NBER WORKING PAPER SERIES

DO HIGH GRADING STANDARDS
AFFECT STUDENT PERFORMANCE?

David N. Figlio
Maurice E. Lucas

Working Paper 7985
http://www.nber.org/papers/w7985

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2000

Thanks to the School Board of Alachua County for providing the confidential data used in this project. 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.

Do High Grading Standards Affect Student Performance?
David N. Figlio and Maurice E. Lucas
NBER Working Paper No. 7985
October 2000
JEL No. I2

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. We find that these effects are not uniform: High-achieving students apparently benefit most from high standards when in a relatively low-achieving class, and low-achieving students benefit most from high standards when in a relatively high-achieving class.

David N. Figlio
Department of Economics
University of Florida
Gainesville, FL 32611-7140
and NBER
figliodn@dale.cba.ufl.edu

Maurice E. Lucas
School Board of Alachua County
620 E. University Avenue
Gainesville, FL 32601

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) 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-achieving students experiencing the largest benefit (at least in reading) from high standards. However, we find that the estimated average differences between high-achieving and low-achieving students mask important distributional effects of high standards. Specifically, we find that initially low-achieving students benefit most from high standards when their classmates are high-achieving, while initially high-achieving students benefit most from high standards when their classmates are low-achieving. 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 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 record.

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. 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:

$$\Delta \text{test}_{itsy} = \alpha_i + \gamma \text{standards}_t + \theta X_{iy} + \xi_s + \epsilon_{itsy} ,$$

where Δtest 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 standards represents the level of grading standards (calculated as described below in section 2.2) of teacher t. We control for all time-invariant student characteristics with the fixed effect α, and control for all factors invariant within a given school with the fixed effect ξ. 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, γ, 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. Alternative specifications of the above regression employ disciplinary problems as the dependent variable, or replace student fixed effects with family-level fixed effects. In this latter case, X includes a wider set of student level covariates, including base-year test score, race, ethnicity, and sex, that are subsumed into the student-level fixed effect in our primary specification.

2.2. Measuring grading standards

We adopt three alternative measures of teacher-level grading standards, though all are similar in nature to the definition 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 is calculated as follows:

$$\text{standards}(1)_t = \sum_i \sum_y (\text{FCAT}_{ity} - \text{grade}_{ity})/n,$$

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.[1] The higher the value of standards(1), 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.[2] 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 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

[1] 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.

[2] Our results are invariant to changing the ways in which pluses and minuses are treated.

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:

$$\text{FCAT}_{ity} = \delta_t + \beta \text{grade}_{ity} + \epsilon_{ity} ,$$

where all notation is as before. The second measure of standards (standards(2)), then, is the retained estimated teacher-level fixed effect δ_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[3] 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 miniumum 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 79 percent attained level 2 or below.

2.3 Patterns in teacher-level grading standards

[3] 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.

The above-mentioned comparisons provide a first piece of evidence that teachers vary considerably in their grading standards, even within a single school district. This subsection provides evidence on three other important patterns seen with respect to teacher-level grading standards. The first pattern present in the data is that the within-school variation in teacher-level grading standards is almost as great as the population variation in grading standards. As seen in Table 2, in the 1997-98 school year, 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.[4] 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.

The aforementioned evidence suggests that there exist considerable differences within a school in the level of teacher grading standards. However, our identification strategy relies on individual teachers' standards being relatively invariant over time. Table 3 explores the degree to which this is the case. In Table 3 we stratify the set of teachers into thirds in each academic

[4] In Table 2, our method of calculating teacher grading standards only uses data for the year in question. That is, a teacher teaching in both 1997-98 and 1998-99 is assigned separate measures of standards for the purpose of this exercise.

year, for the purpose of measuring the toughest, average, and lightest graders in each year. In the top panel of Table 3 we observe that 69 percent of teachers ranking in the bottom third of standards level in 1997-98 remained in the bottom third, while only 14 percent transitioned to the top third. Among the teachers ranking in the top third of standards in 1997-98, 87 percent remained in the top third in 1998-99, and none fell to the bottom third of standards. All told, 67 percent of the teachers are located on the diagonal of this transition matrix (where 33 percent would be chance) and only 8 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 3 present the results of analogous transition matrices, in which, in turn, teachers taught a higher-achieving class in 1998-99 than in 1997-98 (middle panel) and teachers taught a lower-achieving class in 1998-99 than in 1997-98. Class achievement 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. Seventy percent of low-standards teachers whose class improved in initial ability from year to year remained in the lowest third, and 93 percent of high-standards teachers facing the same improvement remained in the highest third. Seventy-five percent of low-standards teachers whose class declined in initial ability from year to year remained in the lowest third, and 82 percent of high-standards teachers facing the same decline remained in the highest third. 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 achievement 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.

Are grading standards merely reflective of some observed teacher qualification level? To 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.[5] As can be seen in Table 4, 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. Therefore, the fact that we do not observe grading standards levels varying across measured teacher attributes often found to affect student outcomes suggests that the standards effects that we observe are not likely due to unmeasured teacher quality.

[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.

Additional support for this sentiment is shown below, in the results section. We observe that the effects of measured grading standards are highly nonlinear--that is, the estimated effects of grading standards vary interactively between the type of student and setting in which the student is taught. While this observed pattern of results could itself still be consistent with an explanation that unmeasured teacher quality actually drives our results, it is at least as plausible to conclude that the teacher quality measure that drives the results is the teacher's level of standards.

2.4. Teacher-level grading standards and student class assignment

One threat to identification of standards effects concerns the potentially nonrandom assignment of students to teachers. If, for instance, higher-ability students (or students who differ along any other dimension) are more likely to select into high-standards teachers' classes, then the estimated effects of grading standards may be biased. While this concern is much greater in cross-section than when controlling for student-level effects, the potential problem remains even in a student fixed effects model.

The results presented in Table 5 demonstrate why identification from cross-sectional variation, at least in the current context, is likely to lead to faulty inference. We observe that our four dependent variables differ significantly between relatively high-standards teachers and relatively low-standards teachers in the expected manner: high standards are associated with increased test score gains and lower disciplinary problems, in cross-section. However, we also observe that student characteristics tend to differ systematically across standards groups. Specifically, we observe that high-standards teachers have students significantly less likely to be black, significantly more likely to be gifted, and significantly less likely to be free-lunch eligible or learning disabled. The fourth column in Table 5 demonstrates that while the racial and economic differences decline considerably when we compare teacher-level grading standards within the same school, the differences remain statistically significant (as do the other differences) and, in fact, there is also a small but statistically significant difference among the sexes as well. Therefore, we observe in cross-section that high standards are associated with better outcomes, but we also observe in cross-section that high standards are associated with different types of students as well.

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. Table 6 demonstrates that 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. Since randomization suggests that 50 percent of these students would transition from one group to the next, 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. In sum, the evidence suggests that many students experience potentially large changes in grading standards from year to year. We seek to exploit this variation in our empirical analysis.

3. Empirical results

Our empirical results are presented in Table 7. The first row of Table 7 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 Table 5 that these results should not be taken to represent causal effects of grading standards. In the second row of Table 7 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.

The fourth row of Table 7 presents the results of our primary specification--the model

[6] Our empirical results presented below are quite similar if we restrict our analysis to non-gifted students. These results are available on request from the authors.

[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.

with student and school fixed effects. 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 larger in magnitude, and discipline problem results that are smaller in magnitude, than those drawn from models without student fixed effects. These estimated mean effects are also on the cusp of statistical significance, with p-values ranging from 0.08 to 0.14. While not overwhelmingly significant, these results together suggest a pattern of positive mean effects of grading standards on student outcomes.

This conclusion is bolstered by the results presented in row 5 of Table 7, 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. We find that when we use this within-family identification strategy, our results are quite similar to those using the within-student identification strategy, with the estimated effects of the discipline variables being more significant than they are using the within-student identification strategy.

The final two rows of Table 7 present results of model specifications analogous to row 4, that is, using student-level fixed effects, 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, when we employ our alternative measures of grading standards. In sum, our general conclusion from Table 7 is that grading standards have modest effects, on average, on student test scores and discipline problems.

3.1. 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.[8] Therefore, in Tables 8A and 8B (for mathematics and reading, respectively) we revise our primary model (Table 7, row 4) 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 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

[8] They also find that minority students are harmed by grading standards because standards are estimated to reduce minority high school graduation rates.

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 Tables 8A and 8B are models that interact grading standards with the class's average third grade mathematics (or reading) score.[9] 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 achieving *classes* may fare somewhat better with higher standards than with lower achieving classes, although these distinctions are not statistically significant at conventional levels.

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 Tables 8A and 8B address this question. Specifications 3M and 3R examine the differential effects of grading standards on initially above-average students as 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 only statistically significant in the case of mathematics. 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 two or three percent level, depending on the test

[9] 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.

score considered. In other words, low-achieving students differentially benefit from high standards when they are in a high-achieving class, and high-achieving students differentially benefit from high standards when they are in a low-achieving 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-achievers in high-achieving classes and high-achievers in low-achieving classes the most.[10] These results are clearest when presented in Table 9, which translates these point estimates into predicted years of test score gains[11] associated with increased standards at different points of the student ability-class ability continuum. The results are large in magnitude as well as statistical significance: 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 half a year or more of reading test score gains. 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-achievers in low-achieving classes than

[10] The models also include a two-way interaction between class average and own score, which is omitted from the table.

[11] We measure a "year of test score gain" as the average gain from one year to the next in Alachua County Public Schools. Because Alachua County gain scores tend to be larger than the national average, these are more conservative estimates of "years of gain" than are those based on national grade equivalents.

for their counterparts in high-achieving classes. Likewise, low-achievers in high-achieving classes are at relatively more "risk" of receiving a low grade than are low-achievers in low-achieving classes. Hence, it seems sensible that high standards that lower the "safety" for high-achievers in low-achieving classes may generate more effort and greater learning, as might high standards that increase the "risk" for low-achievers in high-achieving classes. While this is by no means a definitive explanation of our empirical findings, it is a plausible explanation.

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. 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.

References

Becker, William and Sherwin Rosen. 1990. “The Learning Effect of Assessment and Evaluation in High School.” Discussion paper 90-7, Economics Research Center, NORC.

Betts, Julian. 1995. “Do Grading Standards Affectr the Incentive to Learn?” Working paper, University of California-San Diego.

Betts, Julian. 1998. “The Impact of Educational Standards on the Level and Distribution of Earnings.” *American Economic Review*, 266-275.

Betts, Julian and Jeff Grogger. 2000. “The Impact of Grading Standards on Student Achievement, Educational Attainment, and Entry-Level Earnings.” NBER working paper 7875, September.

Costrell, Robert. 1994. “A Simple Model of Educational Standards.” *American Economic Review*, 956-971.

Goldhaber, Dan and Dominic Brewer. 1997. “Why Don’t Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity.” *Journal of Human Resources*, 505-523.

Hanushek, Eric. 1986. “The Economics of Schooling.” *Journal of Economic Literature* 1141-1177.

Lillard, Dean and Philip DeCicca. Forthcoming. “Higher Standards, More Dropouts? Evidence Within and Across Time.” *Economics of Education Review.*

Moulton, Brent. 1986. “Random Group Effects and the Precision of Regression Estimates.” *Journal of Econometrics*, 385-397.

Rivkin, Steven, Eric Hanushek, and John Kain. 1998. “Teachers, Schools, and Academic Achievement.” NBER working paper 6691, August.

Table 1: Distribution of letter grades and FCAT Scores

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

Assigned letter grade	FCAT level (5=highest; 1=lowest)				
	level 5	level 4	level 3	level 2	level 1
A+/A/A-	0.09	0.41	0.34	0.11	0.06
B+/B/B-	0.01	0.10	0.28	0.31	0.30
C+/C/C-	0.00	0.02	0.12	0.25	0.62
D+/D/D-	0.00	0.02	0.06	0.16	0.76
E/F	0.00	0.00	0.00	0.08	0.92

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

Assigned letter grade	FCAT level (5=highest; 1=lowest)				
	level 5	level 4	level 3	level 2	level 1
A+/A/A-	0.12	0.53	0.30	0.05	0.00
B+/B/B-	0.02	0.19	0.43	0.28	0.08
C+/C/C-	0.00	0.04	0.23	0.31	0.42
D+/D/D-	0.00	0.03	0.11	0.21	0.65
E/F	0.00	0.00	0.00	0.13	0.87

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

Assigned letter grade	FCAT level (5=highest; 1=lowest)				
	level 5	level 4	level 3	level 2	level 1
A+/A/A-	0.04	0.24	0.40	0.19	0.13
B+/B/B-	0.00	0.03	0.18	0.34	0.45
C+/C/C-	0.00	0.00	0.05	0.20	0.75
D+/D/D-	0.00	0.00	0.00	0.11	0.88
E/F	0.00	0.00	0.00	0.00	1.00

Table 2: Within-school versus population variation in teacher-level grading standards

School year	1997-98	1998-99
Population standard deviation in grading standards	0.68	0.79
Mean within-school standard deviation in grading standards	0.60	0.72
Between-school standard deviation in grading standards	0.30	0.28

Table 3: Persistence of grading standards across years

I. Full population of teachers: fraction of teachers transitioning to each standards group (row percentages)

"Standards third" in 1997-98 academic year	"Standards third" in 1998-99 academic year		
	Bottom third of standards	Middle third of standards	Top third of standards
Bottom third of standards	0.69	0.17	0.14
Middle third of standards	0.18	0.45	0.36
Top third of standards	0.00	0.13	0.87

Fraction on diagonal: 0.67 Fraction transitioning from top to bottom, or vice versa: 0.08

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 (row percentages)

"Standards third" in 1997-98 academic year	"Standards third" in 1998-99 academic year		
	Bottom third of standards	Middle third of standards	Top third of standards
Bottom third of standards	0.70	0.15	0.15
Middle third of standards	0.06	0.56	0.38
Top third of standards	0.00	0.07	0.93

Fraction on diagonal: 0.73 Fraction transitioning from top to bottom, or vice versa: 0.08

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 (row percentages)

"Standards third" in 1997-98 academic year	"Standards third" in 1998-99 academic year		
	Bottom third of standards	Middle third of standards	Top third of standards
Bottom third of standards	0.75	0.17	0.08
Middle third of standards	0.25	0.38	0.38
Top third of standards	0.00	0.18	0.82

Fraction on diagonal: 0.65 Fraction transitioning from top to bottom, or vice versa: 0.05

Table 4: Teacher-level grading standards and observed teacher quality measures

Teacher characteristic	Mean of characteristic for above-median standards teacher	Mean of characteristic for below-median standards teacher	Difference
Years of experience	19.66	16.97	2.69
Teacher attended a selective undergraduate institution	0.90	0.94	-0.04
Teacher attended a very selective undergraduate institution	0.53	0.67	-0.14
Teacher has a masters degree	0.65	0.40	0.25**

Notes to Table 4: Differences denoted by ** are significant at the five percent level; differences denoted by * are significant at the ten percent level.

Table 5: Descriptive statistics: Means of dependent variables and selected student characteristics for different levels of measured grading standards

Variable	Mean for students with tougher-than-median teachers	Mean for students with easier-than-median teachers	Difference	Difference (when looking only at within-school differences in standards
Change in math score	16.82	14.87	1.95**	1.84**
Change in reading score	17.40	14.84	2.57**	1.14**
At least one disciplinary infraction	0.15	0.24	-0.09**	-0.07**
At least one severe infraction	0.13	0.23	-0.09**	-0.07**
Black	0.22	0.43	-0.21**	-0.06**
Hispanic	0.04	0.04	0.00	0.01
Female	0.53	0.51	0.01	0.02*
Gifted	0.24	0.08	0.16**	0.16**
Free lunch	0.40	0.58	-0.18**	-0.05*
Learning disabled	0.03	0.12	-0.09**	-0.10**

Notes to Table 5: Differences denoted by ** are significant at the five percent level; differences denoted by * are significant at the ten percent level.

Table 6: Students' propensity to have two successive teachers in same standards class

Group of students	Fraction of students with below-median standards teachers in grade 4 continuing to have below-median standards teachers in grade 5	Fraction of students with above-median toughness teachers in grade 4 continuing to have above-median toughness teachers in grade 5
Full sample	0.57	0.54
Black students	0.52	0.56
White students	0.56	0.53
Gifted students	0.43	0.67
Not gifted students	0.55	0.51
Free-lunch eligible	0.55	0.58
Not free-lunch eligible	0.52	0.51

Notes to Table 6: The grading standards classes described above are school-specific.

Table 7: Estimated effects of teacher grading standards on student outcomes

	Dependent variable			
	Change in ITBS math test scores	Change in ITBS reading test scores	At least one disciplinary infraction	At least one severe disciplinary infraction
(1) No covariates included	2.817 (p=0.000)	2.754 (p=0.000)	-0.124 (p=0.000)	-0.120 (p=0.000)
(2) Controlling for race, ethnicity, sex, free lunch status, gifted status, disability	1.583 (p=0.005)	1.875 (p=0.000)	-0.029 (p=0.043)	-0.028 (p=0.035)
(3) Also including school fixed effects	1.912 (p=0.006)	2.026 (p=0.001)	-0.053 (p=0.000)	-0.055 (p=0.000)
(4) Also including school and **student** fixed effects	3.251 (p=0.114)	5.414 (p=0.081)	-0.023 (p=0.103)	-0.020 (p=0.139)
(5) Also including school and **family** fixed effects	2.860 (p=0.118)	4.227 (p=0.100)	-0.028 (p=0.071)	-0.028 (p=0.064)
(6) School and **student** fixed effects: using FIXED EFFECT measure of standards	3.135 (p=0.122)	5.976 (p=0.025)	-0.028 (p=0.043)	-0.026 (p=0.051)
(7) School and **student** fixed effects: using "GRADE B" measure of standards	2.423 (p=0.098)	4.226 (p=0.052)	-0.024 (p=0.020)	-0.020 (p=0.046)

Notes to Table 7: 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 8A: Differential effects of high grading standards on mathematics test scores (all using *student* fixed effects model, akin to Row 4, Table 7)

	Dependent variable: change in **math** score				
Specification	(1M)	(2M)	(3M)	(4M)	(5M)
Students included in regression	All	All	Above average math in grade 3	Below average math in grade 3	All
Grading standards	3.544 (p=0.00)	4.427 (p=0.00)	4.309 (p=0.02)	4.612 (p=0.00)	4.592 (p=0.00)
Grading standards x 3rd grade math score	1.555 (p=0.15)				0.388 (p=0.77)
Grading standards x class average 3rd grade math score		1.992 (p=0.13)	-5.132 (p=0.09)	4.141 (p=0.02)	0.042 (p=0.98)
Grading standards x class average x own score					-2.517 (p=0.09)

Notes to Table 8A: Each column represents a separate regression. Robust p-values are in parentheses beneath point estimates.

Table 8B: Differential effects of high grading standards on reading test scores (all using *student* fixed effects model, akin to Row 4, Table 7)

	Dependent variable: change in **reading** score				
Specification	(1R)	(2R)	(3R)	(4R)	(5R)
Students included in regression	All	All	Above average reading in grade 3	Below average reading in grade 3	All
Grading standards	5.882 (p=0.00)	6.964 (p=0.00)	11.073 (p=0.00)	5.640 (p=0.01)	8.048 (p=0.00)
Grading standards x 3rd grade reading score	3.005 (p=0.01)				2.162 (p=0.16)
Grading standards x class average 3rd grade reading score		3.549 (p=0.02)	-2.229 (p=0.46)	5.462 (p=0.03)	1.182 (p=0.56)
Grading standards x class average x own score					-3.901 (p=0.01)

Notes to Table 8B: Each column represents a separate regression. Robust p-values are in parentheses beneath point estimates.

Table 9: Estimated effects of a one standard deviation increase in grading standards, for students and classes at different points in the initial ability distribution (derived from Table 8, specifications 5M and 5R)

I. Years of growth in **mathematics** performance attributed to increased toughness

3rd grade mathematics performance of student:	Average 3rd grade mathematics performance of children in classroom			
	1.5 s.d. below mean	0.5 s.d. below mean	0.5 s.d. above mean	1.5 s.d above mean
1.5 s.d. below	-0.06	0.08	**0.22**	**0.36**
0.5 s.d. below	0.09	**0.14**	**0.18**	**0.23**
0.5 s.d. above	**0.24**	**0.20**	**0.15**	0.11
1.5 s.d. above	**0.39**	**0.26**	0.12	-0.02

II. Years of growth in **reading** performance attributed to increased toughness

3rd grade reading performance of student:	Average 3rd grade reading performance of children in classroom			
	1.5 s.d. below mean	0.5 s.d. below mean	0.5 s.d. above mean	1.5 s.d above mean
1.5 s.d. below	-0.21	0.05	**0.31**	**0.56**
0.5 s.d. below	0.08	**0.20**	**0.31**	**0.43**
0.5 s.d. above	**0.38**	**0.35**	**0.32**	**0.29**
1.5 s.d. above	**0.67**	**0.50**	**0.33**	0.16

Notes to Table 9: Boxes with bold numbers indicate cells statistically distinct from zero at the five percent level or lower. No cells above are significant at the ten, but not five, percent level.