



ELSEVIER

Journal of Public Economics x (2003) xxx–xxx

 JOURNAL OF
 PUBLIC
 ECONOMICS

www.elsevier.com/locate/econbase

4

Do high grading standards affect student performance?

David N. Figlio^{a,b,*}, Maurice E. Lucas^c

^aDepartment of Economics, University of Florida, Gainesville, FL 32611-7140, USA

^bNational Bureau of Economic Research, USA

^cSchool Board of Alachua County, FL, USA

Received 6 December 2001; received in revised form 12 February 2003; accepted 13 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.

© 2003 Published by Elsevier Science B.V.

Keywords: Student performance; Grading standards

JEL classification: I2

*Corresponding author. Department of Economics, University of Florida, Gainesville, FL 32611-7140, USA.

E-mail address: figlio@ufl.edu (D.N. Figlio).

38 1. Introduction

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

46 We know of three empirical studies that examine the effects of standards on
47 student outcomes. Lillard and DeCicca (2003) are not interested in the effects of
48 grading standards per se, but rather on the effects of graduation standards,
49 measured by the number of courses required for graduation. They find that higher
50 graduation standards lead to relatively increased dropout rates. Two current
51 working papers (Betts, 1995; Betts and Grogger, 2000, the latter of which was
52 written simultaneously with this paper) present the only empirical work that, to our
53 knowledge, focuses on *grading* standards. Both papers present cross-sectional
54 evidence on the effects of school-level grading standards (measured by their
55 grade-point average relative to test scores) on the level (Betts, 1995) and
56 distribution (Betts and Grogger, 2000) of student test scores, educational attain-
57 ment, and early labor market earnings. Consistent with the theoretical literature,
58 Betts and Grogger (2000) find significant evidence of differential effects of
59 grading standards, depending on student type.

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

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

78 Third, the existing literature (as well as almost all of the work studying other
79 determinants of student outcomes) focuses on students in upper grades rather than

86 at the elementary level. This is, in some ways, an advantage, because one can then
87 measure educational attainment and follow students into the labor market. But in
88 other ways this is a disadvantage, both because sample attrition is likely to be less
89 of a factor at the elementary level and because one might reasonably expect that
90 the most important grades, in terms of student learning, are the early ones.

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

112 2. Data and methods

113 We analyze confidential student-level data provided by the School Board of
114 Alachua County, Florida for this project. Our data consist of observations on
115 almost every third, fourth, and fifth grader in the school system between 1995–96
116 and 1998–99.¹ Alachua County Public Schools is a relatively large district (by
117 national standards), averaging about 1800 test-taking students per year, per grade.
118 Alachua County is racially heterogeneous, with 60 percent of students white, 34
119 percent African-American, 3 percent Hispanic, and 2 percent Asian. Less than one
120 percent receive services for English as a Second Language. Forty-nine percent of

85

81 _____
82 ¹If a child is retained and repeats a grade, we consider year-to-year changes in test scores within a
83 grade level; in other words, we include grade-retained students in the present analysis. Our results are
84 invariant in general magnitude and statistical significance levels to changes in how we deal with (or
whether we include or exclude) grade-retained students.

122 the student body are eligible for subsidized lunches, 19 percent are identified as
123 gifted, and 8 percent are learning disabled.

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

137 In addition, we observe each student's report card in each year for each subject.
138 Furthermore, we are able to match students to teachers, which is essential, of
139 course, for measuring the effects of grading standards at the teacher level. Student
140 records also record the student's race, ethnicity, sex, disability status, and gifted
141 status, as well as the student's discipline record.

142 We employ four dependent variables of interest. Our primary dependent
143 variables are the change from one year to the next in the student's performance on
144 the Iowa Test of Basic Skills' mathematics or reading assessments. We focus on
145 changes in test scores, rather than levels, so that we can control, at least cursorily,
146 for student-specific trends in test performance over time. In addition, we also use
147 as a dependent variable indicators for whether the student had at least one
148 disciplinary infraction that merited recording, or alternatively, at least one severe
149 disciplinary infraction, in a given year. All told, we employ approximately 7000
150 observations each (for mathematics and reading) of changes in test scores from
151 one year to the next—two sets of year-to-year changes apiece for the two cohorts
152 of students for whom we have three years of data.

153 2.1. Identifying the effects of grading standards

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

$$158 \quad \Delta test_{itsy} = \alpha_i + \gamma standards_i + \phi C_{itsy} + \theta X_{iy} + \xi_s + \varepsilon_{itsy},$$

159 where $\Delta test$ represents the change from one year to the next in student i 's Iowa

161 Test of Basic Skills mathematics (or reading) scaled examination score, and
162 *standards* represents the level of grading standards (calculated as described below
163 in Section 2.2) of teacher t . We identify the parameter γ from students with
164 teachers with measured standards levels in both grades 4 and 5. The use of a
165 first-differenced dependent variable allows us to capture a sort of ‘pre-test’ effect.
166 We control for all student characteristics that are either time-invariant or that trend
167 over time with the fixed effect α , and control for all factors invariant within a
168 given school with the fixed effect ξ . The vector C includes variables representing
169 the composition of the classroom; we control for the fraction white, the fraction
170 free-lunch-eligible, and the average third grade mathematics test score among the
171 students in the classroom in question in year y . The vector X represents the set of
172 student-level variables that change over time. In practice, X includes free lunch
173 status, gifted status, and disability status, all of which can change from year to
174 year. Our parameter of interest is the coefficient on teacher grading standards, γ ,
175 which represents the effects of changing a student from one level of grading
176 standards to another, holding constant all student and school attributes that do not
177 change over time, as well as time-varying student and classroom characteristics.
178 Alternative specifications of the above regression employ disciplinary problems as
179 the dependent variable.

180 We employ a difference specification because there exists very strong evidence
181 that students differ systematically in their *rates* of achievement growth over time,
182 and not merely in their *levels* of achievement. Put differently, students who begin
183 at a high level tend to have test score growth rates that eclipse those who begin at
184 a low level. For instance, in our sample mathematics growth rates for students
185 scoring in the top quartile of the third grade mathematics test score distribution are
186 more than 20 percent greater than mathematics growth rates for students scoring in
187 the bottom quartile of the third grade distribution. The difference in reading test
188 score growth rates between top- and bottom-initial-achievers is smaller—about 10
189 percent—but still present and statistically significant at the one percent level. It
190 turns out that our choice of using a difference specification tends to lead to more
191 conservative estimates of the effects of grading standards on test performance,
192 relative to a ‘levels’ specification, which is sensible considering the apparent
193 non-random assignment of students to teachers of varying grading standards
194 described later in the paper.

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

209 2.2. Measuring grading standards

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

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

$$230 \quad standards(1)_t = \sum_i \sum_y (FCAT_{ity} - grade_{ity})/n,$$

231 where t represents the teacher, i represents the student, y represents the year, and n
 232 reflects the number of student-year pairs faced by the teacher.² The higher the
 233 value of $standards(1)$, the higher the standards, because it suggests that students
 234 require a higher score on the FCAT to achieve any given letter grade. The variable
 235 grade is measured in standard grade-point fashion, with an A earning a score of 4,
 236 a B earning a score of 3, and so on. Pluses earn an additional 0.33, while minuses
 237 lead to a reduction of 0.33.³ Therefore, this measure represents the average gap
 238 between the FCAT score and the teacher-assigned letter grade for each particular
 239 teacher. Since students take the FCAT mathematics examination in fifth grade and
 240 the FCAT reading examination in fourth grade, this measure of grading standards
 241 is calculated using mathematics grades and scores for fifth grade teachers and
 242 using reading grades and scores for fourth-grade teachers. For teachers who
 243 switched between these grades during the years of FCAT administration, this
 244 measure of grading standards is computed using both mathematics and reading

207

205 ²Put differently, n represents the number of students taught by the teacher in the years in which both
 206 FCAT scores and letter grades are observed.

208 ³Our results are invariant to changing the ways in which pluses and minuses are treated.

250 scores, depending on the grade level at the time of FCAT assessment. The benefit
 251 of measuring standards in this way is that it ensures that we will observe standards
 252 measures for both a fourth-grade teacher and a fifth-grade teacher for as many
 253 students as possible. The available evidence suggests that this construction is
 254 reasonable: among the teachers who switched between the two grades over the
 255 course of our sample, the correlation between a teacher's reading standards (in
 256 fourth grade) and mathematics standards (in fifth grade) is nearly 0.80. Put
 257 differently, teachers with high reading standards tend to have high mathematics
 258 standards as well, and vice versa.

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

$$261 \quad FCAT_{ity} = \delta_t + \beta \text{ grade}_{ity} + \varepsilon_{ity},$$

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

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

276 The top panel of Table 1 illustrates that, on average, teachers tend to grade less
 277 stringently than the state standards (as reflected in FCAT scores) indicate that they
 278 should. Only nine percent of students awarded A's by their teachers⁴ attained the
 279 corresponding FCAT level, and in fact, only 50 percent attained even level 4. Only
 280 11 percent of students awarded B's by their teachers attained level 4 or above, and
 281 a mere 39 percent attained level 3 or above. Of the students awarded C's by their
 282 teachers, only 14 percent attained level 3 or above, and only 39 percent attained
 283 level 2 or above. Put differently, 86 percent of 'C students' failed to achieve a
 284 minimum acceptable level of competency (level 3) according to the Florida
 285 standards, and even 61 percent of 'B students' and 17 percent of 'A students'
 286 failed to meet this competency level.

287 The middle and bottom panels of Table 1 show that these patterns appear much

249

246 _____
 247 ⁴For the purposes of presentation in this exercise, we collapse plus and minus grades into a single
 248 letter grade. The grading standards measures all distinguish between plus and minus grades, as
 mentioned above.

289 Table 1
 290 Distribution of letter grades and FCAT scores
 291

292 293 294 295	Assigned	FCAT level (5=highest; 1=lowest)				
	letter grade	Level 5	Level 4	Level 3	Level 2	Level 1
296	<i>I. Overall distribution of FCAT scores, by letter grade</i>					
297	<i>(row percentages are reported)</i>					
298	A+ /A/A-	0.09	0.41	0.34	0.11	0.06
299	B+ /B/B-	0.01	0.10	0.28	0.31	0.30
300	C+ /C/C-	0.00	0.02	0.12	0.25	0.62
301	D+ /D/D-	0.00	0.02	0.06	0.16	0.76
302	E/F	0.00	0.00	0.00	0.08	0.92
303	<i>II. Distribution of FCAT scores, by letter grade,</i>					
304	<i>teachers with above-median standards</i>					
305	A+ /A/A-	0.12	0.53	0.30	0.05	0.00
306	B+ /B/B-	0.02	0.19	0.43	0.28	0.08
307	C+ /C/C-	0.00	0.04	0.23	0.31	0.42
308	D+ /D/D-	0.00	0.03	0.11	0.21	0.65
309	E/F	0.00	0.00	0.00	0.13	0.87
310	<i>III. Distribution of FCAT scores, by letter grade,</i>					
311	<i>teachers with below-median standards</i>					
312	A+ /A/A-	0.04	0.24	0.40	0.19	0.13
313	B+ /B/B-	0.00	0.03	0.18	0.34	0.45
314	C+ /C/C-	0.00	0.00	0.05	0.20	0.75
315	D+ /D/D-	0.00	0.00	0.00	0.11	0.88
316 317	E/F	0.00	0.00	0.00	0.00	1.00

318 different for teachers with relatively high standards (the middle panel) and
 319 teachers with relatively low standards (the bottom panel). Here, we stratify
 320 teachers according to whether they are above or below the district median in
 321 standards, as defined by the first measure described above. Among relatively tough
 322 graders, 65 percent of A students attained level 4 or above while 5 percent attained
 323 level 2 or below. Among relatively light graders, in comparison, only 28 percent of
 324 A students attained level 4 or above while 32 percent attained level 2 or below.
 325 Among relatively tough graders, 21 percent of B students attained level 4 or above
 326 while 36 percent attained level 2 or below. Among relatively light graders,
 327 however, just 3 percent of A students attained level 4 or above while 79 percent
 328 attained level 2 or below.

329 2.3. Patterns in teacher-level grading standards

330 The above-mentioned comparisons provide a first piece of evidence that
 331 teachers vary considerably in their grading standards, even within a single school
 332 district. It turns out that the within-school variation in teacher-level grading
 333 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

Table 2
Persistence of grading standards across years

'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
<i>I. Full population of teachers: fraction of teachers transitioning to each standards group</i>			
Bottom third of standards	0.26	0.07	0.02
Middle third of standards	0.05	0.17	0.10
Top third of standards	0.00	0.08	0.25
Fraction on diagonal: 0.68. Fraction transitioning from top to bottom, or vice versa: 0.02			
<i>II. Teachers whose average class 'quality' (measured by average third grade test scores) improved from 1997–98 to 1998–99: fraction of teachers transitioning to each standards group</i>			
Bottom third of standards	0.21	0.10	0.00
Middle third of standards	0.05	0.21	0.12
Top third of standards	0.00	0.02	0.29
Fraction on diagonal: 0.71. Fraction transitioning from top to bottom, or vice versa: 0.00			
<i>III. Teachers whose average class 'quality' (measured by average third grade test scores) fell from 1997–98 to 1998–99: fraction of teachers transitioning to each standards group</i>			
Bottom third of standards	0.27	0.04	0.02
Middle third of standards	0.04	0.16	0.08
Top third of standards	0.00	0.13	0.24
Fraction on diagonal: 0.65. Fraction transitioning from top to bottom, or vice versa: 0.05			

388 standards level in 1997–98 remained in the bottom third, while only 6 percent
389 transitioned to the top third. Among the teachers ranking in the top third of
390 standards in 1997–98, 77 percent remained in the top third in 1998–99, and none
391 fell to the bottom third of standards. All told, 68 percent of the teachers are located
392 on the diagonal of this transition matrix (where 33 percent would be chance) and
393 only 2 percent of those able to do so transitioned from one corner of this matrix to
394 another from year to year.

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

414 Are grading standards merely reflective of some observed teacher qualification
415 level? To determine the degree to which this is the case, we compare teachers with
416 relatively high (above-median) measures of standards to teachers with relatively
417 low (below-median) measures of standards.⁵ Teachers with relatively high levels of
418 standards are slightly more experienced and are slightly less likely to have
419 attended a selective or highly selective undergraduate institution, though none of
420 these differences are statistically different. One difference that is statistically
421 significant is the fraction of teachers with masters degrees; high-standards teachers
422 are more likely to have masters degrees than are low-standards teachers. While this
423 difference suggests that high-standards teachers are observably different from
424 low-standards teachers in at least one dimension, other evidence suggests that this
425 is one dimension that rarely is found to matter for student achievement (see, e.g.,

387

388 _____
389 ⁵These comparisons are only for teachers still employed by the School Board of Alachua County in
384 2000, almost 85 percent of the teachers in our sample. There is no apparent difference in average
385 standards levels between teachers still employed by the district and teachers no longer employed by the
386 district.

430 Hanushek, 1986). On the other hand, the measured teacher attributes generally
431 found to affect student outcomes the most, the selectivity of teacher undergraduate
432 institutions (Goldhaber and Brewer, 1997), is not different between the standards
433 groups. In models presented below, however, we directly control for these teacher
434 qualification measures to rule out the possibility that observed teacher qualification
435 measures may drive the estimated effects of grading standards on student
436 outcomes. Later we discuss results suggesting that our findings are also unlikely to
437 be driven by one important unmeasured teacher quality dimension.

438 2.4. Teacher-level grading standards and student class assignment

439 One threat to identification of standards effects concerns the potentially
440 nonrandom assignment of students to teachers. In cross-section, high-standards
441 teachers also have students who perform higher and have better disciplinary
442 outcomes. But they also have students who are more likely to be white or gifted,
443 and less likely to be low-income or learning disabled. These differences are
444 present even within a single school. Hence, it is unclear that these outcomes
445 associated with high standards are actually due to the high standards themselves.

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

465 Students are not, however, randomly assigned to classrooms, and high-perform-
466 ing students may systematically select into high-standards teachers' classrooms. If
467 teachers tend to grade on a curve, then teachers who have better students on

429

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

484 average will also be measured as having higher grading standards, regardless of
 485 the teacher's actual standards level. It follows that to the extent to which students
 486 self-select dynamically into classes, the estimated effects of grading standards will
 487 be biased. The direction of this bias is not immediately known, and depends on the
 488 relationship between changes in prior test scores and subsequent classroom
 489 placement.

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

510 The suspicion of understated results is strengthened by the evidence suggesting
 511 the presence of dynamic mean reversion presented in the first two columns of
 512 Table 3. These columns indicate the presence of a strong negative correlation
 513 between test score gains between grades 3 and 4 and test score gains between
 514 grades 4 and 5. Therefore, while subsequent test score gains are correlated with

469 Table 3
 470 Mean changes in grading standard transitions faced by students between grades 4 and 5, by change in
 471 student mathematics performance between grades 3 and 4
 472

473 Student group, based 474 on change in math 475 score between 476 grades 3 and 4	477 Mean change 478 in math score 479 between 480 grades 3 and 4	481 Mean grade 482 in math score 483 between 484 grades 4 and 5	485 Grading standards faced by student 486 in grade 4		
			487 Lowest 488 standard third	489 Middle 490 standard third	491 Highest 492 standard third
493 Lowest third	2.44	18.75	0.32	-0.28	-0.93
494 Middle third	15.31	15.69	0.37	-0.22	-0.84
495 Highest third	28.97	11.21	0.38	-0.09	-0.82

483 Note: teacher grading standards are standardized for the purpose of presentation.

550 initial performance, it is not the case that students who gain more in one year gain
 551 more in every year. Instead, the evidence suggests that students who gain the most
 552 between grades 3 and 4 are assigned a tougher teacher in grade 5 and subsequently
 553 do comparatively poorly, in terms of test score gains between grades 4 and 5. This
 554 relationship should work *against* finding a positive relationship between test score
 555 growth and grading standards in a student fixed-effects model.

556 3. Empirical results

557 Our regression results are presented in Table 4. The first row of Table 4 presents
 558 the results of a model with no covariates included.⁷ We observe large, statistically
 559 significant relationships between grading standards and all four dependent
 560 variables. However, it is clear from the above discussion on selection into

516 Table 4
 517 Estimated effects of teacher grading standards on student outcomes
 518

	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
526 (1) No covariates included	2.817 (<i>P</i> =0.000)	2.754 (<i>P</i> =0.000)	-0.124 (<i>P</i> =0.000)	-0.120 (<i>P</i> =0.000)
528 (2) Controlling for race, ethnicity, sex, free 529 lunch status, gifted status, disability	1.583 (<i>P</i> =0.005)	1.875 (<i>P</i> =0.000)	-0.029 (<i>P</i> =0.043)	-0.028 (<i>P</i> =0.035)
530 (3) Same as (2) but also including school 531 fixed effects	1.912 (<i>P</i> =0.006)	2.026 (<i>P</i> =0.001)	-0.053 (<i>P</i> =0.000)	-0.055 (<i>P</i> =0.000)
532 (4) Same as (3) but also including fraction 533 white, fraction free-lunch-eligible, and average 534 third grade test performance in class	2.544 (<i>P</i> =0.005)	2.482 (<i>P</i> =0.001)	-0.030 (<i>P</i> =0.073)	-0.028 (<i>P</i> =0.081)
535 (5) Same as (4) but also including teacher 536 years of experience, education level, and 537 selectivity of undergraduate institution	2.328 (<i>P</i> =0.022)	2.819 (<i>P</i> =0.001)	-0.035 (<i>P</i> =0.068)	-0.030 (<i>P</i> =0.098)
538 (6) Same as (5) but also including student 539 fixed effects	4.039 (<i>P</i> =0.062)	7.696 (<i>P</i> =0.016)	-0.025 (<i>P</i> =0.198)	-0.011 (<i>P</i> =0.562)
540 (7) Specification (6): using FIXED EFFECT 541 measure of standards	4.214 (<i>P</i> =0.046)	8.131 (<i>P</i> =0.003)	-0.032 (<i>P</i> =0.097)	-0.017 (<i>P</i> =0.345)
542 (8) Specification (6): using 'GRADE B' 543 measure of standards 544	2.964 (<i>P</i> =0.040)	4.674 (<i>P</i> =0.060)	-0.037 (<i>P</i> =0.056)	-0.017 (<i>P</i> =0.208)

545 Notes: each cell represents a separate regression. Robust *P*-values (standard errors corrected for
 546 clustering of observations within classes) are in parentheses beneath point estimates.

549

547 ⁷Here and elsewhere, we adjust our standard errors for within-class clustering. See Moulton (1986)
 548 for an illustration of the importance of adjusting the standard errors in this manner.

569 classrooms that these results should not be taken to represent causal effects of
570 grading standards. In the second row of Table 4 we include the student-level
571 covariates available to us in the data—race, ethnicity, sex, free lunch status, gifted
572 status, and disability status—and find our four results still statistically significant,
573 but considerably diminished in magnitude. The third row adds school-level fixed
574 effects to control for any factors common to all students in a school, leading to
575 similar, but somewhat stronger results.

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

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

600 The final two rows of Table 4 present results of model specifications analogous
601 to row 6, except that we vary the definition of grading standards, as described in
602 Section 2.2 above. We find that our results tend to have similar magnitudes, yet are
603 somewhat more statistically significant (and considerably moreso in the case of
604 discipline) when we employ our alternative measures of grading standards. In sum,

568

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

612 our general conclusion from Table 4 is that grading standards have modest effects,
613 on average, on student test scores and discipline problems. These results are not
614 symmetric, however. In models that distinguish between transitions from relatively
615 easy to relatively tough graders and transitions between relatively tough to
616 relatively easy graders, the results suggest that students benefit more from high
617 grading standards in fourth grade than in fifth grade. The coefficient on the
618 standards measure when the student transitions from a more challenging teacher in
619 fourth grade to an easier teacher in fifth grade is very large and strongly
620 statistically significant, while the coefficient on the standards measure for students
621 who transition from an easier teacher in fourth grade to a more challenging teacher
622 in fifth grade is considerably smaller and statistically insignificant. Whether
623 standards matter more in earlier grades or whether the specific nature of the
624 transition is what matters remains an open question.

625 3.1. One explanation for these findings: home production

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

634 In Spring 2001 we conducted a survey of parents in Alachua County Public
635 Schools to assess the possibility of this second explanation.¹⁰ We surveyed the
636 population of families with students in both fourth and fifth grades, and asked the
637 responsible parent to report on how much time he or she spends weekly helping
638 each of the two children with their homework. Sibling comparisons such as these
639 allowed us to control for factors (e.g., parental motivation) that might be common
640 to both siblings in a household. We found that, holding constant the child's grade
641 level (i.e., fourth or fifth grade), third grade test scores, and the average third grade
642 test score in the child's class, parents systematically spend more time helping the
643 child with the tougher teacher (by our measures) with homework than they do
644 helping the sibling with the easier teacher. The results are statistically significant
645 and large in magnitude: we estimated that a parent of a child with a 25th-percentile
646 teacher (in terms of grading standards)—that is, a relatively tough teacher—would

608

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

609 ¹⁰Survey participants are similar to the school population as a whole, in terms of racial, economic,
610 and gifted composition. We appreciate the helpful comment by Karen Conway that inspired us to
611 conduct this survey.

648 spend 60 percent more time helping that child with homework than he or she
649 would spend with that child's sibling who had a 75th-percentile teacher. These
650 results are not due to the parents reporting that tougher teachers assign more
651 homework—indeed, we estimate that, from parental reports, the typical 25th-
652 percentile teacher assigns only 10 percent more homework than the typical
653 75th-percentile teacher. This is consistent with our findings from personal
654 interviews with principals in the district, who report that teachers within any given
655 grade level in the school work to assign the same amount of homework per week.
656 We have no way of judging whether the homework assigned by tougher teachers is
657 more challenging than that assigned by easier-grading teachers.

658 An additional interesting finding from this survey is that *parents do not perceive*
659 *tougher teachers to be better teachers*. We asked each parent to grade their
660 children's teachers from A to F. While there is relatively low variation in these
661 grades (as two-thirds of the parents assigned grades of A to the teachers), the
662 results suggest that, if anything, parents view tough teachers *less favorably* than
663 they view easier teachers. Using the same within-family comparisons as above, we
664 found that parents were 50 percent more likely to assign a grade of B or below to a
665 25th-percentile teacher than to a 75th-percentile teacher, again controlling for
666 grade level, student third grade test score, and average third grade test score in the
667 class. This result, significant at the 16 percent level, suggests that our measure of
668 grading standards is not merely reflecting some other attribute of a teacher that is
669 viewed as desirable to parents.

670 While these survey findings are not conclusive, they do indicate that high
671 grading standards are likely not merely representing other measures of teacher
672 desirability to parents, and that high grading standards may motivate parents to
673 increase their involvement in their children's education. Both findings bolster the
674 argument that it is high grading standards, rather than some unobservable measure
675 of teacher quality, that is responsible for the observed performance gains.

676 3.2. 'Curve grading' as an alternative explanation

677 The prospect remains that the results described above are deterministically due
678 to the proclivities of teachers to grade on a curve. While the presence of
679 classroom-level student characteristics, including mean initial test score, should
680 tend to dampen this potential effect, one cannot entirely rule out curve grading as
681 an alternative explanation. Table 5, however, makes clear that teachers of different
682 standards levels are likely to assign different grade distributions to their classes,
683 and to students who would be forecast to receive the same grade based only on
684 initial test performance. This table breaks down students by quintile of initial test
685 performance, and teachers by quintile of measured grading standard, and reports
686 the proportion of students in each initial performance group receiving a grade of
687 'A' for each standards group. We observe that, unsurprisingly, the likelihood that
688 one will receive a grade of 'A' increases with initial test performance. However,

690 Table 5
 691 Proportion of students receiving 'A' grade, by third grade mathematics test performance and measured
 692 teacher grading standards
 693

694 695 696 697 698 standards	699 700 701 702 703 704 Quintile of student grade 3 mathematics performance				
	Bottom	2nd	3rd	4th	Top
Lowest standards	0.11	0.19	0.32	0.58	0.85
2nd	0.04	0.14	0.29	0.51	0.77
3rd	0.01	0.09	0.27	0.53	0.74
4th	0.03	0.07	0.22	0.46	0.72
Highest standards	0.01	0.12	0.22	0.46	0.73

708 we also observe that, conditional on initial test performance, students facing more
 709 challenging teachers are less likely to receive an 'A'. Parallel findings emerge in
 710 the similar exercise with regards to the probability of receiving a grade of 'C'.

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

723 3.3. Distributional effects of grading standards

724 While the mean effects of grading standards are important, the theoretical
 725 literature on grading standards suggests that there may be substantial distributional
 726 impacts, with winners and losers associated with higher standards. In addition,
 727 Betts and Grogger (2000), in their empirical study, find evidence of distributional
 728 effects of school-level grading standards, with initially high-performing students
 729 (in tenth grade) benefitting the most (in terms of twelfth grade mathematics test
 730 performance) from high grading standards.¹¹ Therefore, in Table 6 we revise our
 731 primary model (Table 4, row 6) to include an interaction between grading
 732 standards and the student's initial mathematics (or reading, depending on the
 707

705 ¹¹They also find that minority students are harmed by grading standards because standards are
 706 estimated to reduce minority high school graduation rates.

734 Table 6

735 Differential effects of high grading standards on test scores (all using *student* fixed-effects model, akin to row 6, Table 4)

737 738 739 740	Specification	Dependent variable: change in <i>math</i> score					Dependent variable: change in <i>reading</i> score				
		(1M)	(2M)	(3M)	(4M)	(5M)	(1R)	(2R)	(3R)	(4R)	(5R)
741	Students included	All	All	Above	Below	All	All	All	Above	Below	All
742	in regression			average	average				average	average	
743				math in	math in				reading in	reading in	
744				grade 3	grade 3				grade 3	grade 3	
745	Grading	4.619	4.609	4.450	5.088	4.863	7.969	8.794	12.552	8.743	10.253
746	standards	(<i>P</i> =0.00)	(<i>P</i> =0.00)	(<i>P</i> =0.03)	(<i>P</i> =0.00)	(<i>P</i> =0.00)	(<i>P</i> =0.00)	(<i>P</i> =0.00)	(<i>P</i> =0.00)	(<i>P</i> =0.00)	(<i>P</i> =0.00)
747	Grading standards×	1.397				0.055					
748	3rd grade math score	(<i>P</i> =0.19)				(<i>P</i> =0.97)					
749	Grading standards×						2.247				1.250
750	3rd grade reading score						(<i>P</i> =0.07)				(<i>P</i> =0.42)
751	Grading standards×class		2.685	−2.075	3.981	0.773					
752	average 3rd grade math score		(<i>P</i> =0.04)	(<i>P</i> =0.52)	(<i>P</i> =0.02)	(<i>P</i> =0.67)					
753	Grading standards×class						3.527	−1.270	4.860	1.190	
754	average 3rd grade reading score						(<i>P</i> =0.02)	(<i>P</i> =0.70)	(<i>P</i> =0.05)	(<i>P</i> =0.55)	
755	Grading standards×class					−2.262					−3.945
756 757	average×own score					(<i>P</i> =0.09)					(<i>P</i> =0.01)

758 Notes: each column represents a separate regression. Robust *P*-values are in parentheses beneath point estimates.

763 dependent variable) test score. Here, base year test scores are standardized with a
764 mean of zero and standard deviation of one, for ease of interpretation. In these
765 interactive models, an average student in third grade is estimated to benefit
766 strongly (and significantly) from higher grading standards, with above-average
767 initial performers unambiguously benefitting as well. However, since the interac-
768 tions with base year test scores are positive (though not statistically significant at
769 traditional levels in mathematics) it is clear that these positive estimated benefits of
770 grading standards are not uniform for all. Indeed, the results suggest that grading
771 standards are only significantly positive (at the 10 percent level), in the case of
772 math performance, for students whose math scores were nine-tenths of a standard
773 deviation below the mean (or better), and in the case of reading performance, for
774 students whose reading test scores were eight-tenths of a standard deviation below
775 the mean, or better. However, the estimated effects of grading standards are
776 negative for less than one percent of the population, and never statistically
777 significantly negative.

778 The second set of specifications reported in Table 6 are models that interact
779 grading standards with the class's average third grade mathematics (or reading)
780 score.¹² Here, as above, class average test scores are standardized to have a mean
781 of zero and a standard deviation of one, for ease of interpretation. Again, we see
782 that higher ability *classes* may fare somewhat better with higher standards than
783 with lower ability classes.

784 What may be more interesting, however, than how entire classes fare with high
785 grading standards is the distributional effect within a class of high grading
786 standards. Put differently, are the benefits of high standards uniform within a class,
787 or are there winners and losers within the class? Specifications 3M, 3R, 4M, and
788 4R in Table 6 address this question. Specifications 3M and 3R examine the
789 differential effects of grading standards on initially above-average students as the
790 average ability level of the classroom rises. We observe that the effects of grading
791 standards are highest for high-ability students when classroom ability is relatively
792 low, although this differential effect is not statistically significant. Specifications
793 4M and 4R examine the differential effects of grading standards on initially
794 below-average students as the average ability level of the classroom rises. We
795 observe that the effects of grading standards are highest for low-ability students
796 when classroom ability is relatively high, a relationship significant at the three
797 percent, depending on the test score considered. In other words, low-ability
798 students differentially benefit from high standards when they are in a high-ability
799 class, and high-ability students may possibly also differentially benefit from high
800 standards when they are in a low-ability class.

801 Specifications 5R and 5M present similar results in a model in which all
802 students are included in the same regression. The three-way interactions between

762

760 ¹²In specifications in which we interact grading standards with a class average score, we also control
761 for the class average mathematics (or reading) score in third grade.

811 grading standards, class average, and own base year score underscores the above
812 results that standards benefit low-ability students in high-ability classes and
813 high-ability students in low-ability classes the most.¹³ These results are clearest
814 when the point estimates are translated into predicted years of test score gains¹⁴
815 associated with increased standards at different points of the student ability–class
816 ability continuum. We find that the estimated effect of increasing grading standards
817 by one standard deviation is associated with as much as one-third of a year or
818 more of mathematics test score gains, and by as much as two-thirds of a year or
819 more of reading test score gains. For instance, for a student with third grade
820 mathematics performance one-half standard deviation below the mean, the
821 estimated effect of increasing teacher toughness by one standard deviation ranges
822 from 0.07 years of extra growth (in a classroom averaging 1.5 standard deviations
823 below the mean) to 0.28 years (in a classroom averaging 1.5 standard deviations
824 above the mean). For a student with third grade reading performance 1.5 standard
825 deviations above the mean, the estimated effect of increasing teacher toughness by
826 one standard deviation ranges from 0.18 years of extra growth (in a classroom
827 averaging 1.5 standard deviations above the mean) to 0.71 years (in a classroom
828 averaging 1.5 standard deviations below the mean). As mentioned above, this
829 pattern of findings also helps further the conclusion that it is grading standards,
830 and not some other unmeasured form of teacher quality, that is likely to generate
831 our findings.

832 This result has intuitive appeal. Given that the distribution of grades within a
833 class varies much less across classes than does the distribution of performance on
834 external assessments, one can assume that high grades are relatively ‘safe’ for
835 high-ability students in low-ability classes than for their counterparts in high-
836 ability classes. Likewise, low-ability students in high-ability classes are at
837 relatively more ‘risk’ of receiving a low grade than are low-ability students in
838 low-ability classes. Hence, it seems sensible that high standards that lower the
839 ‘safety’ for high-ability students in low-ability classes may generate more effort
840 and greater learning, as might high standards that increase the ‘risk’ for low-ability
841 students in high-ability classes. Feltovich et al. (2002) present theoretical results
842 that are consistent with this story as well. In their study of ‘counter-signaling’
843 behavior, they argue that high standards improve the achievement of students
844 mismatched with the typical ability level of their peers. While this is by no means
845 a definitive explanation of our empirical findings, it is a plausible one.

846 There is additional reason to suspect that this might be the case. As mentioned

806

804 ¹³The models also include a two-way interaction between class average and own score, which is
805 omitted from the table.

807 ¹⁴We measure a ‘year of test score gain’ as the average gain from one year to the next in Alachua
808 County Public Schools. Because Alachua County gain scores tend to be larger than the national
809 average, these are more conservative estimates of ‘years of gain’ than are those based on national grade
810 equivalents.

848 above, teachers maintain drastically different grading standards independent of
849 classroom attributes. Children who rank in the bottom third of the third grade test
850 distribution are three times more likely to earn a grade of C or below with a
851 ‘tough’ teacher (in the top third of the distribution) than with an ‘easy’ teacher (in
852 the bottom third of the distribution. If the average third grade score in the
853 classroom is above the median level, then this difference is more than four times,
854 while if the average third grade score is below the median, this difference is less
855 than two times. The reverse is true for children who rank in the top third of the
856 third grade test distribution: they are three times more likely to earn a grade of C
857 or below with a ‘tough’ teacher than with an ‘easy’ teacher, but this relationship is
858 less than two times in an above-median class and greater than four times in a
859 below-median class. Hence, initially high-ability students are challenged more to
860 get a ‘good grade’ with tough teachers, particularly when they are among the
861 strongest members of a class, and initially low-ability students are also challenged
862 more to get a ‘good grade’ with tough teachers, but particularly when they are
863 among the weakest members of a class.

864 4. Conclusion

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

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

888 Acknowledgements

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

899 References

- 900 Becker, W., Rosen, S., 1990. The learning effect of assessment and evaluation in high school.
901 Discussion paper 90-7, Economics Research Center, NORC.
- 902 Betts, J., 1995. Do grading standards affect the incentive to learn? Working paper, University of
903 California-San Diego.
- 904 Betts, J., 1998. The impact of educational standards on the level and distribution of earnings. *American*
905 *Economic Review*, 266–275.
- 906 Betts, J., Grogger, J., 2000. The impact of grading standards on student achievement, educational
907 attainment, and entry-level earnings. NBER working paper 7875, September.
- 908 Costrell, R., 1994. A simple model of educational standards. *American Economic Review*, 956–971.
- 909 Feltovich, N., Harbaugh, R., To, T., 2002. Too cool for school? Signaling and countersignaling. *Rand*
910 *Journal of Economics*.
- 911 Goldhaber, D., Brewer, D., 1997. Why don't schools and teachers seem to matter? Assessing the impact
912 of unobservables on educational productivity. *Journal of Human Resources*, 505–523.
- 913 Hanushek, E., 1986. The economics of schooling. *Journal of Economic Literature*, 1141–1177.
- 914 Houtenville, A., Smith Conway, K., 2001. Parental effort, school resources and student achievement:
915 why money may not 'Matter'. Working paper, Cornell University.
- 916 Lillard, D., DeCicca, P., 2003. Higher standards, more dropouts? Evidence within and across time.
917 *Economics of Education Review* (forthcoming).
- 918 Moulton, B., 1986. Random group effects and the precision of regression estimates. *Journal of*
919 *Econometrics*, 385–397.
- 920 Rivkin, S., Hanushek, E., Kain, J., 1998. Teachers, schools, and academic achievement. NBER working
921 paper 6691, August.