OPEN ACCESS

University of BRISTOL

Hvidman, U., & Sievertsen, H. (2020). High-Stakes Grades and Student Behavior. *Journal of Human Resources*, *0*(0), [0718-9620R2]. https://doi.org/10.3368/jhr.56.3.0718-9620R2

Early version, also known as pre-print

Link to published version (if available):
10.3368/jhr.56.3.0718-9620R2

Link to publication record in Explore Bristol Research
PDF-document

This is the submitted manuscript (SM). The final published version (version of record) is available online via University of Wisconsin Press at http://jhr.uwpress.org/content/early/2019/09/10/jhr.56.3.0718-9620R2. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

High-Stakes Grades and Student Behavior*

Ulrik Hvidman† and Hans Henrik Sievertsen ‡

Abstract

This paper uses a reform-induced recoding of grades, which caused variation in high school students' grade point average (GPA), to identify students' behavioral responses to changes in high-stakes grades. The results show that students who were downgraded by the recoding performed better on subsequent assessments and were more likely to complete university programs after high school. As the recoding did not convey information about actual academic performance, these results emphasize that changes in incentives are important in understanding students' responses to high-stakes grades. There is no evidence that the recoding algorithm predicts outcomes for non-affected cohorts.

Keywords: student behavior, high-stakes tests, human capital.
JEL: I20, I21, I23, J24.

*We 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.

†Department of Political Science, Aarhus University. E-mail: uhvidman@ps.au.dk

‡University of Bristol & VIVE. Corresponding author, e-mail: h.h.sievertsen@bristol.ac.uk

1 Introduction

Standardized tests have become increasingly prevalent in school systems in many countries in recent decades. These tests convey information that may affect students' educational investment decisions. Given that students have imperfect information on their ability and on how their effort translates into performance, they may use test results to learn about their return to investments in schooling (Stinebrickner and Stinebrickner, 2012, 2014; Bandiera et al., 2015; Diamond and Persson, 2016).

However, in addition to providing information on academic ability, test results and exam scores often carry major consequences for students. These high-stakes exams may affect students' likelihood of graduating high school or being admitted to a (selective) university. For example, exit exams—which students have to pass to earn a high school diploma—have become common (Dee and Jacob, 2011; Murnane, 2013). Moreover, university programs rely heavily on information about educational achievement as a screening tool in the admission process. In the US, for instance, many universities base their admission criteria on standardized tests such as SAT scores. In other countries, such as Denmark, Norway, and Sweden, admission to post-secondary education, especially to universities, is determined predominantly by high school Grade Point Average (GPA). Still, much remains unknown about students' behavioral responses to the actual changes in incentives resulting from high-stakes exam scores.

This paper uses a novel identification strategy to isolate the behavioral effect of receiving lower (higher) exam grades. We exploit a grading reform in Denmark that caused exogenous variation in high school students' GPAs (by recoding all their grades) 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 scheme provided by the Ministry of Education. As they feed

into the calculation of the final GPA, these first-year grades were high stakes.[1] The consequence of the reform was that two students with identical GPAs before the grading reform could have very different GPAs after the reform. We exploit this reform-induced change in grades to identify the students' responses to a change in their GPAs. While the reform changed students' grades, it did not provide any new information about academic performance or ability to the students. Thus, any changes in effort investment in response to the grading reform must reflect the grades *per se* and the change in incentives.

The identifying assumption is that there are no systematic differences between students who were adjusted upwards and downwards 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. Particularly, the shock in grades 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 that were affected by the reform, we find that students who experienced a negative shock to their GPAs due to the reform, received better subsequent grades. Students who were downgraded one standard deviation scored 8% of a standard deviation higher in subsequent assessments. Although part of this effect is explained by teachers compensating unlucky students by inflating their post-reform grades, students who received a negative shock also received better subsequent grades in national standardized exams that are externally evaluated. Furthermore, although the effect is modest in

[1] In Denmark, high school performance almost entirely determines admission to post-secondary schooling and to various academic programs, particularly at universities. Furthermore, access to university majors is based on high school performance, and many selective programs (e.g., psychology or medicine) require high overall average scores for admission. Nevertheless, a small share of post-secondary 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 (e.g., 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 (e.g., work experience). In 2008, 10 percent of enrollments were based on this scheme. Thus, high school grades are particularly important.

magnitude, we find that downgrading made students work less for pay alongside their studies. The decrease in labor supply may suggest that students reacted to the negative shock by reducing time spent on activities other than studying. The GPA shock also had long-run consequences: Students who were downgraded by one standard deviation had a 2 percentage point higher probability of enrolling in and graduating from a university program after high school. We find some evidence of gender and ability differences in how students reacted to the GPA shock, but no clear differences by socio-economic background.

Our study contributes to an emerging literature on test takers' behavioral responses to feedback on educational performance. This literature examines how receiving a certain grade affects student behavior and long-term outcomes. First, some studies examine the role of performance information in students' educational choices and find that the acquisition of new information about academic ability affects beliefs about own ability, study effort, drop-out decisions, and college enrollment (Jacob and Wilder, 2011; Zafar, 2011; Stinebrickner and Stinebrickner, 2012).[2] Second, studies have examined how external factors affecting high-stakes outcomes may have long-run implications for individual human capital accumulation. Apperson et al. (2016), Dee et al. (2016) and Diamond and Persson (2016) study how teacher manipulation of test results affects individual human capital accumulation. In related work, Ebenstein et al. (2016) study how variation in exam scores due to pollution exposure affects post-secondary educational attainment and earnings. These studies offer mixed conclusions. Whereas Ebenstein et al. (2016) and Diamond and Persson (2016) find that those who receive a positive shock have *better* long-run outcomes[3], Dee et al. (2016) show that manipulation in scores have mixed impacts, and Apperson et al. (2016) find that test score manipulation can even have a negative impact.

Overall the findings of this study contribute to this literature by showing that scores that carry

[2]Our work is also related to the literature on how performance labels that do not carry official consequences for students affect their choices of post-secondary education (Papay et al., 2016; Avery et al., 2017; Smith et al., 2017).This literature finds that two students with almost identical raw scores make different educational choices because of discontinuities in the labeling. In other words, the two students know that their academic performance was almost identical, but they nevertheless react to the labels. Whereas Papay et al. (2016) and Smith et al. (2017) examine test labels that do not have major consequences for students, our study focuses on students' responses to high-stakes outcomes.

[3]In a study of Swedish lower-secondary students, Diamond and Persson (2016) finds effects only for high-ability students, who do not have strong incentives to perform better.

official consequences for students affect individuals' behavior during the program, and that these behavioral changes translate into educational outcomes that have long-term consequences. The findings of this paper demonstrate that these incentives 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 test grades should not affect the students' perceptions about their ability or self-confidence. Instead, any changes in effort investment in response to the grading reform reflect the change in incentives, not learning. Thus, the findings suggest an important mechanism through 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.

Our findings are also relevant to policy as they are informative about how students respond to exogenous performance shocks in a high-stakes environment. Although the Danish educational system differs in some respects from educational systems in other European countries and the United States, the combination of compulsory standardized testing and internal evaluations in schools is very similar to the high school exam structure in many other countries. The importance of these high school assessments for entrance into post-secondary schooling in Denmark also closely resembles the high stakes of high school exams in the United States and elsewhere. Thus, the results from this paper appear relevant not only for the Danish setting but for educational systems in many other countries as well.

One important implication of the incentive effect is that variation in grading standards across schools or teachers can have implications for students' future academic performance and long-run human capital accumulation. Exposure to higher standards (i.e. lower grades conditional on performance) will require the student to invest more effort to obtain a certain grade average. This mechanism is in line with existing findings on implications of higher grading standards (Betts and Grogger, 2003; Figlio and Lucas, 2004), which finds that children exposed to harsher teachers

have better subsequent test results.[4] These results may also have implications regarding to returns to school quality. Given that grading tends to be harsher in more selective schools (as suggested by Calsamiglia and Loviglio, 2017), the incentive effects may account for some of the returns to school quality (in terms of future academic performance and future labor market outcomes).

The remainder of the paper is organized as follows. Section 2 discusses the theoretical expectations for behavioral responses to grades. Section 3 provides the institutional background about the Danish educational system and describes the grading reform. Section 4 describes the administrative data. Section 5 discusses the identification strategy and the estimation. Section 6 presents the results, and Section 7 concludes.

2 Grades and student behavior: Learning and incentives

To motivate the empirical analysis, this section presents a model on how performance feedback may influence student behavior and outcomes. The objective of the model is to highlight the difference between a low-stakes setting where responses are driven by a learning effect, and a high-stakes setting where responses are also driven by incentives.[5] First, academic output affects individuals directly, for example through intrinsic motivation, confidence, or status. We consider the following production function for academic output:

$$O(a,s) = (a \times s)^{\theta} \tag{1}$$

[4] Betts and Grogger (2003) exploit variation in standards across high schools and find that students in schools with harsher standards (i.e., lower grades conditional on performance) get better subsequent test scores. The effect is found for the entire distribution, but is strongest for the high achieving students. They find no effect on continued schooling except for a negative effect for black and Hispanic students. Figlio and Lucas (2004) exploit variation grading standards across teachers using data on primary school children in Florida. They show that teachers with harder grading standards are associated with better learning outcomes for their pupils.

[5] Closely related to our theoretical framework, Azmat et al. (2016) propose a model that distinguishes between two theoretical mechanisms. In their model, students may respond to information because individuals have an imperfect knowledge of their own ability or because they have inherently competitive preferences.

were $0 < \theta < 1$, a is ability and s is study effort. Second, the individuals care about chances of admission to college:

$$A(a,s,u) = 1 - e^{-(a \times s + u)} \tag{2}$$

where u is exogenous factors affecting the grade. The only choice variable for the student is effort, which is chosen to maximize the following utility function:

$$U = \omega_O O(a,s) + \omega_A A(a,s,u) - C(s) \tag{3}$$

where $C(s) = \eta s$ is the cost of study effort.[6] The weights, ω_O and ω_A, capture how the individuals weight academic output and admission chances. If for example a student has no plans of continued schooling, but gains confidence and satisfaction from learning (an intrinsic motivation), we have the case that $\omega_O > 0$ and $\omega_A = 0$. On the other hand, we could imagine students who have no intrinsic motivation to study but only care about admission to educational programs because they provide better job chances, i.e., a case with $\omega_O = 0$ and $\omega_A > 0$.

We now use the model to characterize how students may respond to performance signals. Performance signals may either be informative about the level of ability and the production function of academic output or about the exogenous shocks.

The learning effect of grades

Let us first consider the case where grades work as signals of academic ability. To simplify, consider the case $\omega_A = 0$. As grades are low stakes and have no consequences, they only affect individual behavior through the first term, i.e., intrinsic motivation and confidence. Consistent with recent literature (Bandiera et al., 2015; Azmat et al., 2016), the model assumes that effort and ability are complements in producing academic output. However, students have imperfect information

[6] Study effort comprises psychic costs (e.g., stress), direct pecuniary costs (e.g., study material such as books), or indirect pecuniary costs (e.g., foregone earnings on the labor market). To keep the model as simple as possible, we assume constant study effort costs; this assumption is not important for the objective of our.

about their ability (and the production function) and, therefore, about how their effort translates into performance. As students perceive the grade as informational about ability, the complementarity between ability and effort implies that the students learn about how their effort translates into grades, which affect their future choices of study effort. In this simple case, a positive performance signal will make the students want to increase their study effort as they realize the higher payoff of their effort.

The incentive effect of grades

Consider now the case where $\omega_A > 0$. In this setting—which reflects a context in which consequences are attached to grades—the response to a performance signal is less clear. The complementarity between study effort and ability implies that the second derivative with respect to ability is ambiguous for the second term:

$$\frac{\partial^2 A}{\partial s \partial a} = (1 - s \times a) e^{-(s \times a + u)} \lessgtr 0 \tag{4}$$

A positive performance signal means that students have to supply less effort to achieve a given chance of admission (the income effect). However, due to the complementarity between effort and ability, the marginal benefit of supplying more effort is also larger (the substitution effect). How the individual responds to a performance signal is therefore ambiguous.

Our empirical strategy is to study an exogenous performance shock, u, that is known to be unrelated to ability. In this case, the effort response is unambiguous, because the first term of the utility function is unaffected, and the second derivative of the second term is unambiguously negative:

$$\frac{\partial^2 A}{\partial s \partial u} = -a \times e^{-(s \times a + u)} < 0 \tag{5}$$

Because of the income effect, students reduce their effort in response to a positive GPA signal. That is, for a given level of ability, the individual will put in less effort because the chances of college

admission, given effort and ability, have increased. The shock only contains information about exogenous factors. It does not affect the true academic output, nor does it contain new information about the return to effort, but it affects the chances of admission. This effect reflects the fact that feedback on exam grades often carries important consequences for students. For example, exam grades are important for high school students as they may determine whether a student graduates high school with a diploma or gets into a university program. If, for example, the cutoff for college admission is lowered, students respond by lowering their effort. Thus, one could speak of an *incentive effect*.

In sum, there are conceivable arguments supporting the hypothesis that changes in grades may affect subsequent student outcomes. These effects could be driven either by students leaning about their return to effort investment, by changes in students' incentives, or by both.In the next section, we present the empirical setting that enables us to isolate the incentive effect from the learning effect.

3 Background

3.1 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 (i.e., grades zero through nine). Having completed basic schooling, students may continue to a three-year high school program (grades 10 through 12), enroll in vocational training, or enter the labor market. Among the 65 thousand 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").

Despite slightly different curricula, the main objective of all high school programs is to prepare

students for higher education, and they all provide equal access to higher education.[7] 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 (e.g., A-level Danish) as well as a minimum number of A-level courses, and within each program, the students may choose between different tracks (i.e., major area of study).[8] Students request a preferred track when they apply for high school but are allowed to change track 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.

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

Post-secondary schooling is free, and students receive a monthly student grant to pay for living expenses for up to six years of post-secondary schooling. Access to post-secondary schooling and to various academic programs, especially at universities, is almost exclusively determined by the high school GPA. After completing high school, all students who wish to enroll in post-secondary schooling apply through a centralized system with a list of prioritized educational programs. The

[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 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 different tracks (e.g., 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.

Table 1: Implementation of Grading Reform across High School Cohorts

		Grading scale		
Enrolled	Graduated	Year1	Year2	Year3
Aug 04	Jun 2007	13	13	13
Aug 05	Jun 2008	13	7	7
Aug 06	Jun 2009	7	7	7

Notes: 13 = "13 scale"; 7 = "7-point scale"

programs set the number of available slots, N, and the course requirements (for example, economics at the University of Copenhagen requires A-level in mathematics and in Danish and B-level in history and in 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 given an offer.[9] The high school GPA is thus particularly important for students that wish to continue in post-secondary schooling. Moreover, as first-, second-, and third-year grades count in the final overall GPA, stakes are high during all three years.

3.2 The Danish Grading Reform of 2007

Until April 2007 student performance in the Danish school system, from lower secondary schooling to post-secondary 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 7-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 were unaffected by the reform as they had all their assessment 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 2005/06 was assessed

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

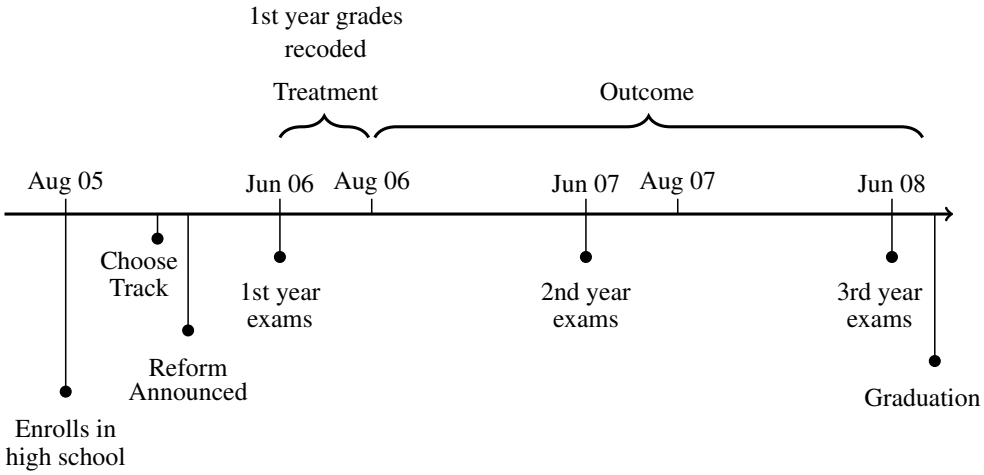


Figure 1: Timeline: Assessment and recoding of grades for students who enrolled in high school in 2005 and graduated in 2008.

on the 13 scale, and coursework completed in the school years 2006/07 and 2007/08 was graded on the new scale. For this cohort, the grades obtained in their first year in accordance with the old scale were subsequently converted to grades in 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 recoding of grades. After the students finished their first year of high school in the summer of 2006, all their first-year grades were converted from the old to the new scale. Thus, their final overall high school GPA was calculated based entirely on grades on the 7-point scale—and only the post-transformation grades were shown on the high school diploma (Appendix Figure A.2 shows a high school diploma for a student from the treated cohort).

Table 2 presents the transformation scheme provided by the Ministry of Education. The first two columns describe the mapping scheme 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 scheme were collapsed to a single new grade. Consider for example a student who only had 8s on the old scheme and another student who only had 9s. Although the latter had higher grades prior to the reform, the two students would have identical grades (i.e., 7s) after the recoding to the new scheme.

Table 2: The Danish Grading System: Transformation from the Old to New Scale

Old 13 scale	New 7-point scale	ECTS	Description
00	-3	F	For a performance which is unacceptable in all respects.
03 5	0	F^+	For a performance which 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. 6 (old) / 2 (new) is the passing threshold.

Second, the distance between the old and the new grades varies along the scale. For example, a 5 on the old scale is "punished" heavily as it is transformed to a 0 (i.e., the difference is five points), whereas a 10 on the old scale is not punished (i.e., the difference is zero points). (Appendix Figure A.4 plots grades on the new scale against the grades on the old scale.) Thus, two students with identical pre-transformation GPAs could have very different post-transformation GPAs because grades were punished differently.

As a result, depending on the composition of grades, the students were either down- or up-graded 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.[10] The number of grades given in the first year of high school depends on the specific high-school track chosen by the student. Students typically receive three to five grades in the first year (pre-transformation) and around 30 grades in years two and three (as shown in Appendix Figure A.5). While more grades imply that more grade combinations can cause a specific pre-transformation GPA, the link between number of grades given on the old scheme and the potential variation in post-transformation GPA is not trivial (as Appendix Figures A.3a and A.3b show). For some pre-transformation GPAs a lower number of grades given is associated with a greater potential post-transformation difference in GPA.

4 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 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 they are unaffected by the change of grading system. The high school data contain information on courses and exam-specific grades. The high school data are merged with administrative data from Statistics Denmark on child background (gender, age, and origin) and with school records on middle school GPA (i.e., 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 Statis-

[10] While the collapsing of grades and the varying distance to the new grades caused noise in the individual students' GPAs, the grading reform also affected the overall level and distribution of grades. Figure A.6 in the Appendix shows the high school GPA distribution for the cohorts graduating in the years 2003 to 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.

[11] As we discuss in Section 6, the pattern in drop-outs appears to be unrelated to the grading reform.

[12] We exclude 695 observations due to missing middle school GPA and 3 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 Denmark. No further data restrictions are imposed. The final sample includes 94 percent of the initial population. Including students with missing observations yields qualitatively identical results. These results are available upon request.

tics 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 post-secondary schooling outcomes and to records on their labor market attachment during high school. We measure the labor supply during high school for the calendar year 2007, which corresponds to the second half of the second year in high school and the first half of the third year in high school.

Table 3: Variable descriptives

	Mean	SD
Age at HS enrollment	16.66	0.67
Female	0.56	0.50
Non-western origin	0.05	0.21
9th grade GPA	0.27	0.85
Parents' years of schooling	14.63	2.01
Parents' income (1,000 Euro)	35.80	24.55
Worked in second year	0.86	0.35
Labour income in second year, (2015 1,000 Euro)	5.75	3.88
Grades recoded	3.41	2.71
Grades given after recoding	29.03	3.24

Notes: Parental characteristics are measured in the calendar year prior to student's high school enrollment. All monetary values are converted to the 2015 price level using the consumer price index.

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 five 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 9th grade cohort. 86 percent of the students worked during high school for an average of nearly 6,000 Euros in gross labor earnings.

5 Identification & estimation

5.1 Empirical strategy

The grading reform constitutes a policy change that allows us to examine whether an observed exogenous change in high-stakes GPA affects students' performance. To illustrate the change caused by the reform, Figure 2 shows a heat-map of students' GPA before and after the recoding of their first-year grades.[13] As Figure 2 illustrates, there is substantial variation in post-transformation GPA for any given level of pre-transformation GPA. For example, if we compare two students with a pre-transformation GPA of 8, one could end up with a post-transformation GPA of about 7, whereas the other could end up with a post-transformation GPA of about 5. The dashed line shows the quadratic fit, which captures the relationship fairly well. To assess the impact of the recoding of grades on subsequent performance, we estimate the following equation with OLS:

$$Y_i = \beta_0 + \beta_1 GPA7_i + f(GPA13_i) + \boldsymbol{\lambda}_s'\boldsymbol{\eta} + \boldsymbol{\delta}'\boldsymbol{X}_i + \varepsilon_i \tag{6}$$

where Y_i is the grade point average of grades given in years two and three (i.e., after the grade transformation), $GPA7_i$ is grade point average of first-year grades *after the recoding* to the 7-point scale, $GPA13_i$ is the grade point average of original first-year grades on the 13-scale *before the recoding* [14], $\boldsymbol{\lambda}$ is a vector of school fixed effects, and $\boldsymbol{X}$ is a vector of individual specific covariates including gender, origin (western or non-western), indicators for being 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, as they are highly predictive of

[13] Due to confidentially issues, we cannot show cells with fewer than three observations. However, the regressions analyses are based on all observations.

[14] We standardize Y_i, $GPA7_i$, and $GPA13_i$ to a mean of zero and a standard deviation of one. In our 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.

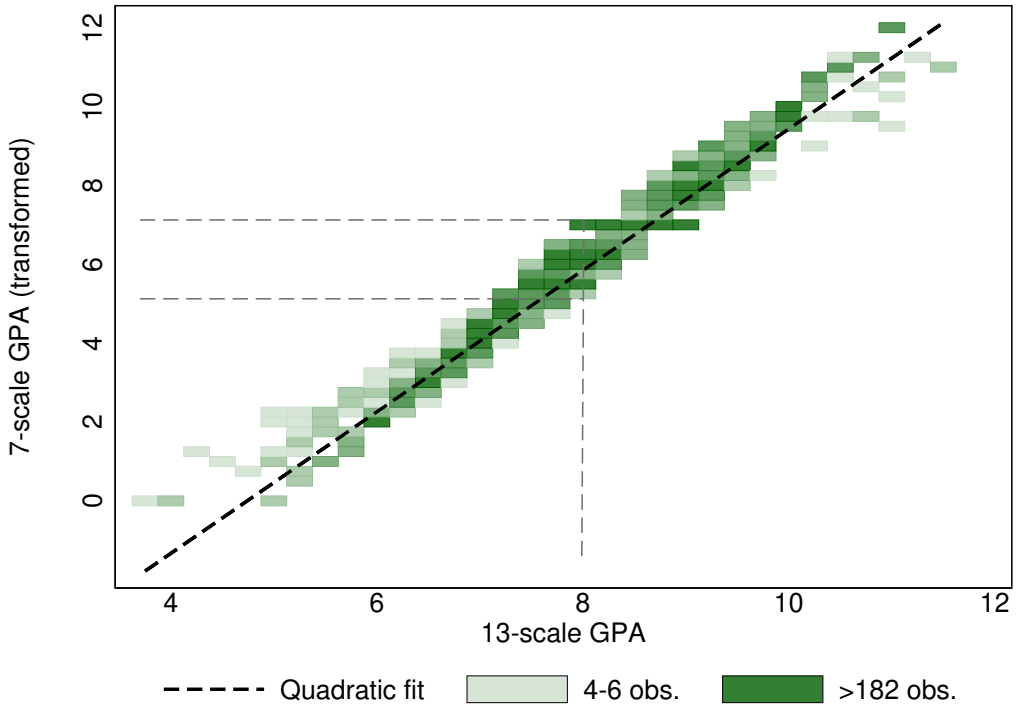


Figure 2: Actual pre- and post-transformation GPA of first-year grades.
Notes: Only combinations with at least three observations are shown.

the students' subsequent performance and educational outcomes. The standard errors are clustered at the school level.

5.2 Strategic responses to the implementation of the grading scheme in Danish high schools?

The key aim of this paper is to study how the change in grades induced by the transformation of first-year grades affects subsequent behavior. The causal interpretation of the GPA shock is based on the assumption that the shock is unrelated to student characteristics that are related to the outcome of interest. There are, however, 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 scheme in early 2006, one concern would be 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 depends on the specific high school track chosen, risk-averse students might have tried to avoid the transformation 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 they had private information on the risk of receiving a grade that was penalized heavily.

Both institutional knowledge and empirical evidence suggests that such strategic behavior was limited. Appendix Figure A.1 shows the Google search term popularity for the new scale for the period 2005 to 2009. The search term "7-trins skala" (English: "7-point scale") gained popularity after July 1, 2006 and 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 (i.e., 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 track after their pre-transformation grades are disclosed.

Apart from the track-specific courses, students can choose a few courses in their second and third years. Students could, therefore, 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 (i.e., 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 did change courses in response to the reform.

Table 4: Regression results, 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 dep. var	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 (6), estimated with ordinary least squares. The dependent variable is denoted in the column header. The grade point averages are standardized to have a mean of zero and a unit standard deviation. We control for first-year GPA before recoding using a second order polynomial. The covariates included are age at high school entry, gender, 9th grade GPA (standardized) origin (indicator for non-western origin), parental education (years of completed education included, average across parents), income (disposable income, average across parents), the number of non-missing parental education and income observations (indicators). All parental variables are measured in the calendar year before the focal individual enrolled in high school. 9th GPA indicates that the sample is split by the median of the students' 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). 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 our data only contains information 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 Appendix Figure A.8. 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 (i.e., before the grade shock occurred) and that there were no changes in dropout levels in year 2 and 3. The change in dropout patterns—with

more students dropping out during their first year—is likely due to the rather 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 prior to enrollment. In 2005 a number of tracks replaced the two-track system and students had to pick 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—it 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 the Table 5 represents an estimate from a regression of the GPA shock on a demographic characteristic. All point-estimates are small and statistically insignificant. The absence of signs that the change in GPA caused by the recoding process is 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.

Table 5: Regression results: Balance of covariates across treatment.

	$\hat{Y}$	9. GPA	Female	Parental education	Parental income
	(1)	(2)	(3)	(4)	(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 dep. var	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 (6), 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 notes for Table 3). Standard errors clustered on the school level in parentheses.

Another concern would be that teachers adjusted their grading behavior prior to the reform.

The internal grading procedure for the classroom grades leaves scope for teachers to manipulate the pre-transformation grades. For example, to help students, teachers could avoid the grades in the first-year exam that were penalized the most in the transformation scheme. Figure A.7(a) in the appendix compares the distribution of first-year grades in the affected cohort with the distribution of grades in the earlier 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[15], some changes in the grade distribution are expected as a result of the high school reform.[16]

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.[17]

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 (e.g., student ability or behavior). As previously shown, the change in GPA caused by the recoding process was not related to observable characteristics, which suggests that the reform did not lead teachers to manipulate certain of their students' scores.

[15] For example, grades in math are usually lower than in other subjects

[16] As the high school reform was implemented nationwide—and all students in our data are affected by the reform—it should not affect our main analysis.

[17] Another way teachers could adjust their grading would be to set the first-year grades by already taking into account the subsequent re-calculation of the grades. To study if this is the case, A.7(b) plots the distribution of grades across cohorts where we re-calculate first-year grades for pre-reform cohorts (i.e., the three cohorts prior to the one affected by the reform) as if the grading reform had been implemented. As A.7(b) shows, the changes in grading pattern that happened in 2005 were rather modest relative to the changes that occurred after the implementation of the reform.

6 Results

6.1 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 year of high school. Table 6 shows the results from estimating the effect of a change in first-year GPA on subsequent grades. The dependent variable is the average of the student's grades in second and third year 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 year 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 first-year GPA perform 8 percent of a standard deviation better in subsequent grades.[18]

Columns (2) and (3) in Table 6 show results from subsample regressions where the sample is split according to the median middle school GPA. While 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 particularly 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 may be more responsive to a change in their GPA. The results presented in columns (4) and (5) of Table 6 show that while the effect of a negative shock is positive for both boys and girls, it is strongest for female students. Finally,

[18] Table A.1 reports the estimates of $GPA13$ and $GPA13^2$. As expected, there is a strong positive association between first-year grades (i.e.,$GPA13$) and second- and third-years grades. Although part of this relationship may be due to the learning effect, a major concern is that the students that receive good grades in first year are likely to be different on 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.

columns (6) and (7) show that there is no clear difference in response by parental background.[19]

To give a sense a of magnitudes. We find that a one standard deviation decrease in grades caused a performance improvement of about eight percent of a standard deviation in subsequent grades. Given that on average three out of 30 grades were recoded (see Table 3), the point estimates should be adjusted by a factor ten, as a rule-of-thumb. In other words, a student who has two grades recoded, and two grades given after the recoding, would be able to compensate for about 80 percent of the shock. Some subgroups were, however, able to fully compensate for the negative shock (i.e. girls: $10 \times 0.106 = 1.06$).

Table 6: The effect of a GPA shock on subsequent grades. Dependent variable: Grades given after transformation (standardized).

		9th GPA		Gender		Parental edu.	
	Main	Low	High	Boys	Girls	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)	(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)
Mean of dep. var	-0.00	-0.54	0.53	-0.09	0.07	-0.16	0.18
P-value		0.03		0.04		0.32	
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 (6), estimated with ordinary least squares. The grade point averages are standardized to have a mean of zero and a unit standard deviation. We control for first-year GPA before recoding using a second order polynomial. The covariates included are age at high school entry, gender, 9th grade GPA (standardized) origin (indicator for non-western origin), parental education (years of completed education included, average across parents), income (disposable income, average across parents), the number of non-missing parental education and income observations (indicators). All parental variables are measured in the calendar year before the focal individual enrolled in high school. 9th GPA indicates that the sample is split by the median of the students' 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-hypotheses that the point-estimates are the same for the two respective subsamples. Standard errors clustered on the school level in parentheses.

In our main analysis, we impose a linear functional form on the relationship between the reform-induced GPA shock and subsequent grade. To test whether the effects are non-linear (e.g., asymmetric for positive and negative shocks), we examine this relationship in a non-restrictive and

[19] Results for low (and high) income parents are very similar to the results for parental education. Results are available upon request.

visual manner. Figure 3 shows the parametric specification (equation 6) as a solid line and plots a histogram that shows 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 non-parametric specification of the relationship between the change in grades and subsequent performance. While the relationship is relatively flat at the lower and upper ends, the linear specification fits the nonparametric pattern fairly well for the range of the GPA shock covering most observations.

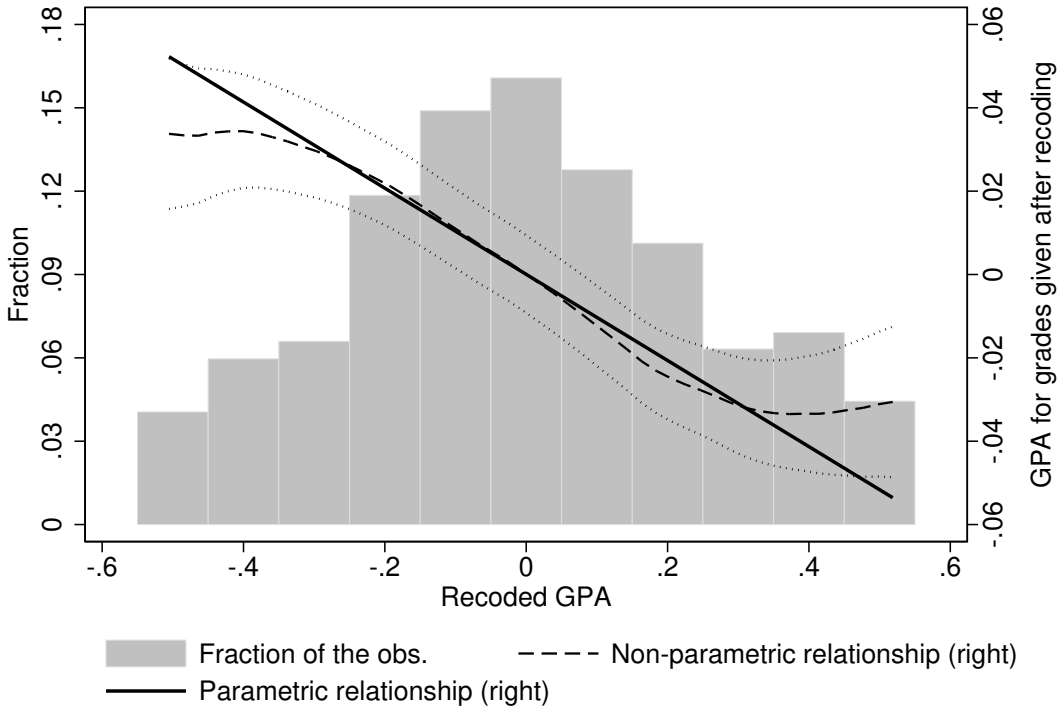


Figure 3: The relationship between the reform-induced GPA shock and subsequent grades.

Notes: The plot shows the relationship between the residuals from regressing respectively the recoded GPA and the GPA for subsequent grades on the all covariates, a second order polynomial in first year grades before recoding and school indicators. The dashed relationship is estimated using the epanechnikov kernel, with the a bandwidth based on the rule-of-thumb bandwidth estimator. The solid line is linear fit using ordinarily least squares. The dotted lines indicate the 95 percent confidence interval. The gray bars show the fraction of the observations (in percent). The figure excludes bottom and top 1 percent of the residuals from the recoded GPA.

To assess the sensitivity of the results presented in Table 6, we conducted a series of robustness checks. Table 7 presents results for a variety of specifications. Column (1) shows the results from estimating a model without covariates. The point estimate is close to the main result that includes

the full set of covariates. In our main analyses we condition on a second order polynomial of the pre-transformation GPA. In columns (2) and (3) we show results from estimating the model in equation (6) using a linear and cubic polynomial for the functional form $f()$. As the table shows, the results are not sensitive to the the choice of functional form.

Furthermore, in the main analyses the model estimates the impact of the change in GPA compared to the cohort GPA change. However, if students do not have access to the nationwide distribution of grades, they may instead compare their GPA change to that of peers at their school. In column (4) we show results from estimating a specification where the pre-recoding GPA is interacted with school indicators. The coefficient is, again, very similar to the main results. Finally, in column (5) we first residualize the recoded GPA (as in Figure 3) and then exclude the individuals in the bottom and top percentile of the residualized GPA. In other words, we exclude the individuals who experienced the largest changes due to the recoding. The coefficient is very close to the main specification.

Table 7: Regression results, alternative specifications. Dependent variable: Grades given after transformation (standardized).

	Without covariates (1)	Linear (2)	Cubic (3)	School specific (4)	No Outliers (5)
Recoded GPA	-0.090	-0.091	-0.082	-0.078	-0.087
	(0.021)	(0.019)	(0.018)	(0.018)	(0.017)
Mean of dep. var	-0.00	-0.00	-0.00	-0.00	-0.00
Observations	26,759	26,759	26,759	26,759	26,299
Clusters	209	209	209	209	209
R^2	0.44	0.60	0.60	0.61	0.59

Notes: The table shows point estimates and standard errors for β_1 in equation (6), estimated with ordinary least squares. See notes for Table 6. Column (1) shows results from estimating a specification without covariates and school fixed effects. Columns (2) and (3) show results from estimating a specifications with respectively linear and cubic polynomials in pre-recoded GPA. Column (4) shows results from estimating a specification where the polynomials in pre-recoded GPA are interacted with school indicators. Column (5) shows results from a specification where we exclude outliers in terms of treatment.

6.2 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 performance. One plausible mechanism would be that teachers systematically manipulate student scores in response to the reform. Earlier literature has focused on how the incentive structures associated with test-based accountability may cause teachers to intentionally manipulate standardized test scores (e.g., Jacob and Levitt, 2003; Neal, 2013). Lavy (2009) finds, on the other hand, that although a teacher incentive program in Israel increased teacher effort, it 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. (2016) 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, they also find that a teacher's propensity to manipulate a student's exam is influenced by the student's prior test scores.

If the teachers know the outcome of the recoding for individual students, they could be more generous to "unlucky" students. If teachers compensate students who are penalized by the grading reform, 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 earlier, each student receives both 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.[20]

Table 8 shows the results from using internal grades only (i.e., teacher evaluations in the second and third year, as well as exams that were partially graded by an internal examiner) and using external grades only (i.e., written exam grades in the second and third year). As Panel A shows, the results for internal grades are positive and precisely estimated. Although smaller in magni-

[20] 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, internal vs. external assessments.

		9th GPA		Gender		Parental edu.	
	Main	Low	High	Boys	Girls	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Dependent variable: internally graded grades							
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)
Mean of dep. var	0.00	-0.52	0.51	-0.11	0.08	-0.16	0.18
P-value (sub groups)			0.02		0.06		0.25
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)
Mean of dep. var	-0.00	-0.51	0.49	-0.00	0.00	-0.13	0.17
P-value (sub groups)		0.24		0.02		0.95	
P-value (int vs. ext)	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 (6), estimated with ordinary least squares. "P-value (int vs. ext)" provides the p-value for a test of equal coefficients on external and internal assessments. See notes for Table 6.

tude, Panel B shows that there are also effects when we use the average of the externally given grades as outcome. However, the difference in the main effects based on the internal and external assessment is significant; a pattern consistent with teachers manipulating scores for students that were unlucky. Nevertheless, the findings suggest that overall improvements in grades also reflects genuine performance improvement.

6.3 Does student labor supply respond to the shock in grades?

Having established that the reaction in terms of subsequent grades reflects improved performance, we now examine whether this is achieved by adjusting effort. Students who increase study effort may have to reduce the time spent on other activities. We assess this in terms of student labor supply during high school. We measure the labor supply for the calendar year 2007, which corresponds to the second half of the second high school year and the first half of the last high school year.

Panel A of Table 9 shows the results from using an indicator for whether the individual worked

for income as the dependent variable. Around 86 percent of the high school cohort worked during high school. There is no evidence of an effect of the GPA shock on this extensive margin of labor supply. However, Panel B of Table 9 shows that students reacted on the intensive margin. Using gross labor income measured in 1,000 euro (2015 level) as the dependent variable, we find that students who received a positive GPA shock increased labor income by, on average, 66 euro (corresponding to an increase of 1.1 percent, evaluated at the sample mean). That is, students who were downgraded due to the reform reduced the time spent on other activities.[21]

Although the subsample analysis in Panel B shows that the coefficients for all subgroups are positive, the labor supply response appears to vary somewhat in magnitude across subgroups. There is a relatively large labor supply response among students with parents with an average length of education below the median, which is in line with the main results that showed that this group also experienced a performance improvement in response to a negative shock. However, there is also some evidence of a larger labor supply response for students with a below-median middle school GPA than for students with an above-median GPA. Although the difference is not statistically significant at the 5 percent level, this pattern is in contrast to the performance response, suggesting that the relationship between educational improvements and time spent on work alongside studying is complex and affected by demographic characteristics. Finally, we do not find that girls and boys differ in their response in terms of their labor supply.

6.4 Effects of a GPA shock on the likelihood of post-secondary schooling

As students reacted to a reform-induced change in their first-year GPAs in terms of subsequent grades, the GPA shock could have long-run effects on human capital accumulation. Table 10 shows the effect of the GPA shock on enrolling in and completing a university degree within six

[21] An alternative conceivable mechanism would be that students who were downgraded responded by improving their educational achievement, which allowed them to work on the side. Moreover, student could potentially shift to better paying jobs in response to the grade shock. Nevertheless, the fact that we only find an effect on the intensive margin (and not the extensive margin) supports the notion that labor supply works as a mediator.

Table 9: Regression results: labor supply mechanisms.

		9th GPA		Gender		Parental edu.	
	Main	Low	High	Boys	Girls	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Dependent variable: labor income > 0							
Recoded GPA	0.002	0.009	-0.003	0.009	-0.001	0.006	-0.004
	(0.008)	(0.011)	(0.012)	(0.012)	(0.010)	(0.013)	(0.011)
Mean of dep. var	0.86	0.86	0.86	0.82	0.89	0.88	0.84
P-value		0.46		0.50		0.56	
B. Dependent variable: labor income (1,000 euro)							
Recoded GPA	0.251	0.409	0.083	0.234	0.236	0.406	0.072
	(0.087)	(0.132)	(0.119)	(0.153)	(0.108)	(0.129)	(0.122)
Mean of dep. var	5.75	6.23	5.29	5.85	5.68	6.23	5.30
P-value		0.07		0.99		0.06	

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

years of finishing high school.[22]

Panel A of Table 10 provides suggestive evidence of an enrollment effect. A one standard deviation decrease in the recoded GPA causes a two percentage point increase in the likelihood of enrolling in a university program. The effects are largest for girls and for individuals with a high GPA in middle school. This is in line with the main effect. Students respond to a negative shock by improving their effort and consequently are more likely to enrol in university.

Panel B shows that these effects also translate into graduation. The effects are even stronger (both larger and more precise). 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. Evaluated at the sample mean of 51 percent this corresponds to an increase of about four percent.

For the overall average relationship, there is no mechanical effect in the sense that the improved performance caused a higher GPA that gives access to more programs, because the student response

[22]We use educational status in 2014, six years after graduation, as this is the last year we observe the educational status.

Table 10: Regression results, long-run effects. Dependent variable: Completed university degree within six years of finishing high school.

		9th GPA		Gender		Parental edu.	
	Main	Low	High	Boys	Girls	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Dependent variable: Enrolled in a university program within three years after HS							
Recoded GPA	-0.017	0.011	-0.027	0.009	-0.036	0.007	-0.037
	(0.010)	(0.015)	(0.015)	(0.016)	(0.013)	(0.016)	(0.013)
Mean of dep. var	0.39	0.24	0.54	0.39	0.39	0.30	0.47
P-value		0.11		0.04		0.04	
B. Dependent variable: Graduated a university program within six years after HS							
Recoded GPA	-0.023	0.004	-0.038	-0.004	-0.040	0.000	-0.042
	(0.010)	(0.016)	(0.014)	(0.017)	(0.013)	(0.018)	(0.013)
Mean of dep. var	0.51	0.35	0.66	0.52	0.50	0.41	0.60
P-value		0.05		0.10		0.05	

Notes: Notes: The table shows point estimates and standard errors for β_1 in equation (6), estimated with ordinary least squares. See notes for Table 6. Higher educational covers 3-4 year programs like teachers college and nursing school. Adv. higher education are typically university programs.

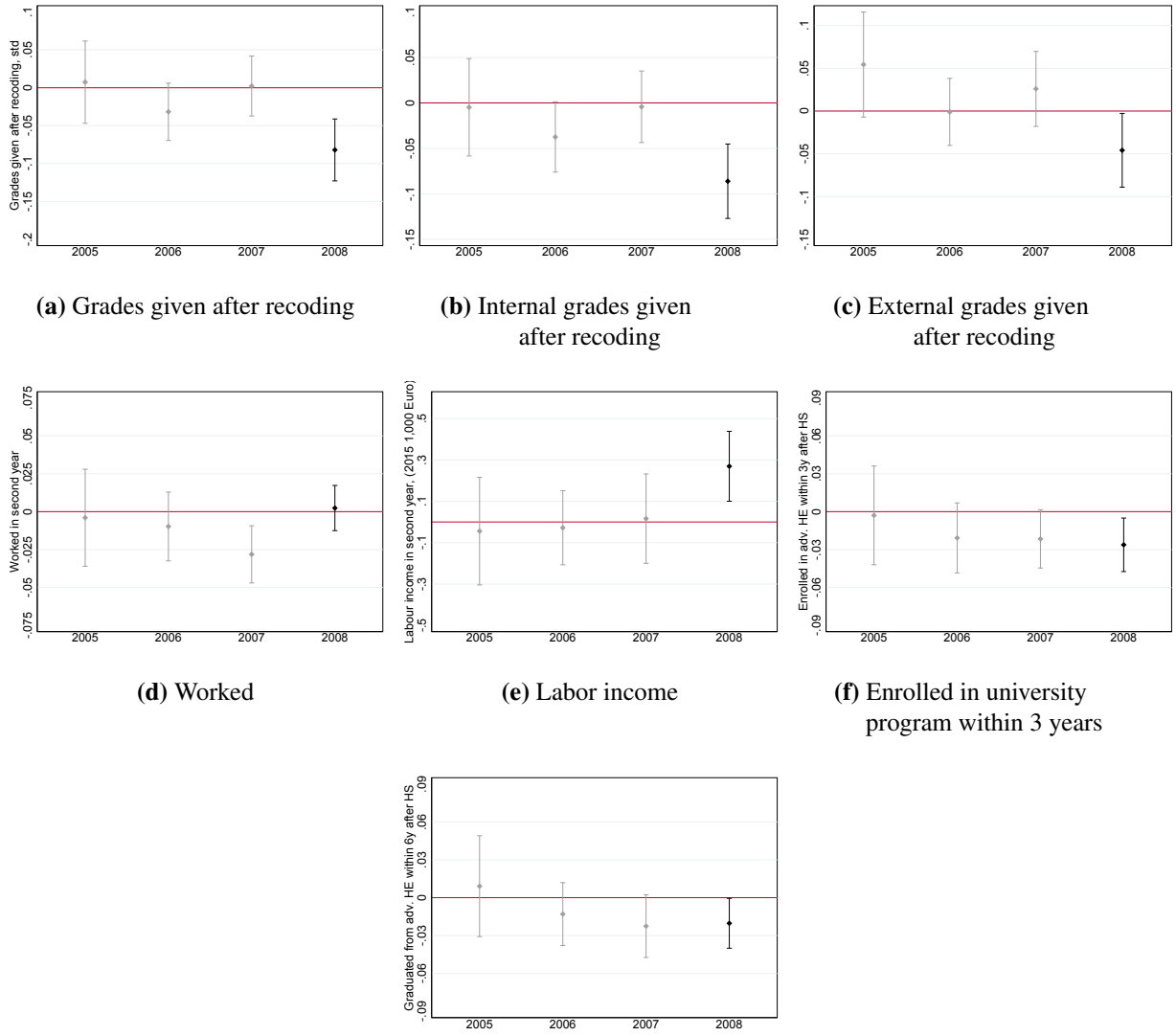
is insufficient to fully compensate for the negative shock. One explanation for this long-run effect could thus be that an increased study effort in response to a GPA change increases the students' exposure to academic material and therefore their aspirations for further education. A long-run effect on university graduation of about four percent, which is driven by behavioral changes, seems substantial. However, for some sub groups (i.e. for girls), the response was sufficient to compensate for the negative shock, which also explains why the long-run university enrollment and graduation effects are strongest for these groups.

6.5 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 earlier , the change in GPA caused by the recoding process is not related to observable characteristics. While this test is informative on whether the reform-induced GPA change is re-

lated to observable characteristics, it cannot inform us on 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 non-reform 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 scheme (i.e., the three cohorts prior to one affected by the reform). We impose the recoding on the first-year grades and proceed as described for the main analysis. As the covariates are not available for all placebo cohorts, these coefficients are all estimated without covariates (but with school fixed effects).

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 4 shows the results across outcomes. Compared to the estimates from main analysis, the estimates of the placebo GPA change are small. Moreover, only 1 out of 21 estimates is statistically different from zero at a 5 percent significance level. Given the large number of tests (and a 95 percent significance level that suggests that in expectation 1 out of 20 independent tests would turn out significant even if there is no effect) one significant estimate is not surprising. 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.



(a) Grades given after recoding

(b) Internal grades given after recoding

(c) External grades given after recoding

(d) Worked

(e) Labor income

(f) Enrolled in university program within 3 years

(g) Graduated from university program within 6 years

Figure 4: Placebo tests. Estimates by high school cohort. 2005-2007 are untreated cohorts and 2008 is the treated cohort. As our data do not include covariates for the 2005-2007 cohorts, all specifications are estimated without covariates, but with school fixed effects.

7 Conclusion

In this paper 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 downgrade of the first-year GPA causes them to do better in their second and third years. The

effects are larger for girls and 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. We also show that students who received a negative GPA shock reduced their labor supply while in high school, which may indicate that one channel through which the effect of the grade shock worked was that students increased their study effort and reduced time spent on other activities.

The findings indicate that students adjust labor supply, subsequent school performance, and college enrollment in reaction to a change in their GPA that is unrelated to their prior performance. The results appear relevant not only for the Danish setting but for educational systems in other countries as well. Although the Danish educational system differs in some respects from educational systems, the importance of exam grades resembles the high stakes of high school exams in other European countries and the United States. The findings may therefore be informative about how students respond to observed external shocks in outcomes of high-stakes assessments (e.g., computer breakdowns during exams or other exam conditions).

More generally, whereas previous literature has focused mainly on the signaling value of grades, our results suggest that the high stakes related to assessments are consequential for student behavior within the educational system. A deeper understanding of how high-stakes feedback on educational achievement affects student behavior remains an important goal for future research.

References

Apperson, J., Bueno, C., and Sass, T. R. (2016). Do the cheated ever prosper? the long-run effects of test-score manipulation by teachers on student outcomes. *mimeo*.

Avery, C., Gurantz, O., Hurwitz, M., and Smith, J. (2017). Shifting college majors in response to advanced placement exam scores. *Journal of Human Resources*, pages 1016–8293R.

Azmat, G., Bagues, M., Cabrales, A., and Iriberri, N. (2016). What you don't know... can't hurt you? a field experiment on relative performance feedback in higher education. *mimeo*.

Bandiera, O., Larcinese, V., and Rasul, I. (2015). Blissful ignorance? a natural experiment on the effect of feedback on students' performance. *Labour Economics*, 34:13–25.

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

Calsamiglia, C. and Loviglio, A. (2017). When having good peers is not good. *MIMEO*.

Dee, T. S., Dobbie, W., Jacob, B. A., and Rockoff, J. (2016). The causes and consequences of test score manipulation: Evidence from the new york regents examinations. *National Bureau of Economic Research Working Paper Series*, (22165).

Dee, T. S. and Jacob, B. (2011). The impact of no child left behind on student achievement. *Journal of Policy Analysis and management*, 30(3):418–446.

Diamond, R. and Persson, P. (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. *National Bureau of Economic Research Working Paper Series*, (22207).

Ebenstein, A., Lavy, V., and Roth, S. (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.

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

Humlum, M. K., Kristoffersen, J. H. G., and Vejlin, R. M. (2014). Timing of College Enrollment and Family Formation Decisions. *IZA Discussion Papers*, (7905).

Jacob, B. A. and Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, pages 843–877.

Jacob, B. A. and Wilder, T. (2011). Educational expectations and attainment. In Duncan, G. J. and Murnane, R. J., editors, *Whither Opportunity? Rising Inequality and the Uncertain Life Chances of Low-Income Children*. Russell Sage Press, New York.

Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. *The American Economic Review*, 99(5):1979–2021.

Murnane, R. J. (2013). Us high school graduation rates: Patterns and explanations. *Journal of Economic Literature*, 51(2):370–422.

Neal, D. (2013). The consequences of using one assessment system to pursue two objectives. *The Journal of Economic Education*, 44(4):339–352.

Papay, J. P., Murnane, R. J., and Willett, J. B. (2016). The impact of test score labels on human-capital investment decisions. *Journal of Human Resources*, 51(2):357–388.

Smith, J., Hurwitz, M., and Avery, C. (2017). Giving college credit where it is due: Advanced placement exam scores and college outcomes. *Journal of Labor Economics*, 35(1).

Stinebrickner, R. and Stinebrickner, T. (2014). Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model. *Journal of Labor Economics*, 32(3):601–644.

Stinebrickner, T. and Stinebrickner, R. (2012). Learning about academic ability and the college dropout decision. *Journal of Labor Economics*, 30(4):707–748.

Zafar, B. (2011). How do college students form expectations? *Journal of Labor Economics*, 29(2):301–348.

Appendices

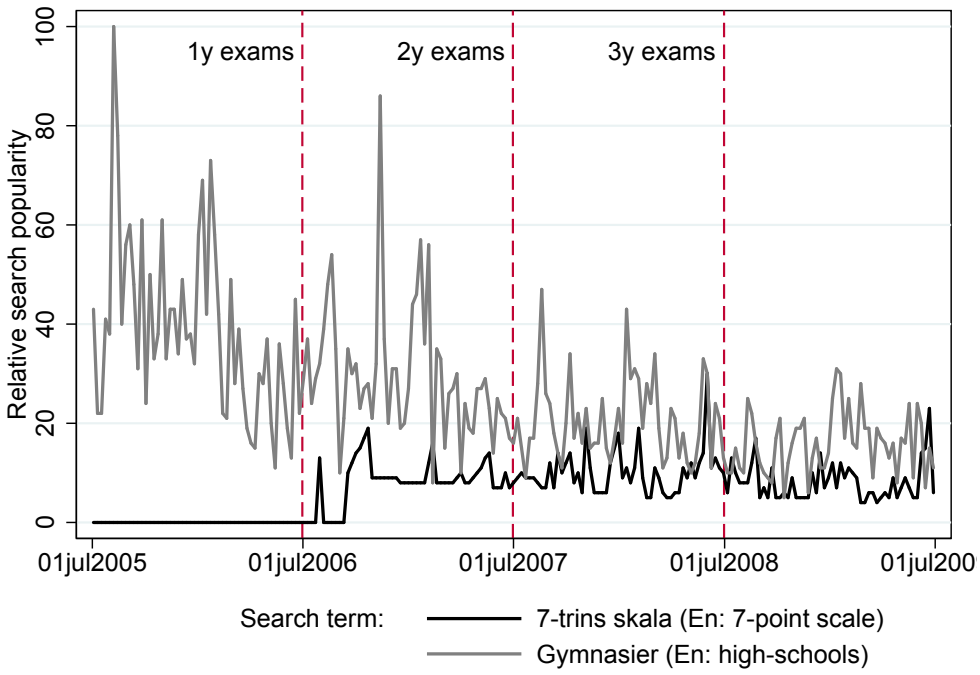


Figure A.1: Google Search Trend 2005-2009. The popularity is measured relative to the most popular search time/term for the period, which is set to 100.

Bevis for Studentereksamen (stx)
Aflagt i henhold til lovgivningen om de gymnasiale uddannelser

Navn: [redacted]
Cpr. nr: [redacted]
Eksamen er afsluttet juni 2008

	Årskarakterer			Prøvekarakterer			Særlige oplysninger		
Fag	Vægt	Karakter	ECTS	Vægt	Karakter	ECTS	Institution	Termin	Merit
Dansk A, mdt.	1	10	B	-	-	-			
Dansk A, skr.	1	10	B	1	10	B			
Engelsk A, mdt.	1	10	B	-	-	-			
Engelsk A, skr.	1	10	B	1	10	B			
Historie A	2	10	B	2	12	A			
Samfundsfag A, mdt.	1	10	B	-	-	-			
Samfundsfag A, skr.	1	12	A	1	10	B			
Spansk A, mdt.	1	10	B	1	12	A			
Spansk A, skr.	1	10	B	-	-	-			
Biologi B, mdt.	0,75	10	B	-	-	-			
Biologi B, skr.	0,75	10	B	-	-	-			
Matematik B, mdt.	0,75	10	B	0,75	12	A			
Matematik B, skr.	0,75	12	A	0,75	7	C			
Fysik C	1	10	B	-	-	-			
Idræt C	1	7	C	-	-	-			
Musik C	1	7*	C	-	-	-			
Naturgeografi C	1	10	B	1	10	B			
Oldtidskundskab C	1	7	C	1	12	A			
Religion C	1	10	B	-	-	-			
Almen studieforberedelse	-	-	-	2	10	B			
Studieretningsprojektet	-	-	-	2	10	B			

Studieretning: Engelsk A, Samfundsfag A, Matematik B

Studieretningsprojekt: Engelsk, Samfundsfag

Almen studieforberedelse: Dansk, Samfundsfag

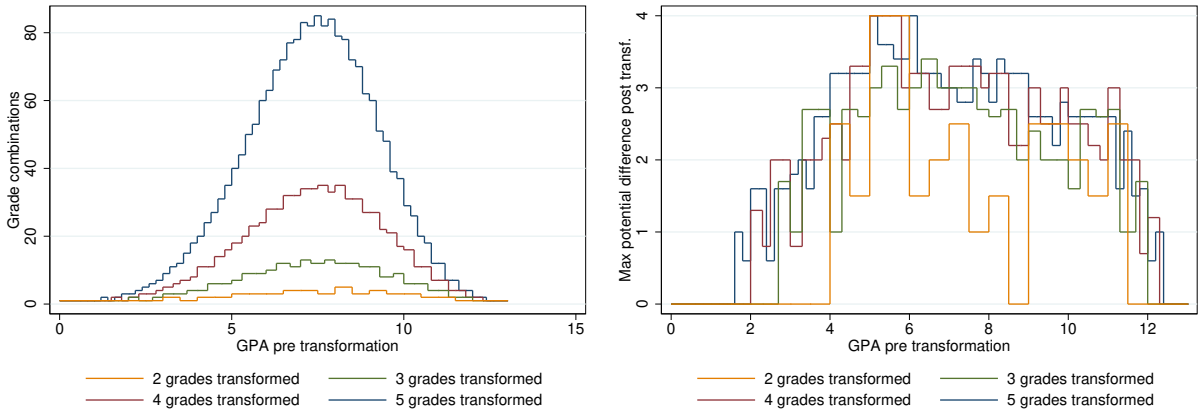
Foreløbigt eksamensresultat: **10,1**
Eksamensresultat: **10,4**

SILKEBORG GYMNASIUM

SILKEBORG GYMNASIUM
Oslovej 10 . 8600 Silkeborg
Tlf. 86 81 08 00 . Fax 86 81 26 06

Silkeborg Gymnasium

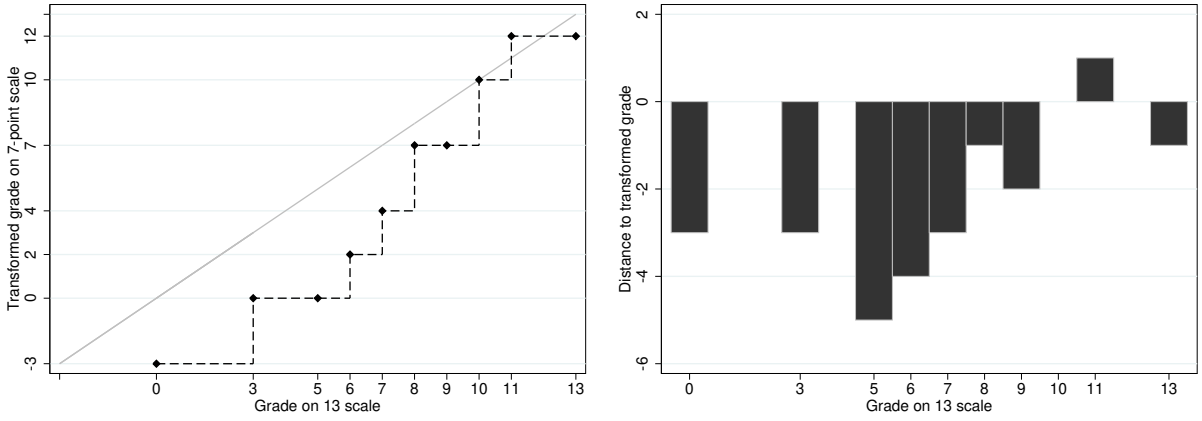
Figure A.2: High School Diploma for the treated cohort (graduates from 2008)



(a) Grade combinations

(b) Maximum difference in post-transformation GPA

Figure A.3: Combinations and maximum difference, given GPA and number of transformed grades.



(a) Mapping between grading scales

(b) Grade distance

Figure A.4: Mapping from 13 scale (old grading scale) to 7-point scale (new grading scale).

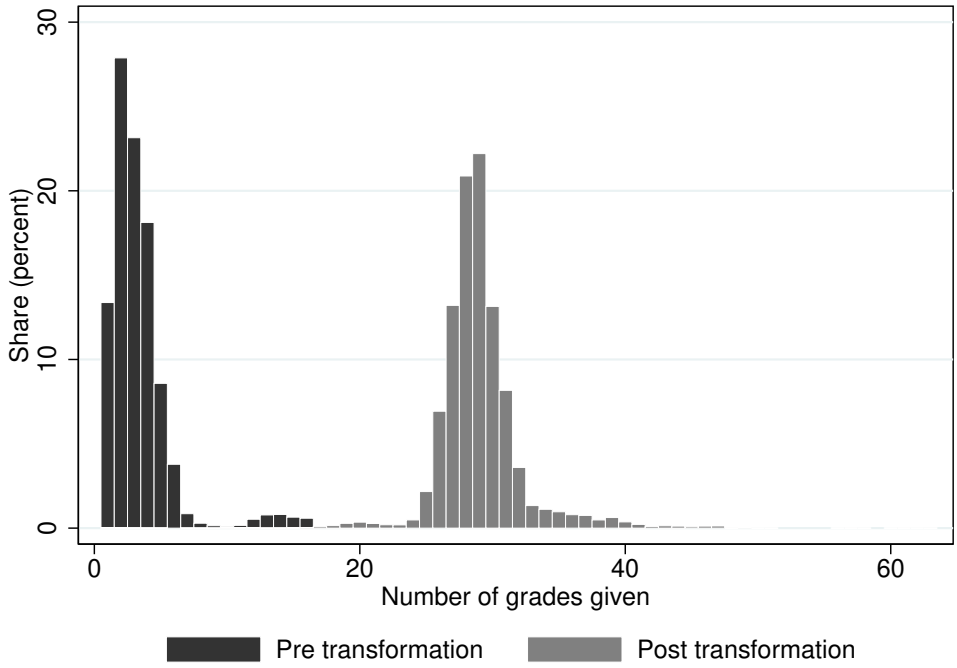


Figure A.5: The number of grades given before and after the transformation.

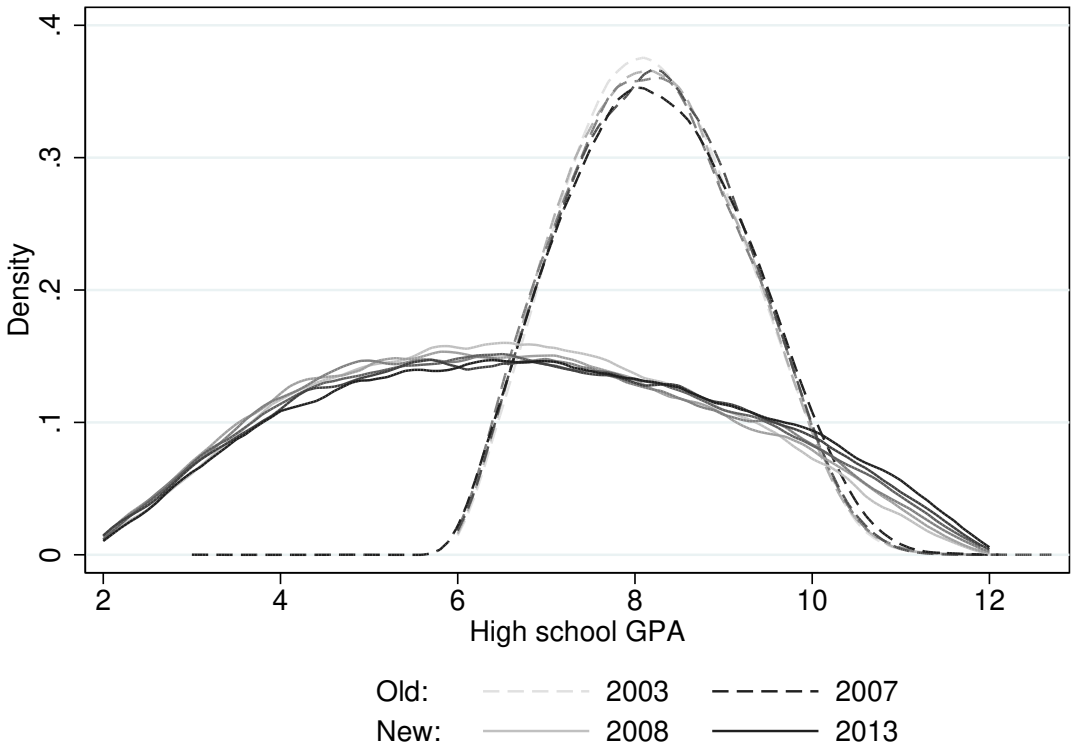
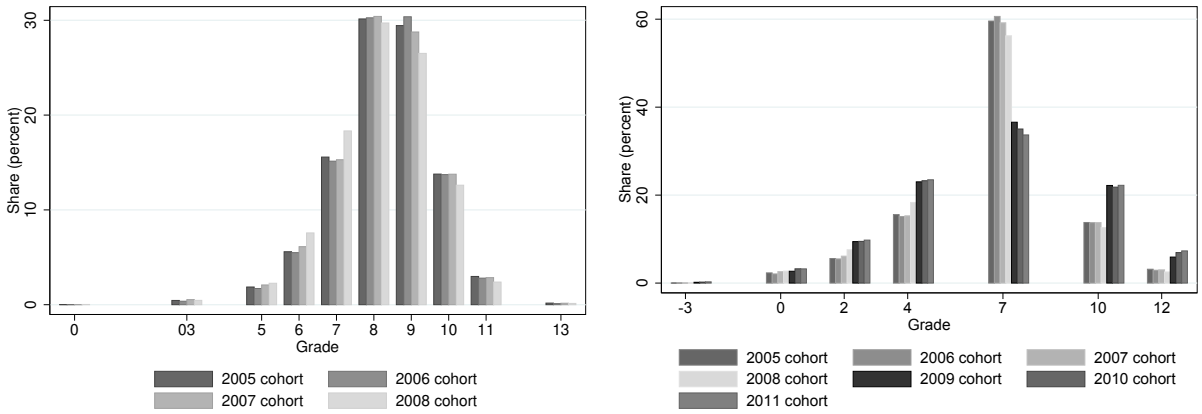


Figure A.6: High school GPA by cohort



(a) 13 scale

(b) Transformed to 7-point scale

Figure A.7: Grading patterns of 1st year grades by high school cohort

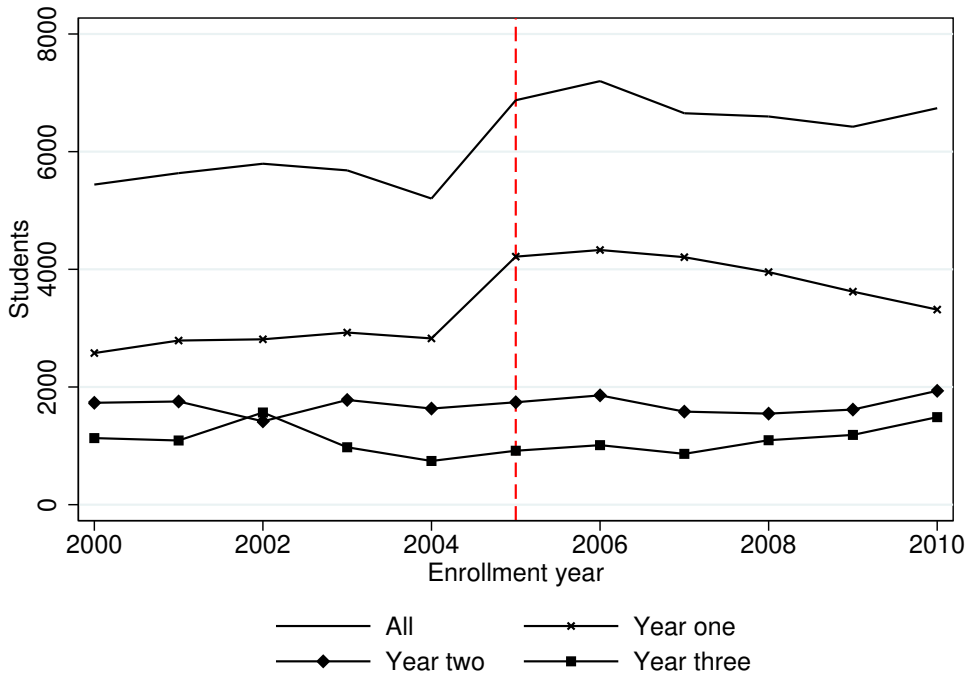


Figure A.8: High school enrollment and dropouts by year of enrollment, divided into groups according to the high school year they dropped out of high school.

Table A.1: The effect of a GPA shock on subsequent grades: Dependent variable: Grades given after transformation (standardized), with coefficients for original GPA.

		9th GPA		Gender		Parental edu.	
	Main	Low	High	Boys	Girls	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)	(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)
Original GPA	0.483	0.458	0.508	0.470	0.489	0.494	0.470
	(0.019)	(0.027)	(0.023)	(0.029)	(0.023)	(0.027)	(0.023)
Original GPA squared	0.052	0.070	0.014	0.059	0.046	0.068	0.040
	(0.004)	(0.006)	(0.006)	(0.005)	(0.004)	(0.005)	(0.005)
Mean of dep. var	-0.00	-0.54	0.53	-0.09	0.07	-0.16	0.18
P-value		0.03		0.04		0.32	
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 (6), estimated with ordinary least squares. The grade point averages are standardized to have a mean of zero and a unit standard deviation. We control for first-year GPA before recoding using a second order polynomial. The covariates included are age at high school entry, gender, 9th grade GPA (standardized) origin (indicator for non-western origin), parental education (years of completed education included, average across parents), income (disposable income, average across parents), the number of non-missing parental education and income observations (indicators). All parental variables are measured in the calendar year before the focal individual enrolled in high school. 9th GPA indicates that the sample is split by the median of the students' 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-hypotheses that the point-estimates are the same for the two respective subsamples. Standard errors clustered on the school level in parentheses.