Test-Score Ceiling Effects and Value-Added Measures of School Quality

Cory Koedel[1], Julian R. Betts[2]
[1]University of Missouri, 327 Professional Building, Columbia, MO 65211
[2]University of California, San Diego, 9500 Gilman Drive #0508, La Jolla, CA 92093

Abstract

This paper uses administrative data to evaluate how test-score ceiling effects influence value-added estimation. There is no evidence of a test-score ceiling in the raw data, which is particularly appealing for this project. Starting with the no-ceiling baseline, we consider the effects of numerous artificially imposed test-score ceilings on school rankings based on value-added. Over a wide range of test-score ceiling severity, school-level value-added estimates are only negligibly influenced by ceiling effects. However, schools' value-added rankings are significantly altered by ceilings equivalent in severity to those found in minimum-competency testing environments.

Key Words: Value-Added Modeling, Test-Score Ceiling Effects

1. Introduction

Performance-based measures of schooling effectiveness are quickly gaining momentum in the United States. The measure of performance that has received the most attention from policymakers of late, and is perhaps the most contentious, is value-added to students' test scores. In this paper, we investigate the extent to which value-added estimates of schooling effectiveness are sensitive to test-score-ceiling effects. Our school-level analysis is extended to look at teacher value-added in Koedel and Betts (2008).

We refer to the tendency for gains in a student's test score to be smaller if the student's initial score is toward the top end of the distribution, simply because the student has little room for improvement given the difficulty level of the test, as a "ceiling effect". Ceiling effects will be most pronounced in minimum-competency or proficiency-based tests, which are being used increasingly across the United States. For example, 22 states nationwide use high school exit exams that are typically pitched at a middle-school or lower high-school level.[1] Furthermore, because federal No Child Left Behind (NCLB) legislation focuses largely on proficiency, mainstream proficiency-based testing is also becoming increasingly common.

The increased focus on proficiency in education coincides with the growing interest from researchers and policymakers in value-added as a tool for measuring schooling performance. The impending collision of ceiling-affected testing instruments with value-added-based evaluations motivates our analysis. Do ceiling effects influence value-added estimation? If so, how important are ceiling effects and how severe must they be to significantly alter value-added results?

We answer these questions using a testing instrument where there is no evidence of a test-score ceiling. Starting with our no-ceiling baseline, we simulate test-score ceilings that vary in severity and evaluate their effects on value-added estimation. Our findings are generally encouraging - over a wide range of test-score-ceiling severity we find that value-added estimates are roughly impervious to ceiling effects. However, ceiling conditions approaching the severity of those found in minimum-competency testing environments noticeably alter value-added results.

[1] The nationwide count applies to 2006 and was calculated based on information in Warren (2007).

2. Measuring Test-Score Ceiling Effects

Test-score ceilings structurally restrict students' test-score gains as test-score levels rise. Because a test-score ceiling directly influences the tool by which value-added is measured, it is intuitive that it will influence results. For example, consider a testing instrument where the top 20 percent of the student population is at or near the maximum possible score. Teachers and schools charged with raising these students' test scores will have little opportunity to add value. Furthermore, they are likely to use advanced curricula that focus at least partly on material that goes beyond the scope of the test, making test-based evaluations uninformative.

In practice it might be quite important whether a district uses a norm-referenced or a criterion-referenced test for the purpose of evaluating schooling effectiveness. A norm-referenced test is a standardized test that is meant to estimate where a student ranks against the test-score distribution of the reference group, typically the national student population. Such a test, if well-designed, should exhibit few ceiling effects because it must include questions with a range of difficulty so that distinctions can be made among students throughout the test-score distribution. Such tests have been in use for many decades.

More recently, partly as a consequence of NCLB, many states are using testing systems designed to measure student understanding of the content standards set by the state's Department of Education. We speculate that these "criterion-referenced" tests are more likely to exhibit ceiling effects, particularly when a state exam is intended, either explicitly or implicitly, to serve as a "minimum-competency" test. For example, in Mississippi the state-level test appears to be targeted at a fairly low level. In 2006-07, 90 percent of fourth-grade students in Mississippi scored at or above the "proficient" level in reading based on the state-level Mississippi Curriculum Test (MCT). However, just 19 percent of these students scored at or above the proficient level on the National Assessment of Education Progress (NAEP).[2,3]

One way to evaluate the impact of ceiling effects on value-added would be to find a population of students that had been tested in several consecutive years using two testing systems – one that lacked a ceiling effect and another that suffered from a ceiling effect. However, it is likely that the different tests in such a scenario would also differ in terms of content, confounding the ceiling effect. A second approach is to use a test that can be demonstrated not to suffer from ceiling effects, and then to simulate test-score ceiling effects using that instrument. This creates a counterfactual of what would have happened had the test been right-censored. We adopt this approach by using Stanford 9 math test scores for fourth grade students in the San Diego Unified School District. The Stanford 9 is a nationally norm-referenced test. For the population we study we find no evidence of a ceiling effect (see below). It thus provides a way of comparing value-added estimates of schooling effectiveness with and without a test-score ceiling.

The first step in our analysis is to provide a reliable measure of test-score ceiling severity. An intuitive approach would be to evaluate the strength of the negative relationship between test-score levels and subsequent test-score gains. However, this approach is problematic because a negative relationship will exist due to regression to the mean even in the absence of a test-score ceiling. Furthermore, in cases where a test-score ceiling does exist, there is no obvious way to dissect the negative relationship between test-score levels and test-score gains to isolate the ceiling effect.

As an alternative, we propose that the distribution of students' test scores can be used to measure test-score-ceiling severity. Specifically, we can use the degree of negative skewness in the test-score distribution as originally suggested by Roberts (1978). We define skewness as the sample analog of $\frac{E\left(x-E\left(x\right)\right)^3}{\left[E\left(x-E\left(x\right)\right)^2\right]^{3/2}} \equiv \frac{\mu_3}{\sigma^3}$, where μ_3 is the third

[2] From the US Department of Education, *Mapping Mississippi's Educational Progress 2008*.

[3] Cullen and Loeb (2004) illustrate another source of ceiling effects that is directly associated with NCLB – reporting requirements that require states to document the percentage of students who are "proficient". Their Figure 12c provides a graphical representation of the mechanical relationship between underlying proficiency levels and growth in proficiency. Clearly, if value-added were estimated based on simple pass-fail measures of student achievement, as emphasized by NCLB, ceiling effects would be severe.

moment about the mean and σ is the standard deviation. One appeal of skewness as a measure of test-score-ceiling severity is that it can be easily compared across testing environments. Furthermore, Koedel and Betts (2008) provide suggestive (although not exhaustive) evidence that skewness is a rather robust measure of test-score-ceiling severity.

Figure 1 displays the frequency distributions of students' lagged (grade 3) and current (grade 4) test scores from our data. As mentioned above, there is no evidence of a test-score ceiling. In fact, the test-score distributions from our sample are skewed mildly *positively*. The figure shows kernel-density plots contrasted with normally-distributed overlays. The skewness in the lagged and current-score distributions in our data are 0.25 and 0.17, respectively. Notice that although both of these distributions are skewed slightly positively, they both closely mirror their normally distributed analogs.

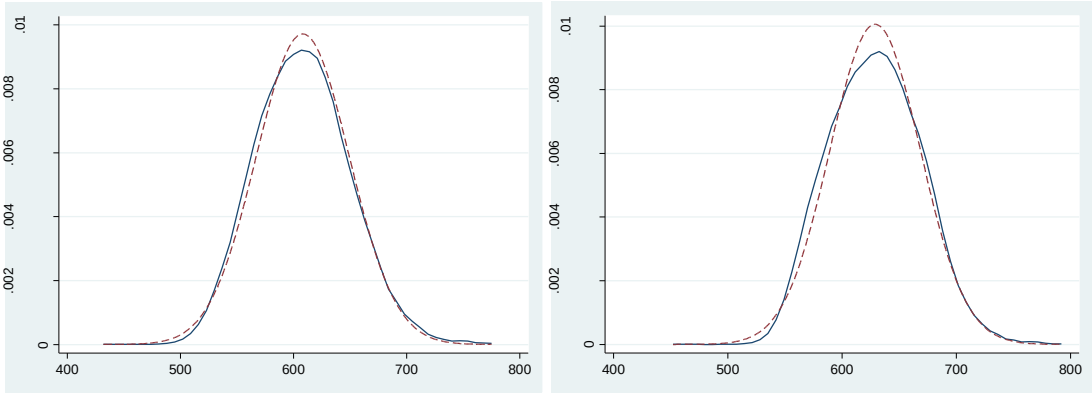

Figure 1: Frequency Distributions of Lagged and Current Math Test Scores from our Data Contrasted with Normal-Distribution Overlays

Notes
LEFT: Kernel-density plot of lagged test scrore distribution – skewness ≈ 0.25
RIGHT: Kernel-density plot of current test scrore distribution – skewness ≈ 0.17
In each graph the solid line represents the distribution of actual scores and the dotted line the normal-distribution overlay. Estimates are calculated using the Epanechnikov kernel with a bandwidth equal to 2.5 percent of the range of test scores.

In our test-score-ceiling simulations, what is the relevant range of skewness to consider? We answer this question using two large-scale, state-level tests: (1) the Texas Assessment of Academic Skills (TAAS) and (2) the Florida Comprehensive Assessment Test (FCAT).[4] The TAAS was administered in Texas from 1991 to 2003 and prior to 1991 was known as the Texas Educational Assessment of Minimum Skills. The minimum-competency-based design of the TAAS makes it a useful test upon which to base our most severe test-score-ceiling simulations. The FCAT was first administered in 1998 in Florida and continues to serve as the state-level standardized test there.

We simulate test-score-ceiling conditions based on the skewness in the test-score distributions of the math portions of the TAAS and FCAT from 2002 and 2007, respectively. Figure 2 shows kernel-density plots of third and fourth-grade mathematics scores on the TAAS compared to normally-distributed overlays based on 2002 test scores (statewide). The skewness in these score distributions are large and negative, at -1.60 and -2.08, respectively. Similar plots for FCAT scores are available in Koedel and Betts (2008). The skewness in the score distributions from the third and fourth-grade FCAT are negative but much milder, at -0.46 and -0.55. For ninth and tenth-grade students, the skewness in the test-score distributions from the FCAT become increasingly negative, at -0.94 and -1.99 respectively.[5]

[4] Statewide distributions of test scores for the TAAS were provided online by the Texas Education Agency (http://www.tea.state.tx.us). FCAT scores were provided by the Florida Department of Education.
[5] Students in Florida must pass the math portion of the tenth-grade FCAT to receive a high-school diploma. This may help to explain the large jump in negative skewness moving from the ninth to tenth-grade version of the exam.

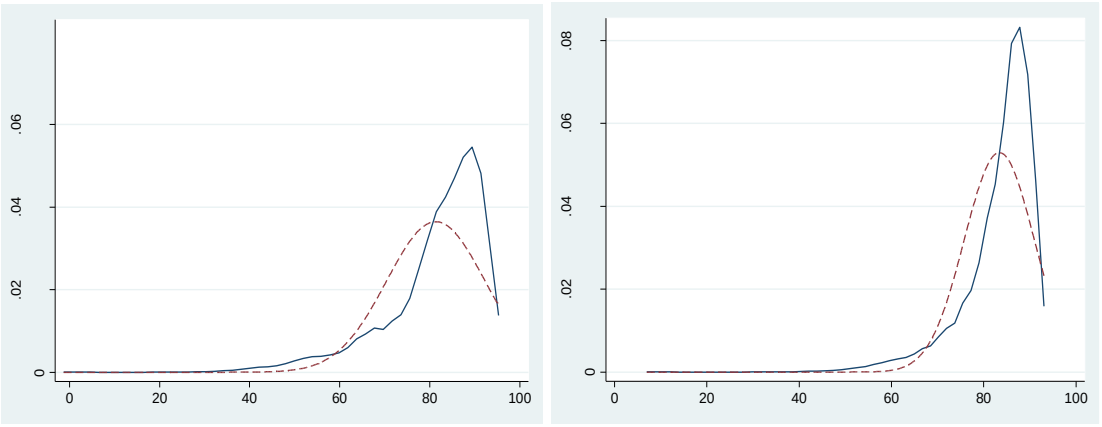

Figure 2: Frequency Distributions of Third and Fourth-Grade Math Scores from the TAAS in 2002 Contrasted With Normal-Distribution Overlays

Notes
LEFT: Kernel-density plot of third-grade test scrore distribution – skewness ≈ -1.60
RIGHT: Kernel-density plot of fourth-grade test scrore distribution – skewness ≈ -2.08
In each graph the solid line represents the distribution of actual scores and the dotted line the normal-distribution overlay. Estimates are calculated using the Epanechnikov kernel with a bandwidth equal to 2.5 percent of the range of test scores.

Starting with our no-ceiling baseline, we create counterfactual testing environments where students' scores are impeded by test-score ceilings of varying severity. Our most severe ceiling simulation is designed to mimic the testing conditions from the fourth-grade TAAS and tenth-grade FCAT. For simplicity, we simulate what we will refer to as "hard" test-score ceilings, where students' scores are restricted at a specific maximum score. An alternative would be to simulate "soft" test-score ceilings that restrict student performance throughout the test-score distribution. For example, students' scores might taper off as they approach a maximum score. Soft test-score ceilings appear to characterize more accurately the true distributions of test-scores in Figure 2. However, there are literally an infinite number of possible soft-ceiling structures that could generate the observed skewness in the TAAS and FCAT distributions, making such an analysis infeasible. Instead, we focus on hard test-score ceilings and compare the results that we obtain from our simulations to a set of results generated using one possible soft-ceiling structure. This analysis is available in Koedel and Betts (2008) and suggests that similarly skewed test-score distributions have similar implications for value-added results, regardless of whether a hard or soft ceiling generates the ceiling effect.[6]

3. Value-Added Model

We estimate school-level value-added using a general value-added model where current test scores are regressed on lagged test scores. It is somewhat common in the literature to use a specific form of the value-added model, the gainscore model, where the coefficient on the lagged test score is forced to one and the lagged-score term is moved to the left side of the equation. Although we do not present results from gainscore models, our findings are nearly identical using the gainscore framework. Results from the gainscore analogs to the below specification are available from the authors upon request.

We use the following model to estimate school-level value-added:

$$Y_{it} = \phi_t + Y_{i(t-1)}\phi_1 + X_{it}\phi_2 + S_{it}\delta + \varepsilon_{it} \quad \textbf{(1)}$$

In (1), Y_{it} is the test score for student i in year t and ϕ_t is a year-specific intercept. X_{it} is a vector of student-specific characteristics including race, gender, parental-education status, English-learner status, the share of days that the

[6] This analysis provides some evidence that skewness is a robust measure of test-score ceiling severity.

student is absent from school and indicator variables for students who switched schools during the fourth grade or were designated as an advanced student. X_{it} also includes measures of the lagged performance of students' classroom-level peers and class-size controls.[7] S_{it} is a vector of school indicator variables where the entry for the school attended by student *i* in year *t* is set to one. The coefficients of interest are in the vector of school effects, δ.

4. Data

We evaluate ceiling effects using administrative data from fourth-grade students in San Diego (San Diego Unified School District) who were in the fourth grade between 1998-1999 and 2001-2002. The standardized test that we use to measure student achievement is the Stanford 9 mathematics test. The Stanford 9 is designed to be vertically scaled such that a one-point gain in student performance at any point in the schooling process is meant to correspond to the same amount of learning. As discussed in Section 2, there is no evidence of a ceiling effect in the test-score data.

Students who have fourth grade test-scores and lagged test-scores are included in our analysis. Our final sample includes test-score records for 30,354 students taught in 116 elementary schools in San Diego. We include students who repeat the fourth grade because our objective is to inform policy and it is unlikely that grade repeaters would be excluded from school evaluations in practice (because of moral hazard concerns). In our original sample of 30,354 students with current and lagged test-score records, just 199 are grade repeaters.

The degree of across-school student sorting will influence the magnitude of the test-score-ceiling effects. At one extreme, random assignment of students across schools will mitigate test-score-ceiling effects insofar as they determine school rankings. At the other, a test-score ceiling where there is strong student-school sorting should lead to a large shift in school rankings based on value-added. One benefit of our analysis is that we can use real student-school matches from a real school district, rather than attempting to simulate the sorting. This is important because there is no consensus in the literature as to how students sort into schools, making it impossible to artificially generate student-school matches. However, if parents, students, teachers and administrators in San Diego act similarly to parents, students, teachers and administrators in other similar school districts, our results will generalize.[8]

We document *observable* student-school sorting in our data by comparing the average realized within-school standard deviation of students' lagged test scores to analogous measures based on simulated student-school matches that are either randomly generated or perfectly sorted. This approach follows Aaronson, Barrow and Sander (2007). Table 1 details our results, which are presented as ratios of the standard deviation of interest to the standard deviation of the test (calculated based on our student sample). Note that while there does appear to be some student sorting across schools based on lagged test-score performance, this sorting is relatively mild.

[7] It is arguable whether these controls should be included in the model given the objective of measuring school effects. The primary results are not sensitive to whether or not these controls are included.

[8] The San Diego Unified School District (SDUSD) is the eighth largest school district in the nation, with considerable student diversity. The one notable difference between SDUSD and some other districts is that there is a larger-than-average share of English learners at SDUSD. For basic demographic information about the student population at SDUSD see Betts, Zau and Rice (2003).

Table 1. Average Within-School Standard Deviations of Students' Period (t-1) Test Scores

	Actual	Random Assignment	Perfect Sorting
Standard Deviations of Lagged Scores	0.90	1.00	0.02

Note: In the "Perfect Sorting" column students are sorted by period (t-1) test-score levels in math. For the randomized assignment, students are assigned to schools based on randomly generated numbers from a uniform distribution. The random assignments are repeated 5 times and the estimate is averaged across all random assignments. The estimates from the simulated random assignments are very stable across simulations.

5. Results

Our test-score ceilings are simulated based on the distribution of students' test scores in the fourth grade. For example, one of our simulations imposes a ceiling where the maximum score is set at the 95th percentile of the fourth grade test-score distribution. Because the Stanford 9 is vertically scaled, this ceiling definition spills over to third grade scores. That is, if a student in the third grade scores above the 95th percentile in the distribution of fourth-grade scores, her third-grade score is set at the maximum. Our approach generates negative skewness in the test-score distributions for each grade. The skewness will be more pronounced in the fourth grade relative to the third grade, and in the third grade relative to the second grade. After imposing each test-score ceiling on our data, we re-standardize students' test scores within grades to have a mean of zero and a variance of one.[9]

We create each test-score ceiling by imposing a maximum possible score that we do not allow students' scores to exceed. We consider test-score ceilings where the maximum score ranges from the 97th percentile to the 33rd percentile of the original distribution of fourth-grade scores. This latter ceiling generates skewness in the current and lagged test-score distributions comparable to skewness from the third and fourth-grade TAAS exams in 2002. For each ceiling simulation, we report the skewness of the generated test-score distributions.

Table 2 shows our results. The first column in the table shows results from the no-ceiling baseline and the seventh column shows results from the most severely skewed simulation. For each ceiling simulation we report the skewness measures and the correlation between schools' ceiling-affected value-added estimates and estimates from the baseline model without ceiling effects.

Table 2 shows that schools' value-added estimates are roughly impervious to test-score-ceiling effects over a wide-range of ceiling conditions. This can be seen by looking at the correlations between the school effects estimated using the actual test-score data and those estimated after the ceilings are imposed. Notice that even the ceiling that affects students' test scores starting at the 75th percentile is largely inconsequential (skewness $\approx$ -0.64). However, value-added results begin to respond to ceiling effects as the ceilings continue to increase in severity. For instance, when the ceiling begins at the 50th percentile of the fourth-grade test-score distribution, the correlation between the school-effect estimates from the actual data and the data with the ceiling imposed falls below 0.90. The correlation drops further when we impose the ceiling at the 33rd percentile, to 0.79. As ceiling conditions approach those found in minimum-competency testing environments, value-added results are non-negligibly altered.

[9] An alternative approach would have been to separately set the ceilings in the 2nd, 3rd and 4th grades such that each ceiling is imposed at the 95th percentile of its respective distribution. However, this approach is inconsistent with the evidence from the TAAS and, more mildly, the FCAT, where later-grade test-score distributions are more skewed.

Table 2. Test-Score-Ceiling Effects on Value-Added Results

	(1)*	(2)	(3)	(4)	(5)	(6)	(7)
Percentile of Fourth-Grade Test-Score Distribution Where Ceiling is Set	99.96	97	95	85	75	50	33
Skewness of Period-t Score Distribution	0.17	-0.02	-0.07	-0.37	-0.64	-1.31	-2.00
Skewness of Period-(t-1) Score Distribution	0.25	0.11	0.07	-0.13	-0.32	-0.83	-1.32
Correlation Between Ceiling-Restricted Value-Added Estimates and Baseline	-	1.00	1.00	0.98	0.96	0.88	0.79

* Column (1) shows results from the no-ceiling baseline. Of course, a ceiling is not "set" here – 0.04 percent of the student population attains the maximum possible score.

Differences in skewness across grades may also affect value-added results. Koedel and Betts (2008) evaluate this possibility in the context of teacher value-added. Their results show that over a wide range of differences-in-skewness across grades, the effect of a given test-score ceiling on value-added results is roughly constant.

6. Concluding Remarks

We evaluate the extent to which test-score-ceiling effects influence estimates of school-level value-added. In the current climate of proficiency-based educational reform, test-score ceilings are likely to be increasingly common. Our findings are generally encouraging – given a wide range of test-score ceiling conditions, some of which might be casually identified as severe, value-added estimates are only negligibly affected. However, researchers and policymakers should be concerned when working in minimum-competency or proficiency-based testing environments. We show that ceiling conditions in such environments can significantly alter school rankings based on value-added.

A more detailed analysis of test-score ceiling effects, focused on how ceiling effects influence teacher value-added, can be found in Koedel and Betts (2008).

Acknowledgements

* The authors thank Andrew Zau and many administrators at San Diego Unified School District, in particular Karen Bachofer and Peter Bell, for helpful conversations and assistance with data issues. We also thank Yixiao Sun, Julie Cullen and Nora Gordon for their useful comments and suggestions, and the Spencer Foundation and the National Center for Performance Incentives for research support. The underlying project that provided the data for this study has been funded by a number of organizations including The William and Flora Hewlett Foundation, the Public Policy Institute of California, The Bill and Melinda Gates Foundation, the Atlantic Philanthropies and the Girard Foundation. None of these entities has funded the specific research described here, but we warmly acknowledge their contributions to the work needed to create the database underlying the research.

References

Aaronson, Daniel, Lisa Barrow and William Sander. 2007. Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25:95-135.

Betts, Julian, Andrew Zau, and Lorien Rice. 2003. *Determinants of Student Achievement, New Evidence from San Diego*. Public Policy Institute of California.

Cullen, Julie Berry and Susanna Loeb. 2004. School finance feform in Michigan: Evaluating Proposal A, in J. Yinger (Ed.), *Helping Children Left Behind: State Aid and the Pursuit of Educational Equity.* Cambridge, MA: MIT Press. pp. 215-250.

Koedel, Cory and Julian R. Betts. 2008. Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation. Working Paper, University of Missouri, Columbia.

Roberts, Sarah Jane. 1978. Test Floor and Ceiling Effects. ESEA Title I Evaluation and Reporting System. Research report, RMC Research Corporation.

Warren, John Robert, *State High School Exit Examinations for Graduating Classes Since 1977*. Minneapolis, MN: Minnesota Population Center, 2007. Available at www.hsee.umn.edu.