VALUE ADDED TO WHAT? HOW A CEILING IN THE TESTING INSTRUMENT INFLUENCES VALUE-ADDED ESTIMATION

Cory Koedel
(corresponding author)
Department of Economics
University of Missouri
118 Professional Building
Columbia, MO 65211
koedelc@missouri.edu

Julian Betts
University of California,
San Diego, and
National Bureau
of Economic Research
Department of Economics
9500 Gilman Drive
La Jolla, CA 92093-0508
jbetts@ucsd.edu

Abstract

Value-added measures of teacher quality may be sensitive to the quantitative properties of the student tests upon which they are based. This article focuses on the sensitivity of value added to test score ceiling effects. Test score ceilings are increasingly common in testing instruments across the country as education policy continues to emphasize proficiency-based reform. Encouragingly, we show that over a wide range of test score ceiling severity, teachers' value-added estimates are only negligibly influenced by ceiling effects. However, as ceiling conditions approach those found in minimum-competency testing environments, value-added results are significantly altered. We suggest a simple statistical check for ceiling effects.

© 2009 American Education Finance Association

1. INTRODUCTION

Teacher performance pay is quickly gaining momentum in the United States. In fact, some districts, and even entire states, are already implementing performance pay programs for teachers that involve sizable public expenditures. For example, the Texas Governor's Educator Excellence Award Programs (GEEAP) allot a large fraction of their combined $330 million annual budget to directly reward classroom teachers based on performance (Podgursky and Springer 2007).

The aspect of teacher performance that has received the most attention from policy makers of late, and is perhaps the most contentious, is value added to students' test scores. While the literature overwhelmingly indicates that there are important differences in teacher quality measured by value added, there is little consensus on the best approach for estimating value added. Furthermore, there is ample evidence that value-added measures of teacher quality are noisy, which creates some concern about the feasibility of using value added for large-scale teacher evaluation.[1] In addition to these unresolved issues, value-added estimates may be sensitive to the quantitative properties of the testing instruments upon which they are based.

This article evaluates the sensitivity of value added to a particularly relevant testing instrument property—the severity of test score ceiling effects. We refer to a "ceiling effect" as 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. Ceiling effects will be most pronounced in minimum-competency or proficiency-based tests, which are being used increasingly across the United States. For example, twenty-two states nationwide use high school exit exams that are typically pitched at a middle school or lower high school level.[2] 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 policy makers in value added as a tool for measuring teacher performance. The impending collision of ceiling-affected testing instruments with value-added-based teacher 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?

[1] See, for example, Aaronson, Barrow, and Sander (2007), Hanushek et al. (2005), Koedel and Betts (2007), and Rockoff (2004). In addition, Rothstein (forthcoming) shows that value-added estimates may be biased by student-teacher sorting.

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

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 teacher value added. 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.

2. TEST SCORE CEILINGS: INTRODUCTION AND MEASUREMENT

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 a large fraction of the student population is at or near the maximum possible score. Teachers teaching these students will have little opportunity to add value to test scores. Furthermore, they are likely to use advanced curricula that focus at least partly on material that goes beyond the scope of the test, making their evaluations based on the test 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 teaching 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 aimed at a fairly low level. In 2006–7, 90 percent of fourth-grade students scored at or above the "proficient" level in reading 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).[3]

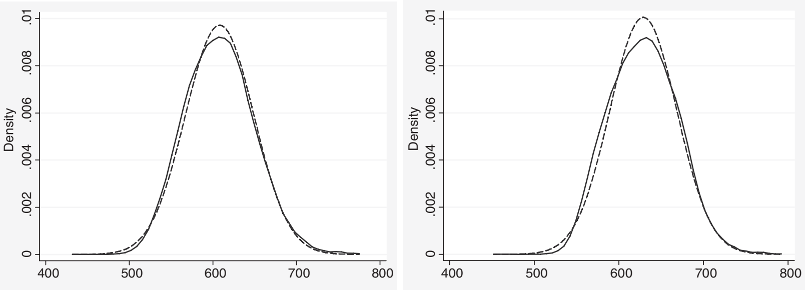
[3] From USDOE (2008). Cullen and Loeb (2004) illustrate another source of ceiling effects that is directly associated with NCLB—reporting requirements that require states to document the

One way to evaluate the impact of ceiling effects on teacher 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 measures of teacher value added 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 in which 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 $\frac{E(x-E(x))^3}{[E(x-E(x))^2]^{3/2}} \equiv \frac{\mu_3}{\sigma^3}$, where μ_3 is the third moment about the mean and σ is the standard deviation. Under the assumption that underlying student achievement in the population is symmetrically distributed, skewness provides an intuitive and straightforward measure of test score ceiling severity. In section 8 below, we provide suggestive (although not exhaustive) evidence that skewness is a robust measure of ceiling severity.

Figure 1 displays the frequency distributions of students' lagged (grade 3) and current (grade 4) math test scores from our data, gathered from the San Diego Unified School District. 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 of the distributions of actual scores contrasted with normally distributed overlays.

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.



Left: Kernel-density plot of lagged test score distribution – skewness ≈ 0.25
Right: Kernel-density plot of current test score 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.

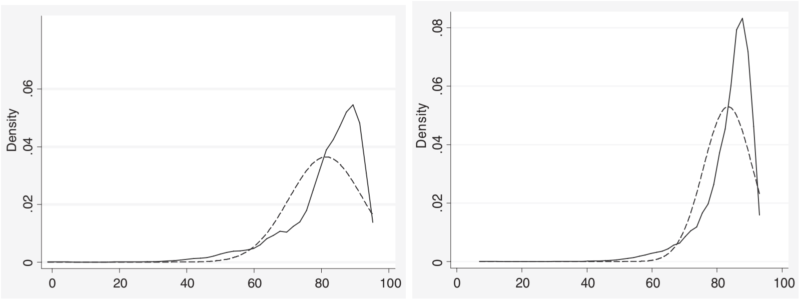
Figure 1. Frequency Distributions of Lagged and Current Math Test Scores from Our Data Contrasted with Normal Distribution Overlays

The skewness in the lagged and current score distributions in our data are 0.25 and 0.17, respectively. Notice that although both these distributions are skewed slightly positively, they both closely mirror their normally distributed analogs.

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: the Texas Assessment of Academic Skills (TAAS) and 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 with normally distributed overlays based on 2002 test scores (statewide). The skewness in these score distributions is large and negative, at −1.60 and −2.08, respectively. Similarly, the top panel of figure 3 shows kernel density plots of third- and fourth-grade mathematics scores on the 2007 FCAT (statewide). The skewness in these score distributions is also negative but much milder, at −0.46 and −0.55. Finally, the bottom panel of figure 3 shows the distributions of scores for ninth- and tenth-grade students on the FCAT in 2007 where the skewness

[4] Statewide distributions of test scores for the TAAS were provided online by the Texas Education Agency (www.tea.state.tx.us). FCAT scores were provided by the Florida Department of Education.

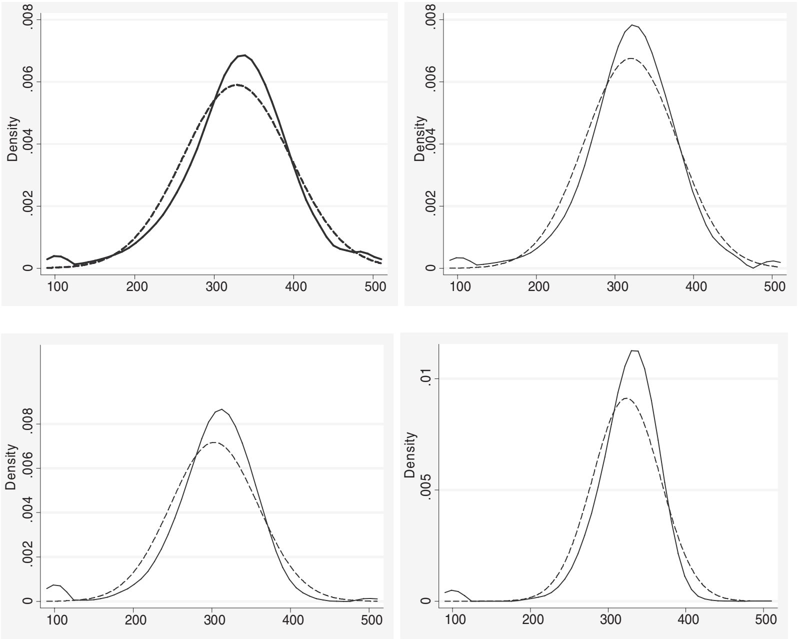


Left: Kernel-density plot of third-grade test score distribution – skewness ≈ –1.60

Right: Kernel-density plot of fourth-grade test score 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.

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



Upper left: Kernel-density plot of third-grade test score distribution – skewness ≈ –0.46

Upper right: Kernel-density plot of fourth-grade test score distribution – skewness ≈ –0.55

Lower left: Kernel-density plot of ninth-grade test score distribution – skewness ≈ –0.94

Lower right: Kernel-density plot of tenth-grade test score distribution – skewness ≈ –1.99

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.

Figure 3. Frequency Distributions of Third-, Fourth-, Ninth-, and Tenth-Grade Math Scores from the FCAT in 2007 Contrasted with Normal Distribution Overlays

in the test score distributions becomes increasingly negative. The ninth- and tenth-grade score distributions from the FCAT have skewness of −0.94 and −1.99, respectively.[5]

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. 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 figures 2 and 3. 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 we obtain from our simulations with a set of results generated using one possible soft-ceiling structure. This analysis is detailed in section 8 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.

Finally, we distinguish two mechanisms by which test score ceiling effects will influence value-added estimation. First, most straightforwardly, ceiling effects represent lost information about student learning. The more severe the test score ceiling, the greater the amount of lost information. Second, ceiling effects will result in model misspecification. A test score ceiling is a data censor, and as such the typical value-added approach is improperly specified in the presence of a ceiling. In practice this is a nontrivial problem because the underlying data-censoring structure will be unknown. Furthermore, the censoring problem is even more complicated in the value-added framework than in the typical dependent-variable censoring problem because lagged test scores will also be censored. In the general value-added approach (where current test scores are regressed on lagged test scores), this means that there will be censoring of an independent variable in addition to the censoring of the dependent variable. Converting to gain scores does not circumvent the problem because censoring will be ill defined—censored gain scores will have

[5] Students in Florida must pass the math portion of the tenth-grade FCAT to receive a high school diploma. It is possible that the exam is aimed at a lower level because of this. In addition, students are allowed to take the test more than once. The distribution of tenth-grade FCAT scores reported in figure 3 is for all tests taken in 2007 (provided by the Florida Department of Education), which will include retaken exams. The retaken exams could either positively or negatively skew the distribution.

zero or near-zero gains, but non-censored scores can also have zero, near-zero, or even negative gains.

The current state of the data-censoring literature in econometrics and statistics is such that there is no solution to the data-censoring problem in this context. Therefore, distortionary test score ceiling effects can be thought of as the product of both of these problems—lost information and model misspecification. For this reason, our primary results are from standard value-added models estimated by least squares. In section 9, we further consider the data-censoring problem and provide some evidence on the extent to which model misspecification alone drives our ceiling effect results.

3. BACKGROUND

Only a fraction of the recent studies measuring teacher value added has considered the potential importance of test score ceiling effects. Furthermore, none have explicitly evaluated the direct implications of ceiling effects for value-added results. Hanushek et al. (2005) provide the most provocative documentation of ceiling effects in the recent value-added literature. These authors estimate value added using the TAAS, where scale scores are such that a gain of zero implies “typical” progress. They divide the exam into ten equal test score intervals and assign each student to one of ten bins based on his or her period (t − 1) test score level. There is a strong negative relationship between students’ period (t − 1) test score levels and period t gains, which is suggestive of a ceiling effect (although mean reversion could also explain the documented relationship). More importantly, approximately two-thirds of the students in their sample are assigned to a bin where the average test score gain is *negative*. Where typical progress is purported to correspond to a gain of zero, and in the absence of a ceiling effect, mean reversion in both directions would suggest that approximately equal shares of students should experience positive and negative gains. That such a large fraction of students shows negative gains suggests that ceiling effects are an important concern. Hanushek et al.’s analysis is one of only a few that carefully consider test score ceiling effects, although a direct analysis of ceiling effects is beyond the scope of their study.

Of the other recent test score–based studies of teacher quality, there is little mention of ceiling effects. Koedel (2009) and Koedel and Betts (2007) acknowledge the potential for test score ceiling effects and report information on the relationship between students’ gains and lagged test score levels. Aaronson, Barrow, and Sander (2007) measure value added using two tests that differ substantially in terms of the distributions of scores, which they thoroughly document, but they do not explicitly consider ceiling effects. Rockoff (2004), who estimates teacher effects outside the value-added framework, reports that 3 to 6 percent of the students in his sample attain the maximum

Table 1. Controls from Value-Added Models

Student-Level Controls (X_{it})	School (and Classroom)-Level Controls (S_{it})
English-learner (EL) status	School fixed effects
Change from EL to English proficient	Classroom-level peer performance in year (t − 1)
Expected and unexpected school changer	Class size
Parental education	Percentage of student body:
Race	by race
Gender	by EL status
Designated as advanced student	by free/reduced price lunch status
Percentage of school year absent[a]	by school changer status

[a]The share of days missed by students is sometimes considered endogenous. Fourth-grade students, however, are not likely to have much influence over their attendance decisions.

possible score, but he does not go into further detail. Lockwood et al. (2007) show that teacher effects are quite sensitive to different testing instruments, but they do not consider the extent to which ceiling effects might be driving their results. Nye, Konstantopoulos, and Hedges (2004) do not discuss test score ceiling effects.

4. VALUE-ADDED MODELS

We estimate teacher value added using three different student achievement specifications. Each specification implies trade-offs in estimation. We focus on the general value-added model (VAM) in which current test scores are regressed on lagged test scores. It is somewhat common in the literature to use a specific form of the VAM, the gain score 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 gain score models, our findings are nearly identical using the gain score framework. Results from the gain score analogs to the specifications below are available from the authors upon request.

The first model that we consider, and the simplest, is a basic VAM that allows for the comparison of teacher effects across schools:

$$Y_{it} = \phi_t + Y_{i(t-1)}\phi_1 + X_{it}\phi_2 + T_{it}\theta + \varepsilon_{it}. \tag{1}$$

In equation 1, Y_{it} is the test score for student i in year t, ϕ_t is a year-specific intercept, X_{it} is a vector of fixed and time-varying student-specific characteristics (see table 1), and T_{it} is a vector of teacher indicator variables where the entry for the teacher who teaches student i in year t is set to one. The coefficients of interest are in the J × 1 vector of teacher effects, θ.

We refer to equation 1 as the basic model. The most obvious omission from the model is school-level information, whether in the form of school fixed effects or time-varying controls. Researchers have generally incorporated this information because of concerns that students and teachers are sorting into schools nonrandomly. This sorting, along with the direct effects of school-level inputs on student achievement (peers, for example), will generate omitted variables bias in equation 1 in the value-added results for teachers.

While the concern about omitted variables bias is certainly relevant, any model that includes school-level information will not allow for a true comparison of teacher effectiveness across schools. For example, if school fixed effects are included in equation 1, each teacher's comparison group will be restricted to the set of teachers who teach at the same school. Furthermore, even in the absence of school fixed effects, the inclusion of school-level controls will restrict teachers' comparison groups to some extent because teachers may sort themselves based on school-level characteristics. If this is the case, controls meant to capture school quality will also partly capture school-level teacher quality, limiting inference from across-school comparisons of teachers.

For most researchers, concerns about omitted variables bias dominate concerns about shrinking teacher comparison groups. This leads to the second model that we consider, the within-schools model, which is more commonly estimated in the literature and includes time-varying school-level covariates and school fixed effects.[6]

$$Y_{it} = \beta_t + Y_{i(t-1)}\beta_1 + X_{it}\beta_2 + S_{it}\beta_3 + T_{it}\gamma + \nu_{it}. \tag{2}$$

In equation 2, S_{it} is a vector that includes school indicator variables and time-varying school-level information for the school attended by student i in year t. The controls in the vector S_{it} are detailed in table 1. The benefit of including school-level information is a reduction in omitted variables bias, including sorting bias generated by students and teachers selecting into specific schools.

Finally, we incorporate student fixed effects into the student achievement specification. This approach is suggested by Harris and Sass (2006), Koedel (2009), and Koedel and Betts (2007):

$$Y_{it} = \alpha_i + \alpha_t + Y_{i(t-1)}\alpha_1 + X_{it}\alpha_2 + S_{it}\alpha_3 + T_{it}\delta + u_{it}. \tag{3}$$

In going from equation 2 to equation 3 we add the student fixed effects, α_i. The inclusion of the student fixed effects also limits the entries in the vector X_{it}

[6] Although teacher effectiveness cannot be compared across schools straightforwardly using value-added estimates from equation 2, this may be acceptable from a policy perspective. For example, policy makers may wish to identify the best and worst teachers on a school-by-school basis regardless of any teacher sorting across schools.

to include only time-varying student characteristics. The benefit of the within-students approach is that teacher effects will not be biased by within-school student sorting across teachers based on time-invariant student characteristics (such as ability, parental involvement, etc.). However, again there are trade-offs. Equation 3 further narrows teachers' comparison groups to those with whom they share students. Thus identification comes from comparing test score gains for individual students when they were in the third and fourth grades. In addition, the incorporation of the student fixed effects makes the model considerably noisier.[7] Finally, the inclusion of the student fixed effects restricts the size of the student population that can be considered because a student record must contain at least three contiguous test scores, instead of just two, to be included in the analysis.[8]

Despite these concerns, econometric theory suggests that student fixed effects will be an effective way to remove within-school sorting bias as long as students and teachers are sorting based on time-invariant characteristics. We estimate the within-students model by first-differencing equation 3 and instrumenting for students' lagged test score gains with their second-lagged levels. This general approach was developed by Anderson and Hsiao (1981) and has recently been used by Harris and Sass (2006), Koedel (2009), and Koedel and Betts (2007) to estimate teacher value added.[9]

Two key issues distinguish the within-students model from the other models that we consider. First, to completely first-difference equation 3 we must incorporate students' lagged teacher assignments, which will appear in the period $(t - 1)$ version of equation 3. That is, we are comparing the effectiveness of students' current and previous year teachers. Second, the requirement that each student record contain three contiguous test scores in the within-students model not only limits the sample size overall but also restricts the student population to less-transient students. Because these students tend to be higher achievers (see, for example, Ingersoll, Scamman, and Eckerling 1989; Rumberger and Larson 1998), a given test score ceiling will have a stronger effect on the restricted student sample. This issue will be revisited when we present our results.

[7] In fact, a test for the statistical significance of the student fixed effects in equation 3 fails to reject the null hypothesis of joint insignificance. However, the test is of low power given the large-N, small-T panel data set structure (typical of most value-added analyses), limiting inference.

[8] Equation 3 also introduces a potential endogeneity concern if teacher assignments are correlated with the time-varying error term component across years. See Rothstein (2008) and Koedel and Betts (forthcoming).

[9] Although all three of these studies use the same basic methodology, Harris and Sass (2006) estimate their model using generalized method of moments, while Koedel (2009) and Koedel and Betts (2007) use two-stage least squares. We use two-stage least squares here.

5. DATA

We evaluate ceiling effects using administrative data from fourth-grade students in San Diego (SDUSD) who started the fourth grade between 1998–99 and 2001–2. We chose the fourth grade because it is an elementary-level grade (so that each student is linked to just one teacher) and because our student fixed effects model requires at least three contiguous test score records per student (students are first tested in the second grade). 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. In our student fixed effects models, we also require students to have second-lagged test scores. For each model, we estimate value added for teachers who teach at least twenty students across the data panel and restrict our student sample to the set of students taught by these teachers.[10] In the models without student fixed effects, we evaluate test score records for 30,354 students taught by 595 teachers. Our sample size falls to 15,592 students taught by 389 teachers in the student fixed effects model. The large reduction in sample size is the result of (1) the requirement of three contiguous test score records per student instead of just two, which in addition to removing more transient students also removes one year cohort of students because we do not have test score data prior to 1997–98 (that is, students in the fourth grade in 1998–99 can have lagged scores but not second-lagged scores) and (2) requiring the remaining students to be assigned to one of the 389 fourth-grade teachers who teach at least twenty students with three test score records or more.[11] 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 teacher 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 student-teacher sorting will influence the magnitude of test score ceiling effects. At one extreme, random assignment of students to teachers will mitigate ceiling effects insofar as they determine teacher rankings regardless of which model from section 4 is used (although ceiling effects may

[10] This restriction is imposed because of concerns about sampling variation (see Kane and Staiger 2002). Our results are not sensitive to reasonable adjustments to the twenty-student threshold.

[11] Only students who repeated the fourth grade in the latter two years of our panel could possibly have had more than three test score records. There are thirty-two students with four test score records in our data set.

Table 2. Average Within-Teacher Standard Deviations of Students' Period (t − 1) Test Scores

		Within Schools		Across District	
	Actual	Random assignment	Perfect sorting	Random assignment	Perfect sorting
Standard deviations of lagged scores	0.81	0.90	0.32	0.99	<0.01

Notes: The numbers above report the average standard deviation of test scores within the classroom for various scenarios, each divided by the overall standard deviation of test scores district wide. In the "perfect sorting" column students are sorted by period (t − 1) test score levels in math, first within school and in the final column across the district. For the randomized assignments, students are assigned to teachers based on randomly generated numbers from a uniform distribution. In the second column, students are not reassigned across schools; in the fourth column, students are reassigned across schools. The random assignments are repeated 25 times, and estimates are averaged across all random assignments and all teachers. The estimates from the simulated random assignments are very stable across simulations.

still lead to an understatement of the importance of teacher quality overall and increase the noise associated with value-added estimation).[12] At the other extreme, a test score ceiling where there is strong student-teacher sorting should lead to a large shift in teacher rankings based on value added.[13]

One benefit of our analysis is that we can use real student-teacher matches from a real school district, rather than attempting to simulate student-teacher sorting. This is important because there is no consensus in the literature as to how students and teachers are actually assigned to one another, making it impossible to artificially generate student-teacher 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.[14]

We document observable student-teacher sorting in our data by comparing the average realized within-teacher standard deviation of students' lagged test scores with analogous measures based on simulated student-teacher matches that are either randomly generated or perfectly sorted. This approach follows Aaronson, Barrow, and Sander (2007). Table 2 details our results, which are presented as ratios of the standard deviation of interest to the total within-grade standard deviation of the test (calculated based on our student sample). Note

[12] If within-teacher student samples are small enough, random assignment will not be sufficient to entirely mitigate ceiling effects on teacher rankings.

[13] In addition to differential student-teacher sorting across districts and schools, there will also be differential sorting across schooling levels. Ceilings will have larger distortionary effects in higher grade levels if student-teacher sorting is stronger.

[14] The 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 SDUSD has a larger than average share of English learners. For basic demographic information about the population of students and teachers at SDUSD see Betts, Zau, and Rice (2003).

that while there does appear to be some student sorting based on lagged test score performance, this sorting is relatively mild.

6. TEST SCORE CEILING SIMULATIONS AND BASIC RESULTS

Our ceiling simulations are 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 restandardize students' test scores within grades to have a mean of zero and a variance of one.[15,16]

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, as well as the ninth- and tenth-grade FCAT exams in 2007.[17] For each ceiling simulation, we report the skewness of the generated test score distributions.

[15] An alternative approach would have been to separately set the ceilings in the second, third, and fourth 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. We do, however, evaluate such a ceiling structure in an omitted analysis and find that altering across-grade differences in skewness has little bearing on our results. This analysis is available from the authors upon request.

[16] Mechanically, the standardization of scores for each grade has no effect on results from the basic and within-schools models. In the within-students model, using within-grade standardized scores reduces the distortionary impacts of the test score ceilings, albeit mildly. This occurs because the first-differenced test scores in the within-students models are scaled by their respective standard deviations before differencing, and the standard deviation of fourth-grade scores is smaller than the standard deviation of third-grade scores. This effectively upweights test scores for students in the current year relative to the lagged year. Because ceilings are defined by skewness in the test score distribution, a larger share of students have above-average test scores as ceiling severity increases across years. In our analysis, the relative upweighting of these scores generated by the standardization appears to partially offset the dampening effect of the test score ceiling. For any test score distortions characterized by increased skewness over time (positive or negative), standardization should be somewhat helpful, although we note that the standardization question is of little practical importance here. Results from models of scaled scores analogous to those from standardized scores models in the within-students analysis are available from the authors upon request. These results suggest even stronger distortionary ceiling effects.

[17] The lagged score distribution is less skewed than the distribution of third-grade scores on the TAAS and more skewed than the distribution of ninth-grade scores on the FCAT.

Table 3. Test Score Ceiling Effects on Value-Added Results: Basic Specification

	(1)[a]	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Percentile of fourth-grade test score distribution where ceiling is set	99.96	97	95	90	85	75	50	33
Skewness of period t score distribution	0.17	−0.02	−0.07	−0.25	−0.37	−0.64	−1.31	−2.00
Skewness of period (t − 1) score distribution	0.25	0.11	0.07	−0.05	−0.13	−0.32	−0.83	−1.32
Correlation between ceiling-restricted value-added estimates and baseline	–	0.99	0.99	0.98	0.97	0.94	0.85	0.77
Estimation error share of variance of teacher fixed effects	0.11	0.11	0.11	0.11	0.12	0.13	0.17	0.24
Unadjusted effect size of teacher quality	0.26	0.25	0.25	0.25	0.25	0.25	0.26	0.26
Adjusted effect size of teacher quality	0.24	0.24	0.24	0.24	0.24	0.24	0.23	0.23

[a] Column 1 shows results from the no-ceiling baseline. A ceiling is not "set" here—0.04 percent of the student population attains the maximum possible score. The last two rows show the unadjusted and adjusted estimates of the number of standard deviations by which student achievement is predicted to change after a one standard deviation increase in teacher quality.

Tables 3, 4, and 5, respectively, show results from the three VAMs discussed above: the basic, within-schools, and within-students models. When the ceilings are imposed, these models are misspecified because the data are censored. Therefore the results from the tables document the combined effects of lost information and model misspecification. Again, because of the complications associated with properly modeling the data censoring given a real-world test score ceiling, these results offer the most pragmatic representation of the influence of ceiling effects. We separately consider the data-censoring problem in more detail in section 9.

Each column in the tables shows results from a different test score ceiling. The ceilings increase in severity moving from left to right, and the first column in each table shows results from our no-ceiling baseline for comparison. The negative skewness measures reported in rows 2 and 3 of each table (and in row 4 in table 5) indicate the degree of ceiling severity. The eighth column of the tables shows results from our most severely skewed simulation, which we refer to as the *minimum-competency equivalent* ceiling. For each ceiling simulation we report three measures of interest in addition to the skewness measures: (1) the correlation between teachers' ceiling-affected value added estimates and estimates from the baseline model without ceiling effects, (2) the estimation error share of the variance of the teacher effects, and (3) the adjusted and unadjusted effect sizes, by which we mean the predicted change

Table 4. Test Score Ceiling Effects on Value-Added Results: Within-Schools Specification

	(1)[a]	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Percentile of fourth-grade test score distribution where ceiling is set	99.96	97	95	90	85	75	50	33
Skewness of period t score distribution	0.17	−0.02	−0.07	−0.25	−0.37	−0.64	−1.31	−2.00
Skewness of period (t − 1) score distribution	0.25	0.11	0.07	−0.05	−0.13	−0.32	−0.83	−1.32
Correlation between ceiling-restricted value-added estimates and baseline	–	0.99	0.99	0.97	0.96	0.93	0.84	0.73
Estimation error share of variance of teacher fixed effects	0.24	0.24	0.24	0.25	0.26	0.28	0.35	0.44
Unadjusted effect size of teacher quality	0.28	0.27	0.27	0.27	0.27	0.27	0.29	0.30
Adjusted effect size of teacher quality	0.24	0.24	0.24	0.24	0.23	0.23	0.23	0.22

[a] Column 1 shows results from the no-ceiling baseline. A ceiling is not "set" here—0.04 percent of the student population attains the maximum possible score. The last two rows show the unadjusted and adjusted estimates of the number of standard deviations by which student achievement is predicted to change after a one standard deviation increase in teacher quality.

Table 5. Test Score Ceiling Effects on Value-Added Results: Within-Students Specification

	(1)[a]	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Percentile of fourth-grade test score distribution where ceiling is set	99.96	97	95	90	85	75	50	33
Skewness of period t score distribution	0.17	−0.10	−0.16	−0.36	−0.49	−0.79	−1.58	−2.39
Skewness of period (t − 1) score distribution	0.25	0.07	0.02	−0.13	−0.22	−0.43	−1.03	−1.62
Skewness of period (t − 2) score distribution	0.15	0.12	0.11	0.07	0.04	−0.04	−0.32	−0.63
Correlation between ceiling-restricted value-added estimates and baseline	–	0.99	0.99	0.97	0.96	0.92	0.80	0.72
Estimation error share of variance of teacher fixed effects	0.33	0.33	0.33	0.34	0.34	0.37	0.45	0.51
Unadjusted effect size of teacher quality	0.29	0.29	0.29	0.29	0.29	0.30	0.32	0.35
Adjusted effect size of teacher quality	0.23	0.24	0.24	0.24	0.24	0.24	0.24	0.25

[a] Column 1 shows results from the no-ceiling baseline. A ceiling is not "set" here—0.04 percent of the student population attains the maximum possible score. The last two rows show the unadjusted and adjusted estimates of the number of standard deviations by which student achievement is predicted to change after a one standard deviation increase in teacher quality.

in student achievement, as a proportion of one standard deviation of test scores, resulting from a one standard deviation increase in teacher quality. The correlations between the ceiling-affected and baseline estimates provide a quick gauge of the distortionary impacts of the ceilings. Teacher effect sizes are commonly used in the literature to evaluate the importance of differences in teacher quality. The unadjusted effect size is just the square root of the raw variance in teacher effects, while the adjusted measure accounts for estimation error in the individual teacher effect estimates. These estimates are reported as ratios of the standard deviation of the teacher effect distribution to the standard deviation of the censored test score distribution for each ceiling simulation. This metric has a straightforward interpretation. For example, the southwest-most entry in table 3, if taken at face value, suggests that a one standard deviation improvement in teacher quality corresponds to a 0.24 standard deviation improvement in test scores. The estimation error shares of the teacher effect variances and the corresponding adjusted variance measures are estimated following Koedel (2009), who separates the variance of the estimated teacher effects into signal and noise components.[18]

The three tables show that teachers' value-added estimates are roughly impervious to test score ceiling effects over a wide range of ceiling severity in each model. This can be seen by looking at the correlations between the teacher 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 ≈ -0.64), as evidenced by the fairly high correlation between teachers' baseline value-added estimates and their value-added estimates from this ceiling simulation. So, for example, policy makers should feel comfortable using FCAT scores from the third and fourth grades, where the skewness in the test score distributions are around -0.5, to measure teacher value added at least insofar as ceiling effects are a concern. 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 teacher effect estimates from the actual data and the data with the ceiling imposed ranges from about 0.85 for the basic and within-schools models to 0.80 for the within-students model. The correlations drop further when we impose the ceiling at the 33rd percentile, with the lowest correlation being 0.72 in the within-students model.

[18] For the within-students model we also report the skewness in the second-lagged test score distribution. In the between- and within-schools models we cluster standard errors at the student level. Because only grade repeaters have more than one record, the clustered standard errors are essentially typical robust standard errors. Our within-students model is estimated using robust standard errors.

As ceiling conditions approach those found in minimum-competency testing environments, value-added results are non-negligibly altered.

Two other observations from tables 3, 4, and 5 are worthy of mention. First, the estimation error share of the variance of teacher effects increases as ceiling severity increases, which surely explains part of the pattern in correlations discussed above. Second, there is a negligible change in the adjusted variance of teacher quality regardless of ceiling severity, which may initially seem counterintuitive. However, note that the test score ceilings are reducing the raw variance of test scores overall and that the teacher effect variance measures are scaled by this underlying variance. That is, although the standard deviation of the teacher effect distribution is reduced when a ceiling is imposed, the standard deviation of the distribution of test scores is also reduced. In fact, our analysis likely understates test score ceiling effects on the measurable variance of teacher quality because it removes variability in test scores more precisely than would be observed in a real-world ceiling.[19]

Finally, note that the test score ceilings induce more skewness in the test score distributions from the within-students sample (table 5) relative to the larger student sample used in the basic and within-schools models (tables 3 and 4). As mentioned in section 4, this is because the restricted student sample used for the within-students model is disproportionately affected by the test score ceiling (that is, the set of students who have three contiguous test scores is higher achieving, on average, than the set of students who have just two test scores). Interestingly, the influence of each test score ceiling on value added is similar across the three models despite the fact that each ceiling is more strongly felt by students in the within-students model. It appears that the stronger skewness in the test score distributions for the restricted student sample is roughly offset by the benefit of looking within students, where ceiling effects will be partially controlled for by the first-differencing procedure.

7. IMPLICATIONS OF MINIMUM-COMPETENCY TESTING FOR VALUE-ADDED ANALYSIS

We further evaluate the sensitivity of teacher value added to the imposition of our most severe test score ceiling, designed to replicate minimum-competency

[19] Our simulations allow students to demonstrate that they are far above the cutoff, and then we restrict their scores ex post. This removes additional variability in test scores that would be found near the highest possible score in a real-world test score ceiling. For example, we might observe a student scoring at the 80th percentile of the actual distribution of test scores and restrict her score to the 50th percentile such that she obtains the maximum possible score in our simulation. However, with a real-world ceiling where she would have to answer every question correctly to score at the maximum, she might bubble in a wrong answer by accident, read a question incorrectly, etc. This would add to the underlying variability in test scores but of course would not be explained by teacher effects.

Table 6. Transition Matrices Documenting the Stability of Teachers' Value-Added Rankings, by Quintile, before and after the Minimum-Competency Equivalent Ceiling Is Imposed

Basic Model

		Ceiling-affected quintile assignments				
		1	2	3	4	5 (best)
Baseline	1	**76**	17	6	1	0
Quintile	2	22	**43**	25	10	0
Assignments	3	2	33	**34**	23	8
	4	0	7	20	**36**	38
	5 (best)	0	2	14	30	**54**

Within-Schools Model

		Ceiling-affected quintile assignments				
		1	2	3	4	5 (best)
	1	**71**	17	8	3	0
Baseline	2	24	**39**	24	11	3
Quintile	3	4	29	**33**	21	13
Assignments	4	1	13	19	**39**	28
	5 (best)	0	2	16	26	**56**

Within-Students Model

		Ceiling-affected quintile assignments				
		1	2	3	4	5 (best)
Baseline	1	**58**	24	13	4	1
Quintile	2	35	**35**	23	6	1
Assignments	3	5	21	**31**	28	15
	4	3	12	17	**37**	32
	5 (best)	0	9	17	25	**49**

Note: Cells report percentage of teachers in each quintile set.

testing conditions, using transition matrices to compare teacher rankings before and after the test score ceiling transformation. The transition matrices provide an alternative documentation of the correlations reported in the final columns of tables 3, 4, and 5.

To construct the transition matrices, we estimate each model before and after the ceiling is imposed. In each case, we keep the vector of estimated teacher effects and rank them from 1 to J, 1 being the lowest and J being the highest. We divide teachers into quintiles based on their value-added rankings, where quintile 5 teachers are those with the highest value added. The transition matrices compare the stability of these quintile assignments before and after the ceiling is imposed. This type of analysis is particularly relevant in the context of teacher accountability. For example, an accountability system might reward the top 20 percent of teachers and sanction the bottom 20 percent as measured by value added. Our results are reported in table 6 for each of the value-added specifications described in section 4.

The vertical dimension of the transition matrices represents teachers' quintile rankings without the ceiling and the horizontal dimension teachers' rankings after the ceiling is imposed. Each cell in table 6 indicates the percentage of teachers who fall into a given quintile set, where a quintile set is defined by the pair of quintile rankings for a given teacher with and without the ceiling (e. g., the set [1,4] would indicate a quintile ranking of 1 in the no-ceiling case and a quintile ranking of 4 after the ceiling is imposed). If ceiling effects did not influence value-added rankings, the diagonal entries in table 6 would all equal 100 percent and the off-diagonal entries would all equal zero.

The transition matrices show that ceiling effects alone can significantly influence value-added rankings. For example, across the three models, just 49–56 percent of the teachers who are identified as being in the top 20 percent of the value-added distribution based on students' actual test scores are also identified as being in this group once test scores are transformed. Furthermore, 14–17 percent of these teachers are pushed below the 60th percentile of the distribution of teacher effects.

In an omitted analysis (available upon request), we also consider whether certain types of teachers are helped or harmed in terms of their value-added rankings by minimum-competency testing. The mechanism through which we might expect an effect is student-teacher sorting within and across schools. For example, if teachers with master's degrees teach a disproportionate share of high-achieving students, their value-added rankings will be more adversely affected by test score ceiling effects. Not surprisingly, we find that more qualified teachers, teachers with higher salaries, and teachers who teach at more advantaged schools are harmed by test score ceiling effects in value-added rankings (the latter result related to the socioeconomic advantage of students across schools, of course, is applicable only in the basic value-added model).

8. ROBUSTNESS OF THE NEGATIVE SKEWNESS MEASURE

In this section we evaluate the robustness of the negative skewness measure by evaluating whether differentially constructed test score ceilings that produce similar negative skewness have similar implications for value-added results. In particular, we construct a set of soft test score ceilings that are designed to replicate the negative skewness in some of the hard-ceiling simulations and look to see if the soft-ceiling design has different implications for value-added results. We stress that our analysis here is far from exhaustive—for any given level of negative skewness in a test score distribution, there are literally an infinite number of soft test score ceiling structures that could generate the skewness. We focus on just one possibility here, creating soft test score

Table 7. Soft-Ceiling Simulations Designed to Mimic Hard Ceilings at the 75th, 50th, and 33rd Percentiles of the Distribution of Fourth-Grade Test Scores

	Soft Ceiling 1	**Soft Ceiling 2**	**Soft Ceiling 3**
Description:	**Mimics the hard ceiling set at the 75th percentile of the fourth-grade test score distribution**	**Mimics the hard ceiling set at the 50th percentile of the fourth-grade test score distribution**	**Mimics the hard ceiling set at the 33rd percentile of the fourth-grade test score distribution**
X_1:	1	1	1
X_2:	1	1	0.60
X_3:	1	0.90	0.40
X_4:	1	0.70	0.20
X_5:	0.90	0.30	0.10
X_6:	0.70	0.10	0.10
X_7:	0.50	0.10	0
X_8:	0.30	0	0
X_9:	0.10	0	0

ceilings using a spline such that for a student with uncensored test score Y_i:

$$
\begin{aligned}
&\tilde{Y}_i = Y_i,\ Y_i \leq S_1 \\
&\tilde{Y}_i = S_1 + X_1(Y_i - S_1),\ S_1 < Y_i \leq S_2 \\
&\tilde{Y}_i = S_1 + X_1(S_2 - S_1) + X_2(Y_i - S_2),\ S_2 < Y_i \leq S_3 \\
&\vdots \\
&\tilde{Y}_i = S_1 + X_1(S_2 - S_1) + X_2(Y_i - S_2) + \ldots + X_n(Y_i - S_n),\ Y_i > S_{n-1}
\end{aligned}
\tag{4}
$$

In equation 4, $\tilde{Y}_i$ is the transformed score for student i, $S_n > S_{n-1} > \ldots > S_1$ where the S_j, $j = 1, \ldots, n$ represent the test score levels at which the n knots appear, and $1 \geq X_1 \geq X_2 \geq \ldots X_{n-1} \geq X_n$, meaning that the test score ceiling is nondecreasing in severity as students' test scores rise. Specifically, we define S_n as the score at the nth decile of the fourth-grade test score distribution for these simulations. For student i, whose score falls between S_2 and S_3, her transformed score can be written (where Y_i is her observed test score):

$$\tilde{Y}_i = S_1 + X_1(S_2 - S_1) + X_2(Y_i - S_2). \tag{5}$$

We generate three soft test score ceilings using this basic structure. These ceilings are designed to produce skewnesss in the distributions of test scores comparable to those from our hard-ceiling simulations imposed at the 75th, 50th, and 33rd percentiles. Table 7 displays the X_n vectors for each of these three ceilings.

Table 8 displays the effects of the three soft ceilings on value-added estimates from each of the three models discussed in section 4. The results are comparable to those in columns 6, 7, and 8 in tables 3, 4, and 5. Although the effects of the soft ceilings are slightly more mild than those from their hard-ceiling counterparts, the results suggest that similarly skewed test score distributions have similar implications for value-added estimation.

9. THE MODEL MISSPECIFICATION PROBLEM

Finally, we explicitly consider the model-misspecification problem, which has partly driven our results thus far. A least-squares approach (and variants thereof), which is typically used in the value-added literature, will be misspecified when there is a test score ceiling because the ceiling acts as a data censor. When ceiling effects are severe, the misspecification problem will be amplified.

In theory one could estimate a censored-data model, such as a Tobit model, to correct this misspecification. However, as a practical matter, there are three complications that arise with respect to resolving the model misspecification problem in the value-added context where a test score ceiling is detected. First, the censor points in a real-world test score ceiling will be unknown; in fact, discontinuous censor points may not even exist. Evidence from Carson and Sun (2007) suggests that misidentifying the censor points will produce substantially biased estimates of the model parameters, meaning that "guessing" at the censor points based on some observed distribution of scores is unlikely to resolve the problem.[20]

A second complication of data censoring in the value-added context is that both current and lagged scores are likely to be censored. In the general VAM, this means that an independent variable will be censored in addition to the dependent variable. The gain score framework does not solve this problem because the censoring in a gain score model is ill defined (censored gains will be zero or near zero, but noncensored gains can also be zero, near zero, or even negative). Although dependent-variable data censoring has received considerable attention in research, there is a much smaller literature that considers independent-variable data censoring. Austin and Brunner (2003) and Austin and Hoch (2004) provide maximum likelihood estimation (MLE) solutions to the independent-variable censoring problem with a known censor point, but their solutions are sensitive to an assumption about the joint

[20] There has been some work in the econometrics literature that looks at data censoring when the censor points are unknown, but this literature is inapplicable to the case of a test score ceiling because a key assumption required to overcome the unknown censoring process is that the censoring is independent of the underlying value of the censored variable (see Chen 2002; Gørgens and Horowitz 1999). This assumption obviously does not apply when the dependent variable, the test score, is subject to a ceiling effect.

Table 8. Soft-Ceiling Results

	Basic Model			Within-Schools Model			Within-Students Model		
Soft Ceiling Number (from table 7)	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
Comparable to hard ceiling imposed at	75th percentile	50th percentile	33rd percentile	75th percentile	50th percentile	33rd percentile	75th percentile	50th percentile	33rd percentile
Share of fourth-grade students at highest score (%)	0.04	22.40	30.10	0.04	22.40	30.10	0.04	22.40	30.10
Skewness of period t score distribution	−0.62	−1.30	−1.96	−0.62	−1.30	−1.96	−0.77	−1.55	−2.30
Skewness of period (t − 1) score distribution	−0.34	−0.84	−1.35	−0.34	−0.84	−1.35	−0.45	−1.04	−1.64
Skewness of period (t − 2) score distribution	NA	NA	NA	NA	NA	NA	−0.07	−0.35	−0.26
Correlation between ceiling-restricted value-added estimates and baseline	0.96	0.88	0.82	0.95	0.86	0.77	0.94	0.84	0.74
Estimation error share of variance of teacher fixed effects	0.12	0.16	0.20	0.27	0.34	0.40	0.36	0.43	0.46
Unadjusted effect size of teacher quality	0.25	0.26	0.26	0.28	0.28	0.30	0.30	0.32	0.35
Adjusted effect size of teacher quality	0.24	0.23	0.23	0.24	0.24	0.23	0.24	0.24	0.26

distribution of the independent variables. Where possible, even these authors strongly recommend circumventing the censoring problem altogether by obtaining uncensored data or, if the sample size permits, restricting the analysis only to uncensored observations.[21]

A third complication in the context of teacher value added is that as the data censoring gets more severe, more and more teachers teach fewer and fewer students whose scores are not censored. At extreme ceiling severities, some teachers do not teach *any* students whose scores are not censored. Clearly, as a larger fraction of the student population's scores are censored, inference for more and more teacher effects becomes unreliable. Thus, where ceiling effects are mild and the misspecification issue has little bearing on the results, a model that appropriately treats the censored data could in principle be informative for most, if not all, of the teacher effects. However, as ceiling effects become increasingly severe, and therefore the data-censoring correction would be most useful, the estimates for more and more teachers become uninformative.

Overall, these three issues suggest that a statistical solution to the misspecification problem, although theoretically possible, is unlikely to be successful. If a severe test score ceiling is detected, the most reasonable solution is to find a different testing instrument. The results from this analysis can be useful for determining whether a test score ceiling is sufficiently severe such that an alternative test should be considered.

Despite these practical difficulties, as a thought experiment it may be of interest to identify the separate impacts of lost information and model misspecification on value-added results. In table 9, we briefly evaluate this question at the level of school effects (for our baseline sample of fourth-grade students) using a basic Tobit model.[22] We focus on school effects to circumvent the problem that at the teacher level, some teachers teach only students with censored scores in the most severe ceiling simulations (where this analysis is most interesting). In all schools, there are at least some students

[21] Whereas thus far we have treated ceiling effects as a "problem" for value-added estimation, an alternative view is that ceiling effects simply signify a shift in the objective function of administrators toward helping students whose scores are not affected by the ceiling. In such cases, modeling student achievement only for students whose scores are below the ceiling, if such a ceiling can be reasonably identified, will be a viable option. However, if school administrators do not want to shift disproportionate weight to low-achieving students in teachers' value-added estimates, the ceiling problem resurfaces.

[22] Unlike ordinary least squares (OLS), heteroskedasticity in the case of Tobit implies inconsistency in the coefficient estimates, and there is substantial heteroskedasticity here. There is some argument in the literature as to how important this is as a practical matter (see, for example, Arabmazar and Schmidt 1981; Brown and Moffitt 1983; Hurd 1979), but in our case a Tobit that directly models the heteroskedasticity in the data performs worse than a simple Tobit. We can only speculate as to the cause in our context—one possibility is that in the heteroskedastic Tobit, the large number of (sometimes imprecisely) estimated heteroskedasticity parameters upon which the parameter estimates of interest are based may be problematic.

Table 9. Test Score Ceiling Effects on Value-Added Results for School-Level Effects ($N = 116$): Tobit versus OLS

	(1)[a]	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Percentile of fourth-grade test score distribution where ceiling is set	99.96	97	95	90	85	75	50	33
Skewness of period t score distribution	0.17	−0.02	−0.07	−0.25	−0.37	−0.64	−1.31	−2.00
Skewness of period (t − 1) score distribution (not censored)	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Correlation between ceiling-restricted value-added estimates estimated by OLS and baseline	–	1.00	1.00	0.99	0.98	0.95	0.86	0.78
Correlation between ceiling-restricted value-added estimates estimated by Tobit and baseline	–	1.00	1.00	1.00	0.99	0.99	0.95	0.90

[a] Column 1 shows results from the no-ceiling baseline.

below the cutoff in all of our simulations. In addition, we avoid the added complication of independent-variable data censoring by censoring only current scores (in practice this has a negligible effect on results). Although our partial censoring approach to estimating school effects is not directly comparable to the preceding analysis, it provides a straightforward setting in which to evaluate separately the information loss and model misspecification components of test score ceiling effects. For brevity, table 9 reports only the correlations of school effects across models. In our school effect models we control for the student-level covariates documented in table 1 (that is, we replace the vector of teacher indicator variables with a vector of school indicator variables in the basic VAM).

Table 9 shows that the Tobit specification improves model performance, and substantially so. For example, even in the minimum-competency equivalent simulation where a significant amount of test score information is lost, modeling the censored data dramatically improves performance. Although the correlation between the baseline school effects and the ceiling-influenced school effects is still far from one in the most severe ceiling simulation, it is much improved (going from 0.78 to 0.90). This exercise suggests that the model misspecification problem is an important contributor to the ceiling effect distortions documented in our primary analysis.

10. CONCLUDING REMARKS

In the current climate of proficiency-based educational reform, test score ceilings are likely to be increasingly common. We evaluate the extent to which ceiling effects influence the estimation of teacher value added. There are two mechanisms by which ceiling effects distort value-added results. First, most straightforwardly, a test score ceiling represents lost information about student learning. Second, a ceiling generally results in model misspecification. Although in theory this latter issue can be resolved by properly modeling the censored data, in practice a statistical solution to the data-censoring problem is unlikely to be feasible.

Our analysis properly treats the test score ceiling problem as a combination of these two distortionary influences. Overall, 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 policy makers should be concerned when working in minimum-competency or proficiency-based testing environments. We show that ceiling conditions in such environments can significantly alter value-added assessments for individual teachers.

The authors thank Andrew Zau and many administrators at the San Diego Unified School District, in particular Karen Bachofer and Peter Bell, for helpful conversations and assistance with data issues. We also thank Dale Ballou, Julie Cullen, Nora Gordon, Shawn Ni, and Yixiao Sun 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.

Anderson, T. W., and Cheng Hsiao. 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Association* 76: 598–609.

Arabmazar, Abbas, and Peter Schmidt. 1981. Further evidence on the robustness of the Tobit estimator to heteroskedasticity. *Journal of Econometrics* 17: 253–58.

Austin, Peter C., and Lawrence J. Brunner. 2003. Type I error inflation in the presence of a ceiling effect. *American Statistician* 57: 97–104.

Austin, Peter C., and Jeffrey S. Hoch. 2004. Estimating linear regression models in the presence of a censored independent variable. *Statistics in Medicine* 23: 411–29.

Betts, Julian, Andrew Zau, and Lorien Rice. 2003. *Determinants of student achievement: New evidence from San Diego*. San Francisco: Public Policy Institute of California.

Brown, Charles, and Robert Moffitt. 1983. The effect of ignoring heteroscedasticity on estimates of the Tobit model. NBER Technical Working Paper No. 27.

Carson, Richard T., and Yixiao Sun. 2007. The Tobit model with a non-zero threshold. *Econometrics Journal* 10: 488–502.

Chen, Songhian. 2002. Rank estimation of transformation models. *Econometrica* 70: 1683–97.

Cullen, Julie Berry, and Susanna Loeb. 2004. School finance reform in Michigan: Evaluating Proposal A. In *Helping children left behind: State aid and the pursuit of educational equity,* edited by John Yinger, pp. 215–50. Cambridge, MA: MIT Press.

Gørgens, Tue, and Joel L. Horowitz. 1999. Semiparametric estimation of a censored regression model with an unknown transformation of the dependent variable. *Journal of Econometrics* 90: 155–91.

Hanushek, Eric, John Kain, Daniel O'Brien, and Steven Rivkin. 2005. The market for teacher quality. NBER Working Paper No. 11154.

Harris, Douglas, and Tim R. Sass. 2006. Value-added models and the measurement of teacher quality. Unpublished paper, Florida State University.

Hurd, Michael. 1979. Estimation in truncated samples when there is heteroskedasticity. *Journal of Econometrics* 11: 247–58.

Ingersoll, Gary M., James P. Scamman, and Wayne D. Eckerling. 1989. Geographic mobility and student achievement in an urban setting. *Educational Evaluation and Policy Analysis* 11: 143–49.

Kane, Thomas, and Douglas Staiger. 2002. The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16: 91–114.

Koedel, Cory. 2009. An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review* 28(6): 682–92.

Koedel, Cory, and Julian R. Betts. 2007. Re-examining the role of teacher quality in the educational production function. Working Paper No. 0708, University of Missouri, Columbia.

Koedel, Cory, and Julian R. Betts. Forthcoming. Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy.*

Lockwood, J. R., Daniel F. McCaffrey, Laura S. Hamilton, Brian Stecher, Vi-Nhuan Le, and Jose Felipe Martinez. 2007. The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement* 44: 47–67.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. How large are teacher effects? *Educational Evaluation and Policy Analysis* 26: 237–57.

Podgursky, Michael J., and Mathew G. Springer. 2007. Teacher performance pay: A survey. *Journal of Policy Analysis and Management* 26: 909–50.

Roberts, Sarah Jane. 1978. Test floor and ceiling effects. ESEA Title I evaluation and reporting system. Mountain View, CA: RMC Research Corporation.

Rockoff, Jonah. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94 (2): 247–52.

Rothstein, Jesse. Forthcoming. Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*.

Rumberger, Russell W., and Katherine A. Larson. 1998. Student mobility and the increased risk of high school dropout. *American Journal of Education* 107: 1–35.

U.S. Department of Education (USDOE). 2008. Mapping Mississippi's education progress 2008. Available www.ed.gov/nclb/accountability/results/progress/ms.html. Accessed 1 July 2009.

Warren, John Robert. 2007. *State high school exit examinations for graduating classes since 1977*. Available www.hsee.umn.edu/. Accessed 1 July 2009.