MATHEMATICA
Policy Research

WORKING PAPER 33

WorkingPAPER

BY ALEXANDRA RESCH AND ERIC ISENBERG

How Do Test Scores at the Floor and Ceiling Affect Value-Added Estimates?

July 2014

ABSTRACT

Some educators are concerned that students with test scores at the bottom or top of the test score distribution will negatively affect the value-added estimates of teachers of those students. At the bottom end of the scale, some students may appear to have learned little, not because they actually have low achievement, but because they do not exert effort on the test and fill in the answer sheet randomly. At the top end, students with very high test scores appear to have "no room to grow," so value-added scores for teachers with high-performing students will be depressed even if those teachers moved their students up. Using data from a large urban district, we find that test scores at the floor contain real information about student performance because students who score at the floor in one year score at or near the bottom of the scale, on average, in other years. For test score ceilings, we find that teachers of high-scoring students are not universally penalized for teaching students at the ceiling. Rather, the lower the ceiling, the more the value-added estimates tend to "shrink" toward the average teacher because test scores at the ceiling contain imprecise information about student achievement.

I. INTRODUCTION

The use of student achievement data to measure teacher effectiveness for annual evaluations has been encouraged by the federal government through initiatives such as Race to the Top and the Teacher Incentive Fund, and adopted by many states and school districts. As a result, many districts have begun to implement value-added models of teacher effectiveness and wrestle with the implications of using student achievement tests, created for a different purpose, to evaluate teachers.

A common question asked by educators on both ends of the evaluation is how test score floors and ceilings—students scoring at the minimum and maximum scores on a given exam—affect value-added estimates. Regarding test score floors, teachers may wonder whether these data actually indicate that a student is a low achiever or whether some middling or high-achieving students might intentionally or inadvertently receive the lowest scores, causing serious repercussions for the value-added estimate of the teacher responsible for that student. As to test score ceilings, teachers may ask whether their growth is limited by teaching high-achieving students who score at the ceiling.

In this paper, we addressed these questions using data from a large urban school district. For test score floors, we examined how students who score at the floor in one year perform in other years. We found that they tend to perform worse than other students in other years as well, suggesting that scores at the floor provide meaningful information. For test score ceilings, we found that less than one percent of students had scores at the ceiling. So we simulated large ceiling effects—as much as 40 percent of students receiving a score at the ceiling—and showed that ceilings cause bias in teacher value added. However, instead of biasing all value-added estimates of affected teachers down (which would support the concern that these teachers' estimates can only decrease), it biases them toward the middle of the distribution of teacher effectiveness. This is not an indication that these teachers are necessarily average teachers, but rather that, given the lack of information on the true achievement levels of many of their students, these teachers cannot be distinguished from the average teacher.

Overall, our findings suggest that test score floors and ceilings do not present an insurmountable obstacle to the use of value-added models for teacher evaluation purposes. These data are imprecisely estimated, and thus not as informative about students' achievement as scores nearer to the center of the distribution of student achievement. In the data we examined, however, test score floors did not appear to grossly misrepresent student achievement levels, and test score ceilings did not universally bias down the value-added estimates of teachers of high-achieving students.[1]

[1] As districts throughout the United States begin to adopt the Common Core, the issues of floor and ceiling effects are likely to remain for students who take the PARCC assessment exams, which give the same set of questions to all students at the same grade level. To the contrary, the Smarter Balanced assessments will be computer adaptive, which should reduce problems associated with floor and ceiling effects.

II. HOW SCALE SCORES ARE DERIVED FROM RAW SCORES

Before examining how test scores at the floor and ceiling relate to value-added estimates, it is important to understand how student test scores are constructed for the official state test used by the district participating in this analysis.[2] There are some common misconceptions about how test score floors and ceilings are derived. To begin, we cover some basics of the tests in question. There are 12 tests, covering math and reading in grades 3 to 8. The majority of questions on this test are multiple choice, with four choices per question. The exact number of questions on a test varies depending on the test's subject and grade level, but there are far fewer than 100 questions per test. Raw scores—based on the number of items answered correctly—are converted to scale scores separately by grade and subject.

It is possible to obtain a scale score at the test score floor while answering some questions correctly. Figure II.1 shows the translation of raw math scores to scale scores for one 60-question test. The translation of raw to scale scores also is similar for other tests. Each line in the figure shows how a raw score, or number of questions answered correctly, on the left side of the figure translates to a scale score on a 0–99 scale on the right side. The values are labeled for the raw scores that end in zero—0, 10, 20, and so forth—and the scale scores that correspond to them. As shown in Figure II.1, raw scores of 0 to 11 questions answered correctly correspond to a scale score of zero.

The method of converting raw scores to scale scores is non-linear; except for students answering few questions correctly, there are larger jumps in scale scores for an additional correct answer at the extremes of the distribution than near its center. For example, moving from 16 answers correct to 17 correct results in a change of three scale score points, while moving from 26 answers correct to 27 results in a change of only one scale score point.

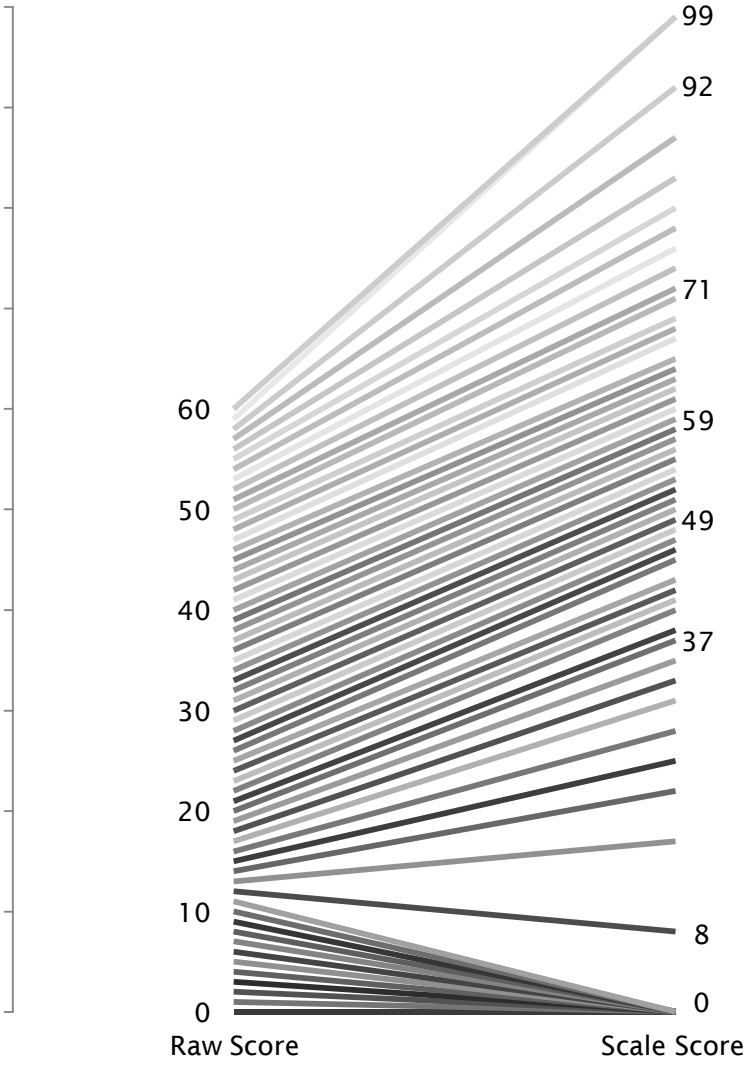
Like most paper-and-pencil tests, this test provides more information about the achievement levels of students in the middle of the distribution, where most of the students score, than for those with very low or high achievement levels. For students in the middle of the achievement distribution, the test contains many items that can be answered correctly by some of these middle performers but not by others. Hence, it is possible to more easily distinguish performance at the middle of the distribution.

The test does not do well in distinguishing students at the very low end of the achievement distribution, however. There are very few items on the test that they can answer correctly without guessing. Students receive a score of zero by performing worse than would be expected from random guessing. In other words, the scale scores are set so that students who answer randomly will tend to get scale scores above zero. Students with very low achievement levels do worse than random guessing. According to a representative from CTB/McGraw-Hill, low performers

[2] We do not refer to the test by name to mask the identity of the large urban district for which we have data. The test is published by CTB/McGraw-Hill.

can do worse than random chance would predict if they get caught by "distracter" answers and/or do not complete all of the questions.

Figure II.1. Translation from Raw Scores to Scale Scores



Source: Authors created figure based on data from CTB/McGraw-Hill.

A similar problem is found for students with very high achievement levels; the highest achievers can answer all questions correctly, so it is not possible to differentiate between them. Like the lowest-achieving students, there is less information available to set their precise scores, so a wider range of achievement is grouped together at a single scale score.

Consequently, the achievement levels of students at the top and bottom of the test score scale are measured with greater error than those of students near the center of the score distribution. Based on item-response theory (IRT), the Standard Error of Measurement (SEM) describes the degree of reliability of test scores, giving an estimate of how much the observed score would vary if the same student took the test multiple times in the absence of floor or ceiling effects. The Conditional Standard Error of Measurement (CSEM) is defined for each

possible scale score. The CSEM is much higher for scores at the extremes of the distribution, implying that the very highest and lowest observed scores are likely to be less informative about current or future student performance than scores in the middle of the test score scale.

The scale scores used for this test are designed so that a student scoring at the floor or ceiling receives an estimate of the median true achievement level of students receiving that score. For example, about half of the students scoring 0 on this test have true achievement above zero and half below zero, but all of them answered fewer questions correctly than the students who received the next lowest scale score (of 8). Similarly, at the top, half have true achievement above 99, and half have true achievement below 99 but none answered fewer questions correctly than those who received a score of 92, the next highest score. Thus, scoring at the floor does not mean that a student's true achievement level is necessarily worse than zero, and scoring at the ceiling does not mean that a student's true achievement level is necessarily higher than 99.

If students with scores of zero have very low achievement levels, the test will work as designed. However, high achievers can also receive a score of zero. For example, high achievers who do not take the test seriously might draw patterns in the bubble sheet rather than attempting to fill in the correct answers. Alternately, some relatively high-achieving students might skip a question early in the test, leading them to fill in answers in the wrong entries for the rest of the test—what is sometimes called an "off-by-one error." These types of zero test scores potentially could be regarded as erroneous scores if students with relatively high achievement levels answer the questions in these ways. Because these tests generally represent "low stakes" for students, there is no individual incentive to perform well.

Concerns about test scores at the ceiling of the test are slightly different, but these test scores present less of a problem for value-added modeling than do scores at the floor. Although any student can score at the floor by choosing not to take the test seriously or being unlucky, only highly skilled students have a reasonable chance of scoring at the ceiling. Perhaps as a consequence, a higher percentage of students score at the floor than at the ceiling on this test. In addition, measurement error is smaller at the upper end of the test score distribution than at the lower end. However, teachers of high-achieving students may express concern that they cannot receive high marks on measures of student learning because their students have "no room to grow."

III. TEST SCORE FLOORS

A. Common Concerns About Test Score Floors Lead to Our Research Question

Some teachers believe that certain students are so far below grade level at pre-test that these teachers are in danger of not receiving credit for producing strong growth in students whose pre-test and post-test scores are both zero. This phenomenon rarely occurs in practice, however, as only one of 500 students has both pre- and post-test scores of zero in our data. In addition, a student performing below grade level may score above zero. For the particular test shown in Figure II.1, the official proficiency cutoff is 43, so any student scoring below 43 is not considered to be proficient.

If, however, a scale score of zero is uninformative, in the sense that many students scoring zero are relatively high-achieving students who do not take the test seriously or make an off-by-one error, this could bias value-added estimates, especially for teachers who teach relatively large numbers of such students. These erroneous scores of zero would affect teachers differently, depending on whether the zeros are on pre-tests or post-tests. A teacher of a student with an erroneous score of zero on only the pre-test would receive a value-added estimate that is too high because the teacher would receive too much credit for raising that student's achievement. A teacher of a student with an erroneous score of zero on only the post-test would receive a value-added estimate that is too low because the observed post-test scores for this student would be low compared to the true achievement level. Therefore, we ask whether students who score at the test floor truly have very low levels of student achievement for their grade.

B. Methods and Data

We examined whether patterns of test scores suggest that scores of zero are informative or anomalous. In a cross-section of student test score data in one year, we could not distinguish zeros for students who actually achieved at the bottom of their class from zeros obtained by students who had higher achievement than was reflected by the number of correct answers. We thus examined student performance across years to see whether those scoring zero in a given year performed differently than other low-achieving students in other years. If students with scores of zero in a baseline year were to score above other low-scoring students on average in later years, this would provide evidence that a score of zero is uninformative and should not be treated as indicating the lowest achievement level among students.

Using test scores from the 2008–2009 school year through the 2011–2012 school year, we examined how the scores of students in the 2008–2009 school year compared to their test scores in later years. We grouped students by their scores in the baseline year, using bins of approximately 5 percentiles each, and compared average scores in later years for those groups. We grouped all scores of zero into a single bin at the bottom of the scale. We divided students into bins based on the percentile rank of their baseline score. Each bin had 5 percent of students, with the exceptions of the first and second bins, which split the lowest 5 percent of students into those with scale scores of zero and those with scale scores above zero but still in the bottom 5 percent. If scores of zero are indicative of performance at the bottom of the achievement scale, we would expect students with baseline scores of zero to continue to score lower on average in later years than those at the 5th or 10th percentile.

We used data from several school years on teachers and students in both traditional public and charter schools in a large urban district. For the analysis of test score patterns, we used test scores for the state assessment for school years 2008–2009 through 2011–2012. The assessment is given in the spring of each school year, so we subsequently refer to each school year by the year in which the assessment occurred. For example, we refer to the 2008–2009 school year as 2009.

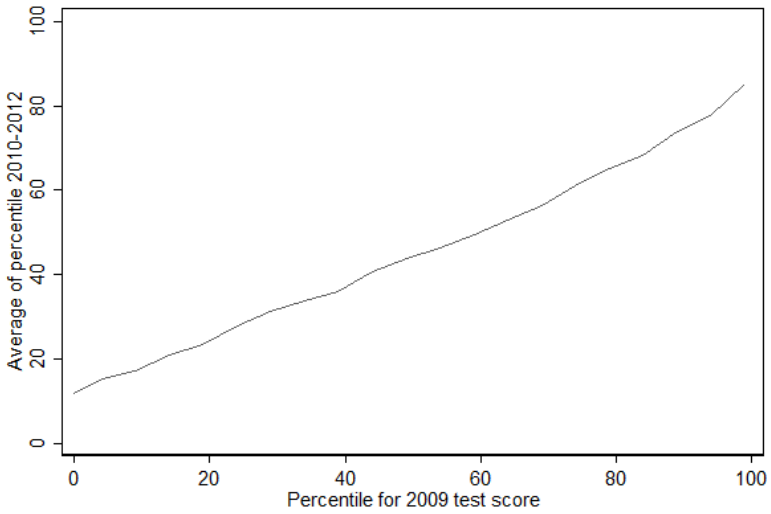
C. Results

During the four school years we examined, about 1.5–2.0 percent of students scored at zero in each year. At the teacher level, this corresponded to less than 1 percent of each teacher's

students on average, with 7 percent of teachers having 5 percent or more of their students at the floor at pre-test.

We found that the pattern over time for students with scores of zero is consistent with the patterns for students with non-zero scores. Figure III.1 presents the relationship between a student's math test score percentile in 2009 and the average math test score percentile in subsequent years. Figure III.2 shows the same relationship for reading test scores. These figures show that there is a positive relationship between scores in 2009 and scores in 2010–2012; if a student was at a higher percentile in the distribution of scores in 2009 (shown on the horizontal axis), that student was likely to be at a higher percentile in the years 2010–2012. Focusing on students near the test score floor, students with a score of zero in 2009 had lower average test scores in years 2010–2012 than those who scored at the 5th percentile in 2009. This suggests that a score of zero does, on average, indicate that a student is performing below others who have scores at the 5th or higher percentiles. As robustness checks of these results, we also examined (1) relationships based on information from each year individually (rather than a multiyear average); and (2) relationships that used each other year (2010, 2011, and 2012) as the base year for sorting students into performance bins. In every case, the patterns were very similar to those shown in Figures III.1 and III.2.

Figure III.1. Average 2010–2012 Test Scores for Students at a Given 2009 Percentile (Math)

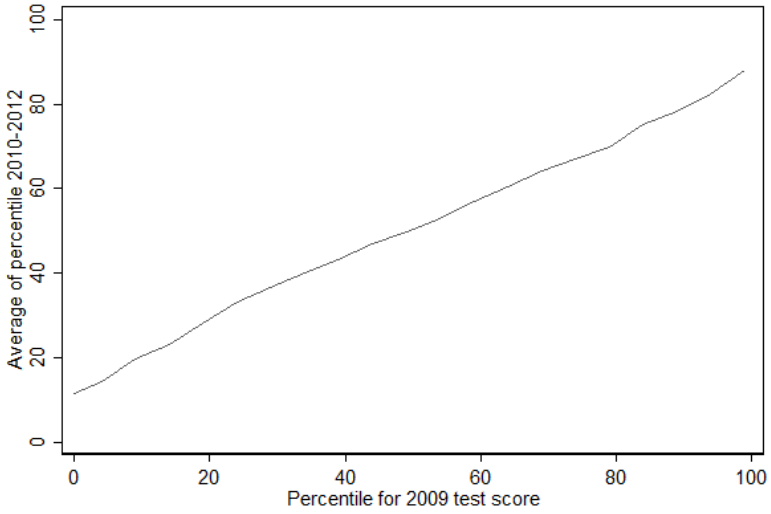


Source: Author calculations based on administrative test data provided by the state education agency for the 2008–2009 to 2011–2012 school years.

Although these patterns suggest that the test scores at zero do represent low-achieving students, there is more measurement error in test scores of zero than in other test scores. Therefore, we calculated the standard deviation of test scores across years for the students within each percentile bin. A higher standard deviation means that students who scored in a given bin had greater variability in their test scores in other years, suggesting that the test score in a particular year may be a less precise measure of their achievement that year. Figure III.3 illustrates the results, showing both the standard deviation of scores by test score bin and the

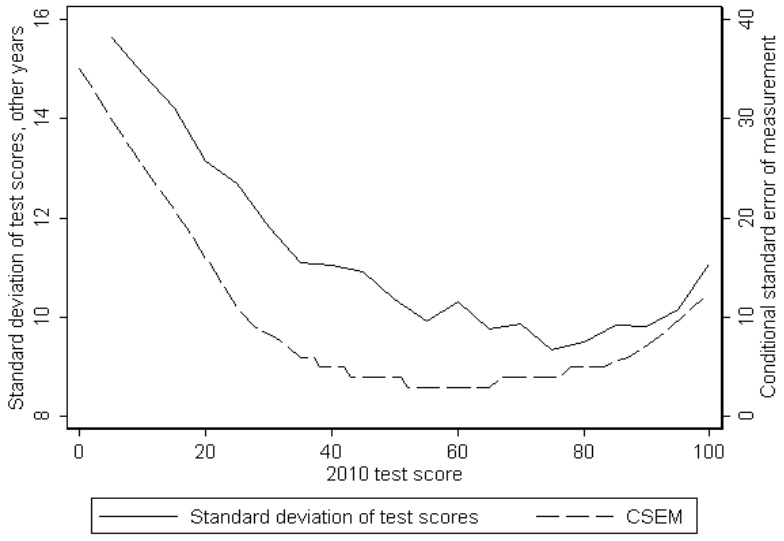
CSEM, as reported by the test publisher. For both statistics, the lowest test scores are associated with the largest values; both the standard deviation of observed scores and the SEM decrease as score scales increase over most of the distribution. Under both measures, there is also an increase in measurement error for test scores at the very top of the distribution. The figure shows these relationships for a particular grade and year, but the patterns are the same in the other years and subjects.

Figure III.2. Average 2010–2012 Test Scores for Students at a Given 2009 Percentile (Reading)



Source: Author calculations based on administrative test data provided by state education agency for 2008–2009 to 2011–2012 school years.

Figure III.3. Standard Deviation of Test Scores in 2010–2012 and Conditional Standard Error of Measurement, by 2010 Percentile Range (Math)



Source: Author calculations based on administrative data provided by the state education agency (standard deviation of test scores) and data from the test publisher (CSEM).

Based on the analyses presented in Figures III.1, III.2, and III.3, we conclude that test scores of zero provide useful information about true student achievement. We found no evidence that students with scores at the floor in one year tend to have unexpectedly high scores in other years. Such a pattern would have suggested that they had not been withholding effort or been unlucky in the year they scored a zero. To the extent that some students are not exerting effort and thus scoring at the floor, it would appear they may do so because they predict they would get few questions correct by applying themselves to the test.

IV. TEST SCORE CEILINGS

A. Concerns About Test Score Ceilings Lead to Our Research Questions

Some observers of value-added methodology have raised concerns about "whether there is considerable room at the top of the scale for tests to detect the growth of students who are already high-achievers" (Baker et al. 2010). Echoing this concern, Diane Ravitch has written in a blog post that "in this age of value-added measurement, when teachers are judged by the rise or fall of their students' test scores, it is very dangerous to teach gifted classes. Their scores are already at the top, and they have nowhere to go, so the teacher will get a low rating" (Ravitch 2014). In more technical terms, teachers of high-achieving students would be at a disadvantage because their value-added estimates would be biased downward. The best they can hope to accomplish is to move students from test score ceiling to test score ceiling, appearing to be an average teacher in the process.

The idea of a necessary downward bias for teachers of high-achieving students may result from an overly literal construction of value added as "growth modeling," in which students who would appear to be at the top of the distribution would have nowhere to grow, putting these teachers at an unfair disadvantage. In fact, however, value-added models are not necessarily growth models, in the sense that they do not impose a coefficient of one on the pre-test score. This relationship is estimated as part of the model, and this coefficient typically is between 0.5 and 0.8. That is, for the average teacher, a high-scoring student would generally score a little closer to the mean at post-test and a low-scoring student would also generally score a little closer to the mean in a positive direction. Thus, by maintaining students at the same number of standard deviations from the test score mean, a teacher can establish himself or herself as a high value-added teacher even if teaching students who are high achieving at baseline. (Although the case is less frequently made that a teacher of low-achieving students would have nowhere to go but up, this is also false, for analogous reasons.)

While imposition of ceiling effects does not systematically reduce value-added estimates for teachers of high-achieving students, it can pull all teacher estimates toward the mean value-added estimate by obscuring the information that can be gained on the performance of students with high test scores. This can harm the value-added estimates of higher-performing teachers of high-performing students and reward lower-performing teachers of high-performing students. For example, imagine a teacher whose students all score 80 on the pre-test who then moves them all up to 90. Assume the coefficient on the pre-test score is 0.8. Abstracting away from student covariates and other features of a value-added model, we would then calculate this teacher's value-added estimate as the average actual post-test score minus the average predicted post-test score, or $90 - (0.8*80) = 26$. But if we impose a test score ceiling at 80, the value-added estimate

is reduced to 80 – (0.8*80) = 16. Now take the opposite case: a teacher whose students start at 90 but end the year at 80. In the absence of a ceiling, the teacher has a value added of 80 – (0.8*90) = 8, which is still above average because actual student test scores were higher than the predicted test scores. However, with a ceiling, this teacher has a value added of 80 – (0.8*80) = 16; this is the same as the first teacher. Teachers of students not subject to ceiling effects are not directly affected, although they are indirectly affected to the extent that value added is always mean zero, so their estimates might decrease or increase as a result of affected teachers' estimates increasing or decreasing.

Therefore, we asked two questions about ceiling effects:

1. For a range of test score ceilings, to what degree do ceiling effects in student test scores affect teacher value-added estimates?
2. Which teachers are affected by student test score ceilings, and how are they affected?

B. Methods and Data

To examine the effect of ceiling effects on value added, we computed value-added estimates for teachers of math and reading in grades 4 through 8. The value-added model we estimated employed features used by value-added models that are used in practice in teacher evaluation systems (Isenberg and Hock 2012; Johnson et al. 2012). For instance, the regression model accounted for student-level background characteristics, including pre-test scores in both subjects. To account for measurement error in pre-tests, we used an errors-in-variables (EIV) approach (Buonaccorsi 2010). This approach incorporates test/retest reliabilities for the test, as reported by the test publisher. This correction effectively increases the expected performance levels of students with high pre-test scores, thus decreasing the value-added estimates for teachers with large numbers of these types of students. We also used the empirical Bayes shrinkage procedure outlined in Morris (1983) to account for imprecise estimates. For details of the value-added model, see the appendix.

We first estimated this value-added model using the actual test scores for students in this large urban district for the 2011−2012 school year. We examined the extent to which test score ceilings affect teachers both on average and in the most extreme cases. To check for whether teachers of high pre-test students are constrained—the contention of some observers—we examined the average value-added estimate of teachers by the quintile of the average pre-test achievement level of their students.

To extend the analysis further, we asked how value-added estimates would change if we were to impose artificial test score ceilings on a sequentially larger number of students. We imposed a ceiling on the top 10 percent, 20 percent, and 40 percent of students, and estimated the value-added models in math and reading for each of these three cases. This method follows the basic approach of Koedel and Betts (2010).

We departed from Koedel and Betts in that we tried to replicate the process of choosing a scale score for the test score ceiling that is consistent with the way in which actual ceilings are set for the test instrument. In particular, instead of giving every student at the ceiling the scale score of the minimum student, we gave them the scale score of the median student in the group,

just as the actual test scores ceilings are set for this test. Thus, half of the students at the ceiling had achievement (as measured by the original test score) higher than the scale score selected to represent the ceiling, and half had achievement lower than the ceiling scale score.

We then examined a number of diagnostics, starting with the correlation in teacher estimates between the original model and the models that impose ceilings on larger numbers of students. We also examined how the standard deviation of teacher effects changes. If an increase in the fraction of students scoring at the ceiling weakens the ability to differentiate among teachers of students with high baseline characteristics, we would expect to see a decrease in the measured standard deviation of teacher effects.

Finally, we examined how the imposition of ceiling effects changes estimates for particular teachers, first by presenting a scatter plot of the original estimates graphed against the new estimates, and then by examining fitted curves for hypothetical teachers whose original estimates were one standard deviation above average (in standard deviations of teacher effectiveness), one standard deviation below average, and average. We examined teachers at three points in the distribution of teacher effectiveness because we expected to find that the value-added estimates of above-average teachers may decrease by the imposition of ceiling effects, the estimates of average teachers would not change, and the value-added estimates of below-average teachers may increase. Furthermore, because we expected the value added of teachers to be more greatly affected when they have more students at the test score ceiling, we examined new estimates as a function of average student pre-test scores. Finally, because ceiling effects in student test scores are predicted to affect value-added scores differently, depending on the original value-added estimate, we interacted student pre-test scores with the original value-added estimate. In particular, to obtain fitted curves of the new value-added estimate, we regressed the new estimate on the original estimate, the average student pre-test score (for the teacher), the average pre-test score squared, the average pre-test score cubed, and three interaction terms of the original estimate with the three functions of the average pre-test score.

We used data on students and teachers in the 2011−2012 school year to estimate value-added models. This analysis included data on approximately 20,000 students. We included elementary and middle school students if they had a test score from 2012 (the post-test) and a test score from the previous grade in the same subject in 2011 (the pre-test). We excluded students from the analysis in the case of missing or conflicting data on school enrollment, test-scores, or student background. We also excluded students who repeated or skipped a grade because they lacked pre-test and post-test scores in consecutive grades and years.

C. Results

When we examined the presence of actual ceiling effects in this particular district, we found them to be rare. On average, for a teacher, only 0.6 percent of students contributing to the value-added model in math scored at the ceiling on the pre-test and 0.4 percent on the post-test. In reading, the percentages were 0.1 percent at pre-test and 0.1 percent at post-test. For a teacher of either subject, the mode was that no students scored at the ceiling, and the maximum was that 14.0 percent of students scored at the ceiling at pre-test in math and 7.0 percent in reading. Fewer than one in 1,000 students scored at the ceiling on both pre-test and post-test in math and no students scored at the ceiling in consecutive years in reading.

Because ceilings that affect few students may have consequences for individual teachers, we also calculated how well the teacher with 14 percent of students scoring at the ceiling on the pre-test could have achieved on value added if these students all scored at the ceiling at the post-test. Had the students scored this high, this teacher would have received the top value-added estimate—75 percent higher than the next-highest teacher in the district. This illustrates that it is theoretically possible for teachers of students with high-achieving baseline test scores to excel, according to value added.

The results from this district may not generalize, however, because few students scored at the test score ceiling; we thus investigated how value-added estimates changed when we imposed ceiling effects that would affect 10, 20, and 40 percent of students.[3] The first set of results is shown in Table IV.1: the correlation between estimates generated under the original model and alternate model decreases as the ceiling is lowered, and the standard deviation of the estimates shrinks as well.

Table IV.1. Comparison of Original Model and Alternate Models with Artificial Ceiling Effects Imposed

	No Additional Ceiling Effects	Top 10 Percent at Ceiling	Top 20 Percent at Ceiling	Top 40 Percent at Ceiling
Math				
Correlation (with a data set with no additional ceiling effects)	1.000	0.988	0.975	0.921
Standard deviation (percentage of standard deviation with no additional ceiling effects)	1.00	0.97	0.94	0.83
Reading				
Correlation (with a data set with no additional ceiling effects)	1.000	0.986	0.963	0.907
Standard deviation (percentage of standard deviation with no additional ceiling effects)	1.00	0.95	0.88	0.81

Source: Author calculations based on data from a large urban school district.

Notes: The standard deviation of teacher effects is presented as a percentage of the standard deviation of teacher effects when no additional ceiling effects are imposed.

The decrease in the standard deviation as more stringent ceiling effects are imposed is evidence of the loss of information on the value-added estimates of teachers with students at the test score ceiling. As it becomes increasingly difficult to distinguish among teachers when we

[3] Koedel and Betts (2010) discuss various tests having large ceiling effects. They point out that these can occur if tests are designed to measure a minimal level of proficiency rather than student achievement throughout a broad range.

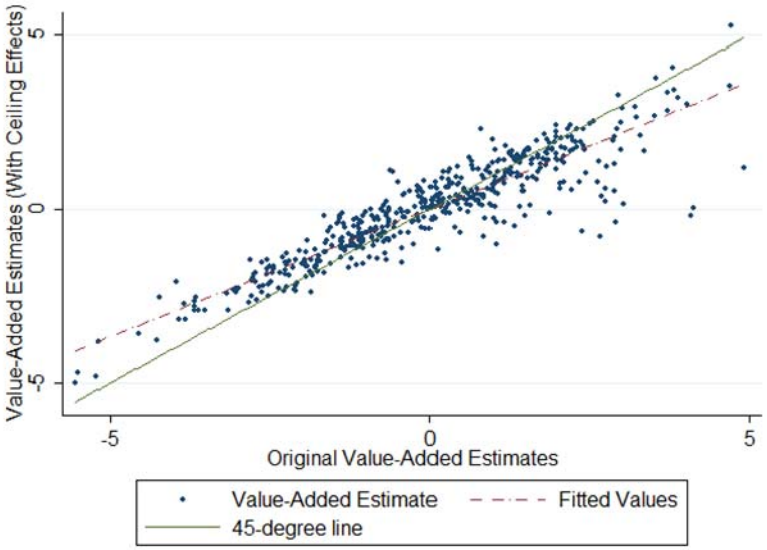
lose precise information about baseline and final achievement for more students, the dispersion of value-added estimates becomes smaller.

The decrease in the correlation as more stringent ceilings are imposed is broadly consistent with results discussed in Koedel and Betts (2010). We should be careful, however, not to interpret correlations above 0.90 as evidence of "little change" in the context of a value-added model. While the overall pattern of results is similar when comparing results based on data with more stringent test score ceilings, there may be policy-relevant changes for particular teachers within an evaluation system if a change would push them over a threshold into a different category for possible outcomes, whether positive or negative.

In fact, many teachers would experience substantial changes to their value-added estimates. The scatter plot shown in Figure IV.1 shows how individual teacher effects change. They compare value-added estimates for teachers of reading using the original data set (on the horizontal axis) compared to using the data set in which 40 percent of students score at the test score ceiling (on the vertical axis). Each dot represents an individual teacher's estimates. The solid line in each figure shows the 45-degree line; teachers below this line received higher value-added estimates using the original data set, and teachers above the line received higher estimates using the data set showing many more students scoring at the test score ceiling. It can be seen from examining this figure that many individual dots are far from the line, indicating teachers who would move up or down by a large margin when substantial ceiling effects occur. A scatter plot for math shows a similar pattern.

We also see in Figure IV.1 that the dotted line, which shows the regression line comparing the two sets of results, is flatter than the 45-degree line, indicating that teachers with low value-added estimates tend to move up when ceiling effects are imposed, and those with high value-added estimates tend to move down. A scatter plot for math shows the same general results as the reading scatter plot. Since value added is by definition at mean zero, there is no net change overall in the value-added estimates.

Figure IV.1. Change in Value Added with Ceiling Effects Added, All Reading Teachers

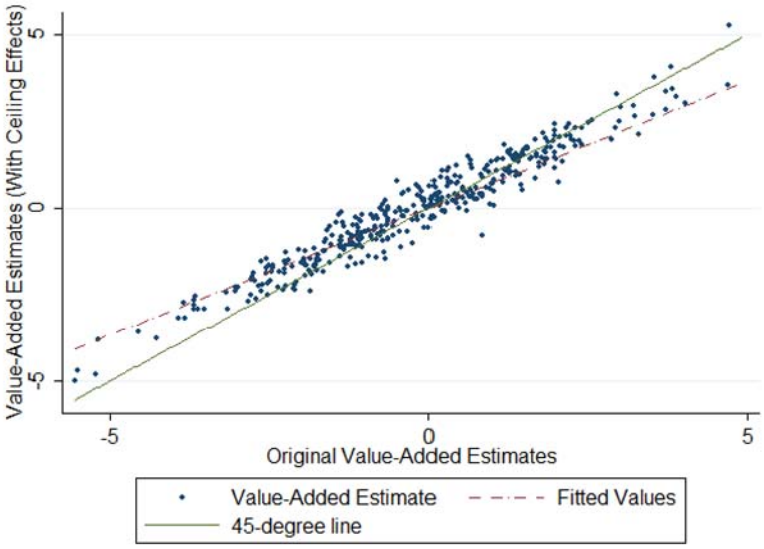


Source: Author calculations based on data from a large urban school district.

Note: Value-added estimates are shown as test score points.

We predicted that the effects would be strongest for teachers of high-achieving pre-test students, who are the most likely to be directly affected by a low ceiling, as we lose more information on their effectiveness than for teachers of students who are lower-achieving at pre-test. To see this effect, we show scatter plots for teachers of students with pre-test scores in the bottom 80 percent of reading teachers (Figure IV.2) and the top 20 percent (Figure IV.3). Although the slopes of the regression lines are similar, most of the extreme cases of teacher effects changing between the two sets of estimates occur when teachers have students in the top 20 percent of the distribution of pre-test achievement. This is consistent with the effects being strongest for these teachers. Results for math followed a similar pattern.

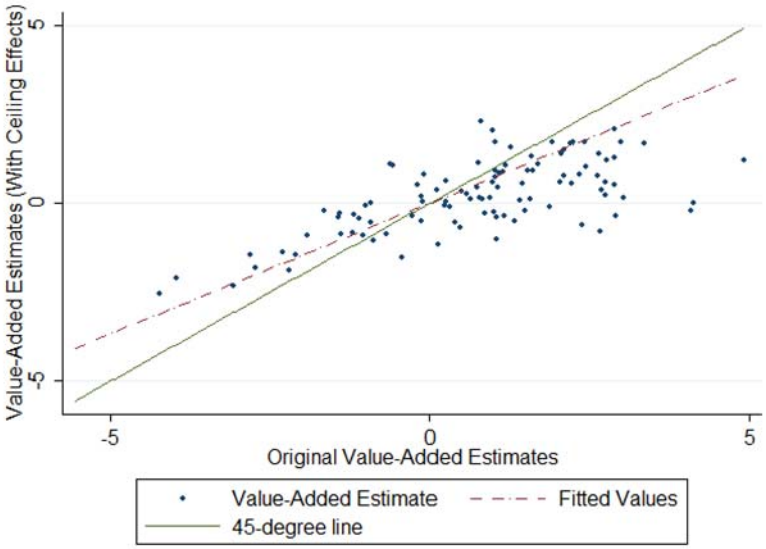
Figure IV.2. Change in Value Added with Ceiling Effects Added, Reading Teachers of Bottom 80 Percent of Test Score Distribution of Students



Source: Author calculations based on data from a large urban school district.

Notes: Ceiling effects have been imposed on the top 40 percent of test takers.
Value-added estimates are shown as test score points.

Figure IV.3. Change in Value Added with Ceiling Effects Added, Reading Teachers of Top 20 Percent of Test Score Distribution of Students



Source: Author calculations based on data from a large urban school district.

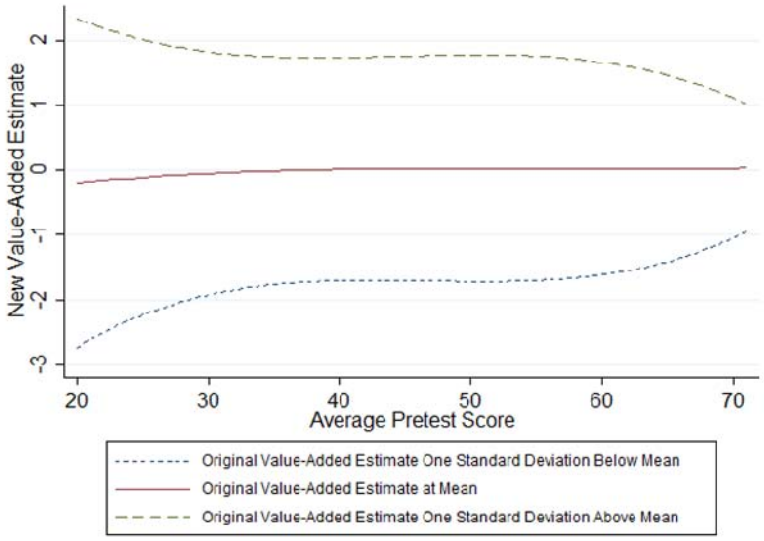
Notes: Ceiling effects have been imposed on the top 40 percent of test takers.
Value-added estimates are shown as test score points.

As a final way of visualizing how ceiling effects affect teachers with varying levels of baseline student test scores and value added (under the original model), we examined figures generated by the regression of the value-added estimate (using the data set with ceiling effects) on the original value-added estimate, functions of the average pre-test score, and the interaction between the two. Figures IV.4 and IV.5 show the results displayed for the new value-added estimate as a function of average pre-test scores for three hypothetical reading teachers: one whose original value-added estimate was one standard deviation above the mean, one who was average, and one whose original value-added estimate was one standard deviation below the mean. The curves we trace out are based on the parameters of the regression model. Figure IV.4 shows results when 10 percent of students score at the test score ceiling, and Figure IV.5 shows results when 40 percent of students do so. The range over which results are shown is from the minimum to the maximum observed average student test score.

Consistent with a small change in the standard deviation and high correlation of original and revised value-added estimates when 10 percent of students score at the test score ceiling, the fitted curves for the relationship between original and revised value added show three curves for the three types of teachers; these are roughly "parallel" throughout much of the range of pre-test scores and then begin to bend slightly toward each other at the top of the range, indicating that, among teachers with the highest pre-test students, some attenuation of the effect occurs whereby relatively effective teachers lose some of their measured effectiveness and relatively ineffective teachers show gains.

However, when 40 percent of students score at the test score ceiling (Figure IV.5), the effects are much stronger. While value-added estimates are fairly stable among all types of teachers of students in the lower half of the distribution of average achievement, at the upper ranges of achievement, there is a strong tendency for teachers of all types to appear average. Value-added estimates collapse toward average among teachers of students with high pre-test achievement levels. This is consistent with the 19 percent decrease in the standard deviation of teacher effects and with the scatter plots, which show the most extreme departures from the 45-degree line to be among teachers of high pre-test students.

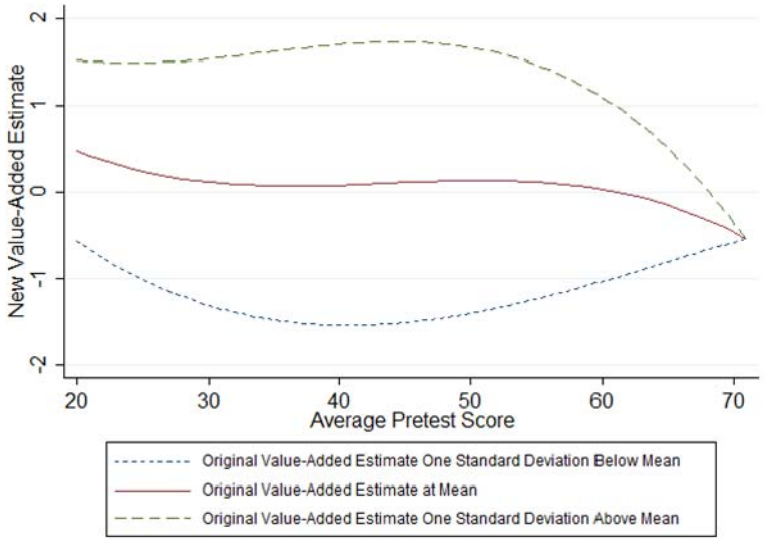
Figure IV.4. Predicted Value-Added Estimates with Modest Ceiling Effects Imposed, Reading Teachers



Source: Author calculations based on data from a large urban school district.

Notes: Ceiling effects have been imposed on the top 10 percent of test takers.
Value-added estimates are shown as standard deviations of teacher effectiveness.

Figure IV.5. Predicted Value-Added Estimates with Severe Ceiling Effects Imposed, Reading Teachers



Source: Author calculations based on data from a large urban school district.

Notes: Ceiling effects have been imposed on the top 40 percent of test takers.
Value-added estimates are shown as standard deviations of teacher effectiveness.

V. CONCLUSION

We have examined test score floors and ceilings to investigate how these phenomena affect value-added models of teacher effectiveness based on data from a large urban school district. Overall, our findings suggest that test score floors and ceilings do not present an insurmountable obstacle to the use of value-added models for teacher evaluation.

Based on our examination of student test scores across years, it appears that the true achievement of students with scores of zero is below that of other very low-scoring students. More precisely, on average, students with scores of zero in a baseline year perform lower in later years than their peers with non-zero baseline scores, including even those at the 5th percentile at baseline.

For test score ceilings, teachers of high-performing students can in fact have high value-added estimates. They are not doomed to be below average by virtue of their students having "no room to grow." Furthermore, a value-added model is fairly robust to test score ceilings affecting around 10 percent of students—more than 10 times the 0.6 percent of students observed to score at the ceiling in the district we studied.

By examining an even more extreme case, in which a larger number of students are affected by ceiling effects, we found that the value-added estimates of teachers of these students move toward the average value-added estimate, regardless of whether they were above-average, average, or below-average instructors. Such a rating may be appropriate in an evaluation context when this many students score at the test score ceilings. Because test score achievement levels of students at the ceiling are very imprecise—a single number is assigned to students potentially representing a large range of actual student achievement—for teachers with many students at the ceiling, value-added estimates of their performance will be similarly imprecise. Policymakers thus would likely be reluctant to take strong actions regarding these teachers, either positively or negatively. Instead, giving these teachers an average value-added estimate is not an indication that their true performance is average but is a reflection of our ignorance of their true performance. In most well-designed teacher evaluation systems, pushing their estimates toward zero would place a teacher in a "no consequences" category for outcomes. While not ideal, under the circumstances of little information gained by virtue of ceiling effects in test scores, this is an appropriate categorization. It is somewhat similar to the way in which empirical Bayes shrinkage, a common final step to the preparation of value-added estimates, moves extreme estimates toward the average in proportion to the precision of the original estimate. In other words, the greater our ignorance of whether the measured effect is close to the teacher's true effect, the less weight we placed on the original estimate and the more we treated the teacher as the average teacher. Again, this is not an ideal circumstance: we would rather have more precise information on the teacher's true effectiveness but, under the circumstances, empirical Bayes shrinkage is an improvement because policymakers can thus avoid assigning inappropriate consequences to some teachers based on imprecise estimates. In the presence of ceiling effects in student test scores, value-added models function in much the same way.

REFERENCES

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

Baker, Eva L., Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd, Robert L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard. "Problems with the Use of Student Test Scores to Evaluate Teachers." EPI Briefing Paper No. 278. Washington, DC: Economic Policy Institute, August 2010.

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

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

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

Johnson, Matthew, Stephen Lipscomb, Brian Gill, Kevin Booker, and Julie Bruch. "Value-Added Models for the Pittsburgh Public Schools." Washington, DC: Mathematica Policy Research, February 2012.

Koedel, Cory, and Julian Betts. "Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation." *Education Finance and Policy*, vol. 5, no. 1, Winter 2010, pp. 54–81.

Lee, P. *Bayesian Statistics: An Introduction*. Second Edition. New York: John Wiley and Sons, 1997.

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

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

Ravitch, Diane. "With VAM: All Teachers of the Gifted Are 'Bad' Teachers." Diane Ravitch's blog. March 24, 2014. Available at http://dianeravitch.net/2014/03/24/with-vam-all-teachers-of-the-gifted-are-bad-teachers/. Accessed June 11, 2014.

APPENDIX: DETAILS OF VALUE-ADDED MODEL AND DATA USED

A. Value-Added Model

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

$$Y_{tig} = \lambda_g S_{ig} + \omega_g O_{ig} + \beta' \boldsymbol{X_i} + \delta' \boldsymbol{T_{tig}} + \varepsilon_{tig}, \quad (1)$$

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 estimated 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.

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. This method of accounting for dosage, including in co-teaching situations, is known as the Full Roster Method (Hock and Isenberg 2012). The coefficient δ represents the teacher effect. Finally, ε_{tig} is the random error term.

Although we estimated value added for teachers with as few as 7 students, we included estimates in our results only for teachers who taught 15 or more students over the course of the year in at least one subject. We did this to be consistent with how value-added estimates might be used in an evaluation system, in which teachers of few students do not necessarily receive a value-added estimate as part of their evaluation.

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

- **Measurement error correction**. Given that measurement error in the pre-test scores attenuates the estimated relationship between the pre-test 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 test. We used an 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 used 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 used the adjusted gains to obtain the teacher effects in the next step.

- **Main regression.** We estimated 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 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 combine them by using a weighted average. We standardized the estimated regression coefficients so that the mean and standard deviation[4] of the distribution of teacher estimates are the same across grades. 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.
- **Empirical Bayes procedure.** We used 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. We calculated the standard error for each shrinkage estimate by using the formulas provided by Morris (1983). As a final step, we removed any teachers with fewer than 15 students from the teacher model and re-center the shrinkage estimates to have a mean of zero.

B. Data on Test Scores

When estimating the effectiveness of teachers, we included elementary and middle school students if they had a test score from 2012 (the post-test) and a test score from the previous grade in the same subject in 2011 (the pre-test). We excluded students from the analysis file in the case of missing or conflicting data on school enrollment, test scores, or student background. We also excluded students who repeated or skipped a grade because they lacked pre-test and post-test scores in consecutive grades and years. This analysis included data on approximately 20,000 students.

Before using the test scores in the value-added model, we created subject- and grade-specific z-scores by subtracting the mean and dividing by the standard deviation within a subject-grade combination. This step allowed us to translate math and reading scores in every grade and subject into a common metric. To create a measure with a range resembling the original test-score-point metric, we then multiplied each test score by the average standard deviation across all grades within each subject and year.

[4] 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.

C. Student Background Data

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

- Pre-test in same subject as post-test
- Pre-test in other subject (we control for math and reading pre-tests regardless of post-test)
- Free-lunch eligibility
- Reduced-price lunch eligibility
- Limited English proficiency status
- Existence of a specific learning disability
- Existence of other types of disabilities requiring special education
- Proportion of days that the student attended school during the previous year

Attendance is a measure of student motivation. We used previous—rather than current-year—attendance to avoid confounding student attendance with current-year teacher effectiveness; that is, a good teacher versus a weaker teacher might be expected to motivate students to attend school more regularly. Attendance is a continuous variable that could range from zero to one. Aside from pre-test variables, the other variables are binary variables taking the value zero or one.

We imputed data for students who were included in the analysis file but had missing values for one or more student characteristics. Our imputation approach used the values of non-missing student characteristics to predict the value of the missing characteristic.[5] For students who did not attend a school in our data for part of the previous year, we used a Bayesian method to impute missing attendance data based on other student characteristics, in addition to attendance during the portion of the year spent in the observed schools.[6] We did not generate imputed

[5] For missing data on free or reduced-price lunch status, we used a slightly more sophisticated imputation procedure because these data are missing for students attending Provision 2 schools, which do not collect information on free and reduced-price lunch status every year. For these students, we used direct certification data to determine their free-lunch status. If a student in a Provision 2 school is not eligible to receive a free lunch via direct certification, we used the student's status from a prior year. If no prior-year data were available for these students, we predicted the value of free and reduced-price lunch status using (1) the percentage of students eligible to receive a free or reduced-price lunch in the school in the last year during which the school sought to collect this information from all students and (2) individual student characteristics correlated with free-lunch status within schools.

[6] We generated a predicted value by using the values of non-missing student characteristics and, for our data, combined this information with the actual attendance data for the part of the year spent in the observed schools. With this method, the more time a student spent in an included school, the more his or her imputed attendance measure relies on actual attendance data from the part of the year spent in a school in our data. Conversely, the less

values for the same-subject pre-test; we dropped from the analysis file any students with missing same-subject pre-test scores.

D. Teacher Dosage

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

We used the confirmed class rosters to construct teacher-student links. If the roster confirmation data indicated that a student had one math or reading teacher at a school, we set the teacher-student weight equal to the school dosage. If a student changed teachers from one term to another, we determined the number of days the student spent with each teacher, subdividing the school dosage among teachers accordingly. When two or more teachers claimed the same students during the same term, we assigned each teacher full credit for the shared students. Finally, similar to tracking time spent at all schools outside the district, we tracked the time a student spent with any teachers not recorded in the confirmed class rosters.

(continued)

time spent in these schools, the more the imputed attendance measure relies on the predicted value. We implemented this approach by using a beta distribution with beta/binomial updating (Lee 1997).

AUTHORS' NOTE

We are grateful to Duncan Chaplin for helpful comments. Emma Kopa, Juha Sohlberg, Dylan Ellis, and Nikhil Gahlawat provided excellent programming support. The paper was edited by Molly and Jim Cameron and produced by Colleen Fitts. The text reflects the views and analyses of the authors alone and does not necessarily reflect views of Mathematica Policy Research. 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 Alexandra Resch, senior researcher, at aresch@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