High-Stakes Grades and Student Behavior

Ulrik Hvidman
Hans Henrik Sievertsen

ABSTRACT

High-stakes exams carry important consequences for the prospects of reaching university. This study examines whether the incentives associated with exam grades affect educational investments. Exploiting a reform-induced recoding of high school students' grade point averages, we identify the effect of high-stakes grades on student behavior. The results show that students who were downgraded by the recoding performed better on subsequent assessments. The increase in academic performance in high school translated into an increased likelihood of university enrollment. As the recoding did not convey information about actual performance, these results emphasize that incentives are important in understanding students' educational investments.

I. Introduction

How much to invest in education constitutes one of the most economically important decisions that people make. Policies that introduce student-based incentives are based on the notion that students do not exert sufficient effort. Previous

Ulrik Hvidman is Head of Research for Governance and Management at VIVE—The Danish Center for Social Science Research. Hans Henrik Sievertsen is a lecturer at the Department of Economics, University of Bristol.
The authors thank Simon Calmar Andersen, Sarah Bana, Kelly Bedard, Paul Bingley, Yuan Cao, Thomas S. Dee, Jens Dietrichson, Jacquie Dodd, Tine Louise Mundbjerg Eriksen, Colin Green, Fanny Landaud, Chang Lee, Alessandro Martinello, Asmus Leth Olsen, Petra Persson, Jesse Rothstein, Kristina Bakkær Simonsen, Dick Startz, Jenna Stearns, Julia Wirtz, and Miriam Wüst for helpful comments and suggestions. The paper also benefited from comments at the 2016 CEN workshop in Copenhagen, the 2016 SFI Advisory Research Board Conference in Copenhagen, the 2016 IWAEE workshop in Catanzaro, the 2016 EALE conference in Ghent, the 2016 DPSA workshop in Vejle, as well as from seminar participants at Aarhus University, University of Bristol, University of Copenhagen, and UC Santa Barbara. Sievertsen acknowledges financial support from the Danish Council for Independent Research through grant DFF: 4182-00200. This paper uses confidential data from Statistics Denmark. The data can be obtained by filing a request to Statistics Denmark through an authorized institution. Contact Sievertsen (h.h.sievertsen@bristol.ac.uk for assistance). For Stata, R, and Python scripts, see: https://github.com/hhsievertsen/high_stakes_grades_and_student_behavior.
[Submitted July 2018; accepted August 2019]; doi:10.3368/jhr.56.3.0718-9620R2
JEL Classification: I20, I21, I23, and J24
ISSN 0022-166X E-ISSN 1548-8004
Supplementary materials are freely available online at: http://uwpress.wisc.edu/journals/journals/jhr-supplementary.html

research has focused primarily on the impacts of monetary rewards (Angrist and Lavy 2009; Angrist, Lang, and Oreopoulos 2009; Fryer 2011; Bettinger 2012; Burgess, Metcalfe, and Sadoff 2016), or studied incentives for lower-performing students through grade retention policies or proficiency exams (Jacob 2005; Dee and Jacob 2006; Reardon et al. 2010).

In this study, we extend the research on student-based incentives by studying whether the high stakes associated with exam scores motivate students to exert more effort. The final high school examination constitutes one of the most important educational landmarks in many countries. Exam scores often carry major consequences for students' likelihood of reaching postsecondary schooling or being admitted to a (selective) university. In the United States and in Europe, university programs rely heavily on information about educational achievement as a screening tool in the admission process.[1] Given the substantial pecuniary and nonpecuniary returns to university education (Oreopoulos and Salvanes 2011), university admission constitutes a particularly strong incentive for students to invest effort and perform well in high school.

However, despite the importance of exam grades for admission to university, much remains unknown about students' behavioral responses to high-stakes exam scores. We use a novel identification strategy to isolate the behavioral response to a change in the incentives associated with high-stakes exam grades. We exploit a grading reform in Denmark that caused exogenous variation in high school students' grade point averages (GPAs) to provide credible estimates of the impact of high-stakes grades on subsequent educational investments. All students who were enrolled in their first year of high school during the implementation had their first-year exam grades recoded to the new scale based on a system provided by the Ministry of Education. Because they feed into the calculation of the final GPA, these first-year grades were high stakes.

In Denmark, high school performance almost entirely determines admission to postsecondary schooling, particularly at universities. Furthermore, access to specific university majors is based on high school performance, and many selective programs (for example, psychology or medicine) require high overall average scores for admission.[2] The consequence of the Danish grading reform was that two students with identical GPAs before the grading reform could have very different first-year GPAs after the reform. We exploit this reform-induced change in grades to identify the students' responses to a change in their GPAs. Although the reform changed students' grades, it did not provide any new information to the students about their academic performance or ability. Thus, any changes in effort investment in response to the grading reform should reflect the grades per se and the change in incentives.

1. In the United States, for example, many universities base their admission criteria on standardized tests (such as SAT scores), and some public universities offer scholarships on the basis of high school performance. In Scandinavian countries, such as Denmark, Norway, and Sweden, admission to postsecondary education, especially to universities, is determined predominantly by high school grade point average (GPA). Other examples of exams that are a prerequisite to matriculate at university include A-levels in the UK, Abitur in Germany, and Bagrut in Israel.

2. Nevertheless, a small share of postsecondary institutions determine their enrollment exclusively based on entry exams or on a combination of high school GPA cutoffs and entry exams. These deviations are typically observed for institutions that offer training in performing arts (for example, music or acting). Moreover, educational programs can decide to enroll a share of the students based on a combination of their GPA and other qualifications (for example, work experience). In 2008, 10 percent of enrollments were based on this scheme. Thus, high school grades are particularly important.

The identifying assumption is that there are no systematic differences between students whose grades were adjusted upward and downward due to the grading reform that would affect future outcomes. Under this assumption, the association between the reform-induced change in GPA and subsequent outcomes has a causal interpretation. We assess and discuss the threats to the identification, and provide evidence of the validity of the design. In particular, the reform-induced change in GPA appears not to be systematically related to observed outcome-relevant traits. Furthermore, falsification tests produce no evidence of performance effects for placebo cohorts that were not affected by the reform.

Using Danish administrative data on the full population of high school students who were affected by the reform, we find that students who experienced a negative shock to their first-year GPAs received better subsequent grades in their second and third years. Students who were downgraded one standard deviation scored 8 percent of a standard deviation higher in subsequent assessments. The number of grades that a student receives that count in the overall GPA increases gradually during the three years of high school. Due to the large number of post-reform grades, the students were, on average able to compensate for 73 percent of the impact of the grading reform on their overall GPA. However, we find considerable gender differences in how students reacted to the GPA shock. A decline in GPA led to performance improvements in subsequent assessments for both genders, but the effect was strongest for girls. Girls were able to almost fully compensate for the reform-induced shock by increasing their effort in their second- and third-year classes, whereas the response for boys was insufficient to offset the shock. We also find some evidence of differences in how students reacted to the GPA shock based on student ability and socioeconomic background.

The achievement effects in high school translate into postsecondary school enrollment and graduation. Students who had their high school GPA downgraded due to the reform had a higher probability of enrolling in and graduating from a university program after high school. The long-run effects on university attendance are driven by girls, who also responded strongest during the final two years of high school. While we find no effects on postsecondary enrollment for boys, we find that girls who received a negative reform-induced GPA change substituted away from shorter postsecondary education programs and into university programs. One explanation for the positive effects on university attendance for girls may be that an enhanced study effort in response to a GPA change increases the students' exposure to academic material, and therefore, their aspirations for further education.

One concern is that the performance effect reflects strategic behavior among teachers within the school rather than genuine improvements in student achievement. To compensate students, teachers could potentially inflate scores in response to the reform. Having access to internal and external evaluations, we are able to distinguish between teachers' grading behavior and true performance improvements. Although the response to the reform-induced GPA change is stronger for internal assessments (potentially due to teachers compensating unlucky students by inflating their post-reform grades), students who received a reform-induced GPA reduction also received better subsequent grades on national standardized exams that are externally evaluated. These findings confirm that the effects reflect actual improvements in student achievements.

These findings contribute to the literature on the role of incentives in education. An extensive literature has focused on test-based accountability and the impacts of teacher incentives (see, for example, Jackson 2010; Fryer 2013; Dee and Wyckoff 2015; Imberman and Lovenheim 2015; Lavy 2015; Deming et al. 2016) and school-based incentives (Dee and Jacob 2011; Jacob 2005; Reback 2008) on student performance. Moreover, recent studies have examined the impacts of monetary and nonmonetary incentives for students (see, for example, Fryer 2011; Bettinger 2012; Levitt et al. 2016; Angrist and Lavy 2009; Angrist, Lang, and Oreopoulos 2009). We add to this literature by providing evidence of the effects of incentives from high-stakes exams on student investments in education for a group of relatively high-achieving students in academic high schools who are at the margin of university enrollment. Interestingly, our finding that the effects of incentives are strongest for girls resembles the results found in Israel and the United States (Angrist and Lavy 2009; Angrist, Lang, and Oreopoulos 2009). Although we are unable to fully explain the gender differences, girls having higher levels of noncognitive skills—as suggested by Jacob (2002)—may be an important mechanism. For example, if girls are more forward-looking—or less prone to present bias—they may be more aware of their desired future educational paths and more willing to adjust their study effort in response to external shocks.

This study also relates to an emerging literature on test-takers' behavioral responses to feedback on educational performance. This literature has focused on how students may use test results to learn about their ability, as well as their return to investment in schooling (Stinebrickner and Stinebrickner 2012, 2014; Zafar 2011; Bandiera, Larcinese, and Rasul 2015).[3] Moreover, recent studies suggest that information on a student's relative rank within a cohort or classroom is particularly important (Murphy and Weinhardt 2014; Elsner and Isphording 2017).[4] Our findings demonstrate that the incentives that are associated with assessments can have important implications for students' human capital accumulation. Specifically, as the reform changed students' grades, but did not provide any new information about academic performance or ability, the change in grades should not affect the students' perceptions of their ability or self-confidence. Instead, any changes in effort investment in response to the grading reform reflect the change in incentives, not new information about how effort translates into academic performance. Thus, the findings suggest an important mechanism through

3. Related studies have examined how performance labels that do not carry official consequences for students affect their choices of postsecondary education (Papay, Murnane, and Willett 2016; Avery et al. 2018; Smith, Hurwitz, and Avery 2017). This literature finds that two students with almost identical raw scores make different educational choices because of discontinuities in the labeling. The present work is also related to the literature on how external factors affecting test outcomes may have long-run implications for individuals' human capital accumulation. Apperson, Bueno, and Sass (2016); Dee et al. (2019); and Diamond and Persson (2016) study how teacher manipulation of test results affects students' human capital accumulation, whereas Ebenstein, Lavy, and Roth (2016) study how variation in exam scores due to pollution exposure affects postsecondary educational attainment and earnings. In contrast to these studies, the students in our study observe the exogenous shock and know their original level. Thus, contrary to previous research, the behavioral responses should not reflect changes in students' self-confidence.

4. Murphy and Weinhardt (2014) and Elsner and Isphording (2017) show that students with the same absolute ability have different subsequent outcomes depending on the ability position relative to their peers. Although the effect of rank is consistent with a model in which students have a desire to perform well relative to others, different mechanisms may explain these findings. For example, Elsner and Isphording (2017) find that students with a higher relative rank have a higher perceived intelligence and higher career expectations, which might translate into more effort in their studies. See also Azmat et al. (2019), who propose a model that distinguishes between two theoretical mechanisms. In their model, students may respond to information because individuals have imperfect knowledge of their own ability or because they have inherently competitive motives.

which exam scores that carry official consequences for students affect their behavior—students may increase their effort in response to negative grades in order to make up for the shock.

The remainder of the paper is organized as follows. Section II discusses the theoretical expectations for behavioral responses to the incentives associated with high-stakes grades. Section III provides the institutional background about the Danish educational system and describes the grading reform. Section IV describes the administrative data. Section V discusses the identification strategy and the estimation. Section VI presents the results, and Section VII concludes.

II. Grades and Student Behavior: Incentives and Learning

Grades provide important feedback on high school students' likelihood of graduation and prospects of attaining postsecondary education or accessing a preferred, selective academic program. The main objective of Danish high school programs is to prepare students for postsecondary education. As most postsecondary programs assign no weight to student essays or recommendation letters in the admission process, high school grades are particularly important and may affect individuals' educational paths, as well as their long-term labor market outcomes.

It is useful to consider exactly how students may respond to changes in their GPA. While the students may use the GPA as a signal about their own ability, it may also affect their incentives because universities use the GPA in the admission process. Thus, one can distinguish between two different mechanisms through which grades can affect behavior: *incentives* and *learning*.

First, due to the high stakes associated with grades, high school students have an incentive to work hard in order to earn a high GPA. Because the final high school GPA is the average across grades given in all three years, a (reform-induced) change in first-year grades causes a change in the expected overall GPA, given the study effort. Therefore, students who experience a negative reform-induced change in their GPA may be motivated to increase their study effort to improve their performance in subsequent assessments and offset the negative reform-induced change.[5] In contrast, if the students experience a positive reform-induced change and their grades are improved (which is equivalent to the threshold for admission to the preferred university major being lowered), the students may respond by reducing their effort. That is, for a given level of ability, an individual is less motivated to put in the effort because the chances of university admission, given effort and ability, have increased. Although the aim of Danish high schools is to prepare students for postsecondary education, not all high school students intend to continue in further education. Thus, the importance of the incentive component of high school grades may vary across students, depending on how they weight academic output and admission chances to educational programs. The incentives

5. We assume that the first-year grades (that is, pre-reform GPA) are the outcome of the student selecting a level of study effort to maximize the chances of university enrollment while trading off the costs of the study effort, such as psychic costs (for example, stress), direct pecuniary costs (for example, study material such as books), or indirect pecuniary costs (for example, foregone earnings in the labor market).

associated with the final exam are particularly strong for students who are determined to go to university or get into a selective university major for which they need a certain GPA.

Second, the GPA also provides information that students may use to learn about how well they have mastered the taught material. Thus, although negative events (for example, exam failure or a low grade) may provide an incentive for students to boost their efforts, some literature suggests that low grades could also have a discouraging effect. According to this line of research, students may have imperfect information about their ability (and the production function), and therefore, about how their effort translates into performance (Bandiera, Larcinese, and Rasul 2015; Azmat et al. 2019). If students perceive grades as informational about ability and about how their effort translates into grades, this could affect the students' future choices of study effort.[6] To the extent that students perceive the grade as informational about innate ability, they may link the signal to perceived competence and self-confidence (Murphy and Weinhardt 2014; Diamond and Persson 2016). Therefore a positive performance signal may make the students want to increase their study effort, as they realize the higher payoff of their effort. Whereas a positive performance signal may increase motivation to the extent that it leads the recipient to feel competent and efficacious, students who receive negative feedback may instead interpret it as implying low self-worth, resulting in lower levels of motivation and effort. Thus, according to this alternative notion, a negative shock in grades may not lead students to increase their workload, but instead can have the opposite effect.

In sum, there are conceivable arguments supporting the hypothesis that changes in grades can affect subsequent student outcomes. These effects could be driven either by students valuing high grades due to their instrumental value or by changes in students' self-confidence. That the students actually perceive the signal as informative about their academic ability is a critical condition for the existence of the learning mechanism. Since Danish high school students observe the reason behind the change in their GPA (the grading reform), and know their initial GPA, the reform-induced change in grades should not affect student behavior through learning about the return to study effort. Although the reform-induced GPA change that we study does not contain information about ability (or about the return to effort), the change affects the chances of postsecondary admission and the incentives to which the students are exposed. In the following section, we describe the details of the empirical setting that enable us to isolate the effect of the incentives associated with grades.

III. Background

A. Secondary Schooling in Denmark

In Denmark, compulsory education begins in August of the calendar year the child turns six and ends after ten years of schooling, that is, Grades 0 (kindergarten) through 9. Having completed compulsory schooling, students may continue to a three-year high school program (Grades 10–12), enroll in vocational training, or enter the labor market.

6. This reasoning is built on the notion that effort and ability are complements in producing academic output. If students perceive grades as informational about ability, the complementarity between ability and effort implies that the students learn about how their effort translates into grades, which affects their future choices of study effort.

Among the 65,000 children who left compulsory schooling in 2005 (the cohort analyzed in this paper), 52 percent continued in high school and 25 percent in vocational training. High school offers different programs: the general upper secondary education program (called "STX"), the higher commercial examination program (called "HHX"), and the higher technical examination program (called "HTX").[7]

Although the programs have slightly different curricula, the main objective of all high school programs is to prepare students for higher education, and all high school programs provide equal access to higher education. The high school programs consist of a wide range of courses on three levels. Level A, the most advanced course level, typically covers all three years. Level B typically covers two years of high school. Level C is typically a one-year course. All students are required to take a number of mandatory courses (for example, A-level Danish), as well as a minimum number of A-level courses, and within each program, students may choose between different tracks (that is, major areas of study).[8] Students request a preferred track when they apply for high school, but are allowed to change tracks within the first six months of their first year. Apart from the mandatory courses and the track-specific courses, students may choose a few optional courses in their second and third years.

Students receive grades during all three years of high school, and the final overall GPA score is calculated as the simple unweighted average of two intermediate average scores. The first is a weighted average of grades for the annual national exams, administered by the Ministry of Education, with independent examiners (that is external to the school). The second intermediate score is a weighted average of classroom grades, determined via an internal assessment by the students' teacher.

Postsecondary schooling is free, and students receive a monthly student grant to pay for living expenses for up to six years of postsecondary schooling. After completing high school, all students who wish to enroll in postsecondary schooling apply through a centralized system with a list of prioritized educational programs. The programs set the number of available slots, N, and the course requirements (for example, Economics at the University of Copenhagen requires A-level courses in mathematics and Danish and B-level courses in history and English). All students who fulfill the course requirements for the prioritized program are ranked according to their high school GPA, and the first N students are offered a place in the program.[9] Thus, the high school GPA is particularly important for students who wish to continue in postsecondary schooling. Moreover, as first-, second-, and third-year grades count in the final overall GPA, stakes are high during all three years.

7. In addition to these three high school types, there are one- and two-year high school programs with specific admission requirements (called "HF"). Whereas students have to enroll in STX, HHX, and HTX programs no later than one year after they finish compulsory schooling, there are no age requirements for HF students, who therefore tend to be older than students in the other programs. In this study, we focus on the three-year programs (STX, HHX, and HTX) because they are very similar in structure, length, and prerequisites, and the implementation of the grading reform was different for the HF programs. The included programs cover about 90 percent of all high school students in Denmark in 2008.

8. As of 2017, the number of tracks (as well as their individual content) is decided centrally by the government. The STX program, for example, has 18 tracks (for example, a math track that consists of A-level math, A-level physics, and B-level chemistry). However, at the time of the implementation of the grading reform, each school decided the number and content of the tracks at their school, which resulted in some variation across schools.

9. For details about the university admission process in Denmark, see Humlum et al. (2014).

Table 1
Implementation of the Grading Reform across High School Cohorts

		Grading Scale		
Enrolled	Graduated	Year 1	Year 2	Year 3
August 2004	June 2007	13	13	13
August 2005	June 2008	13	7	7
August 2006	June 2009	7	7	7

Notes: 13 = 13 scale; 7 = 7-point scale. Year 1, Year 2, and Year 3 refer to the high school year.

B. The Danish Grading Reform of 2007

Until April 2007, student performance in the Danish school system, from lower secondary schooling to postsecondary schooling, was evaluated on a scale from 0 to 13 (called the "13 scale"). In November 2004, the Commission for Examining the Danish Grading Scale recommended the introduction of a new seven-point grading scale from –3 to 12 (called the "7-point scale"). In early 2006, the government decided to introduce the new grading system, and in 2007, the 7-point grading scale replaced the 13 scale grading system.

Table 1 shows how the reform affected students enrolled in a high school program during the implementation. Students who enrolled in August 2004 and graduated in 2007 had all their exams assessed on the old scale. In contrast, the cohort that enrolled in 2006 only received grades on the new scale. For students who enrolled in August 2005 and graduated in 2008, coursework that was completed in the school year 2005–2006 was assessed on the 13 scale, and coursework completed in the 2006–2007 and 2007–2008 school years was graded on the new scale. For this cohort, each grade obtained in their first year in accordance with the old scale were subsequently converted to a grade on the new scale based on a scheme provided by the Ministry of Education.

Figure 1 shows the timeline for the 2008 graduating cohort that was affected by the grade recoding. After the students completed their first year of high school in the summer of 2006, all their first-year grades were converted from the old scale to the new scale. Thus, the students' final overall high school GPA was calculated based entirely on grades on the 7-point scale—and only the post-recoding grades were shown on the high school diploma (Online Appendix Figure 1 shows a high school diploma for a student from the treated cohort). Until 2006, admission to postsecondary schooling was determined on the basis of the final high school GPA on the 13 scale. As of 2007, all students applying for postsecondary education were ranked according to their GPA on the new 7-point grading scale.[10]

10. Therefore students who had received all their grades on the 13 scale had their GPA mapped to a GPA on the 7-point scale based on a system provided by the Ministry for Education. Because this recoding of the overall GPA was monotonic, it did not affect the chances of postsecondary enrollment for the students within the cohorts. Because students with identical high school GPAs on the 13 scale had equal chances of enrollment

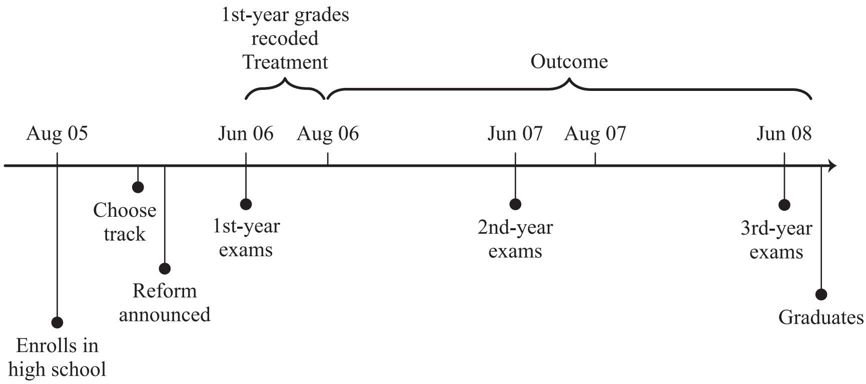

Figure 1

Timeline: Assessment and Recoding of Grades for Students Who Enrolled in High School in 2005 and Graduated in 2008

Table 2 presents the recoding system provided by the Ministry of Education. The first two columns describe the mapping system from the 13 scale to the 7-point scale. There are two important sources of noise in the mapping process. First, because the new grading scale has fewer grades (seven compared to ten), pairs of grades on the old scale were collapsed to a single new grade. Consider, for example, a student who had only 8s on the old scale and another student who had only 9s. Although the latter had higher grades before the reform, the two students would have identical grades (that is, 7s) after the recoding to the new scale.

Second, the distance between the old and new grades varies along the scale. For example, a 5 on the old scale is penalized heavily as it is transformed into a 0 (that is, the difference is five points), whereas a 10 on the old scale is not punished (that is the difference is zero points). Thus, two students with identical pre-recoding GPAs could have very different post-recoding GPAs, because grades were affected differently.

As a result, depending on the composition of grades, students were either down- or upgraded relative to their peers. For example, a student with grades 5, 5, 6, 11, and 13 on the old scale would have their GPA transformed from 8.0 to 5.2, while a student with grades 3, 5, 10, 11, and 11 would have their GPA transformed from 8.0 to 6.8. The collapsing of grades and the varying distance to the new grades caused noise in individual students' GPAs. The grading reform also affected the overall level and distribution of grades. Thus, most students had their GPA downgraded in absolute terms. Online Appendix Figure 2 shows the high school GPA distribution for the cohorts graduating in the years 2003–2013. After the reform, the density in the center of the distribution is lower, and the tails are fatter. The level shift should not affect students' incentives, as the GPA cutoff levels were adjusted mechanically.

after the recoding, everything else equal, we do not exploit this mapping of the overall GPA. Instead, as we explain, we exploit the implementation of the grading scale for the one high school cohort for which individual grades were transformed.

Table 2
Implementation of the Grading Reform across High School Cohorts

Old 13 Scale	New 7-Point Scale	ECTS	Description
00	–3	F	For a performance that is unacceptable in all respects.
03 05	0	F^+	For a performance that does not meet the minimum requirements for acceptance.
6	2	E	For a performance meeting only the minimum requirements for acceptance.
7	4	D	For a fair performance displaying some command of the relevant material but also some major weaknesses.
8 9	7	C	For a good performance displaying good command of the relevant material but also some weaknesses.
10	10	B	For a very good performance displaying a high level of command of most aspects of the relevant material, with only minor weaknesses.
11 13	12	A	For an excellent performance displaying a high level of command of all aspects of the relevant material, with no or only a few minor weaknesses.

Source: The Danish Ministry of Science, Innovation and Higher Education.
Notes: ECTS is the grading system defined by the European Commission. The passing threshold is 6 for the old scale and 2 for the new scale.

The number of grades given in the first year of high school depends on the specific high school track chosen by the student. High-stakes exam grades that count in the final high school GPA are given once a student finishes a class. As students take some courses for more than one year (for example, Danish), they do not get their final grade until they finish the subject. Thus, students typically receive two to five grades in the first year (pre-recoding) and around 30 grades in the second and third years (as shown in Online Appendix Figure 3). Although more grades imply that more grade combinations can cause a specific pre-recoding GPA, the link between the number of grades given on the old scale and the potential variation in the post-recoding GPA is not trivial (as Online Appendix Figures A.4a and A.4b show).

IV. Data

For the analyses, we use administrative data provided by Statistics Denmark that include all students who graduated from a three-year high school program in 2008. As the registers contain information only on individuals who completed high

Table 3
Variable Descriptives

	Mean	SD	P10	P50	P90
Age at high school enrollment	16.66	0.67	15.84	16.61	17.46
Female	0.56	0.50	0.00	1.00	1.00
Non-Western Origin	0.05	0.21	0.00	0.00	0.00
Ninth-grade GPA	0.27	0.85	–0.79	0.26	1.40
Parents' years of schooling	14.63	2.01	12.21	14.25	17.38
Parents' income (€1,000)	35.80	24.55	22.50	33.06	49.59
Grades recoded	3.41	2.71	1.00	3.00	5.00
Grades given after recoding	29.03	3.24	26.00	29.00	32.00
Recoding residual	0.00	0.26	–0.37	–0.01	0.38

Notes: Parental characteristics are measured in the calendar year before students' high school enrollment. All monetary values are converted to the 2015 price level using the consumer price index. P10, P50, and P90 refer to the 10th, 50th, and 90th pseudo-percentiles, respectively. The pseudo-percentile is the mean of the actual percentile and the two values above and below the percentile. Recoding residuals refers to the residuals from regressing the first-year GPA after the recoding on a second-order polynomial of the first-year GPA before the recoding.

school, we do not observe grades for students who dropped out of high school.[11] Furthermore, we exclude 950 students who were not graded on both grading scales, as those students are unaffected by the change in the grading system. The data contain information on courses and exam-specific grades. The high school data are merged with administrative data from Statistics Denmark on the students' background (gender, age, and origin) and with school records on middle school GPA (that is, the exit exams at the end of ninth grade). The final sample consists of 26,760 students.[12] For each student, we record parental characteristics the year before the student enters high school using the income and education registries from Statistics Denmark. We construct a variable for the average parental net income and a variable for the average years of parental schooling. We also link each student to the education registries to measure their postsecondary schooling outcomes.

Summary statistics for key variables are provided in Table 3. There are more girls than boys in the sample. The students are on average 16.7 years old at enrollment, and 5 percent are of non-Western origin. In line with expectations, there is evidence of positive selection into high school. High school students have a middle school GPA that is on average 0.3 standard deviations above the mean for their ninth-grade cohort. Students received, on average, 3.4 grades in their first year and 29.0 grades in their second and third years.

11. As we discuss in Section VI, the pattern in dropouts appears to be unrelated to the grading reform.
12. We exclude 695 observations due to missing middle school GPA and three observations due to incomplete high school records. The most likely reason for a missing middle school GPA is that the students completed lower-secondary schooling outside of Denmark. No further data restrictions are imposed. The final sample includes 94 percent of the initial population. Including students with missing observations yields qualitatively similar results.

V. Identification and Estimation

A. Empirical Strategy

The grading reform constitutes a policy change that allows us to examine whether an observed exogenous change in the high-stakes GPA affects students' performance. To illustrate the change caused by the reform, Panel A of Figure 2 shows the relationship between the students' first-year GPA before and after the recoding of their grades, where both variables have been standardized.[13] As Panel A of Figure 2 illustrates, there is substantial variation in the post-recoding GPA for any given level of pre-recoding GPA.[14] The dashed line shows the quadratic fit, which captures the relationship fairly well. To further illustrate the identifying variation, Panel B plots a histogram of an individual's reform-induced GPA shock (that is, the residuals from regressing the first-year GPA after the recoding on the first-year GPA *before* the recoding).

To assess the impact of the recoding of grades on subsequent performance, we estimate the following equation using ordinary least squares:

$$(1) \quad Y_{i,s} = \beta_0 + \beta_1 GPA7_i + f(GPA13_i) + \boldsymbol{\delta}'\mathbf{X}_i + \boldsymbol{\eta}_s + \epsilon_{i,s}$$

where $Y_{i,s}$ is the grade point average of grades given in the Years 2 and 3 for student i in school s (that is, after the grade recoding), $GPA7$ is grade point average of first-year grades after the recoding to the 7-point scale, and $GPA13_i$ is the grade point average of the original first-year grades on the 13 scale before the recoding. We standardize Y_i, $GPA7$, and $GPA13_i$ to a mean of zero and a standard deviation of one. In the main analysis, we present results using a second-order polynomial for the functional form, $f()$, but as we show, the conclusions are not sensitive to the choice of functional form. $\boldsymbol{\eta}_s$ is a vector of school fixed effects, and $\mathbf{X}_i$ is a vector of individual specific covariates, including gender, an indicator for non-Western origin,[15] indicators for being a first- or second-generation immigrant, age, middle school GPA, average parental income, average parental years of schooling, and indicators for whether parents are observed in the data. We include these covariates to obtain more precise estimates of the impact of the grade shock because these covariates are highly predictive of the students' subsequent educational outcomes. The standard errors are clustered at the school level.

B. Strategic Responses to the Implementation of the Grading Scale in Danish High Schools?

The key aim of this work is to study how the change in grades induced by the recoding of first-year grades affects subsequent behavior. The causal interpretation of the GPA

13. Due to confidentiality issues, we cannot show cells with fewer than three observations. However, the regression analyses are based on all observations

14. Online Appendix Figure A.5 shows the same relationship as Figure 2a, but for unstandardized GPAs. Figure A.5 demonstrates the magnitude of the variation induced by the reform. For example, if we compare two students with a pre-recoding GPA of 8, one could end up with a post-recoding GPA of about 7, whereas the other could end up with a post-recoding GPA of about 5.

15. We use Statistics Denmark classification of countries as Western and non-Western, where Western countries refer to all 28 EU countries, Andorra, Iceland, Liechtenstein, Monaco, Norway, San Marino, Switzerland, Vatican State, Canada, USA, Australia, and New Zealand.

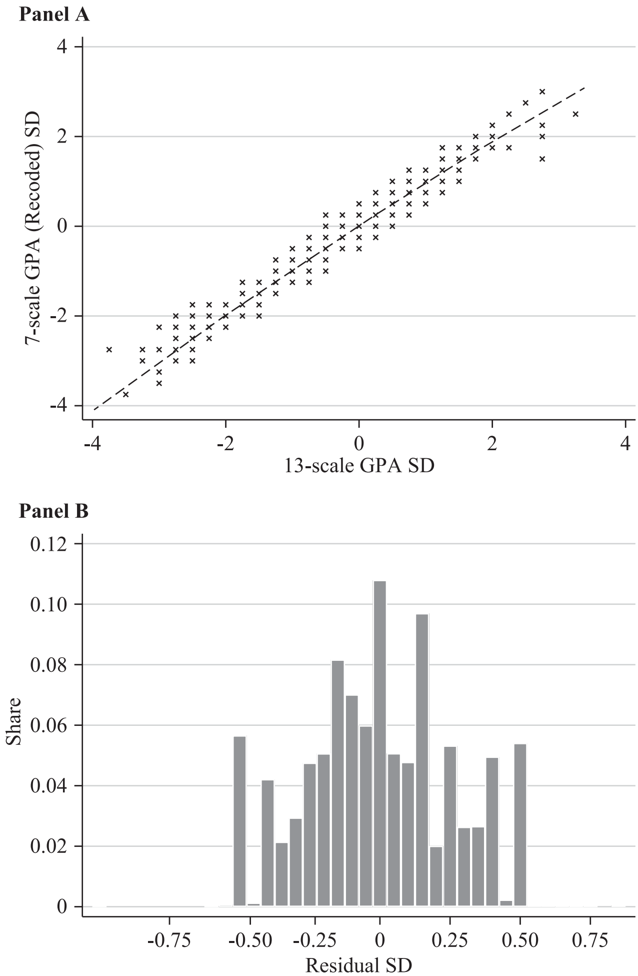

Figure 2

Pre- and Post-Recoding GPA of First-Year Grades

Notes: Only combinations with at least three observations are shown. The fitted line in Panel A is a second-order polynomial. The residuals in Panel B are based on a specification without covariates and fixed effects using a second-order polynomial. The GPA is standardized to a mean of zero and a standard deviation of one.

shock is based on the assumption that the shock is unrelated to student characteristics that are related to the outcome of interest. The identifying assumption is that the effects of other policies that were implemented simultaneously as the grading reform are unrelated to the reform-induced shock from the grading reform. Importantly, as other educational reforms were implemented nationwide, they should not affect students whose GPAs were recoded downward and upward differently—and therefore, not confound the analyses.[16]

However, there are theoretically reasonable ways in which students (and their teachers) could respond to the introduction of the grading reform that would complicate the identification of the effects of the shock in grades. As the government announced the introduction of the new grading system in early 2006, one concern is that high school students responded to this information by changing their choice of tracks in ways that were more advantageous. As the number and composition of grades given in the first year of high school depend on the specific high school track chosen, risk-averse students might have tried to avoid the recoding noise by selecting course tracks that reduced the number of grades that were transformed. Importantly, this would only constitute a problem for identification if groups of students systematically selected tracks to avoid courses for which the students had private information about the risk of receiving a grade that was penalized heavily.

Institutional knowledge and empirical evidence suggest that such strategic behavior was limited. Online Appendix Figure 6 shows the Google search term popularity for the new scale for the period 2005–2009. The search term "7-trins skala" (English: "7-point scale") gained popularity after July 1, 2006, and it maintained a relatively constant level over the remaining period. Although we cannot rule out that students knew about the new scale before the first-year exams (that is, before July 1, 2006), this Google search trend at least suggests that the new grading scale was discussed primarily after its implementation. Importantly, even for well-informed students, challenges and barriers are present. As tracks are selected within the first six months of the first year of high school, this choice is made before the students know about their final first-year grades. Thus, students cannot change tracks after their pre-recoding grades are disclosed.

Apart from the track-specific courses, students can choose a few courses in their second and third years. Therefore, students could respond to the grade shock by taking more (or less) advanced courses. To assess whether students change their course choices in response to the grade shock, Table 4 presents results from models where we regress the number of advanced courses on the grade shock. The students end up with, on average, five A-level courses (Danish and history are mandatory A-levels for all high school types). There is no evidence that the grade shock is related to the level of the courses that the students choose (that is, the number of A- and B-level courses), nor is there any evidence that students were less likely to take A-level mathematics as a result of a negative grade shock. That the grade shock did not affect whether the students take A-level mathematics—which is typically perceived to be one of the most challenging courses—provides suggestive evidence that students did not take less challenging courses because of a negative performance shock. Thus, statistical evidence also speaks to the concern that students changed courses in response to the reform.

16. Specifically, two other significant educational policies were implemented at the same time as the grading reform. First, in 2005 a high school reform involved several changes in the curriculum and the structure of high school programs. The first cohort affected by the reform enrolled in 2005 (that is, graduated in summer 2008). As all students in the data are affected by the reform, it should not confound the results. Second, during the period 2007–2011, a nationwide policy was implemented in the STX program that introduced a mechanical funding system based on the enrollment and the number of students graduating from high school. The reform was introduced simultaneously in all STX schools.

Table 4
Regression Results for Course Selection

	B-Levels (1)	A-Levels (2)	A-Level Math (3)
Recoded GPA	–0.009	0.000	0.006
	(0.019)	(0.011)	(0.015)
Mean of dependent variable	3.49	4.99	0.41
Observations	26,759	26,759	26,759
Clusters	209	209	209
R^2	0.30	0.37	0.18

Notes: The table shows point estimates and standard errors for β_1 in Equation 1, estimated with ordinary least squares. The dependent variable is denoted in the column header. The GPA is standardized to have a mean of zero and a unit standard deviation. We control for the first-year GPA before the recoding using a second-order polynomial. The covariates included are age at high school entry, gender, ninth-grade GPA (standardized), origin (indicator for non-Western origin), parental education (years of completed education, average across parents), income (disposable income, average across parents), and the number of nonmissing parental education and income observations (indicators). All parental variables are measured in the calendar year before the focal individual enrolled in high school. Standard errors clustered on the school level in parentheses.

Dropping out of high school—or switching to another school—is another potential response to the grade shock. As the data contain information only on individuals who completed high school, the design would not provide valid causal inferences if such dropout patterns are related to student outcomes. To assess this threat, we describe the dropout patterns across cohorts in Online Appendix Figure 9. The figure shows that the number of students who dropped out increased considerably for the cohort that enrolled in 2005. Importantly, however, the graph also shows that the increase in dropouts happened during the first year, before the grade shock occurred, and that there were no changes in dropout levels in the second and third years. The change in dropout patterns—with more students dropping out during their first year—is likely due to the comprehensive high school reform that was implemented in 2005. For example, before this reform, the STX program consisted of two tracks: a "Math/Science track" and a "Language track" that students applied to before enrollment. In 2005, a number of tracks replaced the two-track system, and students had to choose their track within six months of enrollment. In sum, as the dropouts mainly increased before the first-year grades were revealed for students, and grades were transformed (and because the increase is a level shift rather than a spike), the increase appears unrelated to the grading reform.

Moreover, if selected groups of students dropped out because of a specific grade shock, we would expect the grade shock to be related to student-specific characteristics. Thus, to further assess the identifying assumption, we study whether the GPA change is related to covariates that are highly predictive of student achievement. We estimate a series of regressions where we use each of the covariates as the dependent variable. Each entry in Table 5 represents an estimate from a regression of the GPA shock on a demographic characteristic. All point estimates are small and not statistically significant. The absence of signs that the change in GPA caused by the recoding process is

Table 5
Regression Results for Balance of Covariates across Treatment

	$\hat{Y}$ (1)	Ninth-Grade GPA (2)	Female (3)	Parental Education (4)	Parental Income (5)
Recoded GPA	–0.004	0.001	–0.016	–0.064	–0.908
	(0.013)	(0.017)	(0.012)	(0.048)	(0.573)
Mean of dependent variable	0.03	0.27	0.56	14.63	35.80
Observations	25,011	26,759	26,759	25,042	26,658
Clusters	209	209	209	209	209
R^2	0.40	0.39	0.07	0.15	0.05

Notes: The table shows point estimates and standard errors for β_1 in Equation 1, estimated without covariates using ordinary least squares. The top row indicates the dependent variables. $\hat{Y}$ is the predicted value from regressing the GPA given after recoding on all covariates included (see the notes for Table 4). Standard errors clustered on the school level in parentheses.

related to observable characteristics strengthens the conclusion that the reform did not lead certain groups of students to drop out and affirms the validity of the design.[17]

Another concern is that teachers adjusted their grading behavior before the reform. The internal grading procedure for the classroom grades leaves scope for teachers to manipulate the pre-recoding grades. For example, to help students, teachers could avoid the grades in the first-year exam that were penalized the most in the recoding system. Figure A.7a in the Online Appendix compares the distribution of first-year grades in the affected cohort with the distribution of grades in the previous cohorts that were not affected by the grading reform. One complication of this analysis is that the high school reform in 2005 affected the curriculum of the high school tracks and the composition of first-year coursework. Given that the grading pattern varies across subjects,[18] some changes in the grade distribution are expected as a result of the high school reform.

Although there are some changes in the grade distribution in 2005, we find no evidence that teachers tried to help students by avoiding the first-year grades that were penalized the most. If the grading reform led teachers to avoid these grades, we would expect fewer 5s, 7s, and 9s and more 6s, 8s, and 10s. However, the treated cohort has more 5s, 6s, and 7s, but fewer 9s and 10s.[19]

17. Another potential source of selection bias is if the recoding led to grade repetition among certain students. Online Appendix Table A.1 shows that the grade shock was not associated with grade repetition in the first year of high school.
18. For example, the grades given in mathematics are usually lower than in other subjects.
19. Another way teachers could adjust their grading would be to set the first-year grades by taking into account the subsequent recalculation of the grades. To study if this is the case, Online Appendix Figure A.7(b) plots the distribution of grades across cohorts where we recalculate first-year grades for the pre-reform cohorts (that is, the three cohorts before the one affected by the reform) as if the grading reform had been implemented. As Figure A.7(b) shows, the changes in the grading pattern that happened in 2005 were modest relative to the changes that occurred after the reform was implemented.

Although we cannot rule out that other types of teacher adjustments took place, the lack of evidence that teachers inflated less-penalized grades is reassuring. Moreover, this would only constitute a challenge to the identification if teachers' propensity to manipulate a student's grade was associated with other student-specific characteristics (for example, student ability or behavior). As previously shown, the change in the GPA caused by the recoding process was not related to observable characteristics, which suggests that the reform did not lead teachers to manipulate certain students' scores.

VI. Results

A. Effect of Shock in First-Year Grades on Subsequent Student Performance in High School

We begin by investigating the effect of the reform-induced change in grades on student performance in the second and third years of high school. Table 6 shows the results from estimating the effect of a change in the first-year GPA on subsequent grades. The dependent variable is the average of the student's grades in the second and third years of high school. Column 1 shows the main effect for the full sample. Students who are downgraded due to the recoding of the first-year grades perform better in the second and third years of high school relative to their peers. The coefficient is precisely estimated and shows that high school students who are downgraded by one standard deviation on their first-year GPA perform 8 percent of a standard deviation better in subsequent grades.[20]

Columns 2 and 3 in Table 6 show results from subsample regressions where the sample is split according to the median middle school GPA. Although we find a small and imprecisely estimated negative effect for the subsample of students with a middle school GPA below the median, we find a larger and statistically significant effect for students with a middle school GPA above the median. The fact that students with an above-median middle school GPA perform better if they receive a negative shock may suggest that high-performing students care in particular about their high school grades. Most admission cutoffs for universities are in the upper part of the high school GPA distribution. Thus, although low-performing students have an incentive to ensure that they end up with a GPA above the proficiency threshold, high achievers appear to be more responsive to a change in their GPA. Moreover, the results presented in Columns 4 and 5 of Table 6 show that although the effect of a negative shock is positive for boys and girls, it is largest for female students. Finally, Columns 6 and 7 show that there is no clear difference in response by parental background.

To get a sense of the magnitude of this response, consider the impact of the grading reform for a student who experienced a negative GPA shock of −0.37 SD (corresponding

20. Online Appendix Table A.2 reports the estimates of the grade point average of the original first-year grades before the recoding, GPA13, and GPA13-2. As expected, there is a strong positive association between first-year grades (that is, GPA13) and second- and third-year grades. Although part of this relationship may be due to the learning effect, a major concern is that the students who receive good grades in the first year are likely to be different in unobserved characteristics from the students who do not, and that these differences may be correlated with performance—a bias that is likely to persist even after we condition on the detailed data from the Danish registers.

Table 6
Regression Results for the Effect of a GPA Shock on Subsequent Grades. Dependent Variable: Grades Given after Recoding (Standardized)

		Ninth-Grade GPA		Gender		Parental Educ.	
	Main (1)	Low (2)	High (3)	Boys (4)	Girls (5)	Low (6)	High (7)
Recoded GPA	–0.079	–0.031	–0.096	–0.041	–0.106	–0.062	–0.091
	(0.017)	(0.025)	(0.021)	(0.027)	(0.021)	(0.024)	(0.022)
p-value		0.03		0.04		0.32	
Mean of dependent variable	–0.00	–0.54	0.53	–0.09	0.07	–0.16	0.18
Fraction recoded	0.10	0.11	0.10	0.10	0.10	0.10	0.1
Observations	26,759	13,218	13,538	11,677	15,080	11,414	13,628
Clusters	209	208	207	207	208	209	208
R^2	0.60	0.39	0.51	0.59	0.62	0.57	0.61

Notes: The table shows point estimates and standard errors for β_1 in Equation 1, estimated with ordinary least squares. The GPA is standardized to have a mean of zero and a unit standard deviation. We control for the first-year GPA before recoding using a second-order polynomial. The covariates included are age at high school entry, gender, ninth-grade GPA (standardized) origin (indicator for non-Western origin), parental education (years of completed education, average across parents), income (disposable income, average across parents), and the number of nonmissing parental education and income observations (indicators). All parental variables are measured in the calendar year before the focal individual enrolled in high school. The ninth-grade GPA indicates that the sample is split by the median of the focal individual's middle school GPA. Parents with high education are parents with an average length of education (years of schooling) above the median (observations with no information on parental years of schooling are not included). p-value provides p-values for the null hypothesis that the point estimates are the same for the two respective subsamples. Standard errors clustered on the school level in parentheses.

to the tenth percentile). On average, 10 percent of all grades were affected by the recoding. Thus, without any behavioral response, the final high school GPA would be 0.037 SD lower due to the grading reform. With the identified behavioral response, the impact of the reform on the overall GPA was –0.01 SD.[21] Therefore, the students were, on average, able to compensate for 73 percent of the impact of the grading reform on the overall GPA.

In the main analysis, we impose a linear functional form on the relationship between the reform-induced GPA shock and the subsequent grades. To test whether the effects are nonlinear (for example, asymmetric effects for positive and negative shocks), we examine this relationship in a nonrestrictive and visual manner. Figure 1 shows the parametric specification (Equation 1) as a solid line and plots a histogram that shows

21. At the tenth percentile, the post-recoding grades were $-0.37 \times -0.08 = 0.03$ SD higher due to the behavioral response, which affected 90 percent of the GPA, leading to a positive impact on the overall GPA of $0.03 \times 0.9 = 0.027$. To calculate the overall impact of the reform on the GPA at the tenth percentile, we add the direct mechanical effect of $-0.37 \times 0.1 = -0.037$ SD.

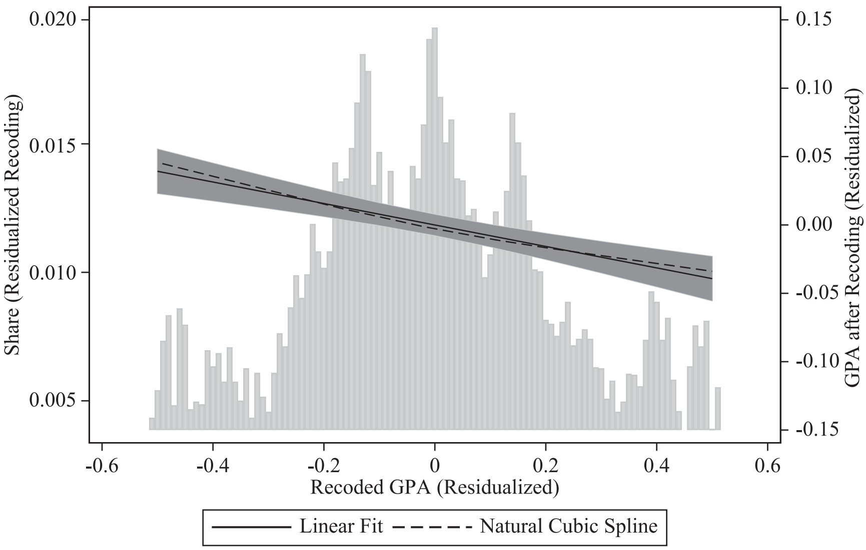

Figure 3

Relationship between the Reform-Induced GPA Shock and Subsequent Grades

Notes: The graph shows the relationship between the residuals from regressing the recoded GPA and the GPA for subsequent grades on all covariates, a second-order polynomial in the first-year grades before the recoding, and school fixed effects. The dashed line shows the natural cubic spline based on three knots. The solid line shows the linear fit using ordinary least squares (corresponding to the estimated relationships presented in Table 6). The gray shaded area shows the 95 percent confidence interval obtained with the delta method. The gray bars show the fraction of the observations (in percent). The graph excludes the bottom and top 1 percent of the residuals from the recoded GPA, but the natural cubic spline and the global linear regression lines are fitted on the full sample. Cells based on less than four observations are not shown.

the distribution of the grade shock. The gradient of the line resembles the negative coefficient from Table 6. The dashed line in Figure 3 shows a nonparametric specification (that is, a natural cubic spline) of the relationship between the change in grades and subsequent performance.[22] The linear specification fits the nonparametric pattern fairly well for the range of the GPA shock covering most observations.

To assess the sensitivity of the results presented in Table 6, we conducted a series of robustness checks. Table 7 presents results for various specifications. As a baseline, Row 1 shows the result from the main specification reported in Table 6. Row 2 presents the results from estimating a model without covariates. Although slightly larger, the point estimate is close to the main result that includes the full set of covariates. In the main analyses we condition on a second-order polynomial of the pre-recoding GPA. In Rows 3 and 4 we show results from estimating the model in Equation 1 using a linear

22. Online Appendix Figure A.8 shows the results from using a local linear regression. Results are qualitatively the same.

Table 7

Alternative Specifications. Dependent Variable: Grades Given after Recoding (Standardized)

	Specification	Coefficient	SE
(1)	Main specification	–0.079	0.017
(2)	No covariates and school fixed effects	–0.090	0.021
(3)	Linear specification	–0.091	0.019
(4)	Cubic specification	–0.082	0.018
(5)	School-specific polynomials	–0.078	0.018
(6)	Subject and level fixed effects	–0.078	0.017
(7)	Including delayed students	–0.064	0.017
(8)	No outliers (top and bottom 1% excluded)	–0.087	0.017

Notes: The table shows point estimates and standard errors for β_1 in Equation 1, estimated with ordinary least squares. See notes for Table 6. Row 1 shows results from estimating a specification without covariates and school fixed effects. Rows 2 and 3 show results from estimating specifications with linear and cubic polynomials, respectively, in pre-recoded GPA. Row 4 shows results from estimating a specification where the polynomials in the pre-recoded GPA are interacted with school indicators. Row 5 shows the result from a specification with controls for the focal individual's number of A-level subjects, the number of B-level subjects, and for whether the focal individual completed A-level mathematics. Row 6 shows the result from a specification that includes students who graduated after 2008. Row 7 shows results from a specification where we exclude the top and bottom 1 percent in terms of the recoding residual.

and cubic polynomial for the functional form *f*(). As the table shows, the results are not sensitive to the choice of functional form.

Furthermore, in the main analyses the model estimates the impact of the change in the GPA compared to the change in cohort GPA. However, if students do not have access to the nationwide distribution of grades, the students may instead compare the change in GPA to that of their peers at their school. In Row 5 we show results from estimating a specification where the pre-recoding GPA is interacted with school indicators. The coefficient is very similar to the main results. To account for the school tracks, we control for the composition of A- and B-level courses that the students take in Row 6. The coefficient is, again, close to the main specification. In Row 7, we include in the sample students who graduated later than 2008, which does not affect the coefficient markedly. Finally, to test how sensitive the model is to outliers, we exclude the individuals who experienced the largest changes due to the recoding in Row 8. We first residualize the recoded GPA[23] and then exclude the individuals in the bottom and top percentiles of the residualized GPA shock in the outcome equation. Excluding these observations does not affect the coefficient considerably.[24]

23. That is, we save the residuals from a regression of the recoded first-year GPA on the first-year GPA before the recoding, the first-year GPA before the recoding squared, the full set of covariates, and school indicators.
24. To assess the effects of a reform-induced GPA change beyond the mean, we also ran a set of quantile regressions. Although the point estimates are slightly larger in the tails, the effect is very homogeneous from the 20th to the 80th percentile. These results are available upon request.

B. Do Teachers Manipulate Scores in Response to the Grading Reform?

Even if a positive effect of a downward GPA shock on subsequent student achievement can be established, it is important to understand the factors driving the improvement in achievement. As discussed in Section II, the reform-induced variation in grades (leading to relative better or worse grades) affected students' chances of attending university. One alternative mechanism is that teachers systematically manipulate student scores in response to the reform. Previous literature has focused on how the incentive structures associated with test-based accountability may cause teachers to intentionally manipulate standardized test scores (for example, Jacob and Levitt 2003; Neal 2013). Lavy (2009) finds, however, that although a teacher incentive program in Israel increased teacher effort, the program did not affect test score inflation. The Danish grading reform did not provide pecuniary rewards to inflate grades for specific students. However, Dee et al. (2019) suggest that even in the absence of incentives, altruism among teachers may be a strong motivation to manipulate scores. In a study of New York City schools, Dee et al. (2019) also find that a teacher's propensity to manipulate a student's exam is influenced by the student's previous test scores.

If teachers know the outcome of the recoding for individual students, the teachers could be more generous to unlucky students. If teachers compensate students who are penalized by the grading reform, then the performance effects reported in the main results section could reflect teacher manipulation rather than true gains in student performance. To assess this explanation, we exploit the variation in how grades are set. As described previously, each student receives exam grades and teacher evaluations based on classroom performance. Whereas the student's own teacher has full discretion regarding the classroom assessment, the written exams are graded by two external examiners. These examiners are teachers from other schools without any knowledge about the students.[25]

Table 8 shows the results from using internal grades (that is, teacher evaluations in the second and third years, as well as exams that were partially graded by an internal examiner) and using external grades only (that is, written exam grades in the second and third years). As Panel A shows, the results for internal grades are positive and precisely estimated. Although the effects are smaller, Panel B shows that there are also effects when we use the average of the externally given grades as the outcome. However, the difference in the main effects based on the internal and external assessments is statistically significant, a pattern consistent with teachers manipulating scores for students who were unlucky. Nevertheless, the findings suggest that overall improvements in grades also reflect genuine performance improvement.

The analyses of heterogeneous effects reveal some interesting differences across subgroups. Girls respond more to the incentives than boys in academic performance measured by externally evaluated exams. Although slightly smaller in magnitude, the response in external grades is also statistically significant for girls, whereas the effect is small and not statistically distinguishable from zero for boys. Moreover, there is some evidence that grades were inflated in particular for students who are high-achieving

25. Whereas examiners are appointed by the Ministry of Education for STX exams, HHX and HTX schools appoint the external examiners themselves.

Table 8
Regression Results for Effects on Internal versus External Assessments

		Ninth-Grade GPA		Gender		Parental Educ.	
	Main (1)	Low (2)	High (3)	Boys (4)	Girls (5)	Low (6)	High (7)
Panel A: Dependent Variable: Internally Graded Assessments							
Recoded GPA	–0.083	–0.032	–0.102	–0.047	–0.108	–0.062	–0.098
	(0.018)	(0.026)	(0.021)	(0.027)	(0.022)	(0.025)	(0.022)
p-value (subgroups)		0.02		0.06		0.25	
Mean of dependent variable	0.00	–0.52	0.51	–0.11	0.08	–0.16	0.18
Panel B: Dependent Variable: Externally Graded Grades							
Recoded GPA	–0.043	–0.012	–0.049	0.005	–0.081	–0.043	–0.045
	(0.018)	(0.024)	(0.025)	(0.028)	(0.023)	(0.023)	(0.024)
p-value (subgroups)		0.24		0.02		0.95	
Mean of dependent variable	–0.00	–0.51	0.49	–0.00	0.00	–0.13	0.17
p-value (internal vs. external)	0.02	0.33	0.01	0.04	0.18	0.39	0.01

Notes: The table shows point estimates and standard errors for β_1 in Equation 1, estimated with ordinary least squares. p-value (internal vs. external) provides the p-value for a null hypothesis of equal β_1 coefficients on external and internal assessments. See notes for Table 6.

and have highly educated parents. These results provide some evidence that teachers were trying to compensate students, particularly children with higher socioeconomic background.

C. Effects of a GPA Shock on the Likelihood of Postsecondary Schooling

Table 9 shows the effect of the GPA shock on enrolling in and completing a university degree within six years of finishing high school. Panel A of Table 9 provides evidence of an enrollment effect. A decrease in the recoded GPA led to an increase in the likelihood of enrolling in a university program. Panel B shows that these effects also translate into graduation. A one standard deviation decrease in the recoded GPA causes a two percentage point increase in the likelihood of graduating from a university within six years of finishing high school. The effects are largest among individuals with a middle school GPA above the median and from a higher socioeconomic background. Moreover, the long-run effects are primarily driven by girls, who also reacted statistically significantly more than boys in their performance during high school.

Table 9
Regression Results for Long-Run Effects

		Ninth-Grade GPA		Gender		Parental Education	
	Main (1)	Low (2)	High (3)	Boys (4)	Girls (5)	Low (6)	High (7)
Panel A: Dependent Variable: Enrolled in a University Program within Six Years after High School							
Recoded GPA	–0.024	0.001	–0.039	–0.007	–0.039	0.005	–0.047
	(0.011)	(0.017)	(0.014)	(0.017)	(0.013)	(0.018)	(0.013)
p-value (subgroups)		0.08		0.13		0.02	
Mean of dependent variable	0.56	0.40	0.71	0.58	0.54	0.45	0.65
Panel B: Dependent Variable: Graduated a University Program within Six Years after High School							
Recoded GPA	–0.017	0.012	–0.027	0.009	–0.036	0.008	–0.037
	(0.010)	(0.015)	(0.015)	(0.016)	(0.013)	(0.016)	(0.013)
p-value (subgroups)		0.09		0.04		0.03	
Mean of dependent variable	0.39	0.24	0.54	0.39	0.39	0.30	0.47

Notes: The table shows point estimates and standard errors for β_1 in Equation 1, estimated with ordinary least squares. See notes for Table 6.

To further explore the gender differences, Figure 4 shows effects on postsecondary education over time separately for girls and boys. The figure presents the results using data on educational status six years after graduation during 2009–2014. Moreover, to study whether the reform-induced recoding affected students' tendency to attend postsecondary education (or whether it merely led students to substitute between different postsecondary educational programs), we distinguish between university and other postsecondary educational programs (for example, teachers college or nursing school).

For girls, there is a positive effect (of a negative shock) on university enrollment during the six years, although the effect is not statistically distinguishable from zero in the first year after high school graduation (the dark line in Figure 4, Panel A). As seen in Figure 4, Panel B, there is an effect on university completion four to six years after high school graduation (such a lag is expected given that Danish bachelor programs are often three-year programs). Interestingly, although somewhat smaller in magnitude, we also find that a negative grade shock led to a decrease in enrollment and graduation from postsecondary educational programs, other than university. These findings suggest that girls who would have attended a shorter postsecondary education program end up in a university program as a result of the grading reform. In contrast, Figure 4, Panels C and D show that effects for boys are small and not distinguishable from zero for all years.

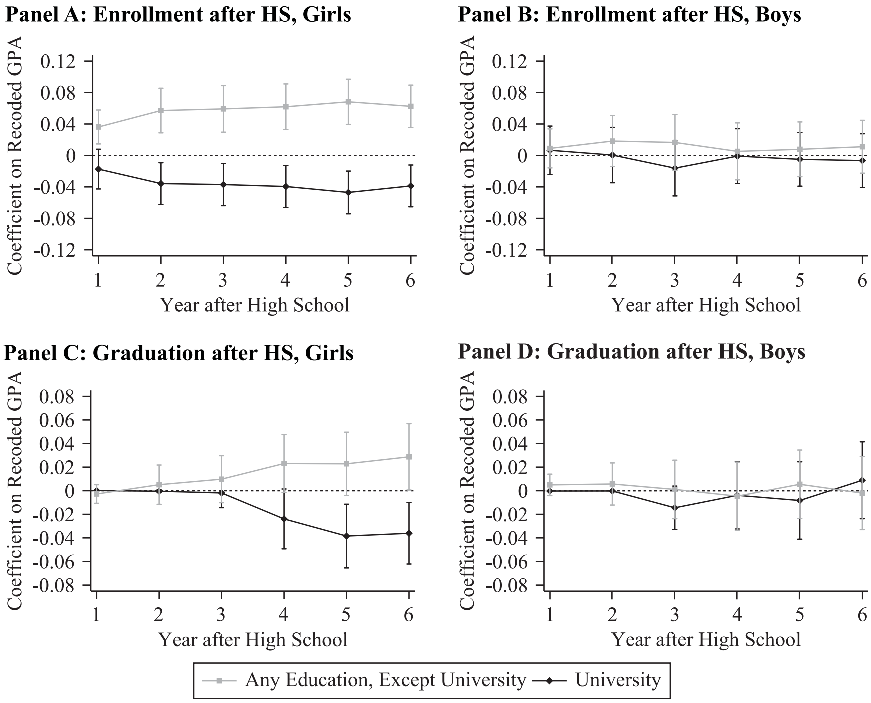

Figure 4

Long-Run Postsecondary Education Enrollment and Graduation Effects

Notes: The graphs show point estimates and standard errors for β_1 in Equation 1, estimated with ordinary least squares. See notes for Table 6. Panels A and B show the coefficients from specifications where the dependent variable is equal to one if the individual was enrolled in an educational program within *y* years after graduation, where *y* corresponds to the value shown on the horizontal axis. Panels C and D show the coefficients from specifications where the dependent variable is equal to one if the individual graduated from an educational program within *y* years after graduation, where *y* corresponds to the value shown on the horizontal axis. Panels A and C are for girls only, and Panels B and D are for boys only. The gray lines show estimates from specifications where educational programs include all programs except university (for example, teachers college or nursing school). The black lines show estimates from specifications where the educational programs include only university programs (as in Table 9).

This pattern in university attendance by gender is in line with the main effects. Girls respond more strongly to a negative shock by improving their effort and, consequently, are more likely to enroll in university. However, different potential mechanisms could explain these findings regarding university attendance for girls. Because enrollment in university is conditional on the final high school GPA, changes in the final GPA have an (mechanical) effect on the number of programs that the student can access. The recoding affected students' expected final GPA. However, as students (particularly girls) also reacted to the reform-induced change in their first-year GPAs in terms of academic performance in subsequent courses, the net effect on the final high school GPA is unclear. An alternative explanation for the long-run effects may be that an enhanced study effort in response to a change in the GPA increases exposure to academic material, and therefore students' aspirations for further education.

To get at the mechanism driving the university enrollment and graduation results for girls, we compare the magnitude of the behavioral response in the second and third years to the initial effect of the recoding in the first year. Following the reasoning in Section A—and by using the estimated coefficient of –0.106 for the girls' behavioral response (from Table 6)—the overall impact for girls of a one standard deviation reform-induced reduction is $0.1-(0.106\times0.9)=0.0046$ SD lower final GPA. In the absence of a behavioral response, the negative impact would have been 0.1 standard deviations (as the reform-affected exams constitute 10 percent of the overall GPA). Thus, girls were able to compensate for more than 95 percent of the shock due to the grading reform. Nevertheless, as girls were not able to compensate fully for the shock, the long-run effects are not merely mechanical and therefore not driven by girls having access to more programs due to a higher GPA.[26] That we still find an effect on postsecondary education attendance for girls suggests alternative mechanisms. One explanation could be that an enhanced study effort increases the students' exposure to academic material, which translates into higher educational aspirations.[27]

D. Falsification Tests

The causal interpretation of the GPA shock is based on the assumption that the shock is unrelated to observed and unobserved characteristics that are related to the outcome of interest. As we showed previously, the change in GPA caused by the recoding process is not related to observable characteristics. Although this test is informative about whether the reform-induced GPA change is related to observable characteristics, the test cannot inform us about how the shock is related to unobservable characteristics. To assess this concern, we run a set of placebo regressions, in which we implement the grading reform on nonreform cohorts and conduct the same analysis as for the main analysis. Specifically, we implement the grading reform on the cohorts of high school students who graduated in 2005, 2006, and 2007 and were graded according to the old scale (that is, the three cohorts before the one affected by the reform). We impose the recoding on the first-year grades and proceed as described for the main analysis.[28]

If the grading shock is unrelated to the outcomes students would exhibit without the shock, one should not expect to see any effect for cohorts unaffected by the reform. Figure 5 shows the results across outcomes. Compared to the estimates from the main analysis, the estimates of the placebo GPA change are small and not statistically

26. If we take the point estimate for boys in Table 3 at face value—and ignore that the coefficient is imprecisely estimated—boys only compensate for 37 percent of the reform-induced shock, as the overall impact for them would be $0.1-(0.041\times0.9)=0.063$.

27. The reaction in terms of subsequent school performance may have been achieved by adjusting effort. Thus, students who increase study effort may have to reduce the time spent on other activities, such as work for pay, alongside their studies. In Online Appendix A we assess this in terms of student labor supply during high school. Although the effect is modest, we find that downgrading made students work less for pay alongside their studies. The decrease in labor supply may provide suggestive evidence that students reacted to the negative shock by reducing time spent on activities other than studying.

28. As the covariates were not available for all placebo cohorts, these coefficients are all estimated without covariates (but with school fixed effects).

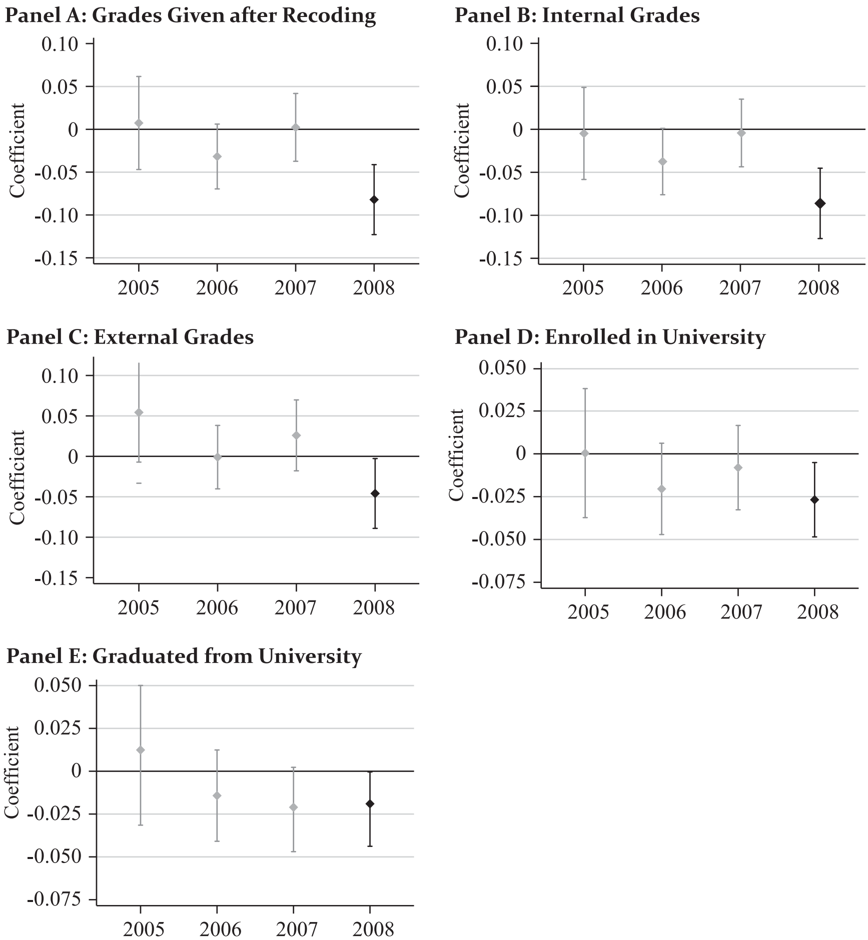

Figure 5

Placebo Tests: Estimates of β_1 *Based on Equation 1, by High School Cohort*

Notes: 2005–2007 are untreated cohorts, and 2008 is the treated cohort. As the data do not include covariates for the 2005–2007 cohorts, all specifications are estimated without covariates, but with school fixed effects.

different from zero at a 5 percent significance level. Two exceptions are the long run effects[29] in Figure 5, Panels D and E, where the coefficients for the placebo cohorts are similar to the treated cohort, but imprecisely estimated. However, when conducting these falsification tests separately for girls and boys, as shown in Online Appendix Figures 10 and 11, the contrast between the placebo cohorts and the treated cohorts becomes much clearer for girls, both in terms of coefficient size and precision. This analysis provides strong evidence that the combination of grades that leads to a downgrade or an upgrade is not related to subsequent performance. These results are reassuring in terms of the causal interpretation of the post-treatment differences in outcomes for the cohort that was affected by the reform.

VII. Conclusion

We present evidence that Danish high school students reacted to a change in their high-stakes GPA that was caused by the implementation of a new grading system. We find that a relative downgrade of the first-year GPA causes students to perform better in their second and third years of high school. Importantly, as the recoding of the grades contained no information about ability, these findings suggest that students respond to changes in incentives. For girls, the shock in first-year grades is almost offset by better academic performance in the second and third years. In contrast, although positive, for boys the response in academic performance was insufficient to make up for the shock. The effects are also larger for students with a middle school GPA above the median.

To address the concern that the effect could be driven merely by teachers manipulating internally assessed grades, we also study standardized national exams that are externally graded, and find that the effects persist. The behavioral response to the negative GPA shock is sufficiently large to have long-run implications—students who received a negative GPA shock to their first-year grades were ultimately more likely to complete a university degree within six years of high school graduation. This effect is driven by girls for whom the negative shock led to better subsequent academic performance, which translated into a higher chance of attending university.

The findings indicate that students adjust subsequent effort and educational choices in response to a change in their GPA that is unrelated to their previous performance. The results appear relevant not only for the Danish setting but also for educational systems in other countries. Although the Danish educational system differs in some respects from educational systems in other European countries and the United States, the importance of high school assessments for entrance into postsecondary schooling in Denmark closely resembles the high stakes of high school exams in many other countries. Therefore, the findings may be informative about how getting into university constitutes an incentive for students to work harder.

Whereas previous literature has focused mainly on pecuniary incentives to motivate student effort, the present results suggest that the incentives related to university enrollment are important for student behavior. A deeper understanding of how incentives affect student behavior remains an important objective for future research.

References

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2009. "Incentives and Services for College Achievement: Evidence from a Randomized Trial." *American Economic Journal: Applied Economics* 1(1):136–63.

Angrist, Joshua, and Victor Lavy. 2009. "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial." *American Economic Review* 99(4):1384–414.

Apperson, J., C. Bueno, and T.R. Sass. 2016. "Do the Cheated Ever Prosper? The Long-Run Effects of Test-Score Manipulation by Teachers on Student Outcomes." Unpublished.

Avery, Christopher, Oded Gurantz, Michael Hurwitz, and Jonathan Smith. 2018. "Shifting College Majors in Response to Advanced Placement Exam Scores." *Journal of Human Resources* 53(4):918–56.

Azmat, Ghazala, Manuel Bagues, Antonio Cabrales, and Nagore Iriberri. 2019. "What You Don't Know…Can't Hurt You? A Natural Field Experiment on Relative Performance Feedback in Higher Education." *Management Science* 65(8):3714–36.

Bandiera, Oriana, Valentino Larcinese, and Imran Rasul. 2015. "Blissful Ignorance? A Natural Experiment on the Effect of Feedback on Students' Performance." *Labour Economics* 34: 13–25.

Bettinger, Eric P. 2012. "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores." *Review of Economics and Statistics* 94(3):686–98.

Burgess, Simon, Robert Metcalfe, and Sally Sadoff. 2016. "Understanding the Response to Financial and Non-Financial Incentives in Education: Field Experimental Evidence Using High-Stakes Assessments." IZA DP 10284. Bonn, Germany: IZA.

Dee, Thomas S., Will Dobbie, Brian A. Jacob, and Jonah Rockoff. 2019. "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations." *American Economic Journal: Applied Economics* 11(3):382–423.

Dee, Thomas S., and Brian A. Jacob. 2006. "Do High School Exit Exams Influence Educational Attainment or Labor Market Performance?" NBER Working Paper 12199. Cambridge, MA: NBER.

———. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management* 30(3):418–46.

Dee, Thomas S., and James Wyckoff. 2015. "Incentives, Selection, and Teacher Performance: Evidence from Impact." *Journal of Policy Analysis and Management* 34(2):267–97.

Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks. 2016. "School Accountability, Postsecondary Attainment, and Earnings." *Review of Economics and Statistics* 98(5):848–62.

Diamond, Rebecca, and Petra Persson. 2016. "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests." NBER Working Paper 22207. Cambridge, MA: NBER.

Ebenstein, A., V. Lavy, and S. Roth. 2016. "The Long-Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution." *American Economic Journal: Applied Economics* 8(4):36–65.

Elsner, Benjamin, and Ingo Isphording. 2017. "A Big Fish in a Small Pond: Ability Rank and Human Capital Investment." *Journal of Labor Economics* 35(3):787–828.

Fryer, Roland G. 2011. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." *Quarterly Journal of Economics* 126(4):1755–98.

Fryer, Roland G. 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." *Journal of Labor Economics* 31(2):373–407.

Humlum, M.K., J.H.G. Kristoffersen, and R.M. Vejlin. 2014. "Timing of College Enrollment and Family Formation Decisions." IZA Discussion Paper 7905. Bonn, Germany: IZA.

Imberman, Scott A., and Michael F. Lovenheim. 2015. "Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System." *Review of Economics and Statistics* 97(2):364–86.

Jackson, C. Kirabo. 2010. "A Little Now for a Lot Later: A Look at a Texas Advanced Placement Incentive Program." *Journal of Human Resources* 45(3):591–639.

Jacob, Brian A. 2002. "Where the Boys Aren't: Non-Cognitive Skills, Returns to School and the Gender Gap in Higher Education." *Economics of Education Review* 21(6):589–98.

Jacob, Brian A. 2005. “Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools.” *Journal of Public Economics* 89(5):761–96.
Jacob, Brian A., and Steven D. Levitt. 2003. “Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating.” *Quarterly Journal of Economics* 118(3):843–77.
Lavy, Victor. 2009. “Performance Pay and Teachers’ Effort, Productivity, and Grading Ethics.” *American Economic Review* 99(5):1979–2021.
Lavy, Victor. 2015. “Teachers’ Pay for Performance in the Long-Run: Effects on Students’ Educational and Labor Market Outcomes in Adulthood.” NBER Working Paper 20983. Cambridge, MA: NBER. https://doi.org/10.3386/w20983
Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff. 2016. “The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance.” *American Economic Journal: Economic Policy* 8(4):183–219.
Murphy, Richard, and Felix Weinhardt. 2014. “Top of the Class: The Importance of Ordinal Rank.” CESifo Working Paper Series 4815. Munich, Germany: CESifo.
Neal, Derek. 2013. “The Consequences of Using One Assessment System to Pursue Two Objectives.” *Journal of Economic Education* 44(4):339–52.
Oreopoulos, Philip, and Kjell G. Salvanes. 2011. “Priceless: The Nonpecuniary Benefits of Schooling.” *Journal of Economic Perspectives* 25(1):159–84.
Papay, John P., Richard J. Murnane, and John B. Willett. 2016. “The Impact of Test Score Labels on Human-Capital Investment Decisions.” *Journal of Human Resources* 51(2):357–88.
Reardon, Sean F., Nicole Arshan, Allison Atteberry, and Michal Kurlaender. 2010. “Effects of Failing a High School Exit Exam on Course Taking, Achievement, Persistence, and Graduation.” *Educational Evaluation and Policy Analysis* 32(4):498–520.
Reback, Randall. 2008. “Teaching to the Rating: School Accountability and the Distribution of Student Achievement.” *Journal of Public Economics* 92(5–6):1394–415.
Smith, Jonathan, Michael Hurwitz, and Christopher Avery. 2017. “Giving College Credit Where It Is Due: Advanced Placement Exam Scores and College Outcomes.” *Journal of Labor Economics* 35(1):67–147.
Stinebrickner, Ralph, and Todd Stinebrickner. 2014. “Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model.” *Journal of Labor Economics* 32(3):601–44.
Stinebrickner, Todd, and Ralph Stinebrickner. 2012. “Learning about Academic Ability and the College Dropout Decision.” *Journal of Labor Economics* 30(4):707–48.
Zafar, Basit. 2011. “How Do College Students Form Expectations?” *Journal of Labor Economics* 29(2):301–48.