Great Expectations

The Impact of Rigorous Grading Practices on Student Achievement

By Seth Gershenson

FEBRUARY | 2020

Foreword and Executive Summary by
Amber M. Northern and Michael J. Petrilli

THOMAS B. FORDHAM INSTITUTE
ADVANCING EDUCATIONAL EXCELLENCE

ABOUT THE FORDHAM INSTITUTE

The Thomas B. Fordham Institute promotes educational excellence for every child in America via quality research, analysis, and commentary, as well as advocacy and exemplary charter school authorizing in Ohio. It is affiliated with the Thomas B. Fordham Foundation, and this publication is a joint project of the Foundation and the Institute. For further information, please visit our website at www.fordhaminstitute.org. The Institute is neither connected with nor sponsored by Fordham University.

SUGGESTED CITATION FOR THIS REPORT

Gershenson, Seth. *Great Expectations: The Impact of Rigorous Grading Practices on Student Achievement*. Washington, DC: Thomas B. Fordham Institute (February 2020). https://fordhaminstitute.org/national/research/great-expectations-impact-rigorous-grading-practices-student-achievement.

ACKNOWLEDGMENTS

This publication was made possible by a grant from the Carnegie Corporation of New York and support from our sister organization, the Thomas B. Fordham Foundation. The statements made and views expressed are solely the responsibility of the author.

We are extremely grateful to Seth Gershenson of American University, with whom Fordham has now had the pleasure of collaborating three times (including recent studies about grade inflation and student-teacher “race match”). We express our gratitude as well to Fordham’s associate director of research Adam Tyner, who managed the project and provided critical feedback and edits throughout the research process.

David Figlio of Northwestern University and Cory Koedel of the University of Missouri ably served as external reviewers and provided valuable input on drafts. At Fordham, we thank Chester E. Finn, Jr. for reviewing drafts, Victoria McDougald for overseeing media relations, Olivia Piontek for handling funder communications, and Pedro Enamorado for managing report production. Fordham research intern Tran Le assisted Fordham staff in organizing and conducting interviews with teachers, as well as provided other supports at various stages in the process. Kudos also to William Buttaggi for developing the report’s layout and design and Pamela Tatz for copy editing the report.

CONTENTS

FOREWORD AND EXECUTIVE SUMMARY

By Amber M. Northern and Michael J. Petrilli

Many moons ago, one of us was a first-year teacher of high school English in a rural district that loved its football. It was the close of the first term as I sat at my desk after school in a musty classroom trailer—there being no room in the main building for rookies—trying to finish up my grades in time for report cards. I heard the steps creak and looked up to find Mr. Simpson, the brawny football coach who'd been at the school since forever, interrupting my calculator punching (this was the Mesozoic era before online gradebooks). Honestly, my first thought was one of alarm, as I was pretty sure Coach Simpson hadn't said one word to me all year. Yet here he was after hours in my low-rent trailer with no one else around.

Thankfully, he was feeling just as awkward about his appearing as I was, so he got straight to the point: "I understand that Darrell is just a couple points shy of getting a C this semester in your class. I know we don't want to give him a grade that he doesn't deserve, but I was just wondering whether you had any make-up work for him to do in study hall, or maybe he could re-take that last test again after he studies some more." My perplexed expression must have signaled that I was no whiz in reading between the lines, so he continued: "I'd *personally* make sure that he does *every* assignment in study hall and gives it his best shot. I hate to ask and all, but he's one of our best players, and we're really going to need him for the playoff game next Friday."

So there it was. You see, our high school back then had a policy that athletes had to keep a C average in every class or couldn't play sports—at least not until they got their grades back up.

I wish I could say that I promptly told Mr. Simpson to take a hike. After all, Darrell had been slacking all semester, barely getting by, though he was clearly able to do much more than execute a perfect pass. I should have responded, "I'll let both Darrell and myself down if I communicate to this young man that I'm willing to accept less from him than I know he's capable of!"

Sadly, I didn't say any of that. I was a twenty-something newbie trying to keep my head above water in a sink-or-swim school. So I dutifully agreed to pull together Darrell's study-hall packet and breathed a sigh of relief as Coach Simpson shut the door behind him.

We tell you this true tale because it underscores the motivation for this report and the complexity surrounding the issues it delves into: high school grading practices and how they intersect with teachers' expectations for their students, and the impact they have on student outcomes.

The limited prior research on this topic shows that instructors who recognize and believe in their students' potential—and maintain high expectations for them—significantly boost the odds that their students will go on to complete high school and college. That's what American University's Seth Gershenson and his colleagues found in a previous study that used teacher survey data to define expectations.

Another way to define expectations is to measure how teachers approach grading—specifically, whether they subject students to more or less rigorous grading practices. A lone study conducted sixteen years ago by David Figlio and Maurice Lucas in one Florida County found positive academic and behavioral impacts for nearly all students from elementary teachers' high grading standards. Still, given the central role of grading in U.S. schools, we know shockingly little about how it impacts a child's future, particularly grading standards in middle and high school.

Dr. Gershenson's existing work on teacher expectations, as well as his prior study for us on grade inflation in high school, made him an ideal partner to tackle this neglected area of research. Like him, we were interested in whether a teacher's approach to grading students' work affected their outcomes in the short and long terms, and whether those standards differed by teacher, student, and school characteristics.

Specifically, Gershenson investigated the following questions:

- How do the grading standards of an Algebra I teacher affect students' content mastery, as measured by their performance on the end-of-course exam (a short-term outcome)?
- Do the grading standards of an Algebra I teacher impact students' performance in subsequent math courses like geometry and Algebra II and their likelihood of graduating from high school (longer-term outcomes)?
- Does the impact of an Algebra I teacher's grading standards vary by pupil, school, or teacher characteristics? And what school and teacher characteristics predict teachers' grading standards?

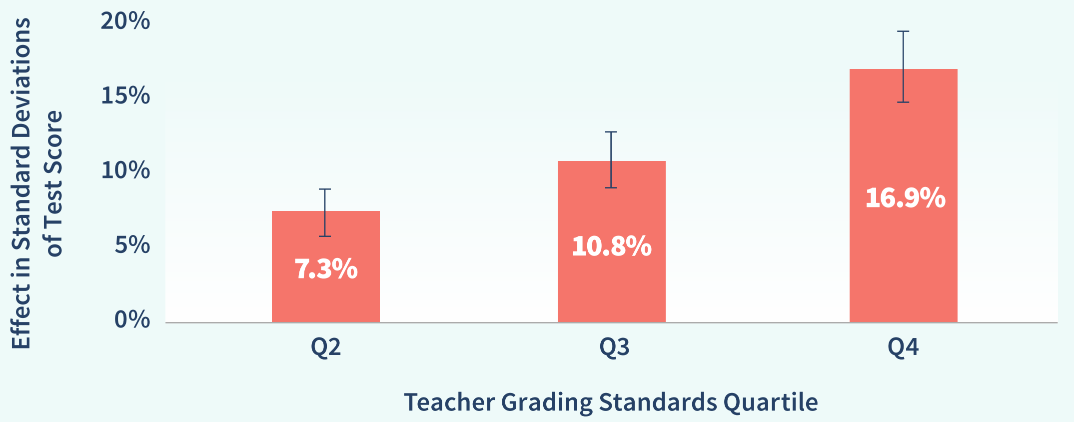
His data come from the grading standards of eighth and ninth grade Algebra I math teachers in North Carolina public schools. Algebra I is ideal for this purpose, as it was a state graduation requirement for the eleven-year period that the study covers (2006–2016), and it also had an end-of-course (EOC) test in those years. Having both course grades and EOC scores allowed Gershenson to define teachers' grading standards in a straightforward manner: Teachers who inflate grades—meaning they assign good grades to students who perform relatively poorly on the EOC–exhibit *low standards*, while teachers who assign lower grades than we might expect given students' scores exhibit *high standards*. Gershenson compared students of teachers with higher grading standards to their peers who have teachers with lower grading standards but still take the same course (Algebra I), in the same school, in the same grade, in the same year.[1]

FINDING 1: Students learn more from teachers who have higher grading standards.

Teachers are categorized into four evenly-sized groups based on their grading standards, where group 1 has the lowest standards and group 4 has the highest. The bottom quartile of the grading standards distribution (group 1) is the reference group to which other groups of teachers are compared, and is therefore omitted in the figures.

Figure ES-1 shows that teachers in the top quartile increase student EOC scores by a whopping 16.9 percent of a standard deviation (SD) over those of their counterparts in the bottom quartile. Even instructors in the middle of the grading standards distribution are significantly more effective than those with the lowest grading standards.

FIGURE ES-1. Students learn more from teachers with higher grading standards.



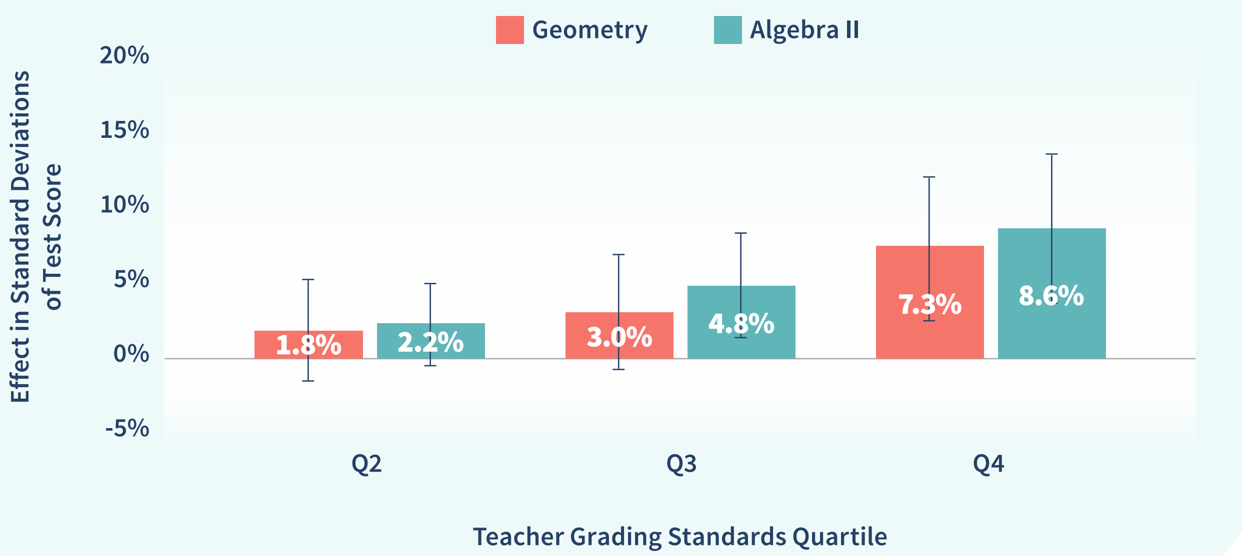
Note: The bars represent the estimated effects of having a teacher in the second, third, or fourth quartile of the grading-standards distribution on Algebra I standardized test scores. Estimates come from the baseline value-added model. The bottom (lowest standards) quartile is the omitted reference group. Error bars represent 95 percent confidence intervals.

[1] Obviously the comparison of middle to high schools is not apples-to-apples, as eighth-grade middle school students are taking Algebra I a year early and therefore might be higher performers or better test takers, on average, than high school Algebra I students. Hence, these types of within-school/grade/year comparisons ensure that these concerns do not confound the results.

FINDING 2: Teachers with higher grading standards improve their students' performance in subsequent math classes up to two years later.

Figure ES-2 reports estimates of Algebra I teachers' grading standards on their students' performance on Geometry and Algebra II scores in subsequent years. Once again, higher grading standards consistently lead to higher achievement. Predictably, since these tests are in somewhat different subjects and are taken one and two years later, the effect on these longer-range outcomes is smaller than the short-term effects above. Still, relative to teachers with the lowest grading standards, students of those with the highest standards performed notably better a year later in geometry (7.3 percent of a SD) and two years later in Algebra II (8.6 percent of a SD).

FIGURE ES-2. Teachers with higher grading standards boost students' subsequent math performance.

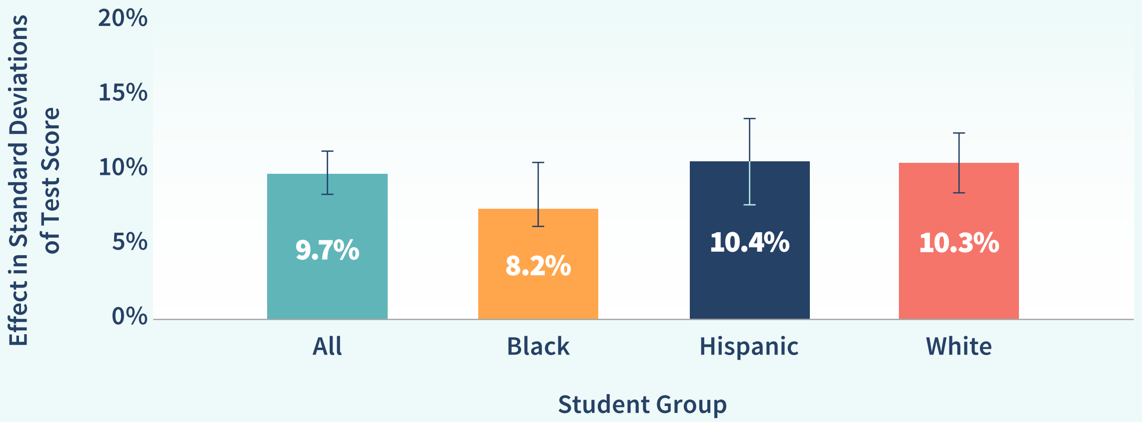


Note: The bars represent the estimated effects of having a teacher in the second, third, or fourth quartile of the grading-standards distribution on Algebra I standardized test scores. Estimates come from the baseline value-added model. The bottom (lowest standards) quartile is the omitted reference group. Error bars represent 95 percent confidence intervals.

FINDING 3: Teachers with higher grading standards significantly improve the learning outcomes of all student subgroups.

Figure ES-3 reports estimates of a simplified version of the baseline model. Here we compare teachers in the top 75 percent of the grading standards distribution to the easiest-grading quartile of teachers, which allows us to focus on one effect per pupil group. Overall, we see that having a teacher among those with higher grading standards improves achievement by about 8 to 10 percent of a standard deviation. In short, white, black, *and* Hispanic students all benefit from exposure to higher grading standards.

FIGURE ES-3. Students of all racial/ethnic groups learn more from teachers with high grading standards.

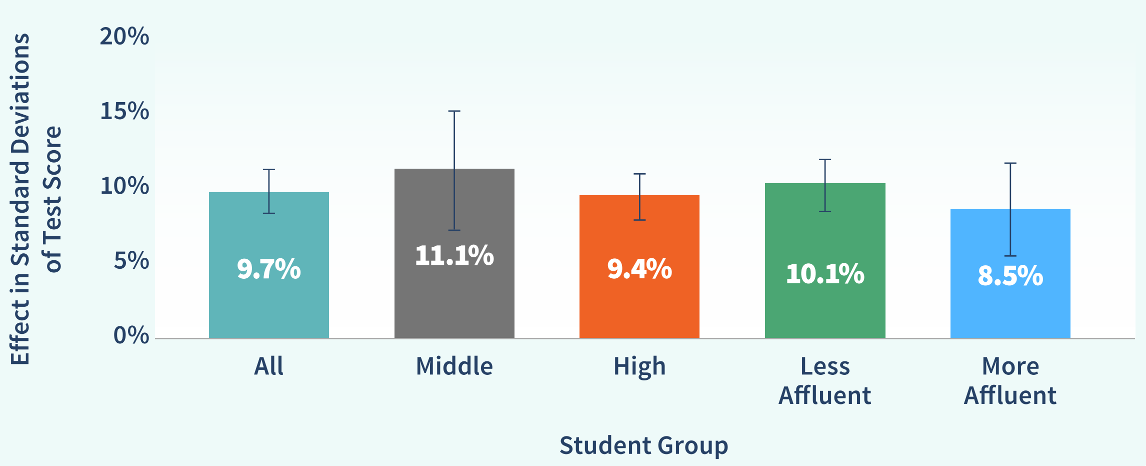


Note: The results are estimated using a version of the baseline value-added model separately by student subgroup, where the reported estimates are of the effect of the teacher being in the top three quartiles of the standards distribution. Error bars represent 95 percent confidence intervals.

FINDING 4: Teachers with higher grading standards significantly improve student learning in all types of schools.

Figure ES-4 shows estimates of the simplified model separately by school level and socioeconomic makeup, once again comparing the effect of teachers in the top 75 percent of the grading-standards distribution to the bottom 25 percent. Importantly, the effects of high grading standards are similar in size in all school types—middle and high school, rich and poor—suggesting that high grading standards are universally beneficial.

FIGURE ES-4. Students in schools with different grade levels and different socioeconomic makeups learn more from teachers who have higher grading standards.



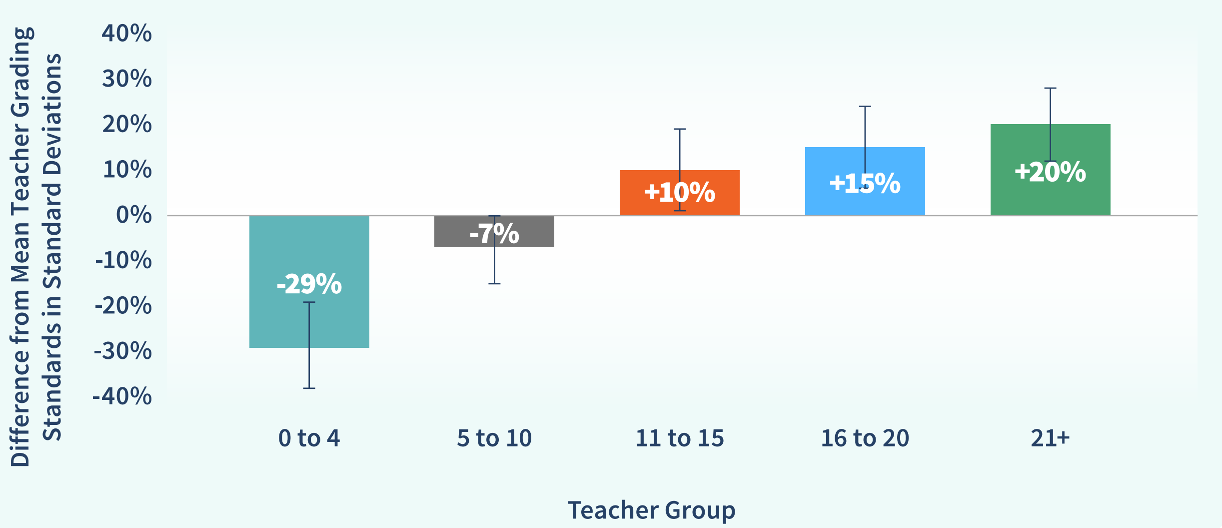
Note: The results are estimated using a version of the baseline value-added model separately by school type, where the reported estimates are of the effect of the teacher being in the top three quartiles of the standards distribution. Error bars represent 95 percent confidence intervals. The schools are considered to be "more affluent" when fewer than 50 percent of students are eligible for free or reduced-price lunches and "less affluent" when more than 50 percent of students are eligible.

FINDING 5: Teachers who attended selective undergraduate institutions, hold graduate degrees, and have more experience tend to have higher grading standards.

Teacher grading standards vary based on some teacher characteristics. To simplify presentation, the average raw measure of grading standards is standardized and reported for each group below, rather than splitting teachers into quartiles. Higher numbers reflect higher grading standards.

Teachers who attended selective undergraduate schools tend to have higher standards, although that difference is not statistically significant. In addition, teachers with graduate degrees have grading standards that are about 19 percent of a standard deviation higher than teachers without them (not shown). Finally, Figure ES-5 shows that as teachers gain experience, their grading standards generally rise as well. Although some of the differences across groups are not statistically significant, there's a clear trend: Teachers increase their grading standards as they remain in the profession, particularly during their first fifteen years. For example, new teachers have grading standards that are 29 percent of a standard deviation lower than average, while the most experienced teachers have standards 20 percent of a SD higher than average.

FIGURE ES-5. Teachers who have more years of experience tend to have higher grading standards.



Note: The outcomes may not average to zero because different groups are of different sizes and data are missing for some teachers. Error bars represent 95 percent confidence intervals.

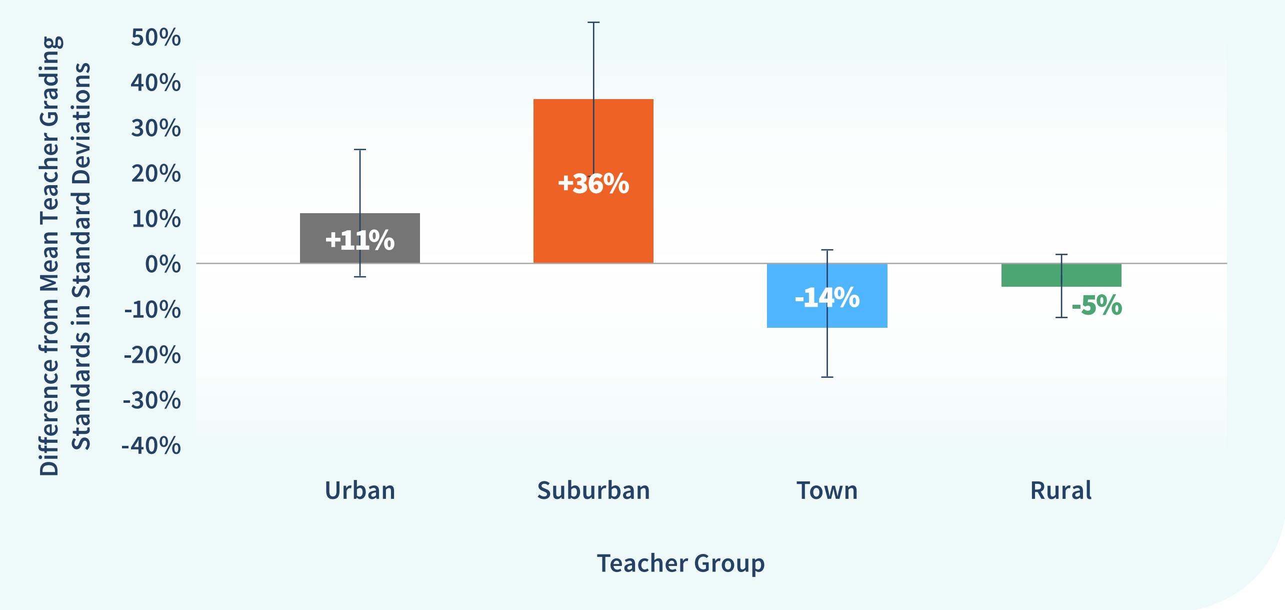
FINDING 6: Grading standards tend to be higher in suburban schools, middle schools, and schools serving more advantaged students.

Above, we presented the *impact* of teachers who hold rigorous grading standards on students in different types of schools. Here we examine what the *average* grading standards look like in different types of schools. (This is a descriptive exercise, as it does not control for the variety of confounding variables that are controlled for in Findings 1–4.) As indicated, higher numbers reflect higher grading standards.

Grading standards are highest in suburban schools and lowest in schools in small towns and rural areas (Figure ES-6). Suburban schools have grading standards that average more than one-third of a SD higher than the mean and about half a SD higher than those in small towns.

Grading standards in middle schools are about 17 percent of a standard deviation higher than in high schools, though it's important to note that students who take Algebra I in middle school are generally higher performers than those who take it later. Finally, more affluent schools have grading standards that are, on average, more than one-third of a standard deviation tougher (not shown). This is troubling, as it provides more evidence of the "soft bigotry of low expectations" for disadvantaged students attending high-poverty schools.

FIGURE ES-6. Grading standards tend to be higher in suburban schools.



Note: The outcomes may not average to zero because different groups are of different sizes and data are missing for some teachers. Error bars represent 95 percent confidence intervals.

The study includes Dr. Gershenson's take on what these findings mean for policy and practice—a couple of which we underscore and extend here.

First, we should use information about grading practices to improve instruction. Teachers are not to be blamed for having low grading standards when many of them don't know where to set the bar for high-quality student work. This is not a major focus of most teacher preparation or professional development. Education Trust has provided a valuable service by taking a closer look at what teachers assign students and asking whether those tasks reflect today's higher academic standards. Similar questions should be raised about teacher grading practices.

Likewise, schools and districts would do well to share with teachers how their grading standards compare to the standards of other instructors teaching the same subjects and at the same grade levels. Teachers need to know whether their expectations fail to match—or possibly surpass—those of their colleagues. Educators might be more willing to aim higher if they knew they were off target and taught how to get closer to the bullseye.

And second, let's not forget that none of this is possible without an external measure of student performance. The simple definition of grading standards used in this report can be easily calculated by schools and districts—but only if they have a summative test, such as an end-of-course exam. The current angst about over-testing has likely resulted in the recent dip in such tests administered across the country.

Let's be honest. Most of us want teachers to have high expectations when it comes to grades, but we're gradually making it harder, not easier, for teachers to do that. Case in point: More than one thousand colleges and universities have adopted test-optional admissions policies, arguing that college entrance exams provide an unfair advantage to middle- and high-income students. Not requiring them, they say, expands access for poor students and students of color.

There's plenty of debate about whether that's true, but one thing seems clear: Grade-point averages will now matter even more, so it is key that they be accurate representations of a student's academic performance. The current push for test-optional college admissions makes it that much more difficult for high school teachers, who now face even greater pressure to be easy graders to help their students get into selective colleges.

That's a big problem. Think about it: If there's pressure on teachers from one of their own to inflate a grade for a kid to play in next week's football game, just imagine what that pressure looks like from his parents to get him into a good college.

INTRODUCTION

Teachers matter enormously. Good ones can improve everything from a student's content knowledge, attendance habits, and noncognitive skills to long-run outcomes such as earnings and college entry.[1] Lack of access to effective teachers is a primary reason that many schools serving disadvantaged communities struggle to improve student outcomes. Indeed, a consensus has emerged that providing effective teachers to all of our students is a necessary step toward ensuring that our public schools truly provide equal opportunities for every pupil.

Yet despite that certainty, we know too little about the teacher characteristics and practices that boost student outcomes, and this lack of knowledge weakens our ability to identify, prepare, deploy, retain, and promote effective teachers.[2] It also hinders our ability to help current teachers improve.

One thing we know for sure, however, is that teacher expectations make a difference. Those who recognize and believe in their students' potential—and hold high expectations for all their students—significantly increase the odds that those children will go on to complete high school and college.[3]

> **"Those who recognize and believe in their students' potential—and hold high expectations for all of their students—significantly increase the odds that their students go on to complete high school and college."**

One way that teachers convey their expectations to students, sometimes directly and sometimes indirectly, is through the grades they assign.[4] Students can respond to this information by recalibrating their own expectations and beliefs about what's possible, reengaging with school, and putting forth greater effort. Accordingly, low grading standards pose a grave threat to the performance and evaluation of U.S. public schools that ultimately jeopardizes the competency of high school and college graduates who are entering the workforce. Assigning good grades for mediocre work signals to students that excellent work is beyond their reach. This is the "soft bigotry of low expectations" of which President George W. Bush warned: When students who have not mastered the material receive passing marks anyway, they can become complacent and fail to reach their full potential. Lax grading is a pernicious practice that provides students and parents with a false sense of security and accomplishment that might prevent them from trying harder, learning more, and maximizing their own future prospects in the "real world."

The common term for lowered teacher standards is "grade inflation," where teacher-assigned course grades overstate students' actual mastery of skills and knowledge. That grade inflation is pervasive in U.S. high schools is evidenced by rising GPAs even as SAT scores, ACT scores, National Assessment results, and other measures of actual academic performance have held stable or fallen.[5] The result is that a "good" grade is no longer a clear marker of solid knowledge and skills. In North Carolina, for example, more than one-third of B students fail to earn a score of proficient on their end-of-course (EOC) Algebra I exam, and more than half of B students fall short of North Carolina's "college- and career-ready" standard.[6] Similar stories are playing out in colleges and universities, where at least some of the recent increase in completion rates is likely due to softened standards.[7]

The question we should be asking ourselves, then, is just how much talent and potential is America squandering when its schools and teachers fail to uphold high standards? And to what extent do different expectations and grading standards exacerbate academic achievement gaps? These are important questions, as teachers often have considerable discretion in determining and assigning individual students' grades. Teachers who inflate grades—those who assign better grades to students than their actual performance warrants—can be thought of as having *low standards*, while teachers who assign lower grades than expected can be thought of as having *high standards*. This report investigates the extent to which teachers' grading standards affect student success. This is a fundamentally important question with implications for all teachers and education policymakers.

The current study addresses this question by examining the grading standards of high school math teachers in North Carolina's public schools, focusing on Algebra I teachers and their students. Why we chose this focus is straightforward: For eleven consecutive and recent years, the Tarheel State required its high school students to take Algebra I and to sit for a standardized EOC exam in that subject. It's possible to match how North Carolina students fared on that exam with the Algebra I grades on their high school transcripts. This allows for the creation of year-specific measures of grading standards for a large number of teachers during that period.

Specifically, this report investigates the following questions:

1. How do the grading standards of an Algebra I teacher affect students' content mastery, as measured by their performance on the EOC exam?

2. Do the grading standards of an Algebra I teacher impact students' performance in subsequent math courses like geometry and Algebra II and their likelihood of graduating from high school?

3. Does the impact of an Algebra I teacher's grading standards vary by student, school, or teacher characteristics? And what school and teacher characteristics predict teachers' grading standards?

To answer these questions, this report uses administrative data for all eighth- and ninth-grade Algebra I students in North Carolina's public schools from 2006 to 2016. We turn to our discussion of the research methods next.

WHAT DO WE KNOW ABOUT GRADING PRACTICES AND STUDENT OUTCOMES?

There is surprisingly little empirical evidence to back up the intuition that high grading standards boost student learning and long-run success. The best evidence to date comes from a 2004 study by David N. Figlio and Maurice E. Lucas of all third through fifth graders in Florida's Alachua County over four years.[8] Having a classroom teacher with high grading standards improves academic performance in math and reading and reduces disciplinary infractions, effects that were largest for high-achieving students.[9] Importantly, however, the effects of high grading standards were positive for nearly all students and never harmful. In families with multiple children, parents reported on surveys that they spent significantly more time helping the child whose teacher was a "tougher grader." These results suggest that the effect of higher grading standards operates partly through increased parental involvement and student effort.[10]

Other studies have found that college students study less—by as much as 50 percent—when they expect the average class grade to be an A rather than a C.[11]

Julian R. Betts and Jeff Grogger used data from a nationally representative survey of tenth graders to show that higher school-level grading standards, defined as schools' average gap between GPAs and standardized test scores, boost achievement.[12] As in the Florida study, the effects are universally positive and largest among high achievers. When disaggregating the data by race, however, they found that the educational attainment of black and Hispanic students might fall as a result of higher grading standards, although the negative correlation identified for these students is imprecisely estimated and not necessarily causal.[13]

Indeed, the same higher grading standards might improve some students' outcomes while harming the outcomes of others. For example, consider two classmates whose teacher has high grading standards and who both receive a C on their midsemester report cards. If the students have different temperaments or innate ability levels, one student might be motivated to improve her study habits while the other takes this same information as a signal that the subject is too difficult for her and further disengages from school.[14]

Indirect evidence also suggests that grading practices matter. Many studies have investigated the extent to which teachers hold gender and race biases and how such biases affect student outcomes. These biases have been gauged by comparing students' outcomes on tests that are scored "blind," meaning the scorer cannot identify the students, and "nonblind," meaning the scorer knows the students' characteristics.[15] Female students, for example, benefit in the semester that they're exposed to a "pro-female" biased teacher, but they also excel for years afterwards and are more likely to pursue the subject.[16] Part of what is happening is that teachers communicate their expectations for students via course grades.[17] This communication suggests that the discretionary component of course grades influences subsequent student outcomes.[18]

RESEARCH METHODS

As noted above, North Carolina has reliable records for both course grades and standardized EOC test scores for Algebra I over the 2006–16 period. Thus, the focus is Algebra I classrooms that had a single teacher for the entire academic year. Specifically, there are full data on about 8,000 unique Algebra I teachers who taught about 350,000 unique eighth- and ninth-grade Algebra I students.[19] These data are summarized in *Appendix B*, Table B-1.

Having both course grades and EOC scores allows us to define teachers' grading standards in an intuitive way. For each teacher in the state, the average EOC score of that teachers' students who received a B in the course is computed. For example, suppose that the average test score of the students who received a B from Ms. Apple was 80 points, while the average test score of students who received a B from Ms. Orange was 90 points. This difference implies that Ms. Orange has higher grading standards than Ms. Apple, because Ms. Orange's students learned more in the course of earning that B. Teachers can then be sorted by this measure and classified in the bottom 25 percent as the easiest graders, the top 25 percent as the toughest graders, and so on.

> **Ms. Orange has higher grading standards than Ms. Apple, because Ms. Orange's students learned more in the course of earning that B.**

This calculation is the preferred measure of grading standards because it is transparent, easy to compute, and easy to understand. Moreover, as Figlio and Lucas point out, it is also unlikely to be influenced by classroom composition, as there are B students in most classrooms but not necessarily A or C students. That said, *Appendix A* describes an alternative definition of grading standards that uses all students' grades and test scores, not just those of the B students. As in the Florida study, the results turn out to be similar regardless of how grading standards are defined.

The next challenge is to isolate the causal effect of teachers' grading standards on student outcomes. That's challenging because students are not randomly assigned to teachers, so we might worry that—for example—concerned parents or principals might ensure that certain children are assigned to teachers with higher grading standards. If so, we'd be unable to distinguish the effect of the grading standards from the effect of those parents and principals.

To control for such confounding factors, the analyses adjust for the student's performance in the previous year.[20] Indeed, there is compelling evidence that, conditional on performance on the previous year's end-of-grade exam, classroom assignments are approximately random.[21] Intuitively, this approach is also how researchers estimate the impact of any schooling input or teacher characteristic: This study simply treats the teacher's grading standards as an observed teacher characteristic that is included as a potential determinant of student outcomes in the value-added model. Details of the value-added model's specification and estimation are described in *Appendix A*.[22]

A related concern is that teachers with strict grading standards differ from teachers with more relaxed standards in other ways, too. If so, we'd again be unable to differentiate the effect of grading standards from the effects of these other differences. Once again, this concern is eliminated by adjusting for other observed teacher characteristics that are known to improve student test scores, such as years of experience and the selectivity of undergraduate institutions.[23] Similarly, models adjust for the demographics and past performance of the teachers' current students, as a student's classmates might jointly influence the teachers' behavior and the student's outcomes. The main results are basically unchanged when these adjustments are made, which is again consistent with the Florida results, and suggest that the impact of teachers' grading standards on student outcomes is real and not conflated with other characteristics of the teachers who have high standards.

Finally, we might worry that it is school culture, district policies, or principal effects that drive both teacher grading standards and student outcomes. Once again, however, this concern can be ameliorated using a regression adjustment. Specifically, the models in this analysis control for school-by-year-by-grade indicators in the value-added model. This means that the students of teachers with higher grading standards are being compared to their peers who have teachers with lower grading standards but still take the same course (Algebra I) in the same school, in the same grade, and in the same year. In addition to school-grade-year specific factors, this adjustment controls for any school-specific influences that might otherwise confound the estimates.[24]

"Students of teachers with higher grading standards are being compared to their peers who have teachers with lower grading standards but still take the same course (Algebra I) in the same school, in the same grade, and in the same year."

As part of this project, Fordham Institute staff conducted open-ended interviews with teachers about their own experiences with grading practices and their opinions about what grades communicated. Although those conversations did not inform the data analysis, they provide an on-the-ground perspective of the issues at stake in this report. The interviewed teachers were located in several regions of the country and included both women and men who taught English, math, and history in middle and high schools. Excerpts of these conversations are included in the *Teacher Voices* sections throughout this report.

TEACHER VOICES: WHAT DOES AN "A," "B," OR "C" ACTUALLY SIGNIFY?

According to our interviews, teachers may base their grades on a mix of both results and pupil effort, and the meaning of each letter grade is neither clear nor consistent for these teachers. Some hold onto the traditional view that "C is average," while others indicate that a B is the new average (B is, in fact, the median grade in Algebra I in this study). Teachers indicated that part of the inconsistency is based on how grading practices have changed over time, and some expressed conflicting feelings or regret that standards seemed to be falling.

"I kind of think a B is almost like the new average. Where it's like B is average, and if you're A, it's above average, and if it's a C, then it's like, nobody likes the C anymore."

"An A is the student who shows the disposition to . . . redo assignments over and over until they get it. They're practicing at school; they're doing their homework at home. A B student is a student, in my mind, who will try their best they can at school, and that's it—'That's all I can do with my school time. I dedicated the 89 minutes of block schedule every other day, and I did my best with that time.'"

"Whatever [the students] get from the lesson, whatever sticks in their head, whatever notes they took, that's the only time that they actually look at their notes, and then they just take the test and hope they did well. If they didn't, then they didn't. That's a C student."

"[A]n A means that the student knows the material or they're a reflective learner, and they reflect on the learning process. A B is that they're nearing mastery, or maybe some of my students who had a B are happy just—it's about an 80 percent, they're happy with not making any effort to reflect on their mistakes."

"A is for those kids who get it, who need extra work that is pushing them, those extending activities—that should be our A students. And our B students would be the kids that get it. They need a little bit of assistance, maybe some reminders, but they pretty much get it. B are those kids that you can catch one on one, do some small groups, and try to push them into that next proficient level."

"An A in my classroom, unfortunately, means that they probably turned everything in. It's not necessarily A work, but that's what it means because if you put out a rubric, and you're like, 'Okay they've met these standards, but it's absolutely awful work,' you have to give the A. . . . [I]f you have a C in my class, that means I was struggling to get any work from you, and I went through your notebook, found anything that I could possibly grade, and put it in the grade book so that you would pass."

"Right now, we're having these conversations as well—'Shouldn't an A be outstanding?' I don't know because what if they're just really good at it, and they actually don't even try to get better? Does that kid deserve an A and an outstanding? Or does the kid deserve an A and a satisfactory?"

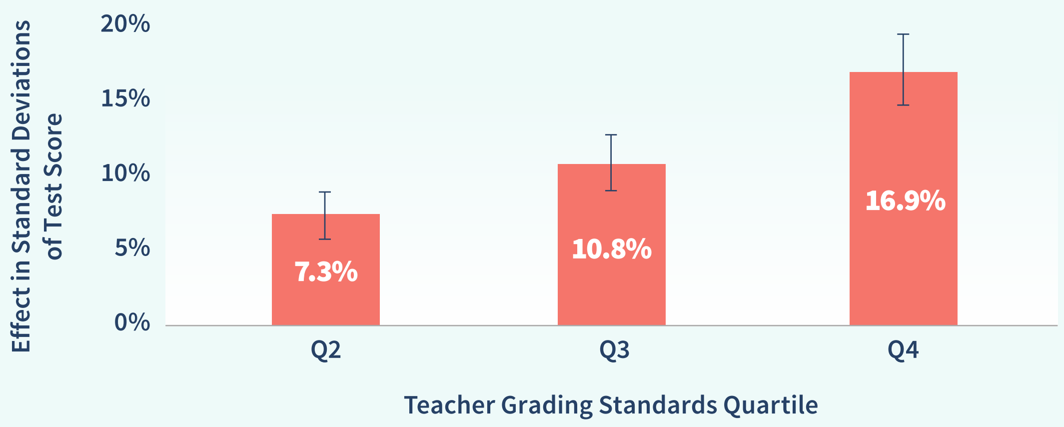
THE EFFECT OF HIGHER GRADING STANDARDS ON STUDENT LEARNING

FINDING 1: STUDENTS LEARN MORE FROM TEACHERS WHO HAVE HIGHER GRADING STANDARDS.

Teacher grading standards are identified in semesters other than the ones where student outcomes are measured to estimate an effect of grading standards on outcomes.[25] Further, the main analysis categorizes teachers into four evenly sized groups (that is, quartiles) based on their grading standards, where group 1 has the lowest standards and group 4 has the highest. Group 1 teachers, the bottom quartile of the grading-standards distribution who have the lowest standards, are the reference group to which other groups of teachers are compared.

Figure 1 shows that teachers in the top quartile increase their students' test scores by a whopping 16.9 percent of a test-score standard deviation over those of their counterparts in the bottom quartile. Even teachers in the middle of the grading standards distribution are significantly more effective than teachers with the lowest grading standards (to put these differences in perspective, they are larger than the impact of 12 student absences or replacing an average teacher with one from the eightieth percentile of the teacher-effectiveness distribution).

FIGURE 1. Students learn more from teachers with higher grading standards.



Note: The bars represent the estimated effects of having a teacher in the second, third, or fourth quartile of the grading-standards distribution on Algebra I standardized test scores. Estimates come from the baseline value-added model. The bottom (lowest standards) quartile is the omitted reference group. Error bars represent 95 percent confidence intervals.

The relationship between grading standards and student performance holds for teachers in all four quartiles of grading standards: At each point in the distribution, the effect of higher standards is larger than that of the lowest standards, and the effects increase in magnitude as grading standards rise.[26] The effect sizes are also large. Compared to teachers with the lowest grading standards, those in the top three quartiles boost student performance by about 10 percent of a test-score standard deviation (put differently, this effect is akin to replacing an average teacher with one at the eightieth or eighty-fifth percentile of the effectiveness distribution[27]). No matter how you slice it, stricter grading standards appear to have a sizable impact on student performance.

TEACHER VOICES: WHAT IS THE CONNECTION BETWEEN GRADES AND EXPECTATIONS?

Our interviewees indicated that having high expectations for students means holding all students to the same standard and personally believing that they can meet it. Teachers communicate these high expectations by clearly stating what students need to accomplish to attain a certain grade and explicitly connecting academic work to future success in the outside world.

"I think that the best way to communicate high expectations is to directly tell the students what the expectations are as far as the grades, to provide rubrics as much as possible based on the assignments so that it's never a mystery to students as to why they get the grade that they do."

"Let's say I give my students an assignment, and I want them to use some sort of a punctuation mark, and they use it wrong. So, 'That's nice that you wrote the whole paper, but the punctuation mark—which is what I was looking for—is not right, so zero. Go do it! Go fix it! Here's the resources on how you can do it. Then bring it back to me, and then you'll get the grade.'"

"It's, 'Do you really believe that as a teacher, every kid can get better at reading?' And some people in my team, they just don't. They have completely taken out books because [the students] don't read, they don't get it, they won't get it, they're not capable of it."

"I actually expect quite a bit from my students, even though I don't grade them really hard, but I talk a lot about the future with them. I'm like, 'My reading class is not about you passing eighth-grade reading. My reading class is about you getting the skills to do well in your freshman English class and your senior English class and when you get out into the real world.'"

"I also believe that when you pass students, that's lowering your expectations. That's just saying that because they had a rough year or because they have a tough home life—I came from a tough home life, and I . . . became this pity case, and then they passed me. That's lowering expectations, and that's not right for students."

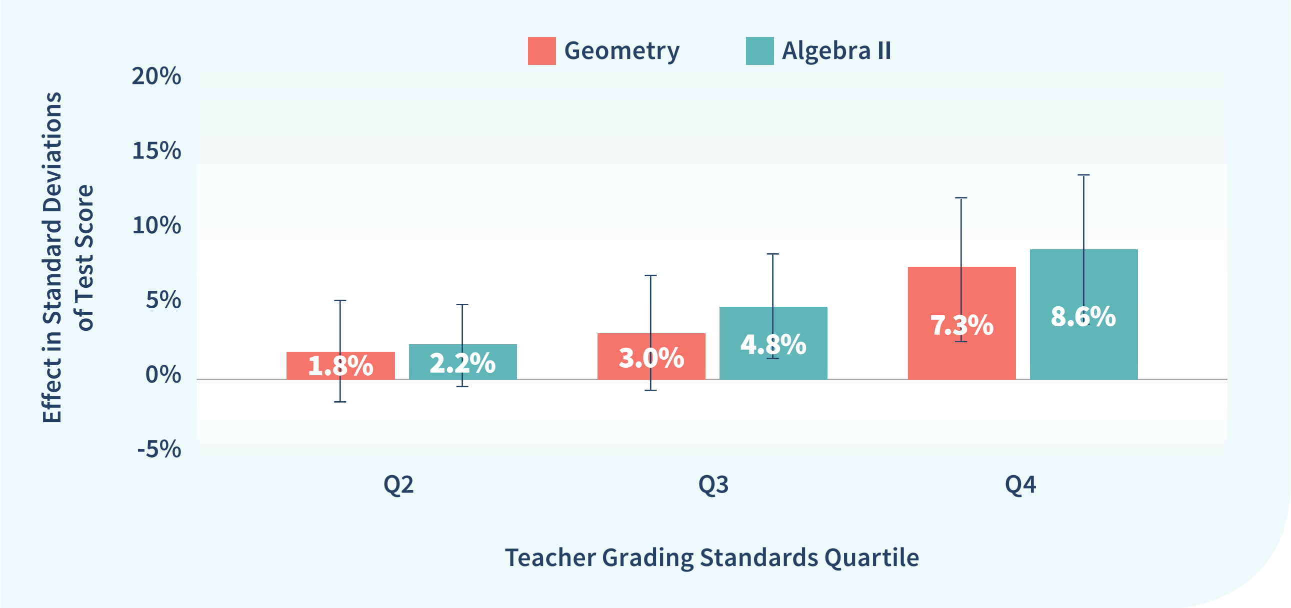
"We teach students that it's okay to make those mistakes as long as you're continually learning, and you can show growth over time. That would mean new and different evidence of a standard. If you can do that, and you're comparing that to what the standard is, that is a skill that you're going to need for life, not just for schooling. And then you get out of school, and you have a job—we want our students to be able to be self-directed and great communicators, be able to say, 'Hey, what are you expecting from me?' to their boss and then be able to compare and say, 'How do I know what I don't know?' or 'How do I know that I'm on where I need to be?'"

FINDING 2: TEACHERS WITH HIGHER GRADING STANDARDS IMPROVE THEIR STUDENTS' PERFORMANCE IN SUBSEQUENT MATH CLASSES UP TO TWO YEARS LATER.

Beyond the effect on EOC scores in the same year the student has a given teacher, longer-term effects can be estimated by examining students' performance on EOC exams in Geometry and Algebra II courses for school years 2006–09.[28]

Figure 2 reports estimates of Algebra I teachers' grading standards on their students' performance on those subsequent math exams. The patterns look much like those in Figure 1: Higher grading standards consistently lead to higher achievement. Predictably, because these tests are in somewhat different subjects and are taken one and two years later, the effect on these longer-range outcomes are smaller than the contemporaneous effects shown in Figure 1.[29] Still, relative to the teachers with the lowest grading standards, students of teachers with the highest grading standards performed significantly better a year later in Geometry (7.3 percent of a standard deviation) and two years later in Algebra II (8.6 percent of a standard deviation).

FIGURE 2. Teachers with higher grading standards boost students' subsequent math performance.



Note: The bars represent the estimated effects of having a teacher in the second, third, or fourth quartile of the grading-standards distribution on Algebra I standardized test scores. Estimates come from the baseline value-added model. The bottom (lowest standards) quartile is the omitted reference group. Error bars represent 95 percent confidence intervals.

It is intuitive that the persistent effects of Algebra I teachers' grading standards are larger in Algebra II than in Geometry, as the course content is more closely aligned. This lends additional support to a causal interpretation of the main result: Higher grading standards improve student learning.

Finally, it is worth asking whether grading standards affect longer-run measures of educational attainment—specifically, high school completion. In fact, we find no effect on high school graduation, perhaps because students taking Algebra I early or on time are already unlikely to drop out[30] (recall that this study did not examine students who took Algebra I later, in tenth grade or beyond). There is, however, suggestive evidence that grading standards may positively influence students' postsecondary intentions (see sidebar *Grading Practices and Postsecondary Intentions,* page 24).

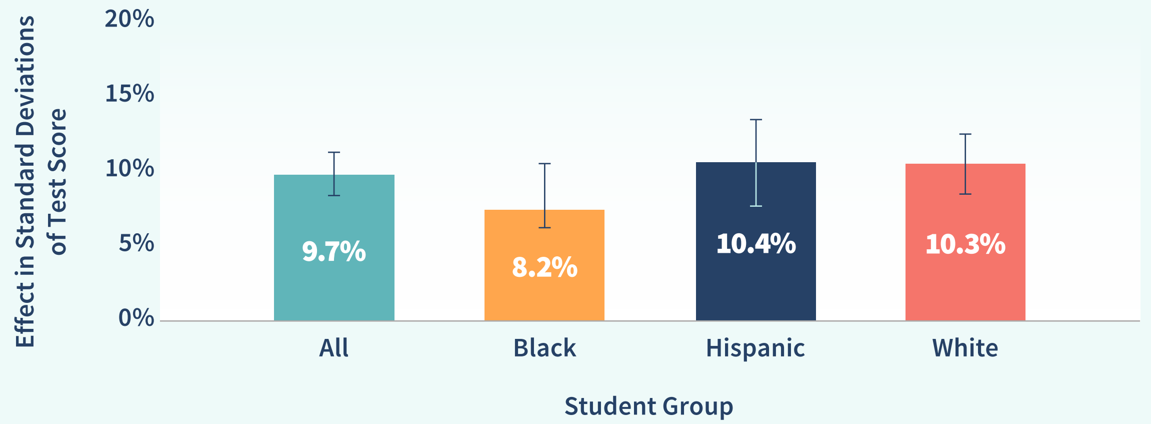
FINDING 3: TEACHERS WITH HIGHER GRADING STANDARDS SIGNIFICANTLY IMPROVE THE LEARNING OUTCOMES OF ALL EXAMINED STUDENT SUBGROUPS.

Figures 3 and 4 report estimates of a simplified version of the baseline model by student subgroups, including race/ethnicity, gender, and previous academic performance. Here we compare teachers in the top 75 percent of the grading-standards distribution to the easiest-grading quartile, which allows us to focus on one effect per subgroup.[31] Overall, we see that having a teacher with higher grading standards boosts achievement by about 10 percent of a standard deviation, similar to what we saw in Figure 1.

> **All of the student groups analyzed benefit from exposure to higher grading standards.**

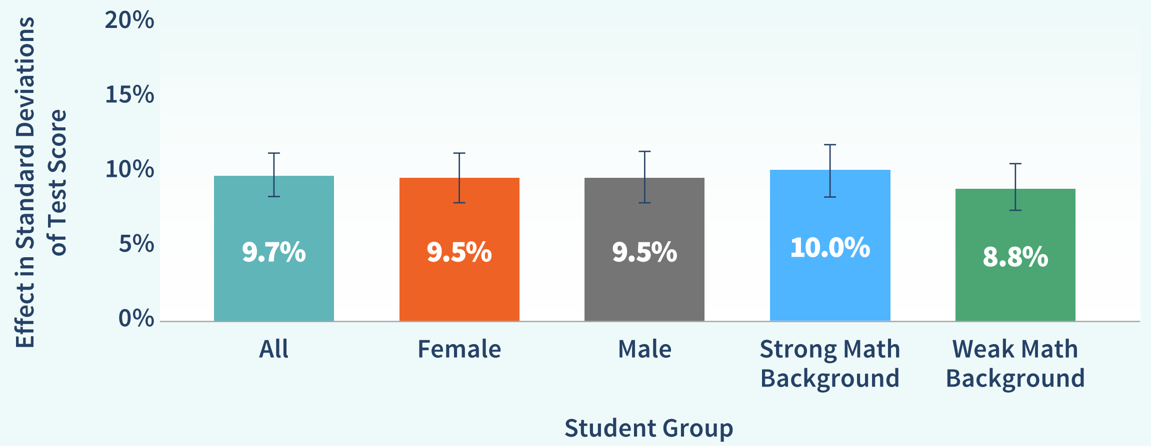
The effect of having a teacher in the upper three quartiles is strongly statistically significant, and similar in size, for each subgroup: It ranges from about 8 to 10 percent of a test-score standard deviation. Some subtle differences emerge, but they are not statistically significant. In other words, all of the student groups analyzed benefit from exposure to higher grading standards. Once again, this finding is largely consistent with the Florida study and alleviates the concern that some students—for example, low performers—could be harmed by higher grading standards.

FIGURE 3. Students of all racial/ethnic groups learn more from teachers with high grading standards.



Note: The results are estimated using a version of the baseline value-added model separately by student subgroup, where the reported estimates are of the effect of the teacher being in the top three quartiles of the standards distribution. Error bars represent 95 percent confidence intervals.

FIGURE 4. Students learn more from teachers with high grading standards regardless of their gender or previous academic background.



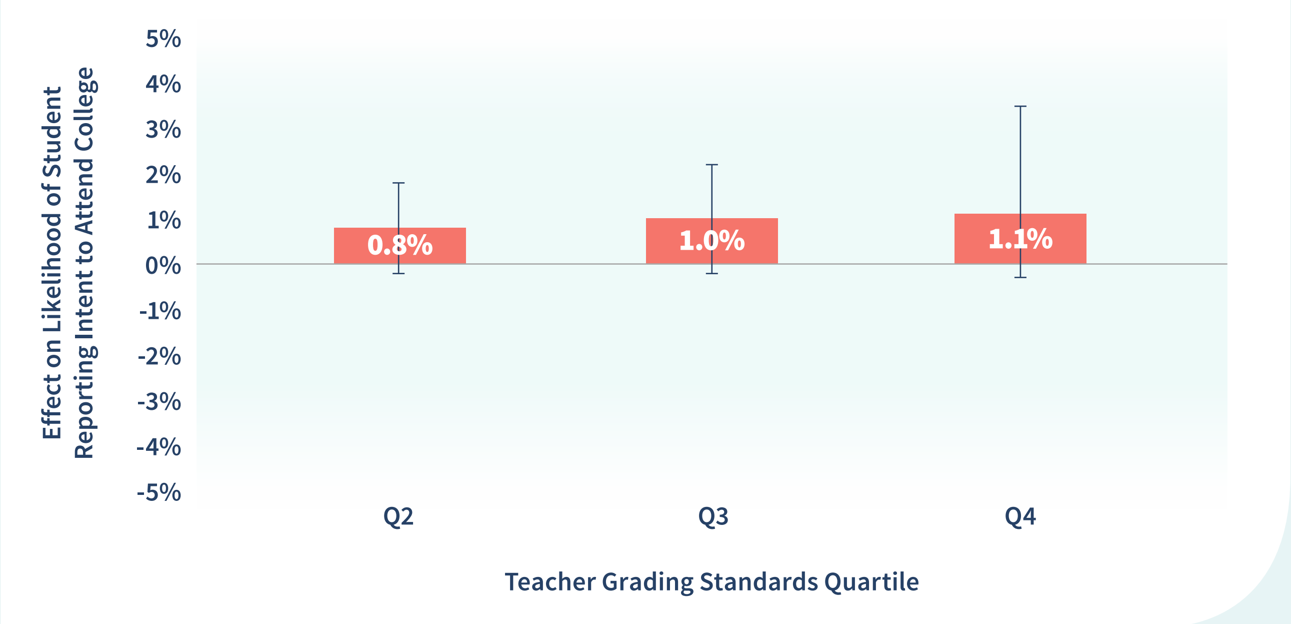
Note: The results are estimated using a version of the baseline value-added model separately by student subgroup, where the reported estimates are of the effect of the teacher being in the top three quartiles of the standards distribution. Error bars represent 95 percent confidence intervals. Designations of strong and weak math background are assigned based on students' end-of-grade math assessment scores in previous years.

GRADING PRACTICES AND POSTSECONDARY INTENTIONS

In addition to math outcomes, higher grading standards could also influence how students perceive their own academic ability and their future academic plans. To examine whether grading standards change students' post–high school plans, we reviewed responses to a survey administered to North Carolina seniors in the years covered by this study. That survey asked students what their plans were after graduation (work, college, military, and so on).[32]

There is suggestive evidence that higher grading standards raise students' intent to attend a four-year college or university, though these results are generally statistically insignificant. Teachers with grading standards in the top three quartiles increase students' stated college intent by about one percentage point, although this result is only marginally statistically significant ($p < 0.10$).

FIGURE SB-1: Having an Algebra I teacher in the top three quartiles of grading standards is weakly correlated with intent to attend a four-year institution.



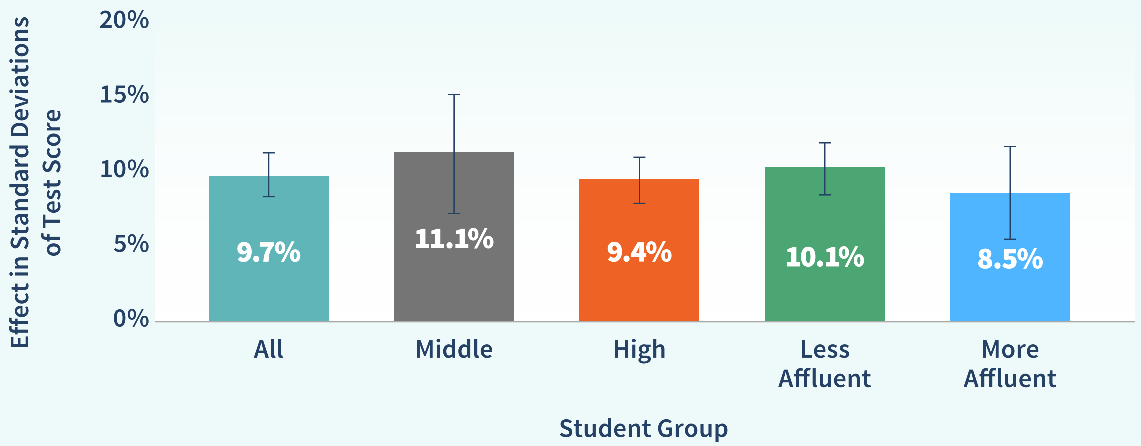
Note: The bars represent the estimated effects of having an Algebra I teacher in the second, third, or fourth quartile of the grading-standards distribution on reported intent to attend a four-year college immediately after high school. Estimates come from the baseline value-added model. The bottom (lowest-standards) quartile is the omitted reference group. Error bars represent 95 percent confidence intervals. These estimates are reported in Appendix B*, Table B-3.*

This finding is not as robust as the other results, but it shows at least that higher expectations do not negatively impact postsecondary plans, and it may suggest that the benefits of high grading standards change students' outlooks and attitudes toward school beyond their math classrooms.[33]

FINDING 4: TEACHERS WITH HIGHER GRADING STANDARDS SIGNIFICANTLY IMPROVE STUDENT LEARNING IN ALL TYPES OF SCHOOLS.

The impact of grading standards might vary by school type for several reasons. For example, school climate might influence how students respond to standards. Figures 5 and 6 report estimates of a simplified baseline model separately by seven school types (based on affluence, location, and grade levels). We see that the effects of high standards are all positive and similar in size in all school types, suggesting that high grading standards are universally beneficial. Specifically, we report the effect of teachers in the top 75 percent of the grading-standards distribution relative to the easiest-grading teachers.[34]

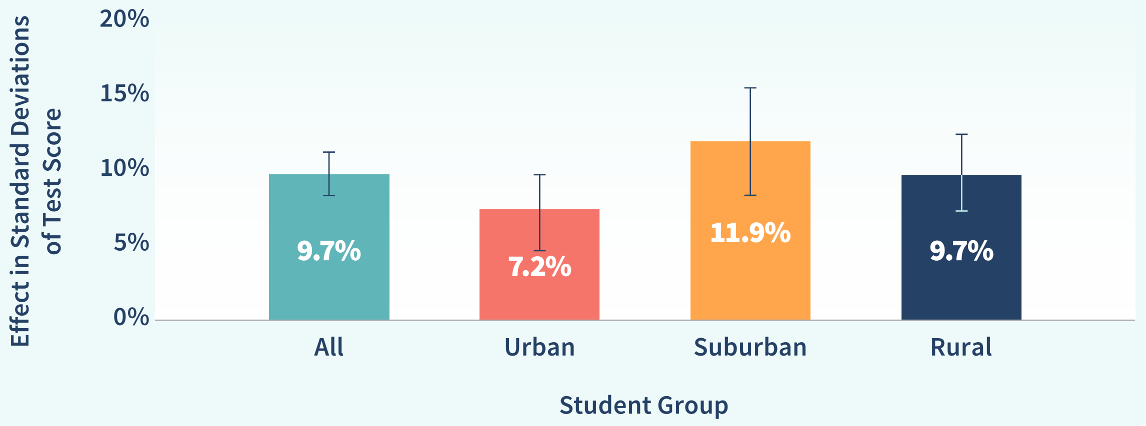
FIGURE 5. Students in schools with different grade levels and different socioeconomic makeups learn more from teachers who have higher grading standards.



Note: The results are estimated using a version of the baseline value-added model separately by school type, where the reported estimates are of the effect of the teacher being in the top three quartiles of the standards distribution. Error bars represent 95 percent confidence intervals. The schools are considered to be "more affluent" when fewer than 50 percent of students are eligible for free or reduced-price lunches and "less affluent" when more than 50 percent of students are eligible.

The effect of high grading standards varies from 7.2 to 11.9 percent of a test-score standard deviation (Figure 6). The biggest difference is by locale, where the effect of high grading standards is five percentage points larger in suburban schools than in urban schools, although not statistically significant.

FIGURE 6. Students in different types of communities learn more from teachers who have higher grading standards.



Note: Results are estimated using a version of the baseline value-added model separately by school type, where the reported estimates are of the effect of the teacher being in the top three quartiles of the standards distribution. Error bars represent 95 percent confidence intervals.

Despite subtle differences between middle and high schools and schools serving more- and less-advantaged student populations, Figures 5 and 6 show that high standards are beneficial in all types of schools. Together with the student subgroup analyses reported in Figures 3 and 4, this result reinforces the idea that high grading standards are a universally beneficial teacher behavior.

TEACHER VOICES: DO TEACHERS FEEL PRESSURE TO CONFER CERTAIN GRADES?

Our teacher interviewees often reported pressure from others to confer higher grades or to pass students. This pressure can come from administrators who want to lower failure rates and raise graduation rates, from parents who insist that their children deserve an A, or from students who understand the stakes associated with higher grades. The extent of these pressures may vary depending on how close-knit the community is and/or parents' socioeconomic status, among other factors.

"It just kind of seems like now there's so much pressure on teachers to, frankly, inflate the grade to help out a student because I know I've been told, 'Hey, Mr. [name], I got a B in your class, it's the only class I ever got a B in, and it might cost me the scholarship.'"

"It's hard because we live in such a small community out here. Everybody knows everybody, and I don't know if it's an intentional pressure, but the pressure that, 'If I say no to this person, will it bite me later on? Will I need something some day?' And they'll be like, 'Well, you didn't help my kid get his A.'"

"There's times, too, where I get emails from parents that are just like, 'I demand to know the rationale behind my kid's score. I demand to know why my kid got a B. My kid has an A in all the other classes but yours. You need to do the right thing and change their grade to an A.'"

"We actually get chastised if anybody even fails our classes. If you have a kid failing, the teacher's the one that's in trouble, not the kid. It's the teacher's fault."

"We just end up, as teachers, it's easier—and this is awful to say—it's easier just to pass the kid than to actually really give valid feedback, if that makes sense. But none of that part's in writing. This is just what we get as teachers. This is how we feel."

"It just seems to be maybe pressure from like society or like the unknown of, 'What will happen if I don't do this?' And the fact that teachers are teaching in isolation, so when you're pressured to do those things, you're kind of on your own, or you feel like it's not a team [that's] going forward. It's just you out there, and so you could get in trouble if you don't do what they're asking you to do."

"When we have these conversations at an in-service or at a staff meeting, I hear teachers ask questions like, so if a student is proficient when they begin, shouldn't I wait and give them a proficient score at the end of the year because we need to show growth?"

WHAT TYPES OF TEACHERS HAVE HIGH STANDARDS?

FINDING 5: TEACHERS WHO ATTENDED SELECTIVE COLLEGES, HOLD GRADUATE DEGREES, AND HAVE MORE EXPERIENCE TEND TO HAVE HIGHER GRADING STANDARDS.

Grading standards vary with teacher characteristics. Here, the average raw measure of grading standards is standardized (presented in standard deviations) and reported for each group rather than splitting the teachers into quartiles (see the *Research Methods* section and *Appendix A* for more information on how grading standards are calculated). Higher numbers reflect higher grading standards.

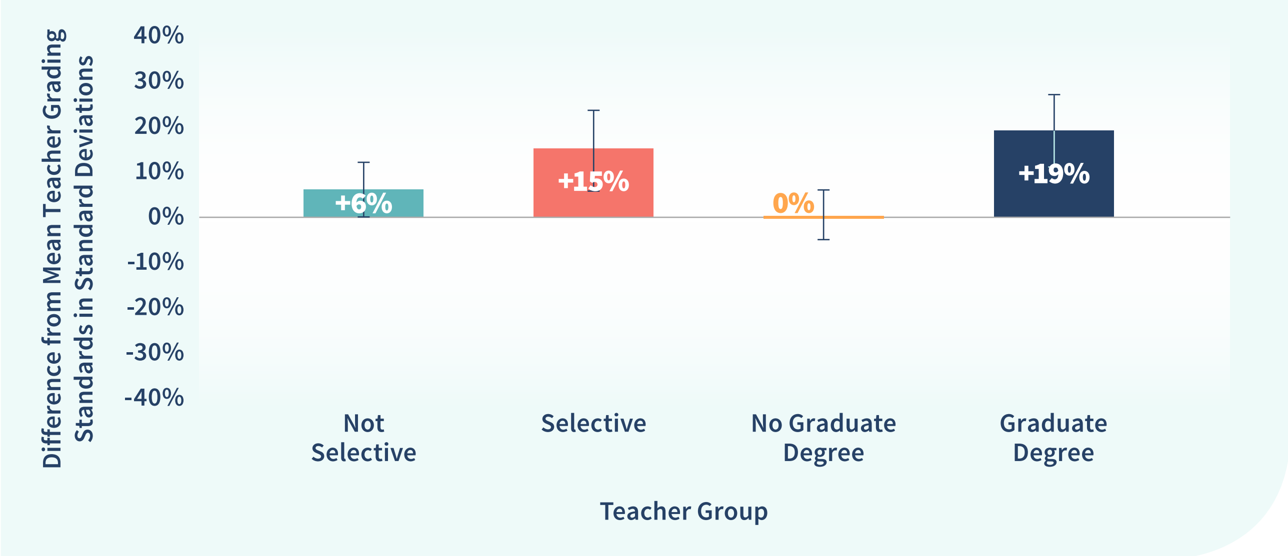
The first characteristic is the selectivity of the teacher's undergraduate college or university, with selectivity defined using ratings from Barron's *Profiles of American Colleges*.[35] Teachers who attended selective schools have higher grading standards than those who did not (see Figure 7). This difference represents about nine percent of a standard deviation in grading standards, although it is not statistically significant.

We see a somewhat larger difference when comparing teachers with graduate degrees to those without (also Figure 7). The former have grading standards that are about 19 percent of a standard deviation stricter than the latter. These results suggest that one's own experience in more challenging academic environments or with higher-quality postsecondary instructors may promote higher standards in grading practices.

Grading standards also differ, on average, by teacher gender. Female teachers tend to have higher grading standards than their male counterparts (see Figure 8). Female teachers have grading standards that are, on average, about 18 percent of a standard deviation higher than those of male teachers.

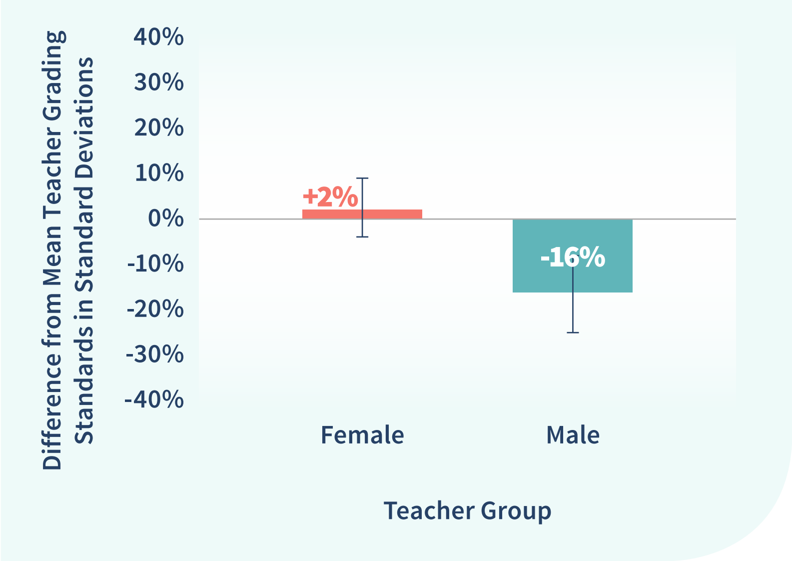
"Female teachers have grading standards that are, on average, about 18 percent of a standard deviation higher than those of male teachers."

FIGURE 7. Teachers who attended selective undergraduate institutions and who have graduate degrees tend to have higher grading standards.



Note: The outcomes may not average to zero because different groups are of different sizes and data are missing for some teachers. Error bars represent 95 percent confidence intervals.

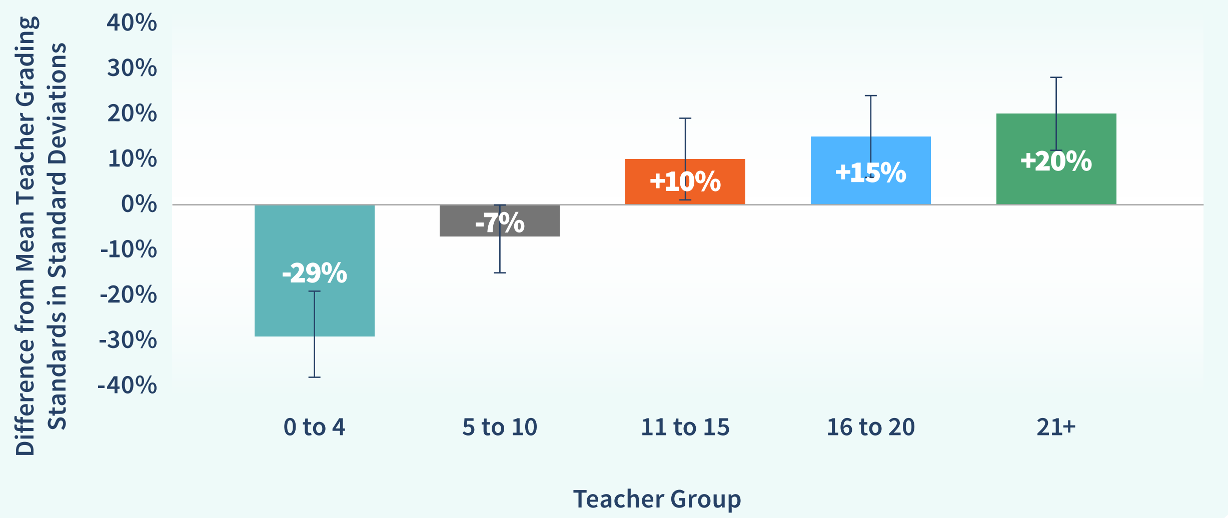
FIGURE 8. Female teachers tend to have higher grading standards than male teachers.



Note: The outcomes may not average to zero because different groups are of different sizes and data are missing for some teachers. Error bars represent 95 percent confidence intervals.

Finally, Figure 9 shows that as teacher experience increases, grading standards rise as well. Although some of the differences across groups are not statistically significant, there is a clear trend that teachers increase their grading standards as they remain in the profession, particularly during their first 15 years. New teachers have grading standards that are 29 percent of a standard deviation lower than average, while veteran teachers are consistently stricter than average. Importantly, this finding suggests that grading standards are malleable, which is relevant to policy makers and school leaders who are considering grading-standard interventions and associated policy changes. However, finding that more years of experience is correlated with higher standards could also be a result of a cohort effect, as more experienced teachers entered the profession in earlier times, when grading practices may have been different.[36]

FIGURE 9. Teachers who have more years of experience tend to have higher grading standards.



Note: The outcomes may not average to zero because different groups are of different sizes and data are missing for some teachers. Error bars represent 95 percent confidence intervals.

FINDING 6: GRADING STANDARDS TEND TO BE HIGHER IN MIDDLE SCHOOLS, SUBURBAN SCHOOLS, AND SCHOOLS SERVING MORE ADVANTAGED STUDENTS.

Figures 10 and 11 report average teacher grading standards separately by school type. This is necessarily descriptive, as teachers are not randomly assigned to schools and school-level factors might differently influence grading standards and EOC exam scores. For example, the comparison of middle to high schools is not an apples-to-apples comparison, as the eighth-grade middle school students are taking Algebra I a year early and therefore might be higher performers or better test takers than typical high school Algebra I students.[37]

Once again, the average raw measure of grading standards is standardized (presented in standard deviations) and reported for each group rather than splitting the teachers into quartiles. Higher numbers reflect higher grading standards.

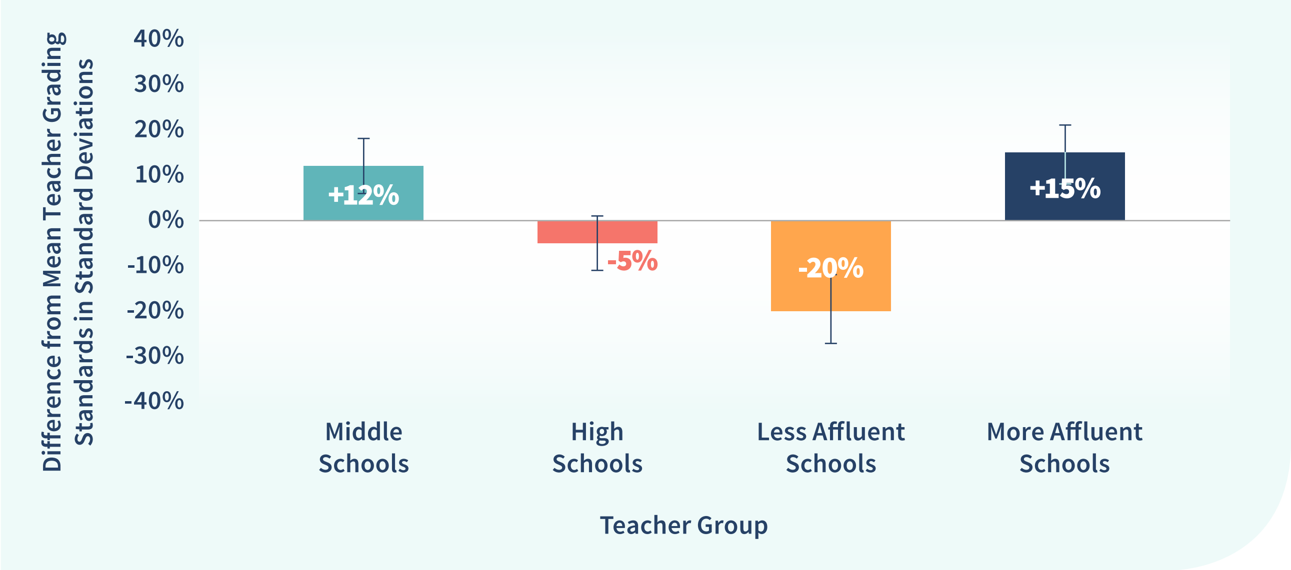
The first comparison is between middle and high schools. Recall that students in the sample took Algebra I in either the eighth or ninth grade, and school cultures likely vary between middle and high schools; moreover, students who take Algebra I in middle school are generally stronger math performers than those who take the course later. As shown in Figure 10, grading standards are markedly higher in middle schools on average. Specifically, they are about 17 percent of a standard deviation higher than grading standards in high schools.

The difference between less and more affluent schools is even more dramatic, with significantly lower standards in the less affluent schools (also Figure 10). More affluent schools have grading standards that are, on average, more than one-third of a standard deviation stricter (35 percent). This is troubling, as it provides more evidence of the “soft bigotry of low expectations” for relatively disadvantaged students and is yet another example of how such schools compound the disadvantages of their students.

> **“More affluent schools have grading standards that are, on average, more than one-third of a standard deviation stricter.”**

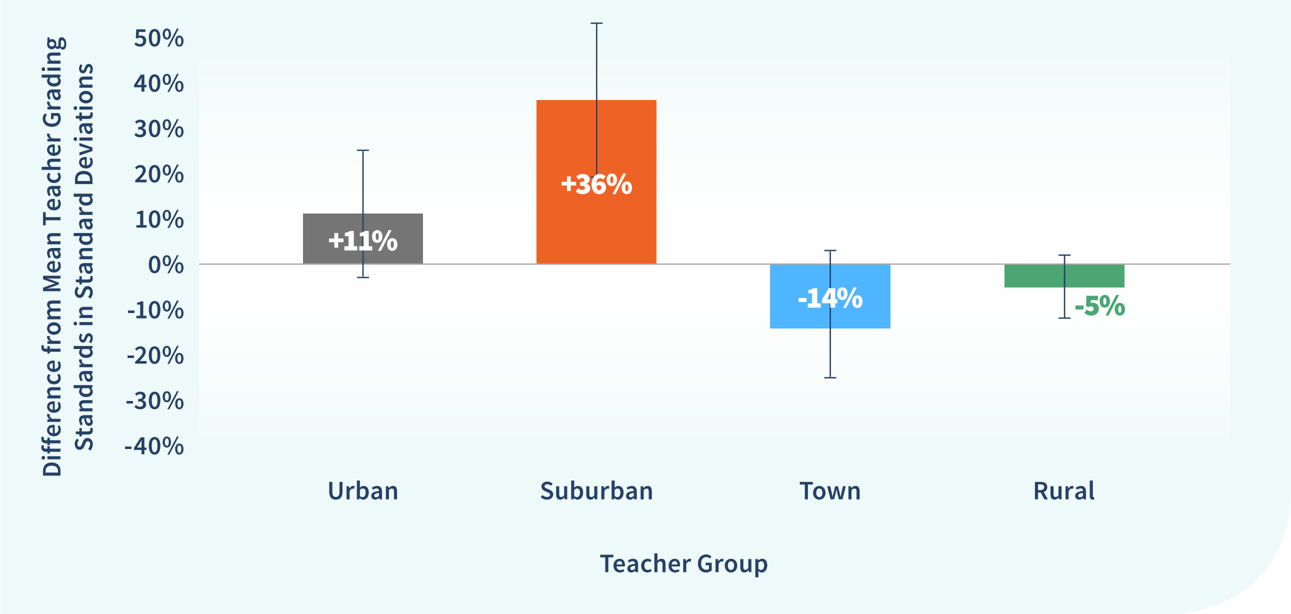
Finally, we see in Figure 11 that grading standards are highest in suburban and urban schools and lowest in schools in small towns and rural areas. Suburban schools have grading standards that are, on average, more than one-third of a standard deviation (36 percent) higher than the mean and about half a standard deviation (50 percent) higher than those in small towns, where grading standards are lowest.[38]

FIGURE 10. Grading standards tend to be higher in middle schools and schools serving more advantaged students.



Note: The outcomes may not average to zero because different groups are of different sizes and data are missing for some teachers. Error bars represent 95 percent confidence intervals. Schools are considered to be "more affluent" when fewer than 50 percent of students are eligible for free or reduced-price lunches and "less affluent" when more than 50 percent of students are eligible.

FIGURE 11. Grading standards tend to be higher in suburban schools.



Note: The outcomes may not average to zero because different groups are of different sizes and data are missing for some teachers. Error bars represent 95 percent confidence intervals.

TEACHER VOICES: IS GRADING CONSISTENT ACROSS TEACHERS AND SCHOOLS?

Some districts and schools adopt grading scales where a numerical grade range corresponds to a letter grade, while others give teachers full autonomy over their grading practices. Less common, but on the rise, is a standards-based grading system, where grades on a letter or number scale (for example, 1–4) are attached to levels of proficiency on specific standards.

Based on our interviews, it appears that, regardless of whether a school or grade level or department has adopted a standard grading scale, variation in grading practices exists among teachers and within schools. The teacher's desire to defend a grade (or not), her feelings of adequacy as a teacher, her concept of how hard students work, as well as her professional and personal experiences with education—not to mention her personality—can all influence how she chooses to grade student work.

"The teachers who grade a little bit easier, like just easy, they probably struggle with teaching it, and so they feel that they don't want to have to [deal with] complaints, so they just give the kid the grade, so they don't have to hear about it, or they felt guilty that they didn't teach it right."

"When I feel like I can justify every single mark, every grade, it's because we've done this the whole week, we've looked at the rubrics, we've practiced here, we've practiced there, there's no excuse. When I grade easier, I'm lowering my expectations."

"With every teacher having so much autonomy, it becomes fairly complex and difficult because that means at our school that students are expected to remember eight different teachers and eight different grading systems. What an A is in one class may not be the same in another class."

"[That teacher]'s more like, 'They look like they were working all day. I'll just give them a checkmark.' And I'm like, 'You only scored 15 out of 20 on this assignment. Either you need to fix these five or take your 15 out of 20.' So each of us have our own personalities and the way we run our classroom, so those are different for sure."

"I think because we all are individual teachers and we all have very strong feelings, or we've grown up with our own experiences, and so some teachers are influenced by, 'Well, when I grew up, an A was [this], and you had to compete, and you had to be the best to get an A.' And that's their values, their core values."

"I think it really goes to the 'lifelong learner' of certain people. I attend conferences to improve practices—what's the best way to measure this, when is measuring too much, when is it too little—and then some of my colleagues don't do any of that. They've been teaching for a long time; their methods have worked for the last 20, 25, 30 years."

IMPLICATIONS

This study breaks new ground in finding that teacher grading standards positively impact immediate and longer-term outcomes. Building on work showing that grading standards influenced elementary student outcomes,[39] the report not only finds that the importance of high grading standards extends to middle school and high school students but also that the effect of high standards is quite consistent for different types of schools and students. Moreover, higher grading standards are associated with specific teacher characteristics, including years of experience, advanced education, and race/gender—all of which may prompt further investigation.

Three main lessons flow from these findings.

First, recognize that it will take active measures to reach and sustain high grading standards.

Academic department, school, district, and state leaders must work proactively to monitor grading practices and take steps to ensure that teachers are not awarding "easy As." As Eva Moskovitz, CEO of Success Academy, recently put it,

> *When teachers give high grades for mediocre work, no one asks any questions, and they can carry on as before. When they give more realistic grades, they have an obligation to follow up with detailed feedback, more support, and better instruction. It's not surprising then that most—often unconsciously—opt for the first course of action.*[40]

In other words, the incentives work against teachers who maintain high expectations for students and, subsequently, high grading standards. Education leaders must bring to light the issue and place a premium on raising the norm for what it means to demonstrate excellence in student work. Unfortunately, that's not what we're seeing in some districts and schools that have adopted particular grading policies, such as banning a grade of zero or less than a 60, which inflates students' grades and restricts teachers' power to enforce high standards.[41]

The difficulty of maintaining high grading standards means that teacher-preparation programs and institutions also have a role here, especially as newer teachers tend to have the lowest standards. Curricula for preservice training and later professional development should stress the importance of holding students to high standards. Some teacher-training programs are already on board. For example, Teach for America's five-week summer institute includes a module dedicated to the power of holding high expectations for all students.[42] Similarly, the teacher professional-development program "Great Expectations" revolves around six basic principles, one of which exhorts teachers to hold high expectations for all students because "they will respond by reaching upward to achieve them."[43]

> **The mindset that says 'everyone gets a gold star' does more damage than good, and those invested in education would be wise to jettison that view and strive to raise the bar for excellence.**

Second, education leaders and researchers must make clear the damaging consequences of low grading standards, addressing the incentives and lack of training that foster them.

Students assigned to teachers with the lowest grading standards underperformed those taught by teachers with the highest standards by 17 percent of a standard deviation in the current year and seven to nine percent of a standard deviation in subsequent years and math courses. Although the mechanism that leads from higher standards to higher performance is not clear—and may be multiple and overlapping—the imperfect information provided by easy grades can lead to a host of unintended consequences.

Assigned grades that are higher than a student's content mastery can foster a sense of complacency that deters a young person from reaching their full potential. Since parents often look to grades as the most important signal of their children's academic achievement, inflated grades make it difficult for them to understand what challenges their children face and to hold them accountable for their performance. Moreover, socioeconomic gaps in this type of grade inflation can contribute to similar gaps in educational outcomes.

In short, the evidence shows that the mindset that says "everyone gets a gold star" does more damage than good, and those invested in education would be wise to jettison that view and strive to raise the bar for excellence.

Finally, grading standards are a useful measure of one component of teacher effectiveness that schools can use to improve their teaching workforce.

The simple definition of grading standards used in this report is easily calculated by schools and districts—that is, if students take external exams. Because observable markers of effective teaching are in short supply, grading measures of this sort provide schools and districts rich opportunities to identify, learn from, retain, and promote teachers who implement high standards. Such teachers can also be a valuable resource for improving professional development on the topic.

On the flip side, these measures can also be used to identify teachers with low grading standards and provide them ample guidance, opportunities, and resources to bolster both their expectations and classroom practice. It bears repeating that none of this is possible without a robust set of external measures of student learning, such as North Carolina's EOC exams.

Of course, changing grading policies—both implicit and explicit—is easier said than done. Better understanding of why and how grading standards matter and can be improved is important work, not only for education researchers but for all those who work in schools. Hopefully, the results of this report—and the teacher voices highlighted throughout it—will kick off much needed discussions among educators about the pervasiveness of low standards and the benefits of raising them. It will take time, but we must learn how to make high expectations and high grading standards a part of the teaching culture through hands-on teaching, optimized incentives, and stronger professional development.

"We must learn how to make high expectations and high grading standards a part of the teaching culture through hands-on teaching, optimized incentives, and stronger professional development."

APPENDIX A: TECHNICAL APPENDIX

A.1 GRADING STANDARDS

The first order of business is to define and create measures of teachers' grading standards, which we'll call *S*. As described in the main text, our preferred measure is a simple, intuitive metric: The mean EOC score of a teacher's students who receive a B grade. However, there is one caveat to creating this (and any other) measure of *S* not mentioned in the main text: We do not want to include the student's grade and EOC score, or those of her classmates, when computing the *S* faced by the student. That is, *S* is not only teacher specific but also year specific. The reason is that we do not want the current student(s) to influence our measure of *S*, as the EOC scores that determine *S* are also the outcome of interest.

In all that follows, EOC scores are standardized by grade and year to enable cross-year comparisons and *i*-, *j*-, and *t*-index students, teachers, and years, respectively. A subscript $-t$ indicates all years other than *t*, and N_j is the number of students taught by teacher *j*. Our goal is to estimate S_{jt} for all teachers. In any given year *t*, S_{jt} is computed using the grades and EOCs of teacher *j*'s students in all years except for year *t*. For the preferred measure of *S*, we proceed to estimate the conditional mean of teacher *j*'s non-year-*t* students' EOC scores, as follows:

$$\text{1.}\quad \hat{S}_{jt} = \frac{\sum EOC_{ij,-t}}{N_{j,-t}} | Grade_{ij} = B, \forall\, t.$$

Following Figlio and Lucas, we also verify that our main results are robust to using an alternative definition of *S*. Here, we use a regression-based approach to estimate teacher fixed effects that explain the time-invariant teacher-level variation in EOC scores that is not explained by course grades. Formally, for each year *t* we estimate linear regression models of the form

$$\text{2.}\quad EOC_{ij,-t} = S_{jt} + f\left(Grade_{ij,-t}\right) + u_{ijt}, \forall\, t.$$

Here, S_{jt} is a teacher fixed effect. The OLS estimates of S_{jt}, $\hat{S}_{jt}$, are our alternative measure of teachers' grading standards. Once again, higher values of *S* indicate more stringent grading standards, because for a given course grade, a larger value of *S* implies higher average EOC scores for the teacher's students. The only remaining question is how *Grade* should enter equation (2); that is, what is the functional form of f? Because the raw transcript data reports grades as integer values from 50 to 100, we opt for a completely unrestrictive, nonparametric specification of f. Formally, this means that $f(Grade_{ij}) = \sum_{k=50}^{100} \delta_k 1\{Grade_{ij} = k\}$, where $1\{.\}$ is the indicator function. This fully conditions on grades, so that only mean EOC differences by teacher remain. Now that we have our estimates of *S*, we can proceed to estimate the impact of *S* on student outcomes.

A.2 VALUE-ADDED MODELS

The now-standard approach to estimating teacher effects is to include teacher indicators in a simple lag-score value-added model, as the inclusion of the lag score controls for the past inputs received by the student and for the most common types of nonrandom sorting of students to classrooms.[44] Similarly, to estimate the impact of observable teacher characteristics on student outcomes, the teacher indicators are replaced by observed teacher characteristics. This is our approach.

Specifically, we estimate value-added models of the form

$$EOC_{ijstg} = \tau\hat{S}_{jt} + \theta_{stg} + \beta_1 X_i + \beta_2 Z_j + \varepsilon_{ijstg}, \quad (3)$$

where the new subscripts *s* and *g* index the school and grade, respectively, in which the student took Algebra I. The coefficient of interest is τ, which represents the causal effect of standards on student EOC scores (or other education outcomes that can be placed on the left-hand side of equation [3], such as performance in subsequent math EOCs and indicators for high school graduation or college intent). θ is a school-by-grade-by-year fixed effect. The vector *X* includes basic controls for student demographics, as well as the all-important measure of lagged math performance. This is the end-of-grade standardized math score from grade 7 or 8, depending on whether the student took Algebra I in grade 8 or 9. The vector *Z* includes controls for observed teacher characteristics, such as years of teaching experience, the selectivity of the teacher's undergraduate institution, and an indicator for whether or not the teacher holds a graduate degree. We allow experience to enter the model linearly, as recent research suggests that the returns to experience are approximately linear.[45] Selectivity is a dichotomous variable based on Barron's *Profiles of American Colleges*. Specifically, the selective indicator equals one if Barron rated the institution as most or highly selective and zero otherwise. We can test for heterogeneous effects by augmenting equation (3) to include interactions between *X* or *Z* and $\hat{S}$ or by estimating equation (3) separately by subgroup.

Although equation (3) is in many ways a standard sort of value-added model that is now common in education research, there are a few nuances that merit further discussion. First, the hat on *S* in equation (3) indicates that *S*, our independent variable of interest, is itself an estimate generated by either equation (1) or (2). Because it was estimated, the usual OLS standard errors will be too small, as they do not take account of the noise inherent in the estimate. This is analogous to the "problem" encountered when naively plugging the first-stage fitted value into the second stage of an instrumental-variables procedure.[46] Accordingly, we compute correct standard errors using a bootstrap procedure. In all analyses, we use 500 bootstrap replications to generate standard errors that allow for proper statistical inference. Because we want to allow for serial correlation within schools—that is, cluster at the "highest level"—we compute cluster-robust standard errors using a block bootstrap at the school level. That said, 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.

Second, a subtle point about our estimated $\hat{S}$ is that because it is estimated from student data, the precision of this estimate will vary as a function of the number of students taught by each teacher.[47] This imprecision means that there is likely some measurement error in the variable of interest, $\hat{S}$, which will lead to attenuation bias in our estimates of τ. In other words, the inherent noise in the estimates of $\hat{S}$ means that the estimates are biased toward zero and thus represent lower-bound, or conservative, estimates of the true impact of high grading standards.

Third, the school-by-grade-by-year fixed effect (θ) plays an important role in the analysis, as it ensures that we are comparing students who have different Algebra 1 teachers (and thus are exposed to different levels of *S*) but who are taking Algebra I in the same school year, in the same school, and in the same grade. As explained in the main text, this step removes the threat to validity that unobserved school- or district-level policies are jointly influencing both teachers' grading standards and student outcomes and thus confounding our estimate of τ. It also adjusts for the fact that students who take Algebra I in the eighth grade are systematically different from those who take it in the ninth grade. A necessary trade-off of this rich research design is that because it relies on comparisons of teachers in the same school and same grade, the estimate reflects only students in schools with multiple Algebra I teachers. Accordingly, we also estimate models that include the school, grade, and year fixed effects separately and find similar results (see column 4 of Table B-2). This indicates that the results apply to schools that have only one Algebra I teacher in a given grade or year.

These rich fixed effects necessarily subsume the individual school, grade, or year indicators that are commonly included in these types of value-added models. This is not an omission, as these individual effects are explicitly controlled for by θ. A similar point is that the inclusion of θ also makes any type of school- or grade-level controls redundant, as these variables would be colinear with θ. For example, there is no need to control for, say, school size, because it is constant in a given school year and thus captured by θ.

Finally, there is a question of how $\hat{S}$ should enter equation (3) and, in turn, how τ should be interpreted. The measure based on equation (1) is in standardized test-score units, so a one-unit increase in $\hat{S}$ leads to a τ increase in EOC. Still, the practical usefulness of the magnitude of this estimate is not entirely clear, other than to say that if the OLS estimate of τ is positive (negative) and statistically significant, teachers with high grading standards improve (harm) student outcomes. Instead, we sort teachers by their estimated $\hat{S}$ and split them into four evenly sized groups, $\hat{S}_1 - \hat{S}_4$, ranging from the easiest-grading to the hardest-grading quarters of the grading-standards distribution. In other words, we prefer to estimate models such as

$$EOC_{ijstg} = \tau_2\hat{S}_{2jt} + \tau_3\hat{S}_{3jt} + \tau_4\hat{S}_{4jt} + \theta_{stg} + \beta_1 X_i + \beta_2 Z_j + \varepsilon_{ijstg}, \quad (4)$$

where the easiest-grading group of teachers ($\hat{S}_1$) form the omitted reference group. This provides a practically useful estimate that can be explained to policymakers, school leaders, and parents: For example, relative to an easy-grading teacher in the bottom quarter of the standards distribution, a stricter-grading teacher in the top quarter of the grading-standards distribution would boost achievement by τ_4. That said, we also estimate the linear, parametric specification of equation (3) as a robustness check and find qualitatively similar results.

A.3 PREDICTING STANDARDS

To address research question 4, which asks which schools and teachers employ higher grading standards, we simply use the observed values of *Z* (teacher characteristics) and some easily observed school characteristics (*W*, which includes school type, school socioeconomic status, and school location) to predict $\hat{S}$, our estimate of the teacher's grading standards. The simplest way we do this, as shown in Figures 5 through 8, is to compute the mean of $\hat{S}$ separately for teachers and schools with different observable characteristics. However, we also employ a multiple regression approach that allows us to control for all elements of *W* and *Z* simultaneously and also to control for school fixed effects, so that we make within-school assessments of *X*'s ability to predict $\hat{S}$. Once again, we must bootstrap the standard errors to account for the fact that the dependent variable was estimated. In addition to isolating the predictive power of each observable teacher and school characteristic net of the others, this approach provides robust statistical inference. Formally, we estimate

$$\hat{S}_{jst} = \beta_1 W_s + \beta_2 Z_j + \varepsilon_{jst}, \quad (5)$$

where *W* can either represent observed school characteristics or a full set of school fixed effects.

APPENDIX B: SUPPLEMENTAL TABLES

TABLE B-1. Analytic sample means

	Full Sample *N*	All	Q1 (lax standards)	Q4 (high standards)
Outcomes				
Algebra I Score	342,098	0.15	−0.27	0.68
Geometry Score	70,454	0.06	−0.25	0.51
Algebra II Score	71,172	0.07	−0.29	0.51
High School Grad	200,959	1.00	1.00	1.00
Expects College	200,959	0.52	0.41	0.69
Standards				
Mean EOC | B	342,098	0.11	−0.42	0.69
Q2	342,098	0.24	0.00	0.00
Q3	342,098	0.24	0.00	0.00
Q4	342,098	0.28	0.00	1.00
Fixed Effect	342,098	0.03	−0.43	0.53
Q2	342,098	0.18	0.20	0.00
Q3	342,098	0.19	0.06	0.00
Q4	342,098	0.28	0.03	0.87
Student Type				
White	342,098	0.53	0.46	0.60
Black	342,098	0.21	0.28	0.13
Hispanic	342,098	0.10	0.12	0.07
Male	342,098	0.48	0.49	0.47
Teacher Type				
Graduate Degree	342,098	0.37	0.32	0.44
Experience	342,098	16.35	14.77	18.03
Selective College	342,098	0.19	0.16	0.21
School Type				
Middle School	342,098	0.31	0.13	0.62
High School	342,098	0.69	0.87	0.38
Disadvantaged	328,271	0.21	0.31	0.14
Advantaged	328,271	0.79	0.69	0.86
Urban	308,114	0.22	0.23	0.25
Suburban	308,114	0.20	0.13	0.30
Rural	308,114	0.48	0.53	0.39

Note: Sample sizes are reported for the full analytic sample used in the baseline analysis of grading standards' effects on Algebra I scores. The Geometry and Algebra II samples are smaller because they are only available for 2006-2009. Both types of grading standard measures are continuous measures that we break into quartiles (Q). Advantaged schools are schools in which less than half the student body is eligible for free or reduced-price lunch. Regression models dummy out "missing locale" and "missing advantaged" so that we use the full analytic sample of 342,098 students.

TABLE B-2. Effect of grading standards on Algebra I end-of-course standardized test scores

Model	Baseline	No Controls	Teacher Controls	School FE	Parametric Standards	FE Standards
	(1)	(2)	(3)	(4)	(5)	(6)
Quartile 2	0.073	0.075	0.073	0.066		0.013
	(0.008)***	(0.008)***	(0.008)***	(0.008)***		(0.008)
Quartile 3	0.108	0.112	0.107	0.102		0.059
	(0.010)***	(0.009)***	(0.010)***	(0.010)***		(0.009)***
Quartile 4	0.169	0.173	0.169	0.154		0.106
	(0.012)***	(0.012)***	(0.012)***	(0.014)***		(0.011)***
E(EOC|Grade = B)					0.151	
					(0.021)***	
N (Students)	342,098	370,707	342,098	342,098	342,098	349,505
N (S-G-Y FE)	7,145	7,818	7,145		7,145	7,274
N (Schools)	1,215	1,258	1,215	1,215	1,215	1,224

*Note: ***, **, and * indicate statistical significance at the 99, 95, and 90 percent confidence levels, respectively.*

TABLE B-3. Effect of Algebra I teachers' grading standards on subsequent educational outcomes

Outcome	High School Grad	Expects College	Geometry Score Observed	Algebra II Score Observed	Algebra I Score	Geometry Score	Algebra II Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Quartile 2	0.000	0.008	0.009	−0.001	0.064	0.018	0.022
	(0.000)	(0.005)	(0.006)	(0.007)	(0.015)***	(0.017)	(0.015)
Quartile 3	−0.000	0.010	0.006	0.003	0.096	0.030	0.048
	(0.000)	(0.006)*	(0.007)	(0.008)	(0.018)***	(0.020)	(0.017)***
Quartile 4	−0.000	0.011	0.015	0.016	0.168	0.073	0.086
	(0.000)	(0.007)	(0.008)*	(0.009)*	(0.024)***	(0.024)***	(0.025)***
N (Students)	200,959	200,959	82,026	82,026	82,026	70,402	67,093
N (S-G-Y FE)	5,356	5,356	2,516	2,516	2,516	2,393	2,383
N (School clusters)	1,125	1,125	923	923	923	893	884

*Note: ***, **, and * indicate statistical significance at the 99, 95, and 90 percent confidence levels, respectively.*

TABLE B-4. Effect of grading standards on Algebra I scores by student subgroups

Subgroup	White	Black	Hispanic	Male	Female	Above Avg Score	Below Avg Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Quartile 1	−0.103	−0.082	−0.104	−0.095	−0.095	−0.100	−0.088
	(0.010)***	(0.011)***	(0.014)***	(0.008)***	(0.008)***	(0.009)***	(0.008)***
N (Students)	182,478	70,849	33,630	165,189	176,909	175,782	166,316
N (S-G-Y FE)	5,660	4,809	4,384	6,647	6,482	6,297	6,176
N (School clusters)	1,142	1,056	1,022	1,170	1,169	1,155	1,168

*Note: ***, **, and * indicate statistical significance at the 99, 95, and 90 percent confidence levels, respectively.*

TABLE B-5. Effect of grading standards by school subgroups

Subgroup	Middle School	High School	Disadvan-taged	Advan-taged	Urban	Suburban	Rural
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Quartile 1	−0.111	−0.094	−0.085	−0.101	−0.072	−0.119	−0.097
	(0.020)***	(0.008)***	(0.016)***	(0.009)***	(0.013)***	(0.018)***	(0.013)***
N (Students)	109,688	232,410	70,434	257,837	66,747	61,635	147,445
N (S-G-Y FE)	2,913	4,232	1,963	4,698	1,390	1,024	3,112
N (School clusters)	642	573	549	913	210	153	484

*Note: ***, **, and * indicate statistical significance at the 99, 95, and 90 percent confidence levels, respectively.*

TABLE B-6. School- and teacher-level determinants of grading standards (S)

	S			S_1 (Bottom 25%)			S_4 (Top 25%)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Middle			0.379			−0.137			0.191
			(0.027)***			(0.015)***			(0.021)***
Charter			−0.056			0.014			0.101
			(0.137)			(0.066)			(0.093)
Advantaged			0.199			-0.074			0.063
			(0.025)***			(0.015)***			(0.020)***
Urban			0.070			−0.019			0.102
			(0.041)*			(0.025)			(0.036)***
Suburban			0.153			−0.056			0.143
			(0.043)***			(0.026)**			(0.036)***
Rural			−0.015			0.034			0.046
			(0.036)			(0.024)			(0.031)
Grad degree	0.085	0.047	0.080	−0.040	−0.021	−0.039	0.061	0.031	0.057
	(0.021)***	(0.021)**	(0.022)***	(0.012)***	(0.014)	(0.013)***	(0.016)***	(0.019)	(0.016)***
Experience	0.007	0.004	0.005	0.000	0.000	0.000	−0.004	−0.002	−0.004
	(0.001)***	(0.001)***	(0.001)***	(0.001)	(0.001)	(0.001)	(0.001)***	(0.001)***	(0.001)***
Selective	0.010	0.029	0.019	−0.014	−0.016	−0.004	−0.016	−0.020	−0.016
	(0.029)	(0.032)	(0.028)	(0.016)	(0.017)	(0.017)	(0.023)	(0.024)	(0.024)
Missing Selective	0.042	0.065	0.059	−0.028	−0.024	−0.020	0.049	0.039	0.025
	(0.031)	(0.033)**	(0.032)*	(0.016)*	(0.018)	(0.018)	(0.022)**	(0.024)	(0.022)
School FE		X			X			X	
School Controls			X			X			X

*Note: ***, **, and * indicate statistical significance at the 99, 95, and 90 percent confidence levels, respectively.*

ENDNOTES

1. Raj Chetty, John N. Friedman, and Jonah E. Rockoff, “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood,” *American Economic Review* 104, no. 9 (2014): 2633–79; Seth Gershenson, “Linking Teacher Quality, Student Attendance, and Student Achievement,” Education Finance and Policy 11, no. 2 (2016): 125–49, doi:10.1162/EDFP_a_00180; and C. Kirabo Jackson, “What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes,” *Journal of Political Economy* 126, no. 5 (2018): 2072–07.

2. Douglas O. Staiger and Jonah E. Rockoff, “Searching for effective teachers with imperfect information,” *Journal of Economic Perspectives* 24, no. 3 (2010): 97–118, doi:10.1086/699018.

3. Nicholas W. Papageorge, Seth Gershenson, and Kyung Min Kang, “Teacher Expectations Matter,” *Review of Economics and Statistics* (just accepted, 2019): 1–46, doi:10.1162/rest_a_00838; and Seth Gershenson and Nicholas Papageorge, “The Power of Teacher Expectations: How Racial Bias Hinders Student Attainment,” *Education Next* 18, no. 1 (2018): 64–71.

4. Lydia Mechtenberg, “Cheap Talk in the Classroom: How Biased Grading at School Explains Gender Differences in Achievements, Career Choices, and Wages,” *Review of Economic Studies* 76, no. 4 (October 2009): 1431–59.

5. Michael Hurwitz and Jason Lee, “Grade Inflation and the Role of Standardized Testing,” in *Measuring Success: Testing, Grades, and the Future of College Admissions*, ed. Jack Buckley, Lynn Letukas, and Ben Wildavsky (Baltimore, MD: Johns Hopkins University Press, 2017), 64–93.

6. Seth Gershenson, “Grade Inflation in High Schools (2005–2016),” (Washington, D.C.: Thomas B. Fordham Institute, 2018).

7. Jeffrey Denning, Eric Eide, and Merrill Warnick, “Why Have College Completion Rates Increased?” IZA Discussion Paper No. 12411 (2019), https://ssrn.com/abstract=3408309. Alarmingly, on many college campuses, grade inflation is most rampant in the schools of education that train our teachers. Cory Koedel, “Grading Standards in Education Departments at Universities,” *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas* 19, no. 23 (2011): 1–23, doi:10.14507/epaa.v19n23.2011.

8. David N. Figlio and Maurice E. Lucas, “Do High Grading Standards Affect Student Performance?” *Journal of Public Economics* 88, no. 9–10 (2004): doi:1815–34, 10.1016/S0047-2727(03)00039-2.

9. The research design of the Figlio and Lucas study provides the basis for the methods used in the current study. It uses longitudinal data to make within-student comparisons and identifies teachers’ grading standards by comparing the standardized test scores and grades of their students. These approaches are explained in greater detail in this report’s methods section and in the technical appendix. The basic idea is to include the teacher’s grading standards as a measure of teacher quality in models of the education production function, just as we would estimate the effect of teacher experience or class size.

10. The experimental literature on the factors that increase effort is another source of supporting evidence. For instance, a randomized field experiment involving over 1,000 Swedish sixth graders finds that performance on a quiz significantly varies by the (randomly assigned) grading scheme of the quiz. Rank-based grading had the largest effect. Interestingly, girls responded to the promise of a “symbolic prize” (a certificate) for scoring well on the quiz. One might consider course grades as partly symbolic, as there are relatively low stakes for students in grades 8 and 9. And once again, the gender difference suggests that there are differences by gender in the response to high grading standards. See Nina Jalava, Juanna Schrøter Joensen, and Elin Pellas, “Grades and Rank: Impacts of Non-Financial Incentives on Test Performance,” *Journal of Economic Behavior & Organization* 115 (July 2015): 161–96, doi:10.1016/j.jebo.2014.12.004. Research on how grading schemes affect student effort and, in turn, student achievement offers additional indirect evidence that grading standards matter. And there are levers available to policy makers. To wit, a recent change in Chilean colleges from an absolute to relative grading scheme led to dramatic changes in the grade distribution. Valentina Paredes, “Grading System and Student Effort,” *Education Finance and Policy* 12, no. 1 (Winter 2017): 107–28, doi:10.1162/EDFP_a_00195.

11. Philip Babcock, “Real Costs of Nominal Grade Inflation? New Evidence from Student Course Evaluations,” *Economic Inquiry* 48, no. 4 (October 2010): 983–96, doi:10.1111/j.1465-7295.2009.00245.x.

12. Academic achievement is operationalized as assessment outcomes. Julian R. Betts and Jeff Grogger, “The Impact of Grading Standards on Student Achievement, Educational Attainment, and Entry-Level Earnings,” *Economics of Education Review* 22, no. 4 (August 2003): 343–52, doi:10.1016/S0272-7757(02)00059-6.

13. In any case, theoretical models predict that the effects of grading standards may be different for different student groups. See William E. Becker and Sherwin Rosen, “The Learning Effect of Assessment and Evaluation in High School,” *Economics of Education Review* 11, no. 2 (June 1992): 107–18, doi: 10.1016/0272-7757(92)90002-K.

14. Chaiklin, Seth. "The zone of proximal development in Vygotsky’s analysis of learning and instruction." *Vygotsky’s educational theory in cultural context* 1 (2003): 39-64.

15. Victor Lavy, “Do Gender Stereotypes Reduce Girls’ or Boys’ Human Capital Outcomes? Evidence from a Natural Experiment,” *Journal of Public Economics* 92, no. 10–11 (2008): 2083–105.

16. Victor Lavy and Edith Sand, “On the Origins of Gender Gaps in Human Capital: Short- and Long-Term Consequences of Teachers’ Biases,” *Journal of Public Economics* 167 (November 2018): 263–79; and Camille Terrier, “Giving a Little Help to Girls? Evidence on Grade Discrimination and its Effect on Students' Achievement,” CEP Discussion Papers dp1341 (London, UK: Centre for Economic Performance, London School of Economics and Political Science, 2015).

17. Mechtenberg, “Cheap Talk in the Classroom.”

18. Papageorge, Gershenson, and Kang, “Teacher Expectations Matter.”

19. Taking Algebra I in the eighth grade is generally considered to be early in North Carolina, whereas taking Algebra I in ninth grade is considered on time. This means that the students who take Algebra I in eighth grade are typically more advanced students in math.

20. This is commonly known as a “lag-score value-added model.”

21. Raj Chetty, John N. Friedman, and Jonah E. Rockoff, “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates,” *American Economic Review* 104, no. 9 (2014): 2593–632.

22. One point worth noting here, though, is that the lag scores we use to control for Algebra 1 classroom assignments are EOC math scores that are only collected in grades 7 and 8; therefore, our analysis is restricted to students who took Algebra I in grades 8 or 9. The majority (about 70 percent) of students take this course in grade 9, and another 25 percent take it in grade 8. That we remove students who take Algebra I “late” is worth keeping in mind when it comes to the generalizability of these results, even though we include about 95 percent of Algebra I takers.

23. Matthew Wiswall, “The Dynamics of Teacher Quality,” *Journal of Public Economics* 100 (April 2013): 61–78, doi:10.1016/j.jpubeco.2013.01.006; and Jonah E. Rockoff, Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger, “Can You Recognize an Effective Teacher When You Recruit One?” *Education Finance and Policy* 6, no. 1 (Winter 2011): 43–74, doi:10.1162/EDFP_a_00022.

24. The careful reader will note that this requires multiple Algebra I teachers in a given school-grade-year. About 16 percent of students in the analytic sample are therefore excluded because they took Algebra I in a school-year-grade that had only one Algebra I teacher. To verify that these students benefit from high grading standards, too, we reestimate the main model using school fixed effects in place of school-by-grade-by-year fixed effects. Of the students studied, 98 percent took Algebra I in schools that had multiple Algebra I teachers. Importantly, the school FE estimates are qualitatively similar to the main results, which suggests that the results are externally valid for the entire analytic sample. These results are reported in *Appendix B*, Table B-2.

25. See *Appendix A*, subsection A.1, for more about how teacher grading standards are identified.

26. Exact coefficient estimates and standard errors are reported in *Appendix B*, Table B-2. In that table, we also report a number of sensitivity analyses that verify the robustness of the general finding that high grading standards significantly improve student learning: They include exclusion of certain control variables, fixed effects models, and the use of alternative definitions of grading standards. In *Appendix B*, Table B-2, we also report the coefficient estimate from the parametric model that does not break teachers into "grading-standard" quartiles. This point estimate of 0.15 is also strongly statistically significant and indicates that a teacher whose B students score, on average, one standard deviation higher than the B students of another teacher in the same school-grade-year will improve their current students' score by about 15 percent of a test-score standard deviation, on average.

27. Mariesa A. Herrmann and Jonah E. Rockoff, "Worker Absence and Productivity: Evidence from Teaching," *Journal of Labor Economics* 30, no. 4 (October 2012): 749–82, doi:10.1086/666537.

28. The EOC exams in these courses were discontinued, so we focus on the four cohorts for which we have coverage on both exams. Column 5 of Table B-3 estimates the baseline Algebra I model for the 2006–09 cohorts to verify that these cohorts were not systematically different from the later cohorts included in the full sample. These estimates are quite similar to those reported in column 1 of Table B-2, suggesting that the early and late cohorts are not systematically different in how they contemporaneously respond to grading standards.

29. The fact that teachers' effects on student achievement fade out is well documented. One prominent study finds that about 75 percent of teachers' effects fade out within one year; see Brian A. Jacob, Lars Lefgren, and David P. Sims, "The Persistence of Teacher-Induced Learning," *Journal of Human Resources* 45, no. 4 (2010): 915–43, doi:10.3368/jhr.45.4.915.

30. See *Appendix B*, Table B-3.

31. Corresponding point estimates and standard errors are reported in *Appendix B*, Table B-4.

32. All high school seniors in North Carolina's public high schools respond to a survey including the question, "What are your plans after graduating high school?" (with small changes to the wording over the years), and the possible responses are "military," "work," "four-year public college/university," "four-year private college/university," "two-year public college/university," "two-year private college/university," and "unsure."

33. Like the effects on subsequent math scores, this finding, at least, does not contradict a causal interpretation of Finding 1 that grading standards improve student learning.

34. Corresponding point estimates and standard errors are reported in *Appendix B*, Table B-5.

35. A college or university is "selective" if Barron rated the institution as most or highly selective.

36. There is evidence of grade inflation (that is, lowered grading standards) in the postsecondary context, and this personal experience could change graduates' perceptions of grading scales. Other changes across cohorts could also be responsible for different grading practices across cohorts. Charles F. Eiszler, "College Students' Evaluations of Teaching and Grade Inflation," *Research in Higher Education* 43, no. 4 (2002): 483–501.

37. Differences in school facilities (for example, air conditioning) are another example. Such differences might boost EOC scores but not grades, and thus suburban schools that tend to have better infrastructure might have higher EOCs than urban and rural schools. This would result in higher measured grading standards in suburban schools. It is for these reasons that this analysis always makes within-school or within-school/grade/year comparisons when estimating the impact of grading standards: so that these concerns do not confound the results in Findings 1–4. Joshua Goodman, Michael Hurwitz, R. Jisung Park, and Jonathan Smith, "Heat and Learning," *American Economic Journal: Economic Policy* (forthcoming).

38. Multiple regression estimates reported in *Appendix B*, Table B-6, confirm these basic results and generally show that even when holding other teacher and school factors constant, teachers with graduate degrees continue to have significantly higher standards, as do teachers in middle schools, suburban schools, and schools serving a more affluent student population. Interestingly, the college selectivity and teacher-experience results are less pronounced when controlling for other characteristics and seem to drive standards in the middle of the distribution (quartiles 2 and 3).

39. Figlio and Lucas, "Do High Grading Standards Affect Student Performance?"

40. Eva Moskowitz, "'Holding Students to Higher Expectations' Is Harder than It Sounds," LinkedIn, August 28, 2018, https://www.linkedin.com/pulse/holding-students-high-expectations-harder-than-sounds-eva-moskowitz.

41. Tim Waker, "Teachers Divided over Controversial 'No Zero' Grading Policy," *NEA Today*, August 4, 2016, http://neatoday.org/2016/08/04/no-zero-policy-pro-con.

42. Will Dobbie and Roland G. Fryer, Jr., "The Impact of Voluntary Youth Service on Future Outcomes: Evidence from Teach For America," *B.E. Journal of Economic Analysis and Policy* 15, no. 3 (2015): 1031–65; and Ronald F. Ferguson, "Teachers' Perceptions and Expectations and the Black-White Test Score Gap," *Urban Education* 38, no. 4 (2003): 460–507, doi:10.1177/0042085903038004006.

43. "Mission Statement," Great Expectations, https://greatexpectations.org/about-us/mission/home-copy.

44. Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers I."

45. Wiswall, "The Dynamics of Teacher Quality."

46. Jeffrey M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data* (Cambridge, MA: MIT press, 2010).

47. This issue is a source of debate in value-added measures of teacher effectiveness, where the teacher effects themselves are more precisely estimated for teachers of larger classes. Cassandra M.Guarino, et al., "An Evaluation of Empirical Bayes's Estimation of Value-Added Teacher Performance Measures," *Journal of Educational and Behavioral Statistics* 40, no. 2 (2015): 190–222, doi:10.3102/1076998615574771.