DISCUSSION PAPER SERIES

Making the Grade: The Effect of Teacher Grading Standards on Student Outcomes

Seth Gershenson
Stephen B. Holt
Adam Tyner

DISCUSSION PAPER SERIES

Making the Grade: The Effect of Teacher Grading Standards on Student Outcomes

Seth Gershenson
American University and IZA

Stephen B. Holt
SUNY

Adam Tyner
Fordham Institute

SEPTEMBER 2022

ABSTRACT

Making the Grade: The Effect of Teacher Grading Standards on Student Outcomes

Teachers are among the most important inputs in the education production function. One mechanism by which teachers might affect student learning is through the grading standards they set for their classrooms. However, the effects of grading standards on student outcomes are relatively understudied. Using administrative data that links individual students and teachers in 8th and 9th grade Algebra I classrooms from 2006 to 2016, we examine the effects of teachers' grading standards on student learning and attendance. High teacher grading standards in Algebra I increase student learning both in Algebra I and in subsequent math classes. The effect on student achievement is positive and similar in size across student characteristics and levels of ability, students' relative rank within the classroom, and school context. High teacher grading standards also lead to a modest reduction in student absences.

Corresponding author:
Seth Gershenson
American University
4400 Massachusetts Ave NW
Washington, DC 20016
USA

E-mail: gershens@american.edu

1 Introduction

A robust consensus in the economics of education agrees that teachers play a pivotal role in shaping students' educational and life trajectories. Effective teachers improve student learning (Chetty et al., 2014a; Rivkin et al., 2005), attendance (Gershenson, 2016; Liu and Loeb, 2021), non-cognitive skills (Jackson, 2018; Kraft, 2019), and long-run outcomes such as college entry and earnings (Chetty et al., 2014b; Gershenson et al., 2022). However, while there is agreement that access to effective teachers is important, relatively little is known about the pedagogical practices that make teachers more or less effective. This gap in the literature limits our ability to train and identify effective teachers.

Various stakeholders have speculated that high grading standards improve student outcomes. This is an intriguing idea, as teachers' expectations and grading practices are likely malleable (De Boer et al., 2018; Pollio and Hochbein, 2015; Quinn, 2020) and teachers can use grades to convey (un)satisfactory progress (or effort) to students and parents (Mechtenberg, 2009; Brookhart et al., 2016). However, economic theory yields mixed predictions regarding whether, and which, students would benefit from high grading standards, as some students may disengage from school rather than increase their efforts (Becker and Rosen, 1992).[1] Whether and for whom high grading standards are beneficial is therefore an empirical question with scant compelling evidence; the current study addresses these questions.

Figlio and Lucas (2004) provide the best evidence to date: analyzing four years of administrative data from Alachua county schools in Florida, they find that, among 3rd to 5th graders, teachers with high grading standards improved math and reading performance relative to their peers with lower standards and – importantly – that no one is harmed by

---

[1]On the one hand, high grading standards raise the cost of low effort and low performance in school, which might induce both students and parents to become more engaged. Similarly, low grading standards might make it difficult for students to identify areas for improvement or create a false sense of achievement that leads to complacency. More nefariously, students might interpret low standards as their teacher not believing in their ability to succeed, creating a self-fulfilling prophecy in which students fail to believe in their own potential (Papageorge et al., 2020). On the other hand, high grading standards that yield low grades might discourage students, particularly those from historically marginalized backgrounds, and cause them to further disengage from academics.

exposure to high grading standards. The current study formalizes and extends Gershenson (2020), who replicated these basic findings in a different context: 8th and 9th grade Algebra I classrooms in North Carolina public schools. Our findings are consistent with those of Figlio and Lucas (2004) in that students typically benefit from exposure to high grading standards and that no students are harmed. We then extend this literature in three main ways. First, we examine effects of grading standards on achievement in other subjects and on student attendance, which is a measure of school engagement. Second, we investigate how teachers' grading standards vary across subjects and how these alternative definitions of grading standards affect student achievement. Finally, we investigate how grading standards' effects might vary across students of different *relative* ability levels within the classroom.

2 Data and Methods

2.1 Data

We examine the effects of teacher grading standards on student outcomes using administrative data from North Carolina public schools. Specifically, we use student-level data from the North Carolina Education Research Data Center (NCERDC) from 2006 to 2016. The NCERDC data contain transcripts for each student that allow us to link course grades to teacher-student pairings and students' scores on standardized state end-of-course (EOC) tests in math in grades 3 through 8 and in Algebra I, Geometry, and Algebra II.[2] We focus on students who took Algebra I in 8th or 9th grade between 2006 and 2016, as these are the grades in which most students take the course. After removing students with incomplete test-score or transcript data, the main analytic sample contains 365,004 unique students and 4,455 unique Algebra I teachers spread across 1,090 schools and 26,819 Algebra I classrooms.[3]

---

[2]In 2014, North Carolina implemented the Common Core Curriculum and Standards. Under the Common Core Standards, Algebra I aligns approximately with Math I, and Geometry and Algebra II align with Math II and Math III respectively. Unfortunately, NCERDC does not provide end of course test scores for Math II or III, which restricts our Geometry and Algebra II samples to 2006 through 2014.

[3]Our analytic sample differs from (Gershenson, 2020) because we impute years of experience for teachers missing experience information using the number of years we observe them within the North Carolina public

We focus on the effects of teacher grading standards on three outcomes: contemporaneous performance in Algebra I, performance in subsequent math courses (i.e., Geometry & Algebra II), and student attendance. Algebra I, Geometry, and Algebra II achievement are measured by performance on the standardized EOC test in the subject, which we standardize within grade-year to have mean 0 and standard deviation (SD) 1. In North Carolina, all students in these courses take the EOC test developed by North Carolina's Department of Public Instruction in consultation with input from select teachers and aligned with state curricula standards.[4] Our data include all students not absent or otherwise excused from the tests.

While not an outcome of interest themselves, the course grades assigned by teachers are crucial to the analysis, as they are used in the construction of teacher grading standards. Course grades appear in the transcript data in both numeric and letter form. Numeric grades appear on a scale from 50 to 100. Letter grades appear as A, B, C, D, or F with occasional + or − modifiers.[5] We convert letter grades to numeric values using a simple midpoint rule, such that a B is an 85, a B+ is an 88, and so on. We also observe information on teacher demographics (race & gender), years of experience, licensure (regular math instruction licensing or lateral/provisional licensing), and educational attainment. Finally, we observe administrative information on student demographics (race and gender), economic disadvantage, and prior achievement in math and English Language Arts (ELA).[6]

2.2 Measuring Teacher Grading Standards

We follow Figlio and Lucas (2004) in measuring teacher grading standards (S) by comparing the more subjective course grades awarded by teachers to the more objective EOC test scores of their students. Intuitively, if one student scores higher on the EOC test than another,

---

school system.

[4]For more information, see https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/state-tests/end-course-eoc

[5]It is unclear why some schools report letter grades and others report numeric grades. In any case, there are no systematic differences by reporting type in the student socio-demographic composition of schools and separate analyses by reporting type yield similar results (Tyner and Gershenson, 2020).

[6]Economic disadvantage is defined as eligible for the means-tested free or reduced price lunch program.

but both receive the same grade from their teacher, the student with the higher test score experienced higher grading standards in the sense that they achieved greater content mastery while receiving the same course grade. There are a few ways to formally estimate S in practice and both Figlio and Lucas (2004) and Gershenson (2020) find that the estimated effect of grading standards on student achievement is robust to how this is done. We focus on two main ways, each with their own strengths and weaknesses. In both cases, we exclude current students' EOC scores and grades from the computation to yield a year-specific estimate $\hat{S}_{jt}$ for teacher j in year t, which avoids the endogeneity concern of current students affecting their teacher's grading standards.

The first approach is to estimate S as the average EOC score of each teacher's B students.[7] This is a simple and transparent approach to defining grading standards. Moreover, it is appealing because nearly all classrooms (and teachers) will have some B students, while certain tracked classrooms may not have many A or C/D students. Here, a higher $\hat{S}$ (i.e., higher average EOC for B students) means higher grading standards.

The second and preferred approach is a regression-based, value-added measure where S_{jt} is operationalized as a set of teacher-year indicators (fixed effects) in a regression of EOC scores on S_{jt} and course grades, again omitting year-t students from the regression:

$$EOC_{ij,-t} = S_{jt} + f(Grade)_{ij,-t} + \varepsilon_{ij,-t}, \forall t, \tag{1}$$

where i, j, and t index students, teachers, and years, respectively. The S in Equation (1) identifies the residualized teacher-year specific average EOC score of all non-t students assigned to teacher j after accounting for the grade given by teacher j to student i. We estimate Equation 1 both by OLS and using a shrinkage estimator that adjusts for variation across teachers in the number of students taught (Chetty et al., 2014a; Stepner, 2013). Once again, higher values of $\hat{S}_{jt}$ indicate higher grading standards, as students of teacher j demonstrate more mastery of the material on EOC tests than do students of other teachers,

---

[7]Formally, $\hat{S}_{jt} = \frac{\Sigma EOC_{i,j,-t}}{N_{j,-t}} | Grade_{i,j,-t} = B$, for all j, t.

despite receiving the same course grade. This regression-based approach is attractive because it uses data on all students (not just B students) and because the estimates can be shrunken to account for teacher-specific variance in the precision of value-added estimation.

2.3 Identifying the Impact of Teacher Grading Standards

We wish to estimate the causal relationship between students' exposure to high teacher grading standards and their short- and medium-term learning outcomes. Again following Figlio and Lucas (2004), we identify this relationship by including $\hat{S}$ as a teacher characteristic in a standard value-added model of the education production function (Harris and Sass, 2011; Pope, 2019). Specifically, we estimate:

$$Y_{ijgst} = \beta \hat{S}_{jt} + \delta X_{it} + \varphi M_{jt} + \theta_{gst} + \varepsilon_{ijgst}, \tag{2}$$

where i, j, and t again index students, teachers, and years while g and s index grade and school, respectively. In equation (2), X is a vector of controls for student characteristics (race, gender, prior achievement); M is a vector of teacher controls (race, gender, educational attainment, years of experience, license held); and θ represents a grade-by-school-by-year fixed effect that ensures we compare the outcomes of students in the same school and same grade at the same time but who are exposed to different grading standards. The inclusion of students' prior math achievement in X accounts for variation in pre-existing student ability and potential sorting of students to particular classrooms (Chetty et al., 2014a). Because we focus on students taking Algebra I in 8th or 9th grade, we use EOC math exam scores from the prior grade (7th or 8th grade, respectively).[8]

The parameter of interest is β, which represents the effect of a one unit increase in teacher grading standards (measured in EOC SD). In practice, we replace $\hat{S}$ with a set of indicators for which quartile of the distribution of grading standards a teacher falls in, though the

---

[8]North Carolina's Department of Public Instruction requires students to take standardized math and ELA exams at the end of each grade from 4th grade through 8th grade.

main results are robust to using the continuous measure. This makes our estimates of β more readily interpretable and allows for nonlinear effects of teacher grading standards.

Finally, while equation (2) is in many ways a standard value-added model now common in economic research, two nuances merit further discussion. First, the predictor of interest, $\hat{S}$, is an estimate. The usual OLS standard errors of β will therefore be too small. Accordingly, we compute standard errors using a bootstrap procedure (500 replications). We allow for serial correlation within schools by implementing a block bootstrap at the school-grade-year level.[9]

Second, the school-by-grade-by-year fixed effect (θ_{gst}) plays an important role in our analysis, as it ensures that we are comparing students who have different Algebra I teachers with different levels of $\hat{S}$ but are taking Algebra I in the same school year, at the same school, and in the same grade. Comparing students studying the same material, in the same environment, and at the same academic stage ensures that the only variation contributing to β_n comes from exposure to different classrooms and avoids confounders that might jointly affect teacher assignment and EOC scores. A necessary trade-off of this research design is that because it relies on comparisons of teachers in the same school teaching Algebra I for the same grade at the same time, the estimate is identified by students in schools with multiple Algebra I teachers in a given year.[10] Accordingly, we verify that the main findings are robust to replacing θ_{gst} with separate school, grade, and year FE.

2.4 Summary Statistics

Table 1 summarizes the Algebra I teachers and students who comprise the analytic sample. Summary statistics are reported both overall and separately by quartile of the grading standards distribution. Columns 1-3 summarize the teachers, who are about 80% white and 30% male. Teachers in the top quartile (toughest standards) awarded higher course grades and

---

[9]The level of clustering does not matter from a practical standpoint: the main results are strongly statistically significant regardless of whether we cluster by classroom, teacher, school year, or school.

[10]Note that θ_{gst} necessarily subsumes the school, grade, and year fixed effects typically included in value-added models. School- and grade-level variables are similarly redundant.

had students with fewer absences and higher test scores, on average, than their counterparts with lower grading standards. Teachers with higher grading standards were disproportionately white, female, and more experienced. The latter difference suggests that grading standards may be malleable and also highlights the importance of controlling for other teacher attributes in equation (2) in order to isolate the effect of grading standards. Moreover, mounting evidence that teacher effectiveness grows with experience (Wiswall, 2013; Papay and Kraft, 2015) suggests that changing grading standards might be one mechanism through which experience translates into effectiveness.

<<< Table 1 about here >>>

Columns 4-6 of Table 1 summarize students. About half the Algebra I students in the sample are white and the average student received a $B-$ grade. White, higher performing, and more advantaged students were more likely to have a high-standards Algebra I teacher. Differences in the lagged test scores of students assigned to more and less lenient graders further highlight the nonrandom sorting of students and teachers and accentuate the need for the value-added model characterized by equation (2).

3 Results

3.1 Main Results

Table 2 presents estimates of several versions of the baseline specification given in equation (2), where the outcome is contemporaneous performance on the Algebra I EOC exam. Columns 1 and 2 estimate the baseline model, using the preferred VAM-based shrinkage estimates of grading standards. Here, teachers are placed into quartiles of the grading standards distribution. Column 1 omits the student and teacher controls and shows a clear positive gradient in the relationship between grading standards and student achievement: the estimated impact of grading standards is positive, statistically significant, and approxi-

7

mately linear. Column 2 adds the student and teacher controls to the model, yielding nearly identical estimates. This suggests that there is an effect of grading standards on student achievement that is separate from other dimensions of teacher quality, (e.g., experience). It also suggests that the lagged test score does a good job of controlling for nonrandom sorting into classrooms.

Column 3 replaces the grading standard quartile indicators with a single, continuous measure of $\hat{S}$. The estimate here is once again positive and strongly statistically significant. Given the standard deviation of our regression-based measure of $\hat{S}$ (0.37), the results in column 3 suggest that a student taught by a teacher one-standard deviation above the mean in terms of grading standards scores 0.09 standard deviations higher on end of course exams in Algebra I (0.37×0.25). Column 4 returns to the preferred categorical treatment specification but replaces the school-by-grade-by-year FE with separate FE for school, grade, and year. This allows smaller schools to contribute identifying variation to the estimates, and once again the estimates are quite similar to those of the preferred specification. Finally, columns 5 and 6 of Table 2 use alternative definitions of grading standards: the average for B students and the unshrunken VAM estimates, respectively. Once again, all of these different approaches yield similar results: each quartile increase in grading standards improves student achievement by about 5 to 8 percent of a test-score SD.

To place these estimates in context, consider that a meta-analysis of 314 randomized controlled trials of interventions that targeted math skills suggest that the most conservative estimate from Table 2, 16% of a SD in column 6 from moving from the bottom to top quartiles, falls at the 70th percentile of documented effect sizes (Kraft, 2020). A more marginal move from one quartile to another yields an effect of about 5% of a test-score SD, which is near the median of the effect size distribution. Thus these estimates are large enough to be economically meaningful but not so large as to be unbelievable. Indeed, even a modest increase in teacher grading standards, raising standards above the lowest quartile, would be equivalent to replacing an average teacher with a teacher at the 70th percentile of

8

the effectiveness distribution (Herrmann and Rockoff, 2012; Kraft, 2020).

Having seen that grading standards improve current performance, a natural follow-up question is whether those effects persist in subsequent years. We investigate this question by re-estimating the baseline model for the subset of students who took Geometry or Algebra II in 10th or 11th grade (the next two courses in North Carolina's math sequence) using EOC scores in those subjects as the outcome.[11] The treatment of interest remains the Algebra I teacher's grading standards. We also look at effects on contemporaneous attendance and "spillover effects" of teacher j's Algebra I grading standards on other math classes taught contemporaneously by teacher j. These estimates are reported in Table 3.

<<< Table 3 about here >>>

Columns 1 and 2 show significant effects of Algebra I teachers' grading standards on math achievement one and two years later in Geometry and Algebra II, respectively. These estimates generally resemble those previously observed in Algebra I classes, as exposure to higher standards leads to significantly higher test scores in subsequent math courses. These are smaller than the contemporaneous effects on Algebra I test scores, which is consistent with existing evidence of "fade out" over time in teacher effects (e.g., Jacob et al., 2010). However, two subtle differences emerge. First, the effects are no longer approximately linear, as there are diminishing returns to the highest standards. The biggest increase comes from moving from the bottom of the distribution to the third quartile. Second, the effects are about twice as large in Algebra II as in Geometry, despite Geometry typically coming first in the math sequence. While this contradicts the aforementioned evidence of fade out, the reason is likely that the skills and content knowledge needed for success in Algebra I align more closely with those in Algebra II than Geometry. All told, this suggests that the contemporaneous gains in Algebra I learning due to high grading standards were the result

---

[11]Data for Geometry tests are only available for years 2007 to 2010 and Algebra II test data are only available from 2007-2011. Table A.4 in the Appendix estimates our main model of Algebra I teacher grading standards on Algebra I achievement using each subsample of non-missing data on alternative outcomes (i.e., Geometry, Algebra II, and attendance) to demonstrate that the main effects are consistent across these changes in sample due to data availability.

of some real learning that persisted over time and was not exclusively due to cramming for the EOC exam or changing study habits and engagement in a short-lived way.

Increased effort and engagement could still be channels through which grading standards improve student outcomes (Figlio and Lucas, 2004); similarly, decreased effort and engagement could be unintended consequences of high standards (Betts and Grogger, 2003; Lillard and DeCicca, 2001). In columns 3 through 5 of Table 3 we investigate the potential for high grading standards to impact other academic behaviors, such as effort and engagement in school, using student absenteeism as a proxy that is also interesting in its own right: student attendance is an educational input that affects achievement and longer-term outcomes like graduation and college going (Liu et al., 2021) as well as an output affected by teachers and other inputs such as class size (Gershenson, 2016; Liu and Loeb, 2021; Tran and Gershenson, 2021).

We measure absenteeism as the count of total days absent and as separate counts of excused and unexcused absences. Importantly, we find no evidence that high grading standards increase absenteeism. In fact, consistent with earlier results showing increased academic achievement, exposure to high grading standards seems to slightly reduce absences. These effects are driven by a reduction in unexcused absences, as seen in column 5.

Finally, columns 6 and 7 of Table 3 show the "spillover" effect of teacher grading standards into other classes taught by the same teacher. One potential concern is that teachers' grading standards are course- or classroom-specific and not part of their general approach to teaching. We show that among the subset of teachers who teach both Algebra I and Geometry (column 6) or Algebra II (column 7) contemporaneously, teachers with higher grading standards in Algebra I also have a large, positive effect on student achievement in their other math classes. This suggests that teachers with high grading standards tend to maintain those standards in all of the classes they teach.

3.2 Distributional Effects and Heterogeneity

High grading standards may differentially affect some student subgroups. Indeed, one of the main arguments against high grading standards is that low-performing students and those from historically disadvantaged groups may become discouraged and disengaged from school when exposed to high standards and the ensuing lower grades. Murphy and Weinhardt (2020) find that students' ability relative to their peers can significantly affect achievement and confidence. Similarly, relatively strong students may benefit from exposure to high grading standards that keep them engaged and challenged. Consistent with this possibility, Betts and Grogger (2003) found that high-ability students and White students received the largest benefit from high grading standards. Alternatively, Figlio and Lucas (2004) found only modest differences in the effect of grading standards between high- and low-ability students, consistent with the possibility that high grading standards induce more effort among both high achievers to secure a good grade and low achievers to avoid a failing grade. In Norway, Bonesrønning (2004) found small effects on low achieving students but large effects on high achieving students, again consistent with the possibility that high standards induce effort, even among high achievers. We test for heterogeneous effects by estimating equation (2) separately by race, gender, locale, and relative math ability. These estimates and their confidence intervals are plotted in Figure 1.

<<< Figure 1 about here >>>

In Figure 1, we estimate versions of equation (2) that include only an indicator for the teacher being in the top quartile of the grading-standards distribution and thus compare the students of these teachers to those of all others.[12] All told, we estimate the model for 20 distinct subsamples and once using the full analytic sample. It is immediately apparent that there is not much heterogeneity: all of the point estimates are between 0.05 and 0.15, most are in the middle of that range at about 0.10, and all are strongly statistically significant.

---

[12]See Table A.1, Table A.2, and Table A.3 in the appendix for table versions of these estimates.

11

Particularly in terms of demographics, there are no significant differences by race or gender.

There are a few ways to conceptualize relative math ability and how that might moderate the impact of high grading standards. We consider three definitions, which all use the previous year's standardized test scores, and again all yield similar results: within-classroom math EOC ranks, higher math EOC than ELA EOC, and within-state math EOC ranks. For the two rank-based definitions we focus on the top and bottom terciles.[13] These estimates are all qualitatively similar and not statistically distinguishable from one another. The largest practical difference here is between the top and bottom terciles in the statewide ranking, where top-tercile performers benefit about 0.025 SD more from exposure to high grading standards than bottom-tercile students, though again this difference is not statistically significant. The largest difference we see is between middle and high schools, with the effect of grading standards being about 0.03 SD larger in the former. This could be because by definition, students taking Algebra I in middle school are more advanced in the sense that they are taking Algebra I early, and thus react better to high standards. This interpretation is also consistent with the results for high performers. Still, that all students in all types of schools benefit, on average, from exposure to high grading standards is consistent with the arguments posited by (Figlio and Lucas, 2004) and Feltovich et al. (2002) that standards induce effort from both high- and low-ability students within a classroom by increasing the risk of receiving a lower grade. These findings are also consistent with prior empirical evidence (e.g., Betts and Grogger, 2003; Figlio and Lucas, 2004).

4 Conclusion

Teachers' expectations for students can exert tremendous influence on their educational achievement and attainment (Papageorge et al., 2020). One way that teachers operationalize high expectations is via high grading standards. We show that in North Carolina, teachers

---

[13]We use terciles because in a typical class of 20 or fewer students, student groupings become quite small with additional bins.

with high grading standards have large, significant, positive effects on students' performance in Algebra I and in subsequent math courses. Smaller, marginally significant reductions in student absenteeism are observed as well. This is true across all student demographic groups, school types, and achievement levels; no students appear to be harmed by exposure to high standards. The effects on absences and subsequent math classes are consistent with the idea that exposure to high grading standards increases student effort and general engagement with school and the change lasts beyond the current class.

These findings are related to two literatures in the economics of education. First, that high grading standards improve achievement is consistent with evidence from North Carolina that teachers' educational expectations of students improves test scores (Hill and Jones, 2021) and from a nationally representative sample that teacher expectations affect college completion (Papageorge et al., 2020). Similarly, these results align with studies of the long-run effects of exposure to biased teachers, where biases are identified by comparing the grades (scores) assigned to blind and non-blind tests (Lavy and Sand, 2018; Terrier, 2020).

Second, grading standards are fundamentally linked to the phenomenon of grade inflation. Historically, grade inflation (low grading standards) is more prevalent in schools serving economicaly disadvantaged communities (Tyner and Gershenson, 2020). Coupled with our findings, this suggests that disproportionate exposure to lax grading standards contributes to socio-economic disparities in educational outcomes. More broadly, school leaders and policy makers should be concerned that grade inflation seems to be increasing at all types of K-12 schools as well as in colleges and universities (Denning et al., 2022; Tyner and Gershenson, 2020). Besides reducing achievement and effort and obfuscating the areas in which students could strive for improvement, this reduces the signal-to-noise ratio of grades as a signal in the labor market and might influence students' choices about major, graduate school, and occupation.

One possible response, and a way to introduce or increase standards, is to require that students take certain courses or curricula in high school. This has been done in a handful

of states with relatively little net benefit, as any modest achievement gains are often offset by decreases in high school completion rates (Lillard and DeCicca, 2001). Jacob et al. (2017), for example, study the implementation of the Michigan Merit Curriculum and find modest gains in ACT scores coupled with small drops in graduation rates among relatively low performing students. Similarly, many states have included exit exams as a requirement for graduation, which also tend to reduce graduation rates among lower-performing and economically disadvantaged students (Dee and Jacob, 2006; Papay et al., 2021). Because we find that all students benefit from high grading standards, the more moderate policy change of increasing grading standards and relying on continuous rather than binary performance indicators (i.e., course grades rather than a high-stakes pass/fail test) can capture the benefits of high standards for all students, including high-achievers, while avoiding the negative consequences of the latter.

References

Becker, W. E. and S. Rosen (1992). The learning effect of assessment and evaluation in high school. *Economics of Education Review 11*(2), 107–118.

Betts, J. R. and J. Grogger (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review 22*(4), 343–352.

Bonesrønning, H. (2004). Do the teachers' grading practices affect student achievement? *Education Economics 12*(2), 151–167.

Brookhart, S. M., T. R. Guskey, A. J. Bowers, J. H. McMillan, J. K. Smith, L. F. Smith, M. T. Stevens, and M. E. Welsh (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research 86*(4), 803–848.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review 104*(9), 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review 104*(9), 2633–79.

De Boer, H., A. C. Timmermans, and M. P. Van Der Werf (2018). The effects of teacher expectation interventions on teachers' expectations and student achievement: Narrative review and meta-analysis. *Educational Research and Evaluation 24*(3-5), 180–200.

Dee, T. and B. Jacob (2006). Do high school exit exams influence educational attainment or labor market performance?

Denning, J. T., E. R. Eide, K. J. Mumford, R. W. Patterson, and M. Warnick (2022). Why have college completion rates increased? *American Economic Journal: Applied Economics 14*(3), 1–29.

Feltovich, N., R. Harbaugh, and T. To (2002). Too cool for school? signalling and countersignalling. *RAND Journal of Economics*, 630–649.

Figlio, D. N. and M. E. Lucas (2004). Do high grading standards affect student performance? *Journal of Public Economics 88*(9-10), 1815–1834.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy 11*(2), 125–149.

Gershenson, S. (2020). Great expectations: The impact of rigorous grading practices on student achievement. *Thomas B. Fordham Institute*.

Gershenson, S., C. Hart, J. Hyman, C. A. Lindsay, and N. W. Papageorge (2022). The long-run impacts of same-race teachers. *American Economic Journal: Economic Policy*.

Harris, D. N. and T. R. Sass (2011). Teacher training, teacher quality and student achievement. *Journal of public economics 95*(7-8), 798–812.

Herrmann, M. A. and J. E. Rockoff (2012). Worker absence and productivity: Evidence from teaching. *Journal of Labor Economics 30*(4), 749–782.

Hill, A. J. and D. B. Jones (2021). Self-fulfilling prophecies in the classroom. *Journal of Human Capital 15*(3), 400–431.

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy 126*(5), 2072–2107.

Jacob, B., S. Dynarski, K. Frank, and B. Schneider (2017). Are expectations alone enough? estimating the effect of a mandatory college-prep curriculum in michigan. *Educational Evaluation and Policy Analysis 39*(2), 333–360.

Jacob, B. A., L. Lefgren, and D. P. Sims (2010). The persistence of teacher-induced learning. *Journal of Human resources 45*(4), 915–943.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources 54*(1), 1–36.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher 49*(4), 241–253.

Lavy, V. and E. Sand (2018). On the origins of gender gaps in human capital: Short-and long-term consequences of teachers' biases. *Journal of Public Economics 167*, 263–279.

Lillard, D. R. and P. P. DeCicca (2001). Higher standards, more dropouts? evidence within and across time. *Economics of Education Review 20*(5), 459–473.

Liu, J., M. Lee, and S. Gershenson (2021). The short-and long-run impacts of secondary school absences. *Journal of Public Economics 199*, 104441.

Liu, J. and S. Loeb (2021). Engaging teachers measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources 56*(2), 343–379.

Mechtenberg, L. (2009). Cheap talk in the classroom: How biased grading at school explains gender differences in achievements, career choices and wages. *The Review of Economic Studies 76*(4), 1431–1459.

Murphy, R. and F. Weinhardt (2020). Top of the class: The importance of ordinal rank. *Review of Economic Studies 0*, 1–50.

Papageorge, N. W., S. Gershenson, and K. M. Kang (2020). Teacher expectations matter. *Review of Economics and Statistics 102*(2), 234–251.

Papay, J. P. and M. A. Kraft (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics 130*, 105–119.

Papay, J. P., A. Mantil, and R. J. Murnane (2021). On the threshold: Impacts of barely passing high-school exit exams on post-secondary enrollment and completion. *Educational Evaluation and Policy Analysis*, 01623737221090258.

Pollio, M. and C. Hochbein (2015). The association between standards-based grading and standardized test scores as an element of a high school reform model. *Teachers College Record 117*(11), 1–28.

Pope, N. G. (2019). The effect of teacher ratings on teacher performance. *Journal of Public Economics 172*, 84–110.

Quinn, D. M. (2020). Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis 42*(3), 375–392.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

Stepner, M. (2013). Vam: Stata module to compute teacher value-added measures.

Terrier, C. (2020). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review 77*, 101981.

Tran, L. and S. Gershenson (2021). Experimental estimates of the student attendance production function. *Educational Evaluation and Policy Analysis 43*(2), 183–199.

Tyner, A. and S. Gershenson (2020). Conceptualizing grade inflation. *Economics of Educa-*

tion Review 78, 102037.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics 100*, 61–78.

Tables

Table 1: North Carolina Algebra 1 Teacher-Years (8th or 9th grade), 2006-2016

	Teachers			Students		
	All	Bottom Q	Top Q	All	Bottom Q	Top Q
	(1)	(2)	(3)	(4)	(5)	(6)
Test Score	-0.06	-0.50	0.47***	0.06	-0.40	0.55***
	(0.67)	(0.62)	(0.57)	(0.97)	(0.93)	(0.90)
Course Grade	79.07	76.62	82.16***	80.12	77.93	82.94***
	(9.45)	(9.58)	(8.76)	(12.63)	(12.82)	(11.76)
Lagged Math Score	0.14	-0.18	0.55***	0.23	-0.07	0.62***
	(0.76)	(0.73)	(0.74)	(0.93)	(0.90)	(0.89)
Student Absences	6.59	7.86	5.54***	5.98	6.78	5.17***
	(4.72)	(6.10)	(3.91)	(6.90)	(8.26)	(5.55)
White	0.81	0.68	0.93***	0.50	0.37	0.62***
Black	0.15	0.28	0.05***	0.22	0.33	0.12***
Hispanic	0.01	0.01	0.01	0.10	0.12	0.08***
Asian	0.01	0.02	0.01***	0.02	0.02	0.03***
Native American	0.01	0.01	0.01	0.01	0.01	0.01***
Other race	0.01	0.01	0.00***	0.04	0.04	0.03***
Male	0.29	0.33	0.24***	0.49	0.50	0.49***
Lateral license	0.09	0.13	0.06***	-	-	-
Provisional license	0.02	0.04	0.01***	-	-	-
Teacher Experience	9.50	8.47	10.40***	-	-	-
	(5.65)	(5.56)	(5.64)	-	-	-
Advanced degree	0.28	0.24	0.35***	-	-	-
Free or Reduced-price Lunch	-	-	-	0.44	0.56	0.31***
Observations	15,854	4,708	3,497	471,997	118,185	117,713

Note: Columns 1-3 observations are teacher-years and columns 4-6 observations are student-years. Standard deviations in parentheses. $*p < .10, **p < .05, ***p < .01$ for t-test of difference in means between column 3 and column 2 and column 6 and column 5, respectively. Q refers to quartile of VAM-estimated teacher grading standards. Test score refers to the average within teacher-year student score on the North Carolina Algebra I end of course tests (standardized by grade-year). Course grade in columns 1-3 refers to the within teacher-year average numeric grade assigned to students in Algebra I and in columns 5-6 refers to the average numeric grade assigned to students in the analytic sample. The lagged math score in columns 1-3 refers to the within teacher-year average math achievement in the prior year of students assigned to teacher j in year t (standardized by grade-year).

Table 2: Effect of teacher grading standards on Algebra I test scores

	VAM W/ Shrinkage				Avg. "B"	VAM 2
	(1)	(2)	(3)	(4)	(5)	(6)
Bottom Q	(Omitted)					
Lower-middle Q	0.06***	0.06***		0.07***	0.08***	0.06***
	(0.01)	(0.01)		(0.01)	(0.01)	(0.01)
Upper-middle Q	0.11***	0.10***		0.11***	0.14***	0.11***
	(0.01)	(0.01)		(0.01)	(0.01)	(0.01)
Top Q	0.19***	0.18***		0.19***	0.19***	0.16***
	(0.01)	(0.01)		(0.02)	(0.01)	(0.01)
Teacher Standards Score			0.25***			
			(0.01)			
Student X	No	Yes	Yes	Yes	Yes	Yes
Teacher X	No	Yes	Yes	Yes	Yes	Yes
School-grade-year FE	Yes	Yes	Yes	No	Yes	Yes
School FE	No	No	No	Yes	No	No
Adjusted R^2	0.46	0.46	0.46	0.45	0.46	0.46
Observations	471,997	471,997	471,997	471,997	471,990	471,997

Note: Bootstrapped standard errors (500 replications) clustered at the school-grade-year level (columns 1-3 and 5-6) and school-level (column 4) in parentheses; $*p < .10 **p < .05 ***p < .01$. Q refers to quartile of teacher grading standards student i faces in algebra 1. Columns 1 through 4 measure grading standards using a regression-based value-added model with shrinkage to account for classroom size variation across teacher-years. Average "B" in column 5 measures teacher grading standards as the leave-year out average score of students receiving a "B" grade from teacher j. VAM 2 in column 6 measures grading standards using a regression-based value-added model without shrinkage adjustment.

Table 3: Effect of teacher grading standards on Other Outcomes

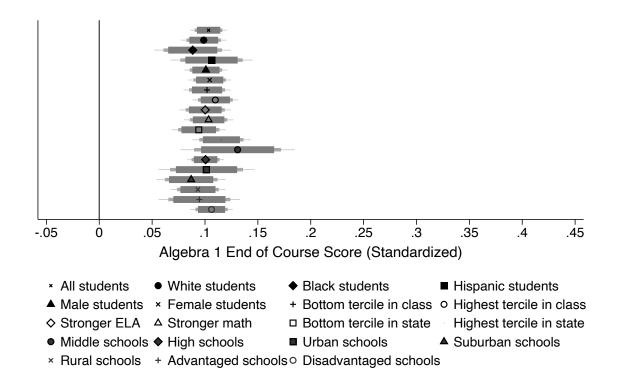|	Geometry (1)	Algebra II (2)	Absences (3)	Excused (4)	Unexcused (5)	Spillover Geometry (6)	Spillover Algebra II (7)
Bottom Q	(Omitted)						
Lower-middle Q	-0.00	0.02	-0.00	-0.01	-0.04*	0.03	0.16***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.08)	(0.05)
Upper-middle Q	0.03*	0.08***	-0.01	-0.03	-0.05*	0.19***	0.10*
	(0.02)	(0.01)	(0.01)	(0.03)	(0.02)	(0.07)	(0.06)
Top Q	0.04**	0.07***	-0.01	-0.05*	-0.06**	0.24***	0.24***
	(0.02)	(0.01)	(0.01)	(0.03)	(0.03)	(0.08)	(0.07)
All controls	Yes	Yes	Yes	Yes	Yes		
Adjusted R^2	0.36	0.27				0.32	0.23
Pseudo R^2			0.19	0.14	0.19		
Observations	48,521	88,084	416,662	104,896	104,847	17,035	41,774

Note: Bootstrapped standard errors (500 replications) clustered at the school-grade-year level (columns 1-2 and 6-7) in parentheses; robust standard errors clustered at the school-level (columns 3-5) in parentheses; $*p < .10$, $**p < .05$, $***p < .01$. Data for Geometry tests only available for years 2007 to 2010; Algebra II test data only available from 2007-2011. Grading standards measured using a regression-based value-added model with shrinkage to account for classroom size variation across teacher-year. In columns 1 to 5, Q refers to quartile of teacher grading standards student i faces in Algebra 1. Columns 3 through 5 present Poisson fixed-effect model estimates. In columns 6 and 7, Q refers to quartile of grading standards of teacher j in Algebra I classes and the sample includes only Algebra II and Geometry classes taught by teacher j in the same school-year as Algebra I classes taught by teacher j.

Figure 1: Effect of Teacher in the Top Quartile of Grading Standards Across Subsamples

Algebra 1 End of Course Score (Standardized)

× All students	● White students	◆ Black students	■ Hispanic students
▲ Male students	× Female students	+ Bottom tercile in class	○ Highest tercile in class
◇ Stronger ELA	△ Stronger math	□ Bottom tercile in state	· Highest tercile in state
● Middle schools	◆ High schools	■ Urban schools	▲ Suburban schools
× Rural schools	+ Advantaged schools	○ Disadvantaged schools	

Note: Figure plots coefficient of teachers in the top quartile of grading standards, measured using a regression-based value-added model with shrinkage to account for classroom size variation across teacher-year, relative to all other quartiles of standards on achievement in Algebra I courses across subgroups of students and school contexts. Estimated using full model with all controls (Column 2 of Table 2). 90%, 95%, and 99% confidence intervals plotted in descending order of line thickness. Bottom and highest terciles in class refers to students in the bottom and top terciles of math achievement in the prior year relative to their peers in the same Algebra I classroom. Stronger ELA (math) refers to students who had higher achievement in ELA (math) relative to math (ELA) in the year prior to taking Algebra I. Bottom and highest terciles in state refers to students in the lowest and highest terciles in terms of math ability statewide by grade-year prior to taking Algebra I. Advantaged schools refers to schools in which > 50% of students are eligible for free or reduced price lunch. Disadvantaged schools refers to schools in which > 50% of students are eligible for free or reduced price lunch.

Appendix A. Supplemental Analyses

Table A.1: Effect of teacher grading standards separately by student race and gender

	White	Black	Hispanic	Male	Female
	(1)	(2)	(3)	(4)	(5)
All Other Q	(Omitted)				
Top Q	0.10***	0.09***	0.11***	0.10***	0.10***
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)
All controls	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.45	0.41	0.45	0.46	0.47
Observations	471,997	237,375	103,759	47,937	232,786
239,211					

Note: Bootstrapped standard errors (500 replications) clustered at the school-grade-year level in parentheses; $*p < .10 * *p < .05 * * *p < .01$. Q refers to quartile of grading standards measured using a regression-based value-added model with shrinkage to account for classroom size variation across teacher-year. Controls include student race, gender, FRPL eligibility, and lagged math achievement; teacher race, gender, qualifications, and experience; and school-by-grade-by-year fixed effects.

Table A.2: Effect of teacher grading standards separately by math ability

	Weak math (1)	Middle math (2)	Strong math (3)	Higher ELA (4)	Balanced (5)	Higher math (6)	Weak state (7)	Strong state (8)
All Other Q	(Omitted)							
Top Q	0.10***	0.09***	0.11***	0.10***	0.10***	0.10***	0.09***	0.12***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
All controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.25	0.25	0.33	0.31	0.55	0.37	0.19	0.26
Observations	186,285	155,379	130,333	132,791	202,427	134,847	163,131	153,857

Note: Bootstrapped standard errors (500 replications) clustered at the school-grade-year level in parentheses; $*p < .10 **p < .05 ***p < .01$. Q refers to quartile of grading standards measured using a regression-based value-added model with shrinkage to account for classroom size variation across teacher-year. Columns 1 through 3 examine the effect of teacher grading standards on students at different terciles of ability relative to their Algebra 1 classmates on entry into the class. Columns 4 through 6 examine students separately by their relative strengths in ELA and math. Controls include student race, gender, FRPL eligibility, and lagged math achievement; teacher race, gender, qualifications, and experience; and school-by-grade-by-year fixed effects.

Table A.3: Effect of teacher grading standards separately by school context

	Middle (1)	High (2)	Urban (3)	Suburban (4)	Rural (5)	More aff. (6)	Less aff. (7)
All Other Q	(Omitted)						
Top Q	0.13***	0.10***	0.10***	0.09***	0.09***	0.09***	0.11***
	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
All controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.43	0.48	0.45	0.47	0.46	0.44	0.48
Observations	121,559	316,737	115,892	137,809	254,141	194,087	278,372

Note: Bootstrapped standard errors (500 replications) clustered at the school-grade-year level in parentheses; $*p < .10 * *p < .05 * * *p < .01$. Q refers to quartile of grading standards measured using a regression-based value-added model with shrinkage to account for classroom size variation across teacher-year. Controls include student race, gender, FRPL eligibility, and lagged math achievement; teacher race, gender, qualifications, and experience; and school-by-grade-by-year fixed effects.

Table A.4: Effect of grading standards on Algebra I by subsample for other outcomes

	Geometry (1)	Algebra II (2)	Absences (3)	Excused (4)	Unexcused (5)	Spillover Geometry (6)	Spillover Algebra II (7)
Bottom Q	(Omitted)						
Lower-middle Q	0.03*	0.03**	0.05***	0.04**	0.04**	0.01	0.01
	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.03)	(0.03)
Upper-middle Q	0.07***	0.09***	0.10***	0.07***	0.07***	0.04	0.05*
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.05)	(0.03)
Top Q	0.15***	0.17***	0.18***	0.16***	0.16***	0.12***	0.13***
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.04)	(0.03)
All controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.43	0.42	0.46	0.46	0.46	0.61	0.61
Observations	48,521	88,084	416,662	104,896	104,847	55,930	75,253

Note: Bootstrapped standard errors (500 replications) clustered at the school-grade-year level in parentheses; $*p < .10$, $**p < .05$, $***p < .01$. Data for Geometry tests only available for years 2007 to 2010; Algebra II test data only available from 2007-2011. Grading standards measured using a regression-based value-added model with shrinkage to account for classroom size variation across teacher-year.