How do Test Scores at the Ceiling Affect Value-Added Estimates?

Alexandra Resch\textsuperscript{a} and Eric Isenberg\textsuperscript{b}

\textsuperscript{a}Mathematica Policy Research, Washington, DC; \textsuperscript{b}American Institutes for Research, Chicago, IL

1. Introduction

Many states and school districts began to use student achievement data to measure teacher effectiveness for annual evaluations, encouraged by the federal government through initiatives such as Race to the Top and the Teacher Incentive Fund, as well as through waivers from the No Child Left Behind Act. Even as the Every Student Succeeds Act removed federal pressure to evaluate teachers in this way, a group of states and districts have continued to use value-added models to measure teacher effectiveness. As a result, they have had to wrestle with the implications of using student achievement tests, created for a different purpose, to evaluate teachers.

One issue that has received considerable attention is the question of whether value added can be fair to teachers of gifted students. 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). Others have made similar arguments (National Research Council 2009; Amrein-Beardsley and Collins 2012; Haertel 2013; Darling-Hammond 2015). The idea that test score ceilings leave teachers of gifted students with nowhere to go has become so well-accepted that the official American Educational Research Association (AERA) statement on value-added models (2015) states “their meaning should be interpreted in the context of an individual teacher’s curriculum and teaching assignments, with cautions issued regarding common interpretation problems, such as ceiling and floor effects of the tests for estimating growth for high- and low-achieving students.”

The issue of ceiling effects was central to the case of Sheri Lederman, a fourth-grade teacher in upstate New York, who sued the state government after receiving the lowest rating on the value added portion of her teacher evaluation. In May 2016, after hearing testimony from a number of expert witnesses, the Supreme Court of New York sided with Lederman, declaring that her rating was “arbitrary and capricious.” The first three of the five reasons that the Court gave for its decision related to ceiling effects:

The Court’s conclusion is founded upon: (1) the convincing and detailed evidence of VAM bias against teachers at both ends of the spectrum (e.g., those with high-performing students or those with low-performing students); (2) the disproportionate effect of petitioner’s small class size and relatively large percentage of high-performing students; (3) the functional inability of high-performing students to demonstrate growth akin to lower-performing students … (Lederman v King 2016).

In more technical terms, teachers of high-achieving students are perceived to be at a disadvantage because their value-added estimates would be biased downward due to the truncation of the dependent variable (the student-level post-test). Supposedly, the best teachers can hope to accomplish under these circumstances is to move students from test score ceiling to test score ceiling, appearing to be an average teacher in the process.

In this article, we ask how a test score ceiling affects the value-added estimates of teachers with many high-achieving students. Our results suggest three arguments against the consensus view:

**ARTICLE HISTORY**

Received September 2016
Accepted November 2017

**KEYWORDS**

Ceiling effect; Education; Measurement error; Teacher evaluation

**CONTACT** Eric Isenberg ejisenberg@air.org American Institutes for Research, 10 S. Riverside Plaza, Suite 600, Chicago, IL 60606-5500.

© 2018 Alexandra Resch and Eric Isenberg. Published with license by American Statistical Association

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Scores at the test score ceiling are not truncated scores but imprecise measures—that is, a test score at the ceiling should not be regarded as the minimum of a range of possible high scores but as the estimate of a true score that could be higher or lower.

The idea that teachers of high-performing students can’t earn a high value-added estimate is false because, for a teacher to receive a high value-added estimate, the students do not have to “grow” more than the students of the average teacher (where growth is the difference between the raw post-test score and raw pre-test score). Rather, the teachers’ students have to outperform their predicted score, which generally allows for all teachers to have an extensive range of possible value-added estimates.

To the extent that imposing a ceiling effect changes teachers’ value-added estimates, it does not universally push down the estimates of teachers of high-achieving students. Rather, for teachers with many students at the ceiling, it causes their value-added estimates to shrink toward the average: teachers with high value-added estimates experience a decrease while teachers with low value-added estimates experience an increase.1

2. Ceilings are not Truncated Scores but are Imprecise Measures of Student Achievement

Before examining how test scores at the ceiling relate to value-added estimates, it is important to understand how student test scores are constructed. We demonstrate this using a typical state test—we will analyze results using data from this state in Section 5.2 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.

The method of converting raw scores to scale scores is nonlinear; 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 15 correct answers to 16 correct answers results in a change of three scale score points, moving from 30 correct answers to 31 results in a change of only one scale score point, and moving from 55 correct answers to 56 again results in a change of three scale score points. This is demonstrated in Figure 1, which 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.

Like most paper-and-pencil tests, this test provides more precision on 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 high (or low) end of the achievement distribution. The highest achievers can answer all questions correctly, so it is not possible to differentiate between them. Without additional data to set their scale scores more precisely, 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. The standard error of measurement 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. Based on item-response theory (IRT), the conditional standard error of measurement (CSEM) is defined for each possible scale score (Embretson and Reise 2000). The

---

1 In an earlier version of this article, we also addressed a separate set of concerns about floor effects, that is, students who score at the minimum of the test score range. In particular, we examined concerns that students at the floor may be high achievers who scored at the floor because they drew patterns on the test answer sheet or mistakenly filled in the wrong rows of bubbles for their answers. We looked at test scores across years for the same students to demonstrate that these test scores at the floor appear to contain valid data on the low achievement levels of students scoring at the floor (Resch and Isenberg 2014).

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 was published by CTB/McGraw-Hill.
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 ceiling (or highest observable scale score) receives an estimate of the true achievement level of students who answered all questions correctly. If there were more questions to distinguish among high achievers allowing for a more precise mapping of raw scores to scale scores (or if an adaptive test were used), some of the students who had received a score of 99 would likely be mapped to scores above 99, while others would be mapped to scores below 99. Thus, scoring at the ceiling does not mean that a student’s true achievement level is necessarily higher than 99. In this sense, the metaphor of a ceiling, where a range of ability levels bump up against a fixed barrier, may be less useful than thinking of the top score as a test score magnet, pulling a range of ability levels into a single estimate.

3. A Large Ceiling Effect Causes a Type Of Shrinkage, not Downward Bias

The idea that teachers of high-achieving students have no room to grow 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 used in districts for teacher evaluations are not growth models, in the sense that they do not consider a student’s true achievement level.

To extend the analysis further, we asked how value-added estimates would change if we were to impose artificial test score ceilings. We examined the extent to which test score ceilings affect teachers both on average and in the most extreme cases, and the range of estimates for teachers of students with high pre-test scores.

We depart from Koedel and Betts (2010) in that we tried to replicate the process of choosing a scale score for the test score ceiling that is closer to the way in which the highest observable scale score is assigned. In particular, instead of giving every student at the ceiling the scale score of the median student in the group. 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 extended beyond the analysis in Koedel and Betts (2010) by examining how the imposition of a ceiling effect effects changes estimates for particular teachers. We present a scatter plot of the original estimates graphed against the new estimates. 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 distinguish between teachers with highest-performing students and other teachers.

We used data on students and teachers from a large urban district 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.

5. Results

We found the actual ceiling effect in this particular district to be small. On average, for a teacher, only 0.6% of students contributing to the value-added model in math scored at the ceiling on the pre-test and 0.4% on the post-test. In reading, the percentages were 0.1% at pre-test and 0.1% 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% of students scored at the ceiling at pre-test in math and 7.0% in reading. Fewer than one in 1000 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. These results are consistent with an examination of five years of test score data from the 2008–09 to 2012–13 school years in 26 medium- to large-sized districts located in 15 states. In that broader data set, fewer than 1% of students had test scores at the ceiling (Isenberg et al. 2016).

To see if ceiling effects were unduly constraining the value added of teachers in our sample, we identified the teacher with the largest percentage of students scoring at the ceiling on the pre-test (a math teacher with 14% of students at the ceiling). Then we calculated how well the teacher could have achieved on value added if these students all scored at the ceiling on the post-test. Had the students scored this high, this teacher would have received the top value-added estimate—75% higher than the next-highest teacher in the district. This illustrates that it is possible for teachers of students with high baseline test scores to receive high value-added estimates.

| Table 1. Comparison of value-added estimates in original model and alternate models with an artificial ceiling effect imposed. |
|------------------|------------------|------------------|------------------|
|                  | No additional ceiling effect | Top 10% at ceiling | Top 20% at ceiling | Top 40% at ceiling |
| **Math**         | **Correlation**   | **Standard deviation** |
|                  | 1.00              | 0.988            | 0.975            | 0.921            |
|                  | **Reading**       | **Standard deviation** |
|                  | 1.00              | 0.986            | 0.963            | 0.907            |

NOTE: The correlations in the second to fourth columns are correlations between value-added estimates using data with a given ceiling imposed and value-added estimates based on the original data set. The standard deviation of teacher effects is presented as a percentage of the standard deviation of teacher effects when no additional ceiling effect is imposed.

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

Given how rare test score ceilings are in practice, to understand how test score ceilings affect value added, we investigated how value-added estimates changed when we artificially imposed a ceiling effect that would affect 10, 20, and 40% of students. The first set of results is shown in Table 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.

The decrease in the standard deviation shows how the value-added estimates of teachers are affected when we impose a more stringent ceiling effect on student test scores. As it becomes increasingly difficult to distinguish among teachers when our measures of the baseline and final achievement of students become coarser, the dispersion of value-added estimates becomes smaller. This result is broadly consistent with results discussed in Koedel and Betts (2010).

We also see, consistent with our prediction, that teachers tend to move toward the middle of the distribution. The scatter plots shown in Figure 2 show how individual teacher effects change. The scatter plots 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% 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. 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 a ceiling effect is imposed, and those with high value-added estimates tend to move down. A scatterplot for math shows the same general results as the reading scatter plot. Since a value-added model like this is by definition at mean zero, there is no net change overall in the value-added estimates.

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. To see this effect, we show scatter plots for teachers of students with pre-test scores in the bottom 80% of reading teachers (Figure 3) and the top 20% (Figure 4), keeping the original regression line on both figures for comparison. As can be seen by comparing the two figures, the largest changes in estimates of teacher effects occur when teachers have students in the top 20% of the distribution of pre-test achievement. Results for math followed a similar pattern.

The change in measured effectiveness of teachers of high-performing students does not imply that they are necessarily average teachers, but rather that, given our ignorance of the true achievement levels of many of their students, these teachers cannot be distinguished from the average teacher. 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 imprecise as well. Notably, even with extreme ceiling effects, while the value added of some teachers who would otherwise have received high scores can be pushed down, they are pushed down to a middle range of estimates, not to the bottom. None of the teachers with value-added estimates in the bottom quintile of the distribution with extreme ceiling effects would have had an above-average value-added estimate in the absence of ceiling effects.

The policy implication is that teachers with a large number of students scoring at the test ceiling are unlikely to be affected by an evaluation system, 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 a well-designed teacher evaluation system, like that of the District of Columbia Public Schools (2013), pushing a teacher’s estimates toward zero would place the teacher in a “no consequences” category for outcomes. While not ideal, in these circumstances this is an appropriate categorization. It is somewhat similar to the way in which empirical Bayes shrinkage, a common final step in the preparation of value-added estimates, is used to handle imprecise estimates that arise from teachers who have a small number of students with valid test score data or many disadvantaged students whose post-test scores are relatively hard to predict (Morris 1983; Herrmann, Walsh, and Isenberg 2016). The empirical Bayes shrinkage step moves extreme estimates toward the mean 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: a school district would rather have more precise estimates of 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 a ceiling effect in student test scores, value-added models function in much the same way.  

6. Conclusion

We have examined test score ceilings to investigate concerns about how this phenomena affects value-added models of

---

4 One corollary of this finding is that computer-adaptive tests like the Smarter Balanced Assessment, which are designed to better pinpoint student achievement in any part of the distribution, should be better able to distinguish teacher effectiveness compared to conventional paper-and-pencil tests.
teacher effectiveness. Our findings suggest that the conventional wisdom about test score ceilings may be skewed, starting with the assumption that a test score ceiling denotes the bottom of the range of a student’s true score. In fact, in the test we examined, published by a major test publisher that has designed dozens of state tests, a score at the test ceiling is assigned to estimate a student's ability in such a way that their true score may be higher or lower than their observed scale score.

A second issue with the conventional wisdom on test score ceilings is that teachers of gifted classes will necessarily be severely penalized by a value-added model. This argument appears to rest on the assumption that value-added models are gains models, implying that teachers need to show above-average gains to demonstrate that they are above-average teachers. In fact, teachers need to beat the average predicted score of their students. In the data we analyzed, the teacher with the largest percentage of students scoring at the test score ceiling was still able to obtain the top value-added estimate in the district.

Finally, to understand how test score ceilings may affect value added when they arise, we examined an extreme case, in which 40% of students were affected by a ceiling effect. 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. So, far from the existence of “convincing and detailed evidence of VAM bias against teachers...with high-performing students” (Lederman v King 2016), in fact we show how test score ceilings—which represent a high degree of uncertainty about student achievement—push value-added estimates not toward the bottom but rather toward the middle of the distribution of teachers. In general, teachers who would have had high value-added estimates are pushed down toward the middle, and teachers who would have had low value-added estimates are pulled up. This should not be interpreted as a statistical judgment that these teachers’ true performance is average but rather indicates ignorance of their effectiveness.

Acknowledgments

We would like to thank Duncan Chaplin, Andrew Swanlund, and two anonymous referees for their comments. We thank Nikhil Gahlawat for providing expert programming.

References

American Educational Research Association (2015), “AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs,” Educational Researcher, 44, 448–452. [1]

American Institutes for Research (2013), Florida Comprehensive Assessment Test (FCAT) 2.0 Value-Added Model: Technical Report 2012–13, Washington, DC: American Institutes for Research. [3]

Amrein-Beardsley, A., and Collins, C. (2012), “The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and Unintended Consequences,” Education Policy Analysis Archives, 20, 1–27. [1]

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., http://www.nap.edu/read/12780/chapter/1#3

Buonaccorsi, J. P. (2010), Measurement Error: Models, Methods, and Applications, Boca Raton, FL: Chapman & Hall/CRC. [3]

Darling-Hammond, L. (2015), “Can Value Added Add Value to Teacher Evaluation?” Educational Researcher, 44, 132–137. [1]

District of Columbia Public Schools (2013), IMPACT: The District of Columbia Public Schools Effectiveness Assessment System for School-Based Personnel 2013–2014, Group 1: General Education Teachers with Individual Value-Added Student Achievement Data, Washington, DC: District of Columbia Public Schools. [5]

Embretson, S. E., and Reise, S. P. (2000), Item Response Theory for Psychologists, Mahwah, NJ: Lawrence Erlbaum Associates. [2]

Haertel, E. (2013), Reliability and Validity of Inferences about Teachers Based on Student Test Scores (Angoff Lecture), Princeton, NJ: Educational Testing Service. Available at https://www.ets.org/Media/Research/pdf/PICANG14.pdf Accessed May 17, 2016. [1]

Herrmann, M., Walsh, E., and Isenberg, E. (2016), “Shrinkage of Value-Added Estimates and Characteristics of Students with Hard-to-Predict Achievement Levels,” Statistics and Public Policy, 3, 1–10. [5]

Isenberg, E., and Walsh, E. (2014), “Measuring Teacher Value Added in DC, 2013–2014 School Year,” Final report submitted to the Office of the State Superintendent of Education for the District of Columbia and the District of Columbia Public Schools, Washington, DC: Mathematica Policy Research. [3]

Isenberg, E., Max, J., Gleason, P., Johnson, M., Deutsch, J., and Hansen, M. (2016), “Do Low-Income Students Have Equal Access to Effective Teachers?” NCEE 2017–4007, Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. [4]

Koedel, C., and Betts, J. (Winter 2010), “Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation,” Education Finance and Policy, 5, 54–81. [3,4]

Lederman v King (2016), Available at http://law.justia.com/cases/new-york/other-courts/2016/2016-ny-slip-op-26416.html. Accessed September 22, 2017. [1,6]

Morris, C. N. (1983), “Parametric Empirical Bayes Inference: Theory and Applications,” Journal of American Statistical Association, 78, 47–55. [3,5]

National Research Council (2009), Letter Report to the U.S. Department of Education on the Race to the Top Fund, Washington, DC: The National Academies Press. Available at http://www.nap.edu/read/12780/chapter/1#3. Accessed May 17, 2016. [1]

Ravitch, D. (2014), “With VAM: All Teachers of the Gifted Are ‘Bad’ Teachers.” Diane Ravitch’s blog, March 24. Available at http://dianeravitch.net/2014/03/24/with-vam-all-teachers-of-the-gifted-are-bad-teachers/. Accessed May 17, 2016. [1]

Resch, A., and Isenberg, E. (2014), “How Do Test Scores at the Floor and Ceiling Affect Value-Added Estimates?,” Mathematica Policy Research Working Paper No. 33, Chicago, IL: Mathematica Policy Research. [2]

Rott, D., Johnson, M., and Gill, B. (2014), Value-Added Models for the Pittsburgh Public Schools, 2012–13 School Year, Cambridge, MA: Mathematica Policy Research. [3]

Value Added Research Center (2014), Hillsborough County Public Schools Value-Added Project: Year 3 Technical Report HCPS Value-Added Models, Madison, WI: Value-Added Research Center at Wisconsin Center for Education Research, University of Wisconsin-Madison. [3]