Do High Grading Standards Affect Student Performance?

David N. Figlio
University of Florida and National Bureau of Economic Research

Maurice E. Lucas
School Board of Alachua County, Florida

Revised: February 2003

Abstract: This paper explores the effects of high grading standards on student test performance in elementary school. While high standards have been advocated by policy-makers, business groups, and teacher unions, very little is known about their effects on outcomes. Most of the existing research on standards is theoretical, generally finding that standards have mixed effects on students. However, very little empirical work has to date been completed on this topic. This paper provides the first empirical evidence on the effects of grading standards, measured at the teacher level. Using an exceptionally rich set of data including every third, fourth, and fifth grader in a large school district over four years, we match students’ test score gains and disciplinary problems to teacher-level grading standards. In models in which we control for student-level fixed effects, we find substantial evidence that higher grading standards benefit students, and that the magnitudes of these effects depend on the match between the student and the classroom. While dynamic selection and mean reversion complicate the estimated effects of grading standards, they tend to lead to understated effects of standards.

Corresponding author: David Figlio, Walter Matherly Professor, Department of Economics, University of Florida, Gainesville, FL 32611-7140, figlio@ufl.edu

JEL Code: I2
Do High Standards Affect Student Performance?

1. Introduction

This paper explores the effects of high grading standards on student test performance in elementary school. While high standards have been advocated by policy-makers, business groups, and teacher unions, very little is known about their effects on outcomes. Most of the existing research on standards (including Becker and Rosen, 1990; Betts, 1998; Costrell, 1994) is theoretical, generally finding that standards have mixed effects on students. However, very little empirical work has to date been completed on this topic.

We know of three empirical studies that examine the effects of standards on student outcomes. Lillard and DeCicca (forthcoming) are not interested in the effects of grading standards per se, but rather on the effects of graduation standards, measured by the number of courses required for graduation. They find that higher graduation standards lead to relatively increased dropout rates. Two current working papers (Betts, 1995; and Betts and Grogger, 2000, the latter of which was written simultaneously with this paper) present the only empirical work that, to our knowledge, focuses on grading standards. Both papers present cross-sectional evidence on the effects of school-level grading standards (measured by their grade-point average relative to test scores) on the level (Betts, 1995) and distribution (Betts and Grogger, 2000) of student test scores, educational attainment, and early labor market earnings. Consistent with the theoretical literature, Betts and Grogger (2000) find significant evidence of differential effects of grading standards, depending on student type.

While the aforementioned papers provide careful and important evidence of the effects of grading standards, there are numerous gaps remaining in this literature. First, the existing
literature does not measure grading standards at the level of the decision-making unit that ultimately sets the standards and assigns grades—that is, at the teacher level. Mounting evidence exists (e.g., Rivkin, Hanushek and Kain, 1998) that the majority of school-level differences in student outcomes are driven by variation in teacher quality, and that there is considerable within-school variation in teacher quality and teacher effectiveness. However, this variation, as well as the ultimate pathway through which even school-level grading standards reach the child, is necessarily masked when relying on school-level variation in policies and practices.

Second, the aforementioned papers rely on cross-sectional variation in school-level standards to address the research question. While this empirical approach is necessary given the data employed, it is easy to conceive of omitted school quality variables that might also be correlated with measured grading standards. In other words, it is impossible to know in cross-section whether the estimated effects of school-level grading standards are in fact due to these standards or to unobserved attributes of the school.

Third, the existing literature (as well as almost all of the work studying other determinants of student outcomes) focuses on students in upper grades rather than at the elementary level. This is, in some ways, an advantage, because one can then measure educational attainment and follow students into the labor market. But in other ways this is a disadvantage, both because sample attrition is likely to be less of a factor at the elementary level and because one might reasonably expect that the most important grades, in terms of student learning, are the early ones.

This paper is the first to address the effects of teacher-level grading standards on student achievement. In addition, it is the first that uses multiple rounds of data on the same student so that the potential for omitted variables bias are much lower than is the case in cross-sectional
analysis. To implement this study, we employ exceptionally detailed data on every third, fourth, and fifth grader in a large school district from the 1995-96 through the 1998-99 school years. Because we observe three years of test data on each student, we can compare two sets of year-to-year test score gains for each student, permitting a tightly-modeled set of within-student comparisons. This same rich data set permits us to measure individual teacher grading standards in several different ways. We find that high teacher grading standards tend to have large, positive impacts on student test score gains in mathematics and reading. In addition, we find that high standards also reduce student disciplinary problems in school. Like Betts and Grogger (2000), we find that high standards differentially affect students, with initially high-ability students experiencing the largest benefit (at least in reading) from high standards. However, we find that the estimated average differences between high-ability and low-ability students mask important distributional effects of high standards. Specifically, we find that initially low-ability students benefit most from high standards when their classmates are high-ability, while initially high-ability students benefit most from high standards when their classmates are low-ability. All results are robust to changes in the definition of teacher-level grading standards.

2. Data and methods

We analyze confidential student-level data provided by the School Board of Alachua County, Florida for this project. Our data consist of observations on almost every third, fourth, and fifth grader in the school system between 1995-96 and 1998-99. Alachua County Public

1If a child is retained and repeats a grade, we consider year-to-year changes in test scores within a grade level; in other words, we include grade-retained students in the present analysis. Our results are invariant in general magnitude and statistical significance levels to changes in how
Schools is a relatively large district (by national standards), averaging about 1,800 test-taking students per year, per grade. Alachua County is racially heterogeneous, with 60 percent of students white, 34 percent African-American, 3 percent Hispanic, and 2 percent Asian. Less than one percent receive services for English as a Second Language. Forty-nine percent of the student body are eligible for subsidized lunches, 19 percent are identified as gifted, and 8 percent are learning disabled.

We observe each third, fourth, and fifth grader’s performance on the Iowa Test of Basic Skills in each year; our only missing observations involve the handful of students who miss the test each year due to illness or other absences, as well as the set of students exempt from test-taking due to a specific disability. In addition, in the last two academic years, we observe each fourth and fifth grader’s performance on the Florida Comprehensive Assessment Test (FCAT). Fourth graders take the FCAT reading assessment, while fifth graders take the FCAT mathematics assessment. Having data on these two different types of examinations is a distinct advantage of conducting this type of research in Florida. The FCAT, which we use to construct our measure of standards, is scored based on the Sunshine State Standards, the same set of curricular standards on which student letter grades in Florida are intended to be based. The ITBS, which we use to construct our dependent variable, is a national test of skills and learning.

In addition, we observe each student’s report card in each year for each subject. Furthermore, we are able to match students to teachers, which is essential, of course, for measuring the effects of grading standards at the teacher level. Student records also record the student’s race, ethnicity, sex, disability status, and gifted status, as well as the student’s discipline we deal with (or whether we include or exclude) grade-retained students.
We employ four dependent variables of interest. Our primary dependent variables are the change from one year to the next in the student’s performance on the Iowa Test of Basic Skills’ mathematics or reading assessments. We focus on changes in test scores, rather than levels, so that we can control, at least cursorily, for student-specific trends in test performance over time. In addition, we also use as a dependent variable indicators for whether the student had at least one disciplinary infraction that merited recording, or alternatively, at least one severe disciplinary infraction, in a given year. All told, we employ approximately 7,000 observations each (for mathematics and reading) of changes in test scores from one year to the next--two sets of year-to-year changes apiece for the two cohorts of students for whom we have three years of data.

2.1. Identifying the effects of grading standards

Our method for identifying the effects of grading standards exploits the fact that we have multiple observations for each student. We measure the effects of grading standards on students’ test performance (or disciplinary problems) by estimating the following equation:

\[
\text{test}_{i,t} = \beta_{0} + (\text{standards}_{t} + \text{NCT}_{t} + 2\text{X}_{i} + \varepsilon_{i,t,y} + \varepsilon_{s},
\]

where \( \text{test} \) represents the change from one year to the next in student \( i \)’s Iowa Test of Basic Skills mathematics (or reading) scaled examination score, and \( \text{standards} \) represents the level of grading standards (calculated as described below in section 2.2) of teacher \( t \). We identify the parameter from students with teachers with measured standards levels in both grades 4 and 5. The use of a first-differenced dependent variable allows us to capture a sort of "pre-test" effect. We control for all student characteristics that are either time-invariant or that trend over time.
with the fixed effect ", and control for all factors invariant within a given school with the fixed effect &gt; The vector C includes variables representing the composition of the classroom; we control for the fraction white, the fraction free-lunch-eligible, and the average third-grade mathematics test score among the students in the classroom in question in year y. The vector X represents the set of student-level variables that change over time. In practice, X includes free lunch status, gifted status, and disability status, all of which can change from year to year. Our parameter of interest is the coefficient on teacher grading standards, (, which represents the effects of changing a student from one level of grading standards to another, holding constant all student and school attributes that do not change over time, as well as time-varying student and classroom characteristics. Alternative specifications of the above regression employ disciplinary problems as the dependent variable.

We employ a difference specification because there exists very strong evidence that students differ systematically in their rates of achievement growth over time, and not merely in their levels of achievement. Put differently, students who begin at a high level tend to have test score growth rates that eclipse those who begin at a low level. For instance, in our sample mathematics growth rates for students scoring in the top quartile of the third grade mathematics test score distribution are more than twenty percent greater than mathematics growth rates for students scoring in the bottom quartile of the third grade distribution. The difference in reading test score growth rates between top- and bottom-initial-achievers is smaller--about ten percent--but still present and statistically significant at the one percent level. It turns out that our choice of using a difference specification tends to lead to more conservative estimates of the effects of grading standards on test performance, relative to a "levels" specification, which is sensible
considering the apparent non-random assignment of students to teachers of varying grading standards described later in the paper.

However, as described later, the fact that initial high performers tend to face greater gains in test scores over our time period than do initial low performers does not intimate that these initial high performers gain more in every year than do initially low performers. There exists regression to the mean in test scores, and students whose scores improve the most from grade 3 to grade 4 tend to be the students whose test scores gain less between grades 4 and 5. The discussion below suggests that in the presence of this regression to the mean, the estimated positive effects of grading standards in a student fixed effects model may be downward-biased.

2.2. Measuring grading standards

We adopt three alternative measures of teacher-level grading standards, though all are similar in nature to the definition also used by Betts and Grogger (2000), in that we compare students’ test performance to their assigned letter grades. To measure grading standards, we compare student letter grades to their score on the relevant FCAT test, a test different from the one used to construct our dependent variable. The FCAT is ideal for measuring standards, because it was designed by Florida officials to measure student performance on the Sunshine State Standards, the same standards that are intended to be the basis for student letter grades and promotion. The FCAT grades student performance on five levels, from 1 (lowest) to 5 (highest), with the thresholds for each performance level designed to correspond with the letter grades A through F. That is, perfect correspondence with the Sunshine State Standards should see a grade of A associated with an FCAT score of 5, a grade of B associated with an FCAT score of 4, and
so forth, with some additional variation introduced due to randomness in test-taking, etc. Our measures of grading standards involve aggregating all FCAT-letter grade comparisons observed for a teacher across the years, to measure time-invariant tendencies of the teacher to grade toughly or lightly, relative to observed student performance on the FCAT.

Our first measure of standards, on which we focus in this paper because it tends to lead to the most conservative results, is calculated as follows:

$$\text{standards}(1)_t = \frac{\sum_{i} \sum_{y} (\text{FCAT}_{iy} - \text{grade}_{iy})}{n},$$

where $t$ represents the teacher, $i$ represents the student, and $y$ represents the year, and $n$ reflects the number of student-year pairs faced by the teacher.\(^2\) The higher the value of $\text{standards}(1)_t$, the higher the standards, because it suggests that students require a higher score on the FCAT to achieve any given letter grade. The variable grade is measured in standard grade-point fashion, with an A earning a score of 4, a B earning a score of 3, and so on. Pluses earn an additional 0.33, while minuses lead to a reduction of 0.33.\(^3\) Therefore, this measure represents the average gap between the FCAT score and the teacher-assigned letter grade for each particular teacher. Since students take the FCAT mathematics examination in fifth grade and the FCAT reading examination in fourth grade, this measure of grading standards is calculated using mathematics grades and scores for fifth-grade teachers and using reading grades and scores for fourth-grade teachers. For teachers who switched between these grades during the years of FCAT administration, this measure of grading standards is computed using both mathematics and

\(^2\)Put differently, $n$ represents the number of students taught by the teacher in the years in which both FCAT scores and letter grades are observed.

\(^3\)Our results are invariant to changing the ways in which pluses and minuses are treated.
reading scores, depending on the grade level at the time of FCAT assessment. The benefit of measuring standards in this way is that it ensures that we will observe standards measures for both a fourth grade teacher and a fifth grade teacher for as many students as possible. The available evidence suggests that this construction is reasonable: among the teachers who switched between the two grades over the course of our sample, the correlation between a teacher’s reading standards (in fourth grade) and mathematics standards (in fifth grade) is nearly 0.80. Put differently, teachers with high reading standards tend to have high mathematics standards as well, and vice versa.

An alternative way of measuring grading standards involves directly regressing FCAT levels against student letter grades:

$$\text{FCAT}_{iy} = *_{t} + \text{grade}_{iy} + \epsilon_{iy},$$

where all notation is as before. The second measure of standards (standards(2)), then, is the retained estimated teacher-level fixed effect $*_{t}$, which reflects the relationship between grade assignment and student FCAT scores that is invariant across students graded by teacher $t$. A higher value of this measure of standards should be interpreted in the same manner as the first standard measure--it requires a greater score on the FCAT for attainment of any given letter grade.

Our third alternative method of measuring teacher-level grading standards (standards(3)) is the simplest to calculate--we measure the average FCAT score of a teacher’s students who were awarded a grade of B. This measure is appealing because it is likely to be the least influenced by class composition. In the tables that follow, we report the results of the first measure of standards because they tend to be the most conservative; results found by employing
the other two measures of standards tend to be stronger and more statistically significant than the results we report.

The top panel of Table 1 illustrates that, on average, teachers tend to grade less stringently than the state standards (as reflected in FCAT scores) indicate that they should. Only nine percent of students awarded As by their teachers\(^4\) attained the corresponding FCAT level, and in fact, only 50 percent attained even level 4. Only eleven percent of students awarded Bs by their teachers attained level 4 or above, and a mere 39 percent attained level 3 or above. Of the students awarded Cs by their teachers, only 14 percent attained level 3 or above, and only 39 percent attained level 2 or above. Put differently, 86 percent of "C students" failed to achieve a minimum acceptable level of competency (level 3) according to the Florida standards, and even 61 percent of "B students" and 17 percent of "A students" failed to meet this competency level.

The middle and bottom panels of Table 1 show that these patterns appear much different for teachers with relatively high standards (the middle panel) and teachers with relatively low standards (the bottom panel). Here, we stratify teachers according to whether they are above or below the district median in standards, as defined by the first measure described above. Among relatively tough graders, 65 percent of A students attained level 4 or above while 5 percent attained level 2 or below. Among relatively light graders, in comparison, only 28 percent of A students attained level 4 or above while 32 percent attained level 2 or below. Among relatively tough graders, 21 percent of B students attained level 4 or above while 36 percent attained level 2 or below. Among relatively light graders, however, just 3 percent of A students attained level 4 or above while 36 percent attained level 2 or below.

\(^4\)For the purposes of presentation in this exercise, we collapse plus and minus grades into a single letter grade. The grading standards measures all distinguish between plus and minus grades, as mentioned above.
or above while 79 percent attained level 2 or below.

### 2.3 Patterns in teacher-level grading standards

The above-mentioned comparisons provide a first piece of evidence that teachers vary considerably in their grading standards, even within a single school district. It turns out that the within-school variation in teacher-level grading standards is almost as great as the population variation in grading standards. In the 1997-98 school year, for instance, the district-wide standard deviation in teacher-level grading standards was 0.68 (measured using the first definition of grading standards), while the mean within-school standard deviation in grading standards was 0.60. The next year, the district-wide variation in standards was slightly greater (a standard deviation of 0.79) and the mean within-school standard deviation in standards was also slightly greater (a standard deviation of 0.72). In both years, the within-school variation is considerably larger than the between-school standard deviation. This provides some corroborative evidence for Rivkin et al (1998), who find that within-school variation in teacher quality exceeds between-school variation in teacher quality in their Texas dataset. This also provides evidence in support of our empirical identification strategy, since we rely on within-school (for the most part) variation in teacher grading standards to identify a standards effect.

Our identification strategy relies on individual teachers’ standards being relatively invariant over time. In Table 2 we stratify the set of teachers into thirds in each academic year, for the purpose of measuring the toughest, average, and lightest graders in each year. In the top panel we observe that 75 percent of teachers (among those present in both years) ranking in the bottom third of standards level in 1997-98 remained in the bottom third, while only 6 percent transitioned
to the top third. Among the teachers ranking in the top third of standards in 1997-98, 77 percent remained in the top third in 1998-99, and none fell to the bottom third of standards. All told, 68 percent of the teachers are located on the diagonal of this transition matrix (where 33 percent would be chance) and only 2 percent of those able to do so transitioned from one corner of this matrix to another from year to year.

It could be the case, however, that some unobserved classroom characteristic that is time-invariant is truly responsible for this transition matrix. To gauge the degree to which this is the case, the middle and bottom panels of Table 2 present the results of analogous transition matrices, in which, in turn, teachers taught a higher-ability class in 1998-99 than in 1997-98 (middle panel) and teachers taught a lower-ability class in 1998-99 than in 1997-98. Class ability here is measured by average third grade test scores, so can be seen as exogenous to a teacher’s standards level. We observe that in both transition matrices, the great majority of cases remain on the diagonals. These transition matrices are virtually unchanged if, say, we require an improvement or a decline to be at least one-quarter of a standard deviation, implying that even large changes in class average initial ability apparently does not affect a teacher’s level of grading standards. In short, teacher-level grading standards remain highly persistent from one year to the next, even when class attributes change. That said, the correlation between a teacher’s change in measured grading standards and change in average third grade test scores is positive and statistically significant. This fact might lead one to suspect that our measures of grading standards are mere artifacts of grading on a curve. We will address this potential concern in considerable detail later in the paper.

Are grading standards merely reflective of some observed teacher qualification level?
determine the degree to which this is the case, we compare teachers with relatively high (above-median) measures of standards to teachers with relatively low (below-median) measures of standards. Teachers with relatively high levels of standards are slightly more experienced and are slightly less likely to have attended a selective or highly selective undergraduate institution, though none of these differences are statistically different. One difference that is statistically significant is the fraction of teachers with masters degrees; high-standards teachers are more likely to have masters degrees than are low-standards teachers. While this difference suggests that high-standards teachers are observably different from low-standards teachers in at least one dimension, other evidence suggests that this is one dimension that rarely is found to matter for student achievement (see, e.g., Hanushek, 1986). On the other hand, the measured teacher attributes generally found to affect student outcomes the most, the selectivity of teacher undergraduate institutions (Goldhaber and Brewer, 1997), is not different between the standards groups. In models presented below, however, we directly control for these teacher qualification measures to rule out the possibility that observed teacher qualification measures may drive the estimated effects of grading standards on student outcomes. Later we discuss results suggesting that our findings are also unlikely to be driven by one important unmeasured teacher quality dimension.

2.4. Teacher-level grading standards and student class assignment

One threat to identification of standards effects concerns the potentially nonrandom

---

5 These comparisons are only for teachers still employed by the School Board of Alachua County in 2000, almost 85 percent of the teachers in our sample. There is no apparent difference in average standards levels between teachers still employed by the district and teachers no longer employed by the district.
assignment of students to teachers. In cross-section, high-standards teachers also have students who perform higher and have better disciplinary outcomes. But they also have students who are more likely to be white or gifted, and less likely to be low-income or learning disabled. These differences are present even within a single school. Hence, it is unclear that these outcomes associated with high standards are actually due to the high standards themselves.

With our identification strategy, however, we do not rely on cross-sectional variation in grading standards but rather on year-to-year changes in the grading standards faced by a student. While there is slight persistence in the grading standards faced by a student, students are nearly as likely to transition to a teacher with a different standards level (measured in halves, within a school) as to remain with a teacher with a similar standards level. Put more concretely, 57 percent of students with below-median teachers (stratified in terms of standards levels within a school) continue to have below-median teachers the next year. An even smaller percentage--54 percent--of students with above-median teachers continue to have above-median teachers the next year. This indicates that year-to-year differences in grading standards are close to random. Similar patterns are observed for most subgroups--blacks and whites are approximately equally likely to transition between groups, as are free-lunch-eligible and ineligible students. The principal outliers are gifted students, who are considerably more likely to transition to a high-standards teacher if they start out with a low-standards teacher, and considerably less likely to transition to a low-standards teacher if they start out with a high-standards teacher, than are non-gifted students.6 But the vast majority of students are almost as likely to transition between low-
standards and high-standards teachers as to persist across years in the same standards group.

Students are not, however, randomly assigned to classrooms, and high-performing students may systematically select into high-standards teachers’ classrooms. If teachers tend to grade on a curve, then teachers who have better students on average will also be measured as having higher grading standards, regardless of the teacher’s actual standards level. It follows that to the extent to which students self-select dynamically into classes, the estimated effects of grading standards will be biased. The direction of this bias is not immediately known, and depends on the relationship between changes in prior test scores and subsequent classroom placement.

The last three columns of Table 3 demonstrates that not only is there a positive correlation between the level of a student’s initial performance and that student’s propensity to transition into a high-standards class from one year to the next, but that there is also a positive correlation between the growth in a student’s test performance and that student’s propensity to move to a more challenging grader in the subsequent year. Put differently, students whose test scores gain the most from grade 3 to grade 4 are more likely to increase the standards level of their teachers from grade 4 to grade 5. Regardless of the level of growth in test scores from one year to the next, students with low-standards teachers are more likely to face higher-standards teachers the following year, and students with high-standards teachers are more likely to face lower-standards teachers the next year. But conditional on the standards level of the teacher in grade 4, the students who gained the most in test performance between grades 3 and 4 were the most likely to face comparatively more challenging teachers in grade 5. These results suggest that students with idiosyncratically strong test performances in grade 4 end up with relatively tough teachers in
grade 5. To the extent to which these idiosyncratic improvements in test performance are random, rather than deterministic, this result indicates that a positive finding of a relationship between grading standards and test score growth from one year to the next is likely *understated* due to this dynamic selection.

The suspicion of understated results is strengthened by the evidence suggesting the presence of dynamic mean reversion presented in the first two columns of Table 3. These columns indicate the presence of a strong negative correlation between test score gains between grades 3 and 4 and test score gains between grades 4 and 5. Therefore, while subsequent test score gains are correlated with initial performance, it is not the case that students who gain more in one year gain more in every year. Instead, the evidence suggests that students who gain the most between grades 3 and 4 are assigned a tougher teacher in grade 5 and subsequently do comparatively poorly, in terms of test score gains between grades 4 and 5. This relationship should work *against* finding a positive relationship between test score growth and grading standards in a student fixed effects model.

3. Empirical results

Our regression results are presented in Table 4. The first row of Table 4 presents the results of a model with no covariates included.\(^7\) We observe large, statistically significant relationships between grading standards and all four dependent variables. However, it is clear from the above discussion on selection into classrooms that these results should not be taken to

\(^7\)Here and elsewhere, we adjust our standard errors for within-class clustering. See Moulton (1986) for an illustration of the importance of adjusting the standard errors in this manner.
represent causal effects of grading standards. In the second row of Table 4 we include the student-level covariates available to us in the data--race, ethnicity, sex, free lunch status, gifted status, and disability status--and find our four results still statistically significant, but considerably diminished in magnitude. The third row adds school-level fixed effects to control for any factors common to all students in a school, leading to similar, but somewhat stronger results.

As mentioned above, one might be concerned that our measures of grading standards are merely reflecting classroom composition. Therefore, in the fourth row of Table 4 we augment the aforementioned specification with controls for the fraction white, fraction free-lunch eligible, and average third grade mathematics test score in the classroom. We observe that the mathematics and reading test score results only grow stronger when we control for classroom composition. On the other hand, while remaining statistically significant at conventional levels, the estimated effects of grading standards on discipline problems fall considerably in magnitude and statistical significance when we control for classroom compositional variables. To test whether the results presented herein are due to excluded teacher characteristics, in the fifth row of Table 4 we add the measured teacher characteristics available in the district data, with no appreciable change in the estimated parameter of interest.

The sixth row of Table 4 presents the results of our primary specification--the model with student and school fixed effects, as well as the classroom compositional variables and observed teacher attributes. Here, observed and unobserved time-invariant student attributes are subsumed in the student fixed effect, and identification is drawn from a student’s changes from year to year in teacher grading standards. We observe test score results that are still larger in magnitude, and discipline problem results that are smaller in magnitude, than those drawn from models without
student fixed effects. The estimated mean effects remain reasonably statistically significant, with p-values from 0.02 to 0.06, in the case of test scores, but are no longer statistically significant in the case of discipline problems.8

The final two rows of Table 4 present results of model specifications analogous to row 6, except that we vary the definition of grading standards, as described in section 2.2 above. We find that our results tend to have similar magnitudes, yet are somewhat more statistically significant (and considerably moreso in the case of discipline) when we employ our alternative measures of grading standards. In sum, our general conclusion from Table 4 is that grading standards have modest effects, on average, on student test scores and discipline problems.

These results are not symmetric, however. In models that distinguish between transitions from relatively easy to relatively tough graders and transitions between relatively tough to relatively easy graders, the results suggest that students benefit more from high grading standards in fourth grade than in fifth grade. The coefficient on the standards measure when the student transitions from a more challenging teacher in fourth grade to an easier teacher in fifth grade is very large and strongly statistically significant, while the coefficient on the standards measure for students who transition from a easier teacher in fourth grade to a more challenging teacher in fifth grade is considerably smaller and statistically insignificant. Whether standards matter more in earlier grades or whether the specific nature of the transition is what matters remains an open

---

8We observe similar patterns in models in which we control for family-level fixed effects rather than student fixed effects. Here, we identify the effects of grading standards using within-family variation in the level of standards faced by siblings. For the purpose of this analysis, we define sibling pairs as two or more students residing at the same address with all known parents in common. When we control for family fixed effects instead of student fixed effects, we find estimated effects that are more statistically significant than those found using the within-student identification strategy.
question.

3.1. One explanation for these findings: Home production

What might generate these positive effects of grading standards? One possibility, of course, is that high standards motivate students to work harder. In such a case, it is sensible to expect that teachers with high standards would bring more out of their students than would teachers with lower standards. A second potential explanation considers student learning gains as being jointly produced between home and school. If parents perceive their children to be struggling at school, they may devote more attention to their children’s schoolwork than they might have if they perceive their children to be performing at a high level.9

In Spring 2001 we conducted a survey of parents in Alachua County Public Schools to assess the possibility of this second explanation.10 We surveyed the population of families with students in both fourth and fifth grades, and asked the responsible parent to report on how much time he or she spends weekly helping each of the two children with their homework. Sibling comparisons such as these allowed us to control for factors (e.g., parental motivation) that might be common to both siblings in a household. We found that, holding constant the child’s grade level (i.e., fourth or fifth grade), third grade test scores, and the average third grade test score in the child’s class, parents systematically spend more time helping the child with the tougher teacher

9Houtenville and Conway (2001), in another context, suggest that parents supply less effort when they perceive schools to be better.

10Survey participants are similar to the school population as a whole, in terms of racial, economic, and gifted composition. We appreciate the helpful comment by Karen Conway that inspired us to conduct this survey.
(by our measures) with homework than they do helping the sibling with the easier teacher. The results are statistically significant and large in magnitude: we estimated that a parent of a child with a 25th-percentile teacher (in terms of grading standards)--that is, a relatively tough teacher--would spend 60 percent more time helping that child with homework than he or she would spend with that child’s sibling who had a 75th-percentile teacher. These results are not due to the parents reporting that tougher teachers assign more homework--indeed, we estimate that, from parental reports, the typical 25th-percentile teacher assigns only 10 percent more homework than the typical 75th-percentile teacher. This is consistent with our findings from personal interviews with principals in the district, who report that teachers within at any given grade level in the school work to assign the same amount of homework per week. We have no way of judging whether the homework assigned by tougher teachers is more challenging than that assigned by easier-grading teachers.

An additional interesting finding from this survey is that parents do not perceive tougher teachers to be better teachers. We asked each parent to grade their children’s teachers from A to F. While there is relatively low variation in these grades (as two-thirds of the parents assigned grades of A to the teachers), the results suggest that, if anything, parents view tough teachers less favorably than they view easier teachers. Using the same within-family comparisons as above, we found that parents were 50 percent more likely to assign a grade of B or below to a 25th-percentile teacher than to a 75th-percentile teacher, again controlling for grade level, student third grade test score, and average third grade test score in the class. This result, significant at the 16 percent level, suggests that our measure of grading standards is not merely reflecting some other attribute of a teacher that is viewed as desirable to parents.
While these survey findings are not conclusive, they do indicate that high grading standards are likely not merely representing other measures of teacher desirability to parents, and that high grading standards may motivate parents to increase their involvement in their children’s education. Both findings bolster the argument that it is high grading standards, rather than some unobservable measure of teacher quality, that is responsible for the observed performance gains.

3.2. "Curve grading" as an alternative explanation

The prospect remains that the results described above are deterministically due to the proclivities of teachers to grade on a curve. While the presence of classroom-level student characteristics, including mean initial test score, should tend to dampen this potential effect, one cannot entirely rule out curve grading as an alternative explanation. Table 5, however, makes clear that teachers of different standards levels are likely to assign different grade distributions to their classes, and to students who would be forecast to receive the same grade based only on initial test performance. This table breaks down students by quintile of initial test performance, and teachers by quintile of measured grading standard, and reports the proportion of students in each initial performance group receiving a grade of "A" for each standards group. We observe that, unsurprisingly, the likelihood that one will receive a grade of "A" increases with initial test performance. However, we also observe that, conditional on initial test performance, students facing more challenging teachers are less likely to receive an "A." Parallel findings emerge in the similar exercise with regards to the probability of receiving a grade of "C".

Table 5 indicates that teachers do not lock-step grade on a curve. Therefore, it should come as little surprise that in regression models (not shown in the paper, but available on request)
controlling for the degree to which teachers grade on a curve, the estimated effects of grading standards on student test scores and disciplinary problems are almost completely unchanged when measures of curve grading are incorporated into the model. In these models, we attempted a variety of methods of capturing curve-grading, including controlling for the ratio of "A" grades to grades of "C" or lower or the variance of the letter grades given to the class, and in no case did the estimated effect of grading standards change meaningfully in magnitude or statistical significance. Therefore, we are more convinced that curve-grading is not the explanation for the findings presented above.

3.3. Distributional effects of grading standards

While the mean effects of grading standards are important, the theoretical literature on grading standards suggests that there may be substantial distributional impacts, with winners and losers associated with higher standards. In addition, Betts and Grogger (2000), in their empirical study, find evidence of distributional effects of school-level grading standards, with initially high-performing students (in tenth grade) benefitting the most (in terms of twelfth grade mathematics test performance) from high grading standards.\textsuperscript{11} Therefore, in Table 6 we revise our primary model (Table 4, row 6) to include an interaction between grading standards and the student’s initial mathematics (or reading, depending on the dependent variable) test score. Here, base year test scores are standardized with a mean of zero and standard deviation of one, for ease of interpretation. In these interactive models, an average student in third grade is estimated to

\textsuperscript{11}They also find that minority students are harmed by grading standards because standards are estimated to reduce minority high school graduation rates.
benefit strongly (and significantly) from higher grading standards, with above-average initial
performers unambiguously benefitting as well. However, since the interactions with base year test
scores are positive (though not statistically significant at traditional levels in mathematics) it is
clear that these positive estimated benefits of grading standards are not uniform for all. Indeed,
the results suggest that grading standards are only significantly positive (at the ten percent level),
in the case of math performance, for students whose math scores were nine-tenths of a standard
deviation below the mean (or better), and in the case of reading performance, for students whose
reading test scores were eight-tenths of a standard deviation below the mean, or better. However,
the estimated effects of grading standards are negative for less than one percent of the population,
and never statistically significantly negative.

The second set of specifications reported in Table 6 are models that interact grading
standards with the class’s average third grade mathematics (or reading) score. Here, as above,
class average test scores are standardized to have a mean of zero and a standard deviation of one,
for ease of interpretation. Again, we see that higher ability classes may fare somewhat better with
higher standards than with lower ability classes.

What may be more interesting, however, than how entire classes fare with high grading
standards is the distributional effect within a class of high grading standards. Put differently, are
the benefits of high standards uniform within a class, or are there winners and losers within the
class? Specifications 3M, 3R, 4M, and 4R in Table 6 address this question. Specifications 3M
and 3R examine the differential effects of grading standards on initially above-average students as

---

In specifications in which we interact grading standards with a class average score, we also control for the class average mathematics (or reading) score in third grade.
the average ability level of the classroom rises. We observe that the effects of grading standards are highest for high-ability students when classroom ability is relatively low, although this differential effect is not statistically significant. Specifications 4M and 4R examine the differential effects of grading standards on initially below-average students as the average ability level of the classroom rises. We observe that the effects of grading standards are highest for low-ability students when classroom ability is relatively high, a relationship significant at the three percent, depending on the test score considered. In other words, low-ability students differentially benefit from high standards when they are in a high-ability class, and high-ability students may possibly also differentially benefit from high standards when they are in a low-ability class.

Specifications 5R and 5M present similar results in a model in which all students are included in the same regression. The three-way interactions between grading standards, class average, and own base year score underscores the above results that standards benefit low-ability students in high-ability classes and high-ability students in low-ability classes the most. These results are clearest when the point estimates are translated into predicted years of test score gains associated with increased standards at different points of the student ability-class ability continuum. We find that the estimated effect of increasing grading standards by one standard deviation is associated with as much as one-third of a year or more of mathematics test score gains, and by as much as two-thirds of a year or more of reading test score gains. For instance,

\[ \text{predicted years of test score gains} = \] 

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{years of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]

\[ \text{predicted years of test score gains} = \]

\[ \text{estimated effect of increased grading standards} \times \text{standard deviation} \]

\[ \text{or as much as} \] 

\[ \frac{1}{3} \times \text{year of test score gain} \] 

\[ \text{or as much as} \] 

\[ \frac{2}{3} \times \text{year of test score gain} \]
for a student with third grade mathematics performance one-half standard deviation below the mean, the estimated effect of increasing teacher toughness by one standard deviation ranges from 0.07 years of extra growth (in a classroom averaging 1.5 standard deviations below the mean) to 0.28 years (in a classroom averaging 1.5 standard deviations above the mean.) For a student with third grade reading performance 1.5 standard deviations above the mean, the estimated effect of increasing teacher toughness by one standard deviation ranges from 0.18 years of extra growth (in a classroom averaging 1.5 standard deviations above the mean) to 0.71 years (in a classroom averaging 1.5 standard deviations below the mean.) As mentioned above, this pattern of findings also helps further the conclusion that it is grading standards, and not some other unmeasured form of teacher quality, that is likely to generate our findings.

This result has intuitive appeal. Given that the distribution of grades within a class varies much less across classes than does the distribution of performance on external assessments, one can assume that high grades are relatively "safe" for high-ability students in low-ability classes than for their counterparts in high-ability classes. Likewise, low-ability students in high-ability classes are at relatively more "risk" of receiving a low grade than are low-ability students in low-ability classes. Hence, it seems sensible that high standards that lower the "safety" for high-ability students in low-ability classes may generate more effort and greater learning, as might high standards that increase the "risk" for low-ability students in high-ability classes. Feltovich, Harbaugh and To (2002) present theoretical results that are consistent with this story as well. In their study of "counter-signaling" behavior, they argue that high standards improve the achievement of students mismatched with the typical ability level of their peers. While this is by no means a definitive explanation of our empirical findings, it is a plausible one.
There is additional reason to suspect that this might be the case. As mentioned above, teachers maintain drastically different grading standards independent of classroom attributes. Children who rank in the bottom third of the third grade test distribution are three times more likely to earn a grade of C or below with a "tough" teacher (in the top third of the distribution) than with an "easy" teacher (in the bottom third of the distribution. If the average third grade score in the classroom is above the median level, then this difference is more than four times, while if the average third grade score is below the median, this difference is less than two times. The reverse is true for children who rank in the top third of the third grade test distribution: they are three times more likely to earn a grade of C or below with a "tough" teacher than with an "easy" teacher, but this relationship is less than two times in an above-median class and greater than four times in a below-median class. Hence, initially high-ability students are challenged more to get a "good grade" with tough teachers, particularly when they are among the strongest members of a class, and initially low-ability students are also challenged more to get a "good grade" with tough teachers, but particularly when they are among the weakest members of a class.

4. Conclusion

This paper provides evidence that students benefit academically from higher teacher grading standards. We find that high standards have mean effects on test score gains and discipline problems that are large in magnitude and modestly statistically significant. In addition, we find evidence of distributional effects of grading standards. While we find support for the notion that high-ability students benefit more than low-ability students from grading standards, we observe that the distributional pattern is more complicated: Initially low-performing students
appear to differentially benefit from high grading standards when the average ability level of the class is high, and high-performing students appear to differentially benefit from high grading standards when the average ability level of the class is low.

It is, however, premature to conclude from this study that high grading standards are unambiguously desirable. We cannot yet speak to the distributional consequences of teacher-level grading standards at the secondary grades, where Betts and Grogger (2000) have found that high school-level grading standards may help some students at the expense of others. In addition, while the present study helps us to better understand the effects of high grading standards at the elementary grades, we do not yet know how to raise the standards of teachers with currently low standards. Moreover, it may still be the case that our measure of teacher grading standards is merely reflective of some other unmeasured teacher attribute. Before we can recommend higher standards as a policy outcome, it is important to understand the distributional consequences at all levels, as well as to know how to implement a policy of high standards.
Acknowledgments

Thanks to the School Board of Alachua County for providing the confidential data used in this project. We appreciate the helpful comments of Karen Conway, Janet Currie, Jeff Grogger, Jon Gruber, Larry Kenny, Jens Ludwig, and Rich Romano, two anonymous referees, and seminar participants at the National Bureau of Economic Research, Duke University, the Universities of Florida and New Hampshire, and the School Board of Alachua County. Figlio appreciates the financial support of the National Science Foundation through grant SBR-9810615. All errors are our own. The views expressed in this paper are those of the authors and not necessarily those of the National Bureau of Economic Research or the School Board of Alachua County.
References


Table 1: Distribution of letter grades and FCAT Scores

I. Overall distribution of FCAT scores, by letter grade (row percentages are reported)

<table>
<thead>
<tr>
<th>Assigned letter grade</th>
<th>FCAT level (5=highest; 1=lowest)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>level 5</td>
</tr>
<tr>
<td>A+/A/A-</td>
<td>0.09</td>
</tr>
<tr>
<td>B+/B/B-</td>
<td>0.01</td>
</tr>
<tr>
<td>C+/C/C-</td>
<td>0.00</td>
</tr>
<tr>
<td>D+/D/D-</td>
<td>0.00</td>
</tr>
<tr>
<td>E/F</td>
<td>0.00</td>
</tr>
</tbody>
</table>

II. Distribution of FCAT scores, by letter grade, teachers with above-median standards

<table>
<thead>
<tr>
<th>Assigned letter grade</th>
<th>FCAT level (5=highest; 1=lowest)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>level 5</td>
</tr>
<tr>
<td>A+/A/A-</td>
<td>0.12</td>
</tr>
<tr>
<td>B+/B/B-</td>
<td>0.02</td>
</tr>
<tr>
<td>C+/C/C-</td>
<td>0.00</td>
</tr>
<tr>
<td>D+/D/D-</td>
<td>0.00</td>
</tr>
<tr>
<td>E/F</td>
<td>0.00</td>
</tr>
</tbody>
</table>

III. Distribution of FCAT scores, by letter grade, teachers with below-median standards

<table>
<thead>
<tr>
<th>Assigned letter grade</th>
<th>FCAT level (5=highest; 1=lowest)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>level 5</td>
</tr>
<tr>
<td>A+/A/A-</td>
<td>0.04</td>
</tr>
<tr>
<td>B+/B/B-</td>
<td>0.00</td>
</tr>
<tr>
<td>C+/C/C-</td>
<td>0.00</td>
</tr>
<tr>
<td>D+/D/D-</td>
<td>0.00</td>
</tr>
<tr>
<td>E/F</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Table 2: Persistence of grading standards across years

I. Full population of teachers: fraction of teachers transitioning to each standards group

<table>
<thead>
<tr>
<th></th>
<th>&quot;Standards third&quot; in 1997-98 academic year</th>
<th>&quot;Standards third&quot; in 1998-99 academic year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bottom third of standards</td>
<td>Middle third of standards</td>
</tr>
<tr>
<td>Bottom third of standards</td>
<td>0.26</td>
<td>0.07</td>
</tr>
<tr>
<td>Middle third of standards</td>
<td>0.05</td>
<td>0.17</td>
</tr>
<tr>
<td>Top third of standards</td>
<td>0.00</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Fraction on diagonal: 0.68  Fraction transitioning from top to bottom, or vice versa: 0.02

II. Teachers whose average class "quality" (measured by average 3rd grade test scores) improved from 1997-98 to 1998-99: fraction of teachers transitioning to each standards group

<table>
<thead>
<tr>
<th></th>
<th>&quot;Standards third&quot; in 1997-98 academic year</th>
<th>&quot;Standards third&quot; in 1998-99 academic year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bottom third of standards</td>
<td>Middle third of standards</td>
</tr>
<tr>
<td>Bottom third of standards</td>
<td>0.21</td>
<td>0.10</td>
</tr>
<tr>
<td>Middle third of standards</td>
<td>0.05</td>
<td>0.21</td>
</tr>
<tr>
<td>Top third of standards</td>
<td>0.00</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Fraction on diagonal: 0.71  Fraction transitioning from top to bottom, or vice versa: 0.00

II. Teachers whose average class "quality" (measured by average 3rd grade test scores) fell from 1997-98 to 1998-99: fraction of teachers transitioning to each standards group

<table>
<thead>
<tr>
<th></th>
<th>&quot;Standards third&quot; in 1997-98 academic year</th>
<th>&quot;Standards third&quot; in 1998-99 academic year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bottom third of standards</td>
<td>Middle third of standards</td>
</tr>
<tr>
<td>Bottom third of standards</td>
<td>0.27</td>
<td>0.04</td>
</tr>
<tr>
<td>Middle third of standards</td>
<td>0.04</td>
<td>0.16</td>
</tr>
<tr>
<td>Top third of standards</td>
<td>0.00</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Fraction on diagonal: 0.65  Fraction transitioning from top to bottom, or vice versa: 0.05
Table 3: Mean changes in grading standard transitions faced by students between grades 4 and 5, by change in student mathematics performance between grades 3 and 4

<table>
<thead>
<tr>
<th>Student group, based on change in math score between grades 3 and 4</th>
<th>Mean change in math score between grades 3 and 4</th>
<th>Mean grade in math score between grades 4 and 5</th>
<th>Grading standards faced by student in grade 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean grade in math score between grades 4 and 5</td>
<td></td>
<td>Lowest standard third</td>
</tr>
<tr>
<td>Lowest third</td>
<td>2.44</td>
<td>18.75</td>
<td>0.32</td>
</tr>
<tr>
<td>Middle third</td>
<td>15.31</td>
<td>15.69</td>
<td>0.37</td>
</tr>
<tr>
<td>Highest third</td>
<td>28.97</td>
<td>11.21</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Note to Table 3: Teacher grading standards are standardized for the purpose of presentation.
### Table 4: Estimated effects of teacher grading standards on student outcomes

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Change in ITBS math test scores</th>
<th>Change in ITBS reading test scores</th>
<th>At least one disciplinary infraction</th>
<th>At least one severe disciplinary infraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) No covariates included</td>
<td>2.817 (p=0.000)</td>
<td>2.754 (p=0.000)</td>
<td>-0.124 (p=0.000)</td>
<td>-0.120 (p=0.000)</td>
</tr>
<tr>
<td>(2) Controlling for race, ethnicity, sex, free lunch status, gifted status, disability</td>
<td>1.583 (p=0.005)</td>
<td>1.875 (p=0.000)</td>
<td>-0.029 (p=0.043)</td>
<td>-0.028 (p=0.035)</td>
</tr>
<tr>
<td>(3) Same as (2) but also including school fixed effects</td>
<td>1.912 (p=0.006)</td>
<td>2.026 (p=0.001)</td>
<td>-0.053 (p=0.000)</td>
<td>-0.055 (p=0.000)</td>
</tr>
<tr>
<td>(4) Same as (3) but also including fraction white, fraction free-lunch-eligible, and average third grade test performance in class</td>
<td>2.544 (p=0.005)</td>
<td>2.482 (p=0.001)</td>
<td>-0.030 (p=0.073)</td>
<td>-0.028 (p=0.081)</td>
</tr>
<tr>
<td>(5) Same as (4) but also including teacher years of experience, education level, and selectivity of undergraduate institution</td>
<td>2.328 (p=0.022)</td>
<td>2.819 (p=0.001)</td>
<td>-0.035 (p=0.068)</td>
<td>-0.030 (p=0.098)</td>
</tr>
<tr>
<td>(6) Same as (5) but also including student fixed effects</td>
<td>4.039 (p=0.062)</td>
<td>7.696 (p=0.016)</td>
<td>-0.025 (p=0.198)</td>
<td>-0.011 (p=0.562)</td>
</tr>
<tr>
<td>(7) Specification (6): using FIXED EFFECT measure of standards</td>
<td>4.214 (p=0.046)</td>
<td>8.131 (p=0.003)</td>
<td>-0.032 (p=0.097)</td>
<td>-0.017 (p=0.345)</td>
</tr>
<tr>
<td>(8) Specification (6): using &quot;GRADE B&quot; measure of standards</td>
<td>2.964 (p=0.040)</td>
<td>4.674 (p=0.060)</td>
<td>-0.037 (p=0.056)</td>
<td>-0.017 (p=0.208)</td>
</tr>
</tbody>
</table>

Notes to Table 4: Each cell represents a separate regression. Robust p-values (standard errors corrected for clustering of observations within classes) are in parentheses beneath point estimates.
Table 5: Proportion of students receiving "A" grade, by third grade mathematics test performance and measured teacher grading standards

<table>
<thead>
<tr>
<th>Quintile of measured teacher grading standards</th>
<th>Quintile of student grade 3 mathematics performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bottom</td>
</tr>
<tr>
<td>Lowest standards</td>
<td>0.11</td>
</tr>
<tr>
<td>2nd</td>
<td>0.04</td>
</tr>
<tr>
<td>3rd</td>
<td>0.01</td>
</tr>
<tr>
<td>4th</td>
<td>0.03</td>
</tr>
<tr>
<td>Highest standards</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Table 6: Differential effects of high grading standards on test scores (all using student fixed effects model, akin to Row 6, Table 4)

<table>
<thead>
<tr>
<th>Specification</th>
<th>Dependent variable: change in <strong>math</strong> score</th>
<th>Dependent variable: change in <strong>reading</strong> score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1M)</td>
<td>(2M)</td>
</tr>
<tr>
<td>Students included in regression</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Grading standards</td>
<td>4.619 (p=0.00)</td>
<td>4.609 (p=0.00)</td>
</tr>
<tr>
<td>Grading standards x 3rd grade math score</td>
<td>1.397 (p=0.19)</td>
<td></td>
</tr>
<tr>
<td>Grading standards x 3rd grade reading score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grading standards x class average 3rd grade math score</td>
<td>2.685 (p=0.04)</td>
<td>-2.075 (p=0.52)</td>
</tr>
<tr>
<td>Grading standards x class average 3rd grade reading score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grading standards x class average x own score</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes to Table 6: Each column represents a separate regression. Robust p-values are in parentheses beneath point estimates.