinrich, and Eric Isenberg. "A Comparison of Two Methods of Modeling Co-Teaching in the DCPS Value-Added Model." Washington, DC: Mathematica Policy Research, 2010.

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, 2011.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." Seattle, WA: Bill and Melinda Gates Foundation, 2013.

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

Mansfield, Richard K. "Teacher Quality and Student Inequality." Working paper. Ithaca, NY: Cornell University, 2012.

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

Sass, Tim, Jane Hannaway, Zeyu Xu, David Figlio, and L. 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.

Wright, S. Paul "An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education." Cary, NC: SAS Institute, Inc., 2010.

APPENDIX A: WHERE THE COLORADO GROWTH MODEL IS USED

Table A.1. States Using the Colorado Growth Model in Teacher Evaluations

State	SGP Measure	Level of Aggregation	First School Year for Evaluations
Colorado	Median	Local education agency decision	2013–2014
Hawaii	Median	Teacher	2014–2015
Massachusetts	Median	Teacher	2013–2014 (some schools starting in 2011–2012)
New Jersey	Median	Teacher	2013–2014
New York	Average	Teacher	2012–2013 (except New York City)
Rhode Island	Median	Teacher	2014–2015
Virginia	Median	Local education agency decision	2012–2013
West Virginia	Median	School	2013–2014

Source: State education agency websites.

APPENDIX B: STATISTICAL DETAILS FOR THE MODELS

Value-Added Model

All details of the value-added model appear in Isenberg and Hock (2011). We repeat the information for the basic regression model here. To calculate teacher value added, we estimated the following regression by subject:

$$(1) \quad Y_{ig} = \lambda_g Y_{i(g-1)} + \omega_g Z_{i(g-1)} + \alpha' X_{1i} + \eta' T_{tig} + \varepsilon_{tig},$$

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 effectiveness in math, Y_{ig} and $Y_{i(g-1)}$ represent math tests, with $Z_{i(g-1)}$ representing 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 X_i denotes control variables for individual student background characteristics, specifically, indicators for eligibility for free lunch, eligibility for reduced-price lunch, English language learner status, special education status, and student attendance in the prior year. The coefficients on these characteristics are constrained to be the same across grades.

The vector T_{tig} contains one indicator variable for each teacher-grade combination. A student contributed one observation to the value-added model for each teacher to whom the student was linked. The contribution was based on a roster confirmation process that enabled teachers to indicate whether and for how long they have taught the students on their administrative rosters and to add any students not listed on their administrative rosters. Students were weighted in the regression according to their dosage, which indicates the amount of time the teacher taught the student.[18] The vector η includes one coefficient for each teacher-grade combination. Finally, ε_{tig} is the random error term. Details on measurement error correction, cross-grade standardization, and empirical Bayes shrinkage appear in Isenberg and Hock (2011).

[18] To estimate the effectiveness of teachers who share students, we used a technique called the full roster method, which attributed equal credit to teachers of shared students. Following this method, each student contributed one observation to the value-added model for each teacher to whom he or she was linked, with students weighted according to the dosage they contributed (Hock and Isenberg 2012).

Colorado Growth Model

We obtained CGM teacher-level measures of effectiveness by calculating the median SGP of students linked to the teacher. In doing so, we followed the Colorado Department of Education (2008). Specifically, we combined SGPs into a teacher value-added measure by calculating the dosage-weighted median for all students linked to each teacher. In a sensitivity analysis, we also calculated the dosage-weighted mean SGP.

We calculated SGPs following the approach described in Betebenner (2007). The CGM estimates SGPs for each student by using quantile regression of the post-test on a flexible function of the history of same-subject pre-tests. The quantile regression provided estimated relationships between the pre-tests and the post-test for any given quantile of the regression residual. Thus, for any given history of pre-test scores, the predicted values from the quantile regression traced out the conditional quantiles of the post-test score. Then, the student's actual post-test was compared to the predicted values conditional on the student's pre-test history. The student was assigned the largest quantile for which the student's post-test exceeded the predicted post-test as his or her SGP.

We estimated unweighted quantile regressions separately by grade level and subject. The pre-test scores were entered into the model as B-spline cubic basis functions, with knots at the 20th, 40th, 60th, and 80th percentiles. These functions allowed for the relationship between pre-tests and the post-test to vary across the range of pre-test scores. Given the maximum of four test scores available for use in DCPS—three pre-tests and one post-test—the conditional quantile model can be expressed as:

$$(9) \quad Q_{Y_{ig}}(q \mid Y_{i(g-1)}, Y_{i(g-2)}, Y_{i(g-3)}) = \sum_{m=1}^{3}\sum_{j=1}^{3} \beta_{gmj}(q)\varphi_{gmj}(Y_{i(g-m)}),$$

where $Y_{i(g-m)}$ is student i's pre-test m grades prior to the post-test Y_{ig} in grade g. The functions φ_{gmj} for j = 1 to 3 give the grade- and pre-test-specific B-spline cubic basis functions. $\beta_{gmj}(q)$ represents the parameters to be estimated in the quantile regression for quantile *q*, which together with φ_{gmj}, describe the conditional polynomial relationship between each pre-test and the post-test for quantile *q*. The relationship was estimated for each of 100 quantiles in increments of 0.5 between the 0.5th quantile and the 99.5th quantile.

To obtain the SGP for each student, the student's post-test was compared to the array of predicted post-test scores for each of the 100 quantiles $\hat{Q}_{Y_{ig}}(q)$ and assigned an SGP of q^* such that q^* is the largest quantile *q* for which the student's post-test exceeded the predicted post-test (that is, $q^* = \arg\max_q \{\hat{Q}_{Y_{ig}}(q)\}$ such that $Y_{ig} \geq \hat{Q}_{Y_{ig}}(q)$).[19] We followed Betebenner (2007) and used the SGP based on the maximum number of pre-tests available for each student. Thus, the SGP for a student with three pre-tests (the maximum available for any student) was based on the results of a quantile regression that included only students with three observed pre-tests. However, the SGP for a student with one observed pre-test was based on a quantile regression that included all students

[19] We used version 7.0 of the Student Growth Percentile package for R statistical software (Betebenner 2007) to implement the quantile regression and obtain SGPs.

with at least one pre-test, though the regression used only information about the most recent pre-test. For students with two observed pre-tests, the quantile regression included only students with at least two observed pre-tests and information on the two most recent pre-tests.

The quantile regression included only pre-tests taken in the years immediately prior to the post-test with no gaps and only for students who progressed a single grade level per year during the period with observed pre-tests. We included all 2010–2011 DCPS students meeting these conditions in the quantile regression regardless of whether they were linked to teachers eligible to receive a value-added estimate. We calculated the CGM evaluation scores only for teachers who received a value-added estimate.

APPENDIX C: ADDITIONAL RESULTS

In a sensitivity analysis, we estimated a version of the CGM that—as in the value-added model we estimated—accounts for only one prior same-subject test score.[20] We found no substantial change in the results; the correlation in estimates of teacher effectiveness across models was 0.99 for both math and reading. We also compared results from the one-prior-test-score CGM to the value-added model and found that they were similar to the comparison of the three-prior-test-score CGM to the value-added model (Table C.1).

Table C.1. How Evaluation Scores Change for Teachers of Disadvantaged Students When Using Alternative Colorado Growth Model Evaluation Scores in Place of a Value-Added Model

	Change in Estimates at High Versus Low Level of Student Characteristics (standard deviations of teacher value added)	
Characteristic	Math	Reading
Panel A: Median Student Growth Percentile, Accounting for One Prior Test Score		
English Language Learner	-0.12*	-0.10*
Learning Disability	-0.07	-0.04
Eligible for Free or Reduced-price Lunch	0.12	0.08
Pre-test Score (same-subject)	0.23*	0.17*
Panel B: Average Student Growth Percentile, Accounting for up to Three Prior Test Scores		
English Language Learner	-0.11*	-0.06
Learning Disability	-0.14*	-0.09*
Eligible for Free or Reduced-price Lunch	-0.03	0.03
Pre-test Score (same-subject)	0.07	0.12

Source: Administrative data from DCPS and the Office of the State Superintendent of Education of the District of Columbia (OSSE).

Notes: The reported estimates give the average difference in evaluation scores for a teacher with more disadvantaged students relative to a teacher with fewer disadvantaged students when switching from a value-added model to the CGM.

The high level of disadvantage is the 90th percentile of the indicated characteristic; the low level is the 10th percentile. The percentiles are reversed for the pre-test score. The high and low levels of disadvantage are those in Table 1.

These regression results are not adjusted to account for other characteristics of students included in the value-added model.

A positive number indicates that a teacher with more disadvantaged students would receive higher evaluation scores from the CGM relative to the value-added model compared to a teacher with fewer disadvantaged students.

*Statistically significant at the 5 percent level.

[20] The value-added model we estimated also accounts for a prior opposite-subject test score.

In a second sensitivity analysis, we compared the CGM to a type of value-added model in which we used the average of the student-level residuals from a value-added regression model to calculate a teacher's value-added estimate (Table C.2).

Table C.2. How Evaluation Scores Change for Teachers of Disadvantaged Students When Using the Colorado Growth Model in Place of an Average Residual Value-Added Model

	Change in Estimates at High Versus Low Level of Student Characteristics (standard deviations of teacher value added)	
Characteristic	Math	Reading
Panel A: Without Adjusting for Other Student Characteristics		
English Language Learner	-0.04	-0.05
Learning Disability	-0.25*	-0.17*
Eligible for Free or Reduced-price Lunch	-0.32*	-0.39*
Pre-test Score (same-subject)	-0.20*	-0.33*
Panel B: Adjusting for Other Student Characteristics		
English Language Learner	-0.02	-0.03
Learning Disability	-0.19*	-0.15*
Eligible for Free or Reduced-price Lunch	-0.26*	-0.50*
Pre-test Score (same-subject)	0.41*	0.30

Source: Administrative data from DCPS and the Office of the State Superintendent of Education of the District of Columbia (OSSE).

Notes: The reported estimates give the average difference in evaluation scores for a teacher with more disadvantaged students relative to a teacher with fewer disadvantaged students when switching from a value-added model to the CGM.

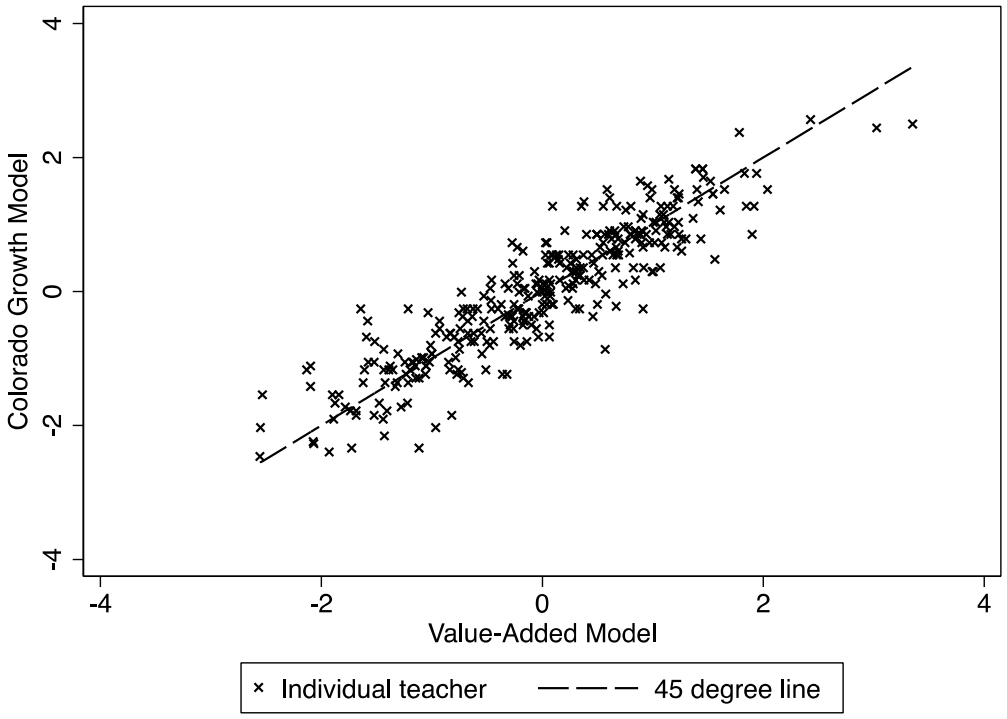
The high level of disadvantage is the 90th percentile of the indicated characteristic; the low level is the 10th percentile. The percentiles are reversed for the pre-test score. The high and low levels of disadvantage are those in Table 1.

A positive number indicates that a teacher with more disadvantaged students would receive higher evaluation scores from the CGM relative to the value-added model compared to a teacher with fewer disadvantaged students.

*Statistically significant at the 5 percent level.

We presented some of our main results only for math. In Figures C.1 and C.2, we present results for reading.

Figure C.1. Colorado Growth Model and Value-Added Evaluation Scores in Reading

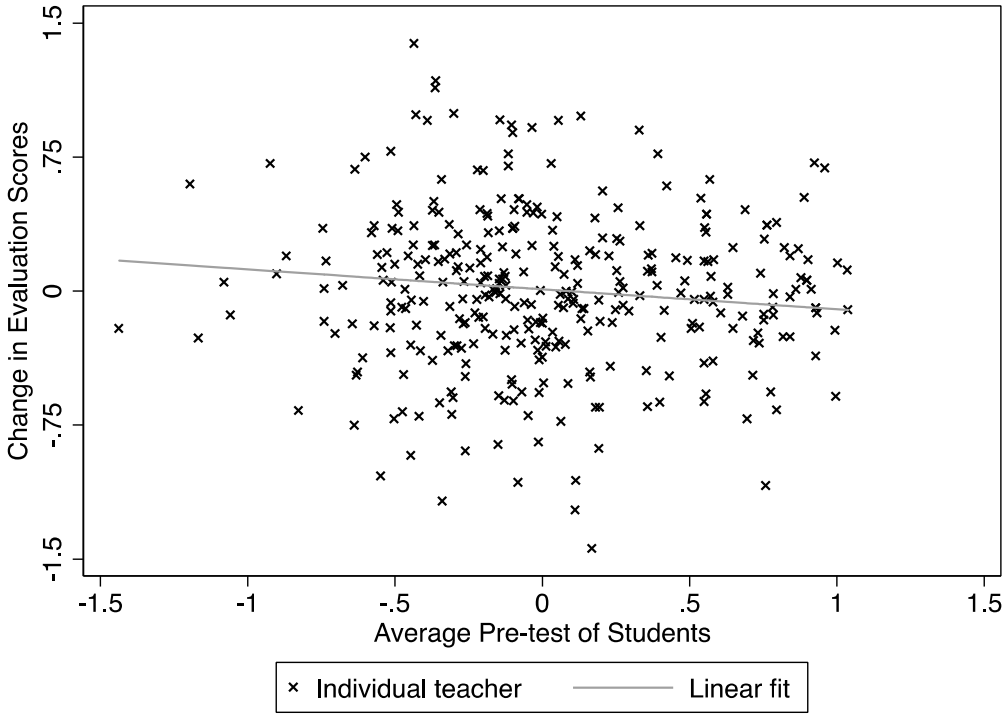


Source: Administrative data from DCPS and the Office of the State Superintendent of Education of the District of Columbia (OSSE).

Notes: The figure includes data for the 340 reading/ELA teachers in grades 4 through 8 with value-added estimates.

The two sets of evaluation scores are scaled to have a mean of zero and a standard deviation of one.

Figure C.2. Change in Reading Evaluation Scores When Using the Colorado Growth Model in Place of the Value-Added Model by Average Achievement of Teachers' Students



Source: Administrative data from DCPS and the Office of the State Superintendent of Education of the District of Columbia (OSSE).

Notes: The figure includes data for the 340 reading/ELA teachers in grades 4 through 8 with value-added estimates.

The change is reported in standard deviations of teacher value added.

A positive change indicates that the teacher would receive higher evaluation scores from the CGM relative to the value-added model.

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 for helpful comments. Juha Sohlberg and Mason DeCamillis provided excellent programming support. The paper was edited by Carol Soble 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 Elias Walsh, researcher, at ewalsh@mathematica-mpr.com, or Eric Isenberg, senior researcher, at ejisenberg@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