Andresen, Martin Eckhoff; Løkken, Sturla Andreas

Working Paper

High school dropout for marginal students: Evidence from randomized exam form

Discussion Papers, No. 894

Provided in Cooperation with:
Research Department, Statistics Norway, Oslo

Suggested Citation: Andresen, Martin Eckhoff; Løkken, Sturla Andreas (2019) : High school dropout for marginal students: Evidence from randomized exam form, Discussion Papers, No. 894, Statistics Norway, Research Department, Oslo

This Version is available at:
https://hdl.handle.net/10419/210958

Statistisk sentralbyrå
Statistics Norway

High school dropout for marginal students: Evidence from randomized exam form

TALL SOM FORTELLER
DISCUSSION PAPERS
894

Martin Eckhoff Andresen and Sturla A. Løkken

Discussion Papers No. 894, February 2019
Statistics Norway, Research Department

Martin Eckhoff Andresen and
Sturla A. Løkken

High school dropout for marginal students: Evidence from randomized exam form

Abstract:
We exploit the assignment of exam form in a high-stakes Norwegian high school exam to estimate the impact of exam form on exam results, later school performance, graduation and longer run outcomes. Results indicate that written exams significantly reduce exam grades and reduce the probability of passing relative to the alternative oral exam, particularly for initially low-performing students. Because passing the exam is mandatory to obtain a high school diploma, this translates into reduced high school graduation rates that remain significant over time, permanently shifting a group of marginal students to drop out of high school entirely. IV estimates on labor market earnings are close to zero, but these results are too imprecise to draw firm conclusions.

Keywords: Exam form, high school dropout, returns to education

JEL classification: I21, I26, J24

Acknowledgements: We thank Lars Kirkebøen, Magne Mogstad, Edwin Leuven, Simon Bensnes, Kjetil Telle, Brita Bye and other colleagues at Statistics Norway for comments and feedback at various stages of this project. Funding from the Norwegian Research Council, grant no. 237840, is gratefully acknowledged.

Address: Statistics Norway, Research Department. E-mail: mrt@ssb.no/sal@ssb.no

Discussion Papers comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no/en/forskning/discussion-papers
http://ideas.repec.org/s/ssb/dispap.html

ISSN 1892-753X (electronic)

Sammendrag

Frafall i videregående skole er en utfordring i de fleste industrialiserte land. I Norge er det bred enighet om at frafall er et samfunnsproblem, og det er stor politisk vilje til å innføre tiltak som kan øke gjennomstrømmingen i videregående skole, slik som dem nylig foreslått av Livsoppholdsutvalget. Statistisk sett vil ungdom som ikke fullfører videregående utdanning med et vitnemål gjøre det dårligere i arbeidsmarkedet og på andre områder senere i livet enn de som fullfører. Denne forskjellen betraktes ofte som den potensielle gevinsten av tiltak som hjelper marginale studenter å få et vitnemål. En alternativ forklaring er at det er andre egenskaper ved disse elevene som både fører til frafall og dårlige utfall senere i livet. I så fall vil slike tiltak ha begrenset effekt ettersom de underliggende faktorene ikke blir addresert, men en slik seleksjon gjør også spørsmålet om effekten av et vitnemål vanskelig å besvare.

I denne studien undersøker vi effekten av å stryke på eksamen i andre klasse på studiespesialiserende videregående skoler på senere skoleresultater, vitnemål og inntekt for studenter som gikk i andre klasse i 2004 - 2007. For å håndtere seleksjonsproblemene nevnt ovenfor utnytter vi to egenskaper ved eksamenssystemet: at det er tilnærmet tilfeldig hvem som kommer opp i muntlig og skriftlig eksamen, og at studenter som kommer opp i skriftlig eksamen får systematisk dårligere karakterer. Dermed har vi et kvasi-eksperiment hvor tilfeldige studenter trekkes til skriftlig eksamen og dermed får dårligere karakterer og økt sjanse for stryk. Studenter som trekker skriftlig eksamen får eksamensresultater som er mer enn 0,6 karakterer dårligere og har 3,3 prosentpoeng høyere sannsynlighet for å stryke enn de som trekker muntlig eksamen.

Vi finner at studenter som stryker grunnet tilfeldig variasjon i eksamensform, har mye lavere sannsynlighet for å få vitnemål. Denne effekten viser seg dessuten å vare over tid, slik at flesteparten av de som ikke får vitnemål grunnet stryk på eksamen heller ikke får det senere i livet. Den tilfeldige trekningen av eksamensform skyver dermed en gruppe med marginale – og relativt svake – elever ut av videregående skole for godt. Av andre langtidsutfall ser vi at det er noe høyere sannsynlighet for å ta høyere utdanning blant dem som står på eksamen. Når det gjelder effekter på inntekt når elevene er 18-27 år gamle er estimatene for upresise til å trekke sterke konklusjoner, men effektene ligger nær null, som kan tyde på at et vitnemål kanskje ikke har så stor effekt for denne gruppen med elever.

Selv om trekning av eksamensform er tilfeldig, vil de negative konsekvensene av skriftlig eksamen kunne oppleves som vilkårlige. Når i tillegg svake studenter blir hardest rammet av denne vilkårligheten, kan en stille spørsmål om dette er en ønsket konsekvens av dagens eksamenssystem.

1 Introduction[1]

In many western countries, combating high school dropout and increasing high school graduation rates have been a long-term political goal. From the Educate America Act of 1994 through OECD's Education at a Glance reports (OECD, 2018) to more recent endeavors such as the Subsistence commission in Norway (NOU2018:13, 2018), policy-makers have proposed various policies and interventions to address this concern (Lamb et al., 2011). These concerns have been particularly strong in countries where dropout rates are increasing, such as the US (Card and Lemieux, 2001; Heckman and LaFontaine, 2010). When these policies are discussed, it is often with an implicit or explicit reference to the sizable differences in later life outcomes such as earnings, subsistence or welfare dependency between high school graduates and dropouts.

These differences can of course not be given a causal interpretation as the effect of finishing high school, as there can be strong selection into and out of high school. A large literature in economics have investigated the rates of return to education, with internal rates of return as large as 50 percent (Heckman et al., 2008) for high school completion, or around 11% per year for basic levels of schooling (Bhuller et al., 2017), suggesting that investments in schooling in general and high school in particular are highly profitable. Nonetheless, in a world with heterogeneous treatment effects, it is unclear how these or other estimates of returns to schooling apply to the marginal students targeted by policies aimed at reducing dropout rates. These students could for instance have lower than average treatment effects of finishing high school because they might learn less during high school even if they complete and because their diplomas will be weak even if they get one. It is unclear how the labor market values these marginal diplomas.

We aim to inform this question by exploiting the quasi-random allocation to exam form in Norwegian high schools. In second grade of high school at age 17 to 18, Norwegian high school students at the academic track are drawn to take one exam in a graduating course, either oral or written. The allocation is supposed to be random given course choices made the previous year, and we document how students allocated to the oral and written exam are statistically indistinguishable based on predetermined performance and family background. Being allocated the oral exam has three important consequences: First, the course teacher is part if the examination committee, and although the external examiner has the final word on the exam grade, this allows the schools to affect the grading. Second, a friendly face in the committee might help the student perform better and the exercise might be simpler than the centralized written exam. Third, some students may just perform better in oral versus written exams.

We show that the potentially positive consequences largely dominate this relationship,

[1]We thank Lars Kirkebøen, Magne Mogstad, Simon Bensnes, Edwin Leuven, Brita Bye, Kjetil Telle and other colleagues at Statistics Norway for comments and feedback at various stages of this project. Funding from the Norwegian Research Council, grant no. 237840, is gratefully acknowledged.

and being assigned the written exam has large negative effects on exam outcomes that we argue are causal, reducing grades and almost doubling the probability of failing. This, in turn, has important consequences for the students as this exam is mandatory in order to finish high school and get a diploma. Using the exam form as an instrumental variable and interacting with predetermined performance, we instrument both for the test score itself and for the event of failing the exam to look at further performance in school, finding large and persistent effects on later graduation. Among the marginal students shifted into failing by our instrument, more than 50% do not get a diploma on time and almost 50% still haven't acquired one at age 27. This could potentially have large consequences for these students, because further progression into tertiary education requires a diploma. Our estimates on longer term labor market earnings, however, are close to zero, indicating that the value of finishing high school might not be very high for these marginal students, but results are too imprecise to draw firm conclusion.

Our paper relates to four strands of the literature. First, as mentioned above, we relate to the large literature on returns to education in general and high school completion and GED certification in particular (Cameron and Heckman, 1993; Clark and Martorell, 2014; Heckman et al., 2014; Heckman and LaFontaine, 2006; Tyler et al., 2000). Second, we estimate effects of passing an exam on future performance in school, which could be though of as working either as a signaling effect within schools (Jacob and Lefgren, 2004; Manacorda, 2012) or through student motivation (Diamond and Persson, 2016). Third, the oral exam allows schools to apply local grading practices to the exam scores, in contrast to the centrally made and graded written exams, which relates our paper to a literature in grade inflation and manipulation of test scores (Andersland, 2017; Apperson and Bueno, 2017; Dee et al., 2016; Diamond and Persson, 2016; Jacob, 2005; Jewell et al., 2013; Lavy and Sand, 2018) Fourth and finally, we relate to a few papers on the effects of various shocks to exams on longer-term outcomes. These include local pollution (Ebenstein et al., 2016), quasi-random variation in preparation time (Bensnes, 2018) and the effect of the exam course itself (Falch et al., 2014).

This paper proceeds as follows: Section 2 presents the institutional setting of Norwegian high schools and exam allocation, Section 3 presents our Norwegian registry data, sample selection procedure and some summary statistics and Section 4 the empirical strategy. Results are found in Section 5, while section 6 concludes.

2 Institutional setting

The Norwegian compulsory education system is operated by the municipalities and consists of primary school (ages 7-12) and lower secondary school (ages 13-15). Children born after 1990 receive ten years of compulsory schooling and enroll in primary school the year they turn 6 years of age, while the children in our sample started school at age 7

and have 9 years of compulsory schooling before high school. The next tier of education is upper-secondary school (ages 16-18) which is operated by the counties and roughly corresponds to high school in the U.S. High schools are predominantly public[2] and free, and more than 95 percent of students enroll in high school the same year they finish compulsory schooling. A school reform in 1994 afforded all students the right to a high school education, but students still had to apply for admission and were allocated to high schools based on ranked preferences and final grades from lower secondary school. Students have a choice between vocational and academic track high schools, but we study only students enrolled in the latter. Academic track high schools last for three years (first, second and, third grade), and qualifies students for tertiary (college) education. For a more in-depth description of the Norwegian school system see Falch et al. (2014).

High school exams are either centrally or locally given. Central exams are prepared by the Norwegian Directorate for Education and Training (UDIR) which is the executive agency for the Ministry of Education and Research and given as nationwide written exams to all students at the same time. All central exams are graded by a randomly assigned external examiner.[3] Locally given exams are formally administered by the county administration, but is in practice partly delegated to the individual high schools. Locally given exams are typically oral exams[4] where the students are assessed by one internal (most often the subject teacher) and one external examiner. Both examiners should in principle agree on the exam grade, but the external examiner has the final word in case of disagreement. Exams are held in late May and June, written exams typically earlier than oral exams. Exams are graded on a numerical scale from 1 to 6, where 1 is a fail grade, 2 is the weakest passing grade, and 6 is the highest grade. The same scale is used on the teacher assessed internal grades. If a student fails the exam, or are sick on exam day, they are still permitted to progress to the next grade, and have the opportunity to re-take the exam in following semester.

Students are at risk of being selected for examination in all graduating courses throughout high school. A graduating course is defined as any course where the internal grade may end up on the diploma. At the end of third grade all students who passed all graduating courses and exams within stipulated time will receive a primary diploma.[5] All high school students must take several mandatory exams. In first grade, 20 percent of students sit for one exam in a graduating course. In second grade, all students sit for one exam in a graduating course. In third grade, all students take four exams: one written exam in Norwegian language, two written exams in graduating courses, and one oral exam in

[2] 8 percent of students attend private high schools (SSB Statistikkbanken)

[3] A subset of the assignments assessed by each examiner are cross-examined.

[4] Other forms of locally given exams either practical or oral-practical. The execution of these exam types are similar to oral exam for all practical purposes, and we group them together and contrast them to the centrally given written exams in the following.

[5] A primary diploma gives students some benefits when applying for colleges straight after high school. Students who change courses, re-take exams or who do not finish on time will get a regular diploma.

a graduating course. We exploit a feature of the second-grade exams: Exam form is as good as **randomly assigned** conditional on predetermined course choices. The process of allocating students to exams first starts in early spring when the county administration sends lists of centrally given exams to all high schools in the county. After the internal grades are determined, school administrators[6] assign students to written or oral examination, and sends a list of student-exam pairs back to the county administration. The county administration checks the list for course-level randomization and availability of external examiners before approving the school's proposal. Exams are then held in late spring or early summer.

At first glance, this process seems to introduce a possibility for school administrators to engage in grade manipulation by allocating students to their best matched exam, both in terms of exam course and type. However, several formal and informal restriction make such manipulation implausible. First of all, since each student must be assigned to exactly one exam, it is not possible for school administrators to simply select the best students for examination. It is not enough to only assess the potential outcomes of the individual students, but also their potential outcomes relative to all other students. Official instruction by the directorate also stipulates that exam allocation must be randomized and that oral exams must make up around 30 percent of the total exams.

Furthermore, school administrators also need to navigate several informal restrictions when assigning students to exams. First, oral exams are costly and there is a fixed cost component to arranging an oral examination. This means that schools not only want to minimize the number students selected for oral examination, but also the number of courses. Second, school administrators typically start out by planning the examination schedule for third grade students which have four exams each. The exam schedule for the second-grade students are then planned later, conditional on available dates, teachers, rooms, etc. Third, school administrators also try to rotate courses and teachers over time so as to even out the burden among the teachers. Finally, in some cases students may have certain course combination that limits the choice set of the school administrators, both when it comes to subject and type of exam.

All these formal and informal restrictions make any attempt by the school administration to inflate (maximize) the school level exam grades by manipulating student-exam assignments very difficult, not to mention the extremely detailed nature of the knowledge of students potential outcomes in various courses and exam types needed in order to efficiently execute such manipulation. In practice, the allocation of students to different exams is more of an accounting exercise where many competing concerns must be balanced, without much room for other motives.[7] This institutional setting provides us with an empirical strategy that may enable us to estimate the effect of exam form on

[6] Usually the principal, or in larger schools, an exam administrator.

[7] We talked to several school administrators who promoted this view.

exam outcomes: Conditional on school, year and subject, the form of examination should be as good as randomly assigned. In section 4, we provide evidence of this conditional randomization by showing that characteristics such as student performance, parent's education and gender are balanced across examination form. An important part of being

assigned an oral exam is that schools have the potential to apply **grade inflation**. As our first stage estimates reveal, students are systematically awarded better grades on oral exams than on written exams. This bias seems to be well-known among students and education professionals alike. Still, little has been done to investigate the causes of this phenomenon. One argument is that students are simply better at oral examinations, and that the examiners help the students display their potential. While this might be part of the story, another explanation is that schools and teachers engage in various forms of grade inflation which diffuse into oral exam scores.

In an analysis of grading practices in Norwegian compulsory schools, Galloway et al. (2014) find evidence for systematic school-wide differences in grading. They document a high degree of within-school correlation in subject grading practices, meaning that schools who systematically inflate grades in one subject often do the same in others as well. Furthermore, the researchers find that schools with low ability students are more likely to inflate the grades which suggests that there is some amount of grade normalization going on in the schools.

School level grade normalization is one potentially important source of grade inflation. All students apply for high school admission using the GPA from lower secondary school, this means the average ability of the student body varies across schools and over time with the admission cut-off. Still, schools tend to use the full range of the grading scale, and there might be strong internal pressure towards grading conformity within schools (from principal, other teachers, parents). If teachers normalize grading behavior to fit the ability distribution of the students, school level grade inflation will be an emergent feature. This means we will see larger grade inflation in schools with low ability student over high ability students. Another potential source of grade inflation is competitive pressure between schools in the same county/municipality. School administrators might encourage grading leniency when competing with otherwise similar schools for the same students.

Teachers might also engage in grade inflation on a individual student level. Teachers develop personal relationships with the students, and in some cases the families, which makes it uncomfortable to award bad grades. Another possibility is that teachers sometimes grade students conditional on person specific background information (i.e. she lost her mother, or her parents separated). This kind of practice is asymmetric in nature since teachers are unlikely to discount student performance based on positive factors. Hence, the effect on grades will be positive.

Since written exams are randomly and anonymously graded by external examiners using the same assessment guidelines and cross-validated between examiners, these grades are the best available measure of student ability. Internal grades are composite measures of student ability as well as school, teacher, and subject level grade inflation. Since oral exams are partly graded by the subject teacher, much of the same teacher and school level grade inflation will be included in the exam scores. This means that being assigned to a written exam implicitly will entail an exam score penalty. We provide evidence for this in Section A in the appendix.

In the following analysis we exploit both the random exam-form assignment and the written exam penalty to investigate the effect of grades on short and long-run outcomes.

3 Data and summary statistics

Our data comes from Norwegian administrative registers for the years 2004 – 2016. All data are hosted at Statistics Norway and contain unique individual identifiers that allow us to link students across registers and connect students to mothers' and fathers' characteristics. From the residency registers, we obtain data on municipality of residence, gender and birth dates. Enrollment data for high school comes from the national education registers, while the results from individual courses and exams are reported annually by each school to VIGO, who administers the IT-systems and databases for the counties operating high schools. Graduation data comes from the national transcript records (*"nasjonal vitnemålsdatabase"*), covering all issued high school diplomas with results in individual courses.

To measure longer-run outcomes, we use application data for tertiary education from the the centralized application procedure (*"Samordna opptak")* and data on the total years of finished education from the national education registers. These contain field of study and type of degree for all finished schooling at Norwegian institutions. Finally, we measure labor supply using yearly income data from tax records.

To **define our sample**, we start out with all youths alive and resident in Norway and turning 18 years of age in the years 2004 to 2007. Of these, we focus on those who start their second year of a three year academic track high school program in the respective year, ruling out redshirters, delayed students or those who start early.[8] We make a few other sample restrictions to ensure that our students are as comparable as possible, as documented in Table 1.

We then link these students to all registered courses with results, allowing us to compute a grade point average for each student based on the teacher-assigned grades before the exam. Next, we link these students to the results from the end-of-year exams. For

[8]Both early and late school start is uncommon in Norway.

Table 1: Sample selection

	2004	2005	2006	2007	Sum	%
Resident in Norway at age 17	55,385	56,933	60,276	62,231	234,825	100 %
Started second year of high school	49,525	50,987	54,396	56,152	211,060	89,9 %
Registered for only one study program	45,710	47,424	50,266	52,884	196,284	83,6 %
Academic track	22,426	23,033	24,573	26,123	96,155	40,9 %
Standard student status	21,928	22,738	24,244	25,439	94,349	40,2 %
Full time student	21,268	22,461	23,865	25,178	92,772	39,5 %
Registered for at least one course	20,110	21,210	22,512	24,467	88,299	37,6 %
...with exam	20,109	21,207	22,508	24,467	88,291	37,6 %
Takes at least one exam	18,495	19,711	19,839	22,747	80,792	34,4 %
At most one exam, valid teacher grade	18,036	19,626	19,807	22,696	80,165	34,1 %

around 7,500 students, or around 8.5% of the remaining sample, we cannot find any exam results and consequently have to drop them from the sample. The main reason for this is that our data lack the results from postponed or re-taken exams due to sickness, no-show or a failed exam. If this is endogenous to being assigned a written exam, this sample selection will potentially bias our estimates, but we return to this issue in section 4, where we show that the groups assigned written and oral exams are balanced with respect to pre-assignment performance and other characteristics. Lastly, we drop a few hundred students that are registered with more than one exam, which could happen if students take courses privately, without registering for instruction. These students are not subject to the standard exam draw.

This leaves us with a final sample of 80,165 students. **Summary statistics** for these students are given in Table 2, both for the sample as a whole and for students assigned oral and written exams separately. Girls are somewhat over-represented due to the sample selection criteria, and students are on average registered for almost 11 exam-eligible courses throughout the year, with no discernible difference between students with oral and written exams.

The students assigned oral exams, however, have somewhat higher GPA and teacher evaluated grade in the course they're assigned exam. This indicates that straightforward comparisons of results for students with written and oral exams will not reflect causal differences of the exam form. Rather, they reflect a mix of the effects of an exam form and the effect of courses with different difficulty having different exams. In practice, the courses chosen will affect both the exam form and likely the results on the exam, as some courses are more difficult than others.

Furthermore, Table 2 reveals large differences in the results of the exam depending on

Table 2: Summary statistics

	Full sample		Written exam		Oral Exam	
Variable	mean	s.d.	mean	s.d.	mean	s.d.
Female	0.54	(0.50)	0.53	(0.50)	0.54	(0.50)
Father's years of education	13.82	(3.51)	13.84	(3.48)	13.76	(3.59)
Mother's years of education	13.61	(3.45)	13.64	(3.42)	13.54	(3.52)
Number of exam-eligible courses	10.85	(1.43)	10.85	(1.42)	10.84	(1.45)
Teacher grade in exam course	3.88	(1.12)	3.82	(1.13)	4.03	(1.09)
Teacher evaluated GPA	4.03	(0.78)	4.05	(0.77)	3.97	(0.79)
Written exam	0.71	(0.46)	1.00	(0.00)	0.00	(0.00)
Grade at exam	3.66	(1.29)	3.42	(1.24)	4.22	(1.23)
Fails exam	0.04	(0.21)	0.06	(0.23)	0.02	(0.13)
Finishes grade 12	0.89	(0.32)	0.89	(0.32)	0.88	(0.32)
Starts grade 13 following year	0.99	(0.10)	0.99	(0.09)	0.99	(0.10)
Finishes grade 13 following year	0.80	(0.40)	0.80	(0.40)	0.79	(0.41)
Number courses in grade 13	9.60	(2.46)	9.61	(2.43)	9.58	(2.51)
Passed courses in grade 13	9.43	(2.59)	9.44	(2.56)	9.39	(2.65)
GPA in grade 13	4.05	(0.86)	4.08	(0.85)	3.99	(0.87)
Obtains primary diploma on time	0.72	(0.45)	0.72	(0.45)	0.70	(0.46)
Obtains any diploma on time	0.73	(0.44)	0.74	(0.44)	0.71	(0.45)
Labor earnings at age 27, 1,000 NOK	382.55	(230.15)	386.40	(231.62)	373.28	(226.30)
N	80,165		56,672		23,493	

exam form. Students assigned a written exams score on average .8 grades lower and have almost 4 percentage points higher probability of failing the exam than students assigned an oral exam. Nonetheless, they have higher probability of both finishing second grade and obtaining a high school diploma on time. Our sample of students take exams in a total of 124 courses over the four year window. 45 of these courses have students sitting for both written and oral exam(s) in the sample window, and the empirical strategy we detail below exploits this variation between students who sit exams in the same course, but with different exam forms.

4 Empirical strategy

The most straightforward way to evaluate the effect of exam form on immediate and medium-run outcomes is to simply compare the exam results and graduation rates of students with an oral and a written exam. This ignores, however, that the assignment of

exam form may be endogenous. Although exam form is supposed to be drawn randomly by school administrators, this is done conditional on course choices made by the students. In particular, courses which more often have written exams may be of a different difficulty than courses with oral exams, and students' may have different strategies for choosing courses, leading simple outcome comparisons of results for written and oral exams to be biased.

To account for this, it is crucial to control for course choices. We do this by controlling flexibly for fixed effects in a two stage least squares setup

$$\begin{aligned} y_{ikt} &= \alpha_k + \pi_t + \kappa_{c(i)} + \beta s_{ikt} + \epsilon_{ikt} \\ s_{ikt} &= \delta_k + \eta_t + \lambda_{c(i)} + \gamma Z_{ikt} + \mu_{ikt} \end{aligned} \tag{1}$$

Where y_{ikt} is an outcome for student i at school k in year t. where α_k and δ_k are school fixed effects and π_t and η_t are year fixed effects\. The fixed effects set $\kappa_{c(i)}$ and $\lambda_{c(i)}$ controls for course choices. In our baseline specification, this contains fixed effects for the exam course of student i, in practice ensuring that we compare students with different exam type in the same course to estimate the effect of exam form. We show in the next section that results are similar if we additionally control for the full menu of potential exam courses. Finally, s_{ikt} is the exam score obtained by student i, which we instrument using the exam form Z_{ikt}, a binary variable equal to 1 if student i was assigned a written exam. The two parameters of interest are γ, reflecting the first stage effect of the exam form on the exam score, and β representing the IV estimate of the effect of the score on the outcome for the compliers.

As with all IV applications, we rely on relevance, exclusion and monotonicity assumptions. The relevance assumption is testable, and we show in the next section that the instrument indeed affects the exam outcome. The exclusion restriction requires the instrument to affect the outcome y only through the treatment s. In order for β to be interpreted as a causal effect of the exam form itself, we need exam form to be as good as randomly assigned given the sets of fixed effects. As discussed in Section 2, the institutional setting provides a promising background for such conditional randomization. Nonetheless, we cannot rule out the endogenous allocation of students to exams, given the discretion afforded to local school administrators when assigning exams. We therefore rely on balancing tests to show that the samples of students assigned different exam forms are statistically indistinguishable when controlling for course choice. Panel A of Table 3 presents balancing tests for the assignment of exam form. In row 1, we condition only on year fixed effects. It is evident that the two samples are not balanced with respect to predetermined variables: Students assigned written exams come from families with higher education, have higher GPAs but, somewhat surprisingly, have lower scores in the course they're assigned an exam than students assigned oral exams. These differences are strongly significant. Moving to row 2, we add school fixed effects to account for any

Table 3: Balancing tests

A: Simple specification (eq. 1)

Specification	teacher grade	GPA	mother has higher education	father has higher education	female
Year fixed effect	-0.209***	0.0874***	0.0133***	0.00649	-0.00573
	(0.0175)	(0.0153)	(0.00479)	(0.00630)	(0.00589)
add school fixed effects	-0.179***	0.107***	-0.0112	0.0160***	-0.0110**
	(0.0163)	(0.0127)	(0.00846)	(0.00470)	(0.00554)
add exam course F.E.	-0.0245	-0.0211	-0.0131	-0.00558	-0.00464
(baseline)	(0.0238)	(0.0151)	(0.0115)	(0.00758)	(0.0102)
add course menu F.E.	-0.0151	-0.0194	0.00649	0.00302	-0.0156
	(0.0241)	(0.0156)	(0.00630)	(0.00952)	(0.0111)
N (spec. 3)	80,138	80,138	80,143	80,143	80,138

B: Interacted specification (eq. 2)

written exam × Teacher grade	GPA	mother has higher education	father has higher education	female
teacher grade 1	-0.142	-0.000582	-0.0227	-0.0800
	(0.117)	(0.0578)	(0.0715)	(0.0651)
teacher grade 2	-0.0457	-0.0124	-0.0455*	0.00287
	(0.0286)	(0.0265)	(0.0264)	(0.0251)
teacher grade 3	-0.0134	-0.0258	0.0136	-0.0145
	(0.0183)	(0.0163)	(0.0148)	(0.0169)
teacher grade 4	-0.0144	0.00196	-0.00248	-0.00326
	(0.0154)	(0.0166)	(0.0166)	(0.0171)
teacher grade 5	0.0470	-0.0157	-0.00114	-0.0212
	(0.0347)	(0.0370)	(0.0391)	(0.0382)
teacher grade 6	0.00127	-0.00314	-0.00693	0.0107
	(0.0145)	(0.0138)	(0.0145)	(0.0150)
N	79,951	79,951	79,951	79,951
F-stat	1.246	0.443	0.708	0.621
p-value	0.282	0.850	0.643	0.714

Notes: Panel A shows balancing tests for the simple specification where the coefficients presented are for a written exam dummy and the balancing variable are shown in the column header, using controls as specified in rows (1) through (4). Panel B shows balancing tests for the same variables in the interacted specification, using specification (3) from panel A. Standard errors are clustered at school ($G = 327$) and robust to heteroskedasticity. Singleton observations are dropped, and the number of observations thus vary somewhat over fixed effects sets in panel A. Reported number of observations are for the baseline specification in row 3. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$

across-school differences in grading practices that correlate with course choices and/or exam form. This does not alleviate the problem, but when adding exam course fixed effects in row 3, all balancing tests shows insignificant differences. This is largely driven by the point estimates dropping, not the standard errors increasing due to the added sets of fixed effects. Our interpretation is that this accounts for the effect of individual course choices: Because some courses are easier than others and the likelihood of being assigned a written exam differ from course to course, unobserved ability will affect both the likelihood of getting a written exam and the performance on the exam. In row 4 we introduce a more robust course menu fixed effects, leading to a large drop in observations, without the balancing tests changing dramatically. Our preferred specification is therefore the one in column 3, including exam course fixed effects.

With a multivalued treatment such as an exam score, it is well known that the estimated coefficient β represents a weighted average of effects for different complier populations defined in terms of counterfactual outcomes (Angrist and Imbens, 1995). In our particular example, the instrument is likely to shift students exam scores down, both from high to somewhat lower grades, but also from passing to failing grades. Without further assumptions, we cannot know the relative size of the threshold specific treatment effects, and large potential treatment effects for example at the fail margin could be masked by no effects for other treatment shifts. In particular, it is tempting to redefine the treatment variable to a binary indicator $P = \mathbb{1}(s_{ikt} \geq 2)$ equal to 1 if the student passes, because the effect of failing a mandatory exam might potentially be large on later outcomes such as high school completion and earnings. In order for the exclusion restriction to be valid for this treatment defintion, we need the instrument to a) only affect the pass/fail margin or b) other instrument-induced shifts in exam scores to have no effect on outcomes (Andresen and Huber, 2018).

We therefore expand our empirical strategy to be able to identify the effect of exam score and passing the exam simultaneously. To this end, we exploit the fact that students with different skills in the exam course, as measured by their teacher assessment in the course given before the exam, are affected on different margins. Specifically, we instrument both for the exam score s and the probability of passing P by interacting the entire specification from before with the internal grade g the student received in the exam course from the teacher prior to the exam form assignment:

$$\begin{aligned} y_{igkt} &= \alpha_{gk} + \pi_{gt} + \kappa_{ig} + \beta_s s_{ikt} + \beta_S P_{ikt} + \epsilon_{ikt} \\ s_{igkt} &= \delta^s_{gk} + \eta^s_{gt} + \lambda^s_{gc(i)} + \textstyle\sum_{g=1}^{6} \gamma^s_g \mathbb{1}(g_{ikt} = g) Z_{ikt} + \mu^s_{ikt} \\ P_{igkt} &= \delta^P_{gk} + \eta^P_{gt} + \lambda^P_{gc(i)} + \textstyle\sum_{g=1}^{6} \gamma^P_g \mathbb{1}(g_{ikt} = g) Z_{ikt} + \mu^F_{ikt} \end{aligned} \tag{2}$$

This specification of the first stages amounts to running the above specification separately within groups of teacher assessed grades and using parametric assumptions on the effects at other margins to be the same to separately identify the effect of passing the exam

and improving grades at other margins . The balancing tests for these six instruments are provided in panel B of table 3. Compared to the simple specification, we cannot use the teacher grade in the exam course itself as a balancing variable, because it is absorbed by our fixed effects, but for the other four outcomes we see only one borderline significant difference for all six instruments and 4 balancing variables, and the joint F-values are small and far from significant.

For the IV estimate to be interpretable as a local average treatment effect, traditional IV analysis rely on a monotonicity (or rather, uniformity) assumption: The instrument(s) cannot affect some people's exam scores positively and others negatively. This is, at first glance, a potential problem in our application. Oral and written proficiency can be seen as skills, and it is plausible that some students perform better at written than oral exams. Remember, however, that oral examination also allows schools and teachers to influence the grades of their students because the internal teacher is one of the examiners, allowing the teacher who knows the student to potentially help the student perform better. Moreover, this allows the school to affect grades directly, independently of the performance of the student, which is impossible for the centrally given and graded written exams. Only students who are relatively much better at written than oral exams, so that this effect dominates the grade inflation practice of the school and any help the teacher may provide during the examination, will be potential defiers in our setup. In Section A in the appendix we provide evidence that supports the monotonicity assumption, and show that any group of potential defiers will likely be relatively small. Furthermore, de Chaisemartin (2017) show that 2SLS still identifies a LATE when the monotonicity assumption is weakened. As long as there exists a sub-group of the compliers that have the same distribution of treatment effects as the defier group and there are more compliers than defiers, the IV still estimate the LATE for the remaining compliers. If the complier group is large relative to the defiers, the IV estimate will closely approximate the LATE for the compliers.

Finally, it is important to emphasize the local nature of our IV estimates. In the presence of treatment effect heterogeneity, our estimates of β will represent the causal effect of exam score or passing the exam on the outcome for the population of compliers: Students who do worse on the exam because they are assigned a written exam. This group may be strikingly different from the average students. In particular, we argue that the compliers on the pass margin are likely to be very weak students. They may or may not respond similarly to the average student who fails an exam. We argue, however, that the compliers to our instruments are a very policy relevant group: They represent marginal students on the margin of dropping out of high school. As such, our results may be particularly interesting for other policy interventions that aim to increase high school completion, as these interventions are likely to affect students similar to our compliers if they work at all.

5 Results

Given the balancing tests reported in the previous section, we believe that exam form is as good as randomly assigned given fixed effects that control for course choice. Next, we present estimates of the immediate effect of exam form on various margins of exam scores. Following this, we present IV estimates of other schooling outcomes in Table 5, and finally long run estimates on high school graduation and labor market earnings in Figure 1.

Panel A of Table 4 shows the direct impact of a written rater than an oral exam on binary indicators of each exam score and on the linear score s_{ikt} in column 7. We see that being assigned a written exam increases the probability of failing the exam by around 3.3 percentage points compared to the oral exam, around 75% of the mean in the sample. Written exams decreases performance for students with all initial levels of performance, even the top grades 5 and 6. On average, students assigned the written exams get around .62 lower grades on the exam than students assigned the oral exams. These effects are strongly significant, as evident by the large F-statistic on the instrument.

Moving to panel B, we use the interacted specification to see how different students are affected differently. Unsurprisingly, the effects on the pass margin is concentrated at the lower end, with students with a teacher assigned grade 2 being 12 percentage points more likely to fail when given a written rather than an oral exam and the corresponding number for grade 3-students being 5.1 percentage points. Even top students with grades 5 and 6, are shifted into passing by the instrument. Moving down the table, the impact for higher performing students are concentrated at the top of the exam score distribution, reducing the probabilities of obtaining the top grades. Taken together using the linear score in column 7, the aggregate effect of a written exam on the linear score is relatively flat at -.5 to -.6 points for students with initial grades from 3 to 6, and somewhat smaller for the weak students. The joint F-statistics for these first stages are shown at the bottom of the table, confirming that our six instruments strongly affect exam scores also in the interacted specification. Although using all these six margins of response is interesting in order to see how the instruments affect exam scores, using six endogenous variables in our IV strategy makes interpretation of the estimates cumbersome. We therefore in the following revert to the specification in eq. 2, where we instrument for a) a dummy for passing the exam and b) the linear exam score itself. This allows us to disentangle the effects on the compliers at the pass margin, where we expect potential impacts on later outcomes to be large, from the effects on compliers on the score margin. Results of this specification is found in Table 5, where outcomes are ordered by timing from top to bottom. Starting in row 1, passing the exam unsurprisingly has a large effect on the probability of successfully finishing second grade, as passing the exam is required to pass. As the students have some opportunities to re-take the exam, the coefficient is not

Table 4: First stage effects of a written exam

A: Simple specification

	Exam score					linear	mean
	≥ 2	≥ 3	≥ 4	≥ 5	$= 6$	score	Z
written	-0.033***	-0.112***	-0.194***	-0.188***	-0.088***	-0.616***	0.707
exam	(0.004)	(0.009)	(0.012)	(0.010)	(0.007)	(0.034)	
F-stat	78.5	155.8	249.6	336.6	144.5	320.3	

B: Interacted specification

teacher grade	Exam score					linear	mean
× written exam	≥ 2	≥ 3	≥ 4	≥ 5	$= 6$	score	Z
1	-0.135	-0.041	0.010	0.001	-	-0.165	0.794
	(0.097)	(0.039)	(0.018)	(0.011)	-	(0.127)	
2	-0.120***	-0.165***	-0.057***	-0.008	0.000	-0.350***	0.771
	(0.017)	(0.028)	(0.016)	(0.006)	(0.000)	(0.049)	
3	-0.051***	-0.205***	-0.237***	-0.090***	-0.009***	-0.593***	0.741
	(0.007)	(0.019)	(0.021)	(0.011)	(0.003)	(0.044)	
4	0.002	-0.015***	-0.132***	-0.323***	-0.213***	-0.681***	0.704
	(0.002)	(0.004)	(0.011)	(0.021)	(0.019)	(0.045)	
5	-0.003**	-0.001	-0.031***	-0.135***	-0.300***	-0.470***	0.660
	(0.001)	(0.005)	(0.009)	(0.026)	(0.041)	(0.064)	
6	-0.012***	-0.103***	-0.270***	-0.232***	-0.055***	-0.672***	0.658
	(0.002)	(0.009)	(0.017)	(0.015)	(0.007)	(0.037)	
F-stat	19.3	38.8	54.5	73.4	34.3	75.4	N
mean dep.	0.955	0.800	0.547	0.279	0.076	3.66	79,951

Notes: Each column shows results from separate regressions of the outcome as specified in the column header on a) a written exam dummy in panel A and b) the written exam dummy interacted with teacher grade in the exam course in panel B. All regressions control for school, year and exam course fixed effects, all interacted with dummies of the teacher grade in panel B. Singleton observations in any regression are dropped in all to keep a consistent sample. Standard errors are clustered by school ($G = 327$) and robust to heteroskedasticity. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$

Table 5: IV effects of exam results on school outcomes

	Outcome	Passes exam		Exam score		N
(1)	Finishes second grade same year	0.774***	(0.146)	-0.000373	(0.00836)	79,951
(2)	Starts third grade following year	0.0314	(0.0867)	-0.00500	(0.00345)	79,951
(3)	Finishes third grade following year	0.369*	(0.190)	-0.0351**	(0.0145)	79,951
(4)	Number of courses following year	0.776	(1.142)	-0.0508	(0.0883)	79,951
(5)	Passed courses following year	1.984*	(1.199)	-0.0577	(0.0894)	79,951
(6)	GPA following year	0.547*	(0.307)	0.00963	(0.0209)	76,838
(7)	GPA in similar courses	0.815**	(0.373)	-0.0236	(0.0265)	71,655
(8)	Primary diploma on time	0.630***	(0.167)	-0.00940	(0.0127)	79,951
(9)	Any diploma on time	0.544***	(0.183)	-0.0120	(0.0130)	79,951

Notes: IV estimates using written exam interacted with teacher grade as the instrument for both the continuous grade on the exam and a dummy for failing the exam. Each row contains results from a separate run of our IV model using the outcome in the row header as dependent variable. Controls include school, year and course fixed effects interacted with teacher grade. Standard errors in parenthesis, clustered at the school level. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$

1, but almost 80 percent of the students induced to fail the exam by the instruments do not finish second grade on time. There is, however, no effect on the probability of starting third grade because having passed second grade is no prerequisite for starting third grade. Compliers are however less likely to complete third grade the following year (row 3). There is a lone significant coefficient for exam score for this variable, indicating a surprising negative impact of better exam scores on the probability of completing third grade the following year.

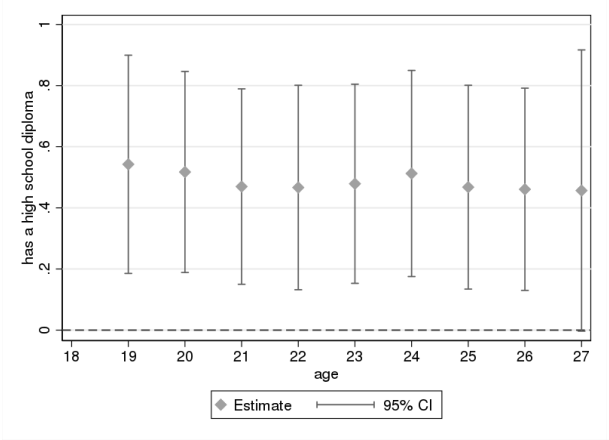
We find no impact on the number of courses the passing students register to, but they pass almost 2 more courses during third grade (this estimate is only significant at the 10 percent level). Their overall GPA on courses in third grade is more than half a grade point higher than failing students, a large effect corresponding to around 65 percent of a standard deviation of this variable. Note, however, that we lose a few thousand students for this outcome because they do not finish any courses so we can calculate their GPA, an event that could potentially be endogenous to failing the exam. We next investigate whether this is concentrated within courses similar to the exam course by calculating the GPA in courses within four rough groups (STEM, language, social sciences and humanities, other) and using the GPA for the course group of the exam course in row 7. The estimate of -.82 is significant and large, indicating stronger effects on courses in this group than other courses, but keep in mind that we lose around 10 percent of the sample for this outcome, which is potentially endogenous. These results could point to effects of passing the exam working through motivation and performance in later school outcomes.

Finally we estimate the effects on graduation in the last two columns. We find strong effects on the probability of obtaining a primary diploma: students who pass because of the exam form are 63 percentage points more likely to obtain a primary diploma the following year, as the possibilities for re-sitting and re-taking exams are limited if you want to obtain a primary diploma. Our estimate on the probability of having any high school diploma on time is smaller than the estimate on having a primary diploma, indicating that some of the compliers on the fail margin manage to redo the exam and get an ordinary diploma on time.

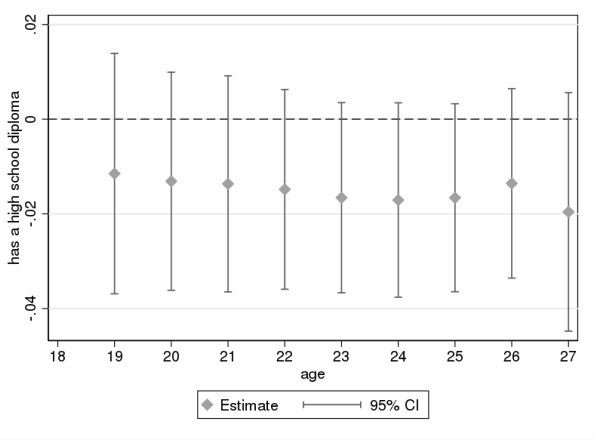
More than 50% of the compliers who fail due to the exam form is however still without a diploma by the year they were supposed to graduate, and we might wonder whether they are just delayed in redoing the courses and exam or whether they have dropped out altogether. To understand this, we estimate the impact on the probability of having any high school diploma for all years we can measure in the graduation data. We use the specification from equation 2 separately by age to see if the effects on graduation fall over time, plotting the results in the top panels of figure 1. The estimates are remarkably stable over time, decreasing only slightly, and students who fail the exam because of the exam type are still almost 50% less likely to have a high school diploma 8 years after they were scheduled to finish high school. This is an indication of clear long-term effects on graduation rates for a group of marginal students: students who manage to get a diploma on time if they're assigned the easier oral exam, but who fail and drop out completely when assigned the written exam. Thus, the group of compliers to this instrument could be considered on the brink of dropping out, and are unable to recover from failing the exam due to the exam form. Although this complier group is unique to the instruments at hand, they might be similar to compliers to other policy relevant instruments that aim to increase high school completion.

A natural next step is to ask how the lack of a high-school diploma affects these marginal students, both in terms of later education and earnings. We do this in the bottom panels of Figure 1, where we use a dummy for whether a person is enrolled in tertiary education and labor earnings over time for each year from 18 (the year of the exam) through 27 as the outcome of our IV model. As before, we keep the specification from eq. 2 and run the model separately by age. Unfortunately, precision is relatively low for this outcome, but there is indication of an increased probability of being enrolled in higher education for the students who pass the exam, although these are significant only at the 10% level at ages 22 and 23.

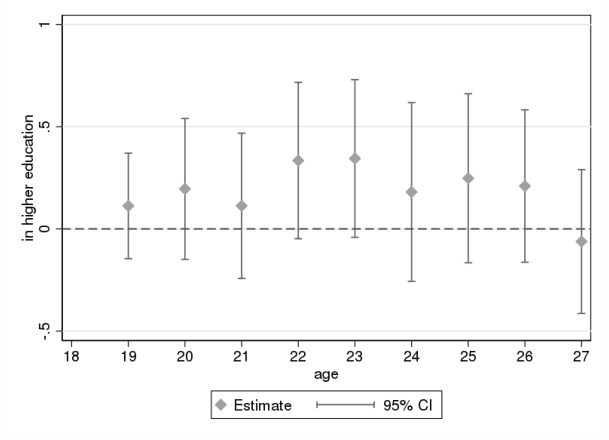
The estimated effects of passing the exam on labor income is plotted in the bottom left panel with associated 95% confidence intervals. In light of the suggestive evidence from above that finishing the exam may affect tertiary education, positive impacts on earnings need not necessarily translate into long-term effects on education, because enrolling in tertiary education means the student will enter the labor market later. The coefficients



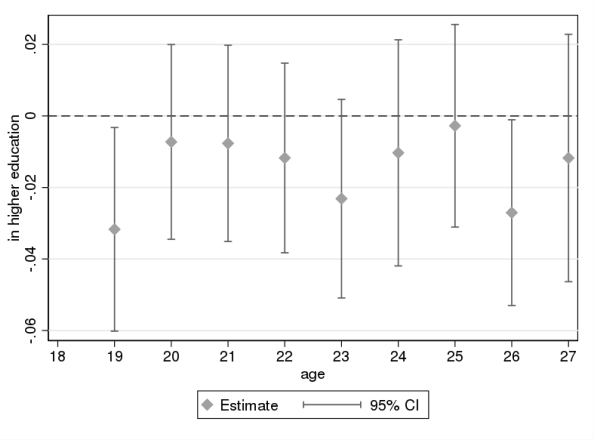
(a) Exam pass: Has a high school diploma



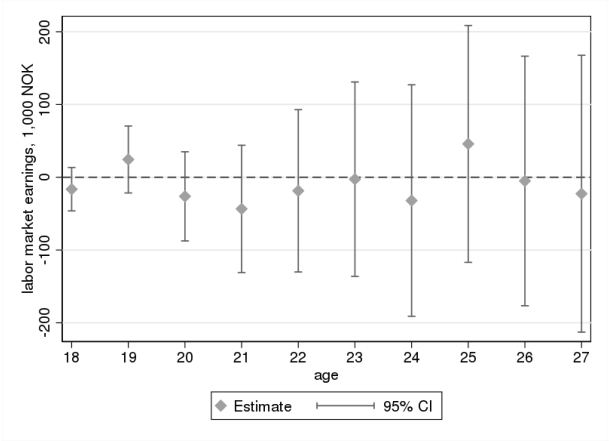
(b) Exam score: Has a high school diploma



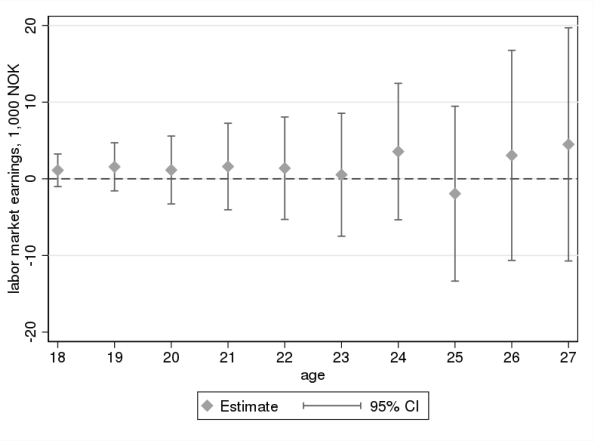
(c) Exam pass: In tertiary education



(d) Exam score: In tertiary education



(e) Exam pass: Labor income in 1,000 NOK



(f) Exam score: Labor income in 1,000 NOK

Figure 1: Long run effects on high school graduation, higher education and earnings

Note: Figures show results of our interacted IV specification on the probability of having any high school diploma, of being in higher education and on labor market earnings over time. Left panels shows the impact of failing the exam, while right panels show the impact of one grade better exam score.

are relatively close to zero, indicating that passing the exam, and perhaps having a high school diploma, does not matter much for the earnings for these marginal compliers, but unfortunately the estimates are too imprecise to draw firm conclusions. Likewise, there is no significant effect of increased exam scores on earnings. Despite the low precision, these results indicate that measures to increase high school completion for these marginal compliers might not be very efficient at raising labor market outcomes for these low-performing students.

6 Conclusion

This paper addresses the impacts of passing a high-stakes exam on future performance in school, graduation, tertiary education and later labor market outcomes. By exploiting the unique quasi-random assignment to exam form in Norwegian high schools, we address the selection into high school completion in order to estimate plausibly causal effects of passing the exam. Furthermore, by instrumenting both for the exam score itself and the event of passing the exam specifically by interacting with pre-exam performance, we isolate a group of marginal students who are shifted out of school altogether by the instrument, shedding light on the response of a group of students at the brink of dropping out of school and potentially informing the effects of other policies aimed at combating high school dropout.

Results indicate that the assignment to a harder written exam indeed shifts otherwise identical students into worse performance on the exam. This includes both the important margin of passing the exam, but also other margins of performance. Because exam is high-stakes and passing the exam is required to eventually graduate, being shifted into failing the exam has large and long-term consequences for the compliers: More than 50% of them still do not hold a high school diploma at age 27, eventhough they are allowed to re-take the exam. We also find evidence of dropout already during high school, and some indication of worse performance in school the following year. Results for later labor market earnings are too imprecise to draw strong conclusions, but estimates are close to zero, indicating that the value of these marginal diplomas on the labor market might not be very large. This analysis thus provides important insights for a policy relevant group of compliers to a range of policies aimed at combating high school dropout.

References

Andersland, L. (2017). The Extent of Bias in Grading. Working Papers in Economics 10/17, University of Bergen, Department of Economics.

Andresen, M. E. and Huber, M. (2018). Instrument-based estimation with binarized treatments: Issues and tests for the exclusion restriction. FSES working paper series no 492.

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442.

Apperson, J. and Bueno, C. (2017). Do students shake it off? evidence from a cheating scandal. Working paper.

Bensnes, S. S. (2018). Scheduled to gain: Short- and long-run effects of examination scheduling. *forthcoming Scandinavian journal of economics*.

Bhuller, M., Mogstad, M., and Salvanes, K. G. (2017). Life-cycle earnings, education premiums, and internal rates of returnhec. *Journal of Labor Economics*, 35(4):993–1030.

Cameron, S. V. and Heckman, J. (1993). The nonequivalence of high school equivalents. *Journal of Labor Economics*, 11(1):1–47.

Card, D. and Lemieux, T. (2001). Dropout and enrollment trends in the postwar period: What went wrong in the 1970s? In *Risky Behavior among Youths: An Economic Analysis*, pages 439–482. National Bureau of Economic Research, Inc.

Clark, D. and Martorell, P. (2014). The signaling value of a high school diploma. *Journal of Political Economy*, 122(2):282–318.

de Chaisemartin, C. (2017). Tolerating defiance? local average treatment effects without monotonicity. *Quantitative Economics*, 8(2):367–396.

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. Working Paper 22165, National Bureau of Economic Research. Forthcoming in American Economic Review: Applied Economics.

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

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.

Falch, T., Nyhus, O. H., and Strøm, B. (2014). Causal effects of mathematics. *Labour Economics*, 31(C):174–187.

Galloway, T. A., Kirkebøen, L. J., and Rønning, M. (2014). Grading practices in norwe-

gian middle schools,. *Statistics Norway Report*, (14).

Heckman, J., Humphries, J., and Kautz, T., editors (2014). *The Myth of Achievement Tests*. University of Chicago Press.

Heckman, J. J. and LaFontaine, P. A. (2006). Bias-Corrected Estimates of GED Returns. *Journal of Labor Economics*, 24(3):661–700.

Heckman, J. J. and LaFontaine, P. A. (2010). The American High School Graduation Rate: Trends and Levels. *The Review of Economics and Statistics*, 92(2):244–262.

Heckman, J. J., Lochner, L. J., and Todd, P. E. (2008). Earnings Functions and Rates of Return. *Journal of Human Capital*, 2(1):1–31.

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5-6):761–796.

Jacob, B. A. and Lefgren, L. (2004). Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. *The Review of Economics and Statistics*, 86(1):226–244.

Jewell, R. T., McPherson, M. A., and Tieslau, M. A. (2013). Whose fault is it? Assigning blame for grade inflation in higher education. *Applied Economics*, 45(9):1185–1200.

Lamb, S., Markussen, E., Teese, R., Sandberg, N., and Polesel, J., editors (2011). *School Dropout and Completion. International Comparative Studies in Theory and Policy*. Springer.

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

Manacorda, M. (2012). The cost of grade retention. *The Review of Economics and Statistics*, 94(2):596–606.

NOU2018:13 (2018). Voksne i grunnskole- og videregående opplæring — finansiering av livsopphold. Technical Report 13:2018, Kunnskapsdepartementet. NOU 2018:13.

OECD (2018). *Education at a Glance 2018*.

Tyler, J. H., Murnane, R. J., and Willett, J. B. (2000). Estimating the Labor Market Signaling Value of the GED. *The Quarterly Journal of Economics*, 115(2):431–468.

A Grade inflation and instrument monotonicity

The monotonicity assumption means that all who are affected by the instrument are affected in the same direction, which in our setting means that no one can get a better exam score by drawing written exam instead of oral exam. In this section we will demonstrate that this assumption is likely to hold for the vast majority of our sample because of two distinct emergent features of the school system. First, there seems to be a oral examination specific exam score bonus that is persistent across several school characteristics. Second, nearly all schools inflate the internal grades and oral exam scores of its students, and that this inflation is factored out when written exams are graded by a random anonymous external examiner. Together, these features give students allocated to written exams a clear disadvantage which overshadow most individual differences in exam taking ability.

Table 6 shows the average grades and exams characteristics in our sample on the school-year level. In all sub-samples, the average oral exam scores are higher than average written exam scores, with the average GPA and internal grade in the exam course in between. While this suggests that students perform better on oral than written exams, we are not comparing apples to apples. Answering such a claim requires a more nuanced approach.

For the monotonicity condition to hold, each student must perform equally or worse if selected for written exam than if they would have if selected for oral exam. However, we can never observe both outcomes for the same student, which means that this assumption cannot be tested directly. Instead, we need to look for evidence that indirectly supports the monotonicity condition. This means that for each exam grade, the student has a corresponding internal grade. **Exam gains** can then be defined as the difference between the exam grade and the internal grade, effectively removing subject, school, and teacher effects from the exam grades. While gains from oral and written exams cannot both be measured for the same student, we can observe the average gains from oral and written exam in the same school each year. The difference between the average exam gains can be interpreted as the school-year level **written exam penalty** and should be negative for school-level compliers. Figure 2a shows that internal grades and oral- and written exam scores all increase with school enrollment size. Nonetheless, the difference between the exam scores and the internal grade does not seem to change much. The written exam score is the best measure we have of student ability, and Figure 2b shows how exam gains vary with average school-level written exam results. Not surprisingly, the best (worst) performing students have the lowest (highest) written exam penalty. But the difference in written exam penalty between the best and the worst performing students are nowhere close to the difference in written exam score. A one grade increase in the average written exam scores results in 0.4 grade reduction in the written exam penalty. This suggests the existence of widespread grade inflation in schools with low-ability students. Gains from

Table 6: School grades

	2004	2005	2006	2007	All schools	Large schools
GPA	4.00	4.03	4.05	4.02	4.03	4.08
Internal grade	3.84	3.88	3.91	3.88	3.88	3.95
Oral exam score	4.13	4.20	4.29	4.19	4.20	4.29
Written exam score	3.52	3.43	3.46	3.34	3.43	3.53
Oral exam gain	0.19	0.23	0.23	0.19	0.21	0.21
Written exam gain	-0.27	-0.40	-0.39	-0.49	-0.39	-0.36
Written exam penalty	-0.46	-0.63	-0.62	-0.68	-0.60	-0.57
School-level compliers	0.92	0.97	0.92	0.97	0.95	0.97
Number of exams	98.5	107.8	108.4	120.3	109.4	152.6
Number of oral exams	29.5	31.9	31.1	35.1	32.0	46.1
Number of written exams	69.0	75.9	77.3	85.2	77.3	106.5
Share oral exams	0.30	0.30	0.28	0.29	0.29	0.31
N schools	288	296	293	287	1,164	278
N students	18,053	19,626	19,807	22,696	80,182	40,754

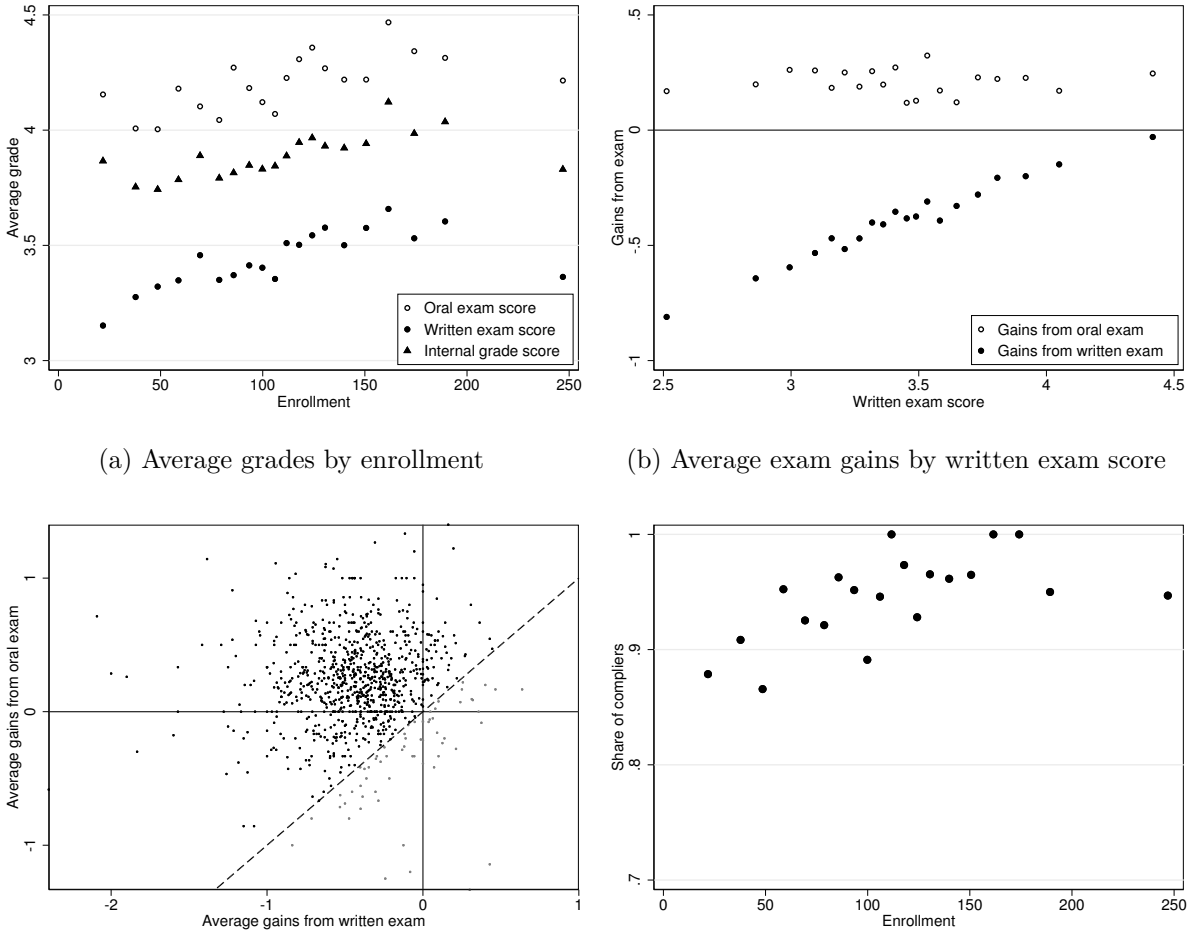
oral exam, on the other hand, seems to be stable across schools with different levels of written exam results. This suggests that school and teacher level grade inflation constitute a large part of the written exam penalty we observe, but in addition there also exist an oral exam-specific benefit. Figure 2c plots the gains from oral exam against the gains from written exams at the school-year level, where compliers are above and to the left of the dashed 45 degree line. Most of the observations (around 74 %) are concentrated in the second quadrant where the gains from oral exams are positive and the gains from written exams are negative.

However, school-level compliance does not directly imply individual-level compliance, but it makes defiers less likely. Some students might naturally be more adept at answering written exams than oral exams, but this will not necessarily threaten the monotonicity assumption. As long as any idiosyncratic advantage individual students might have for writing exams does not strictly dominate the benefits the student gets from oral examination, such as help from the subject teacher and exposure to grade inflation, the potential outcome of an oral exam will still be higher than for a written exam and monotonicity holds. On the other hand, school-level defiance does not directly imply that the students

are defiers, but it makes individual-level compliers less likely.

On average, students who draw written exams are graded almost 0.4 grade points lower than their own internal grade in the same subject, and 95.8 percent of students attend schools where the average gains of written exam are negative. For oral exams, students are on average graded more than a 0.2 grade points higher than the internal grades and the average gains are higher than written exam gains for 96.3 percent of schools. This translates to an average written exam penalty of -0.6 grade points.

In rare cases (less than 5 percent) the average gains from written exams will exceed the gains from oral exams. We argue that this is partly the result of random sampling noise. The number of exams can be quite low in some schools in some years, especially for the number of oral exams, which increase the probability that random noise from student-exam allocation or exam day conditions will dominate the written penalty. This interpretation is supported by the decline in the share of schools with non-negative written exam penalty as the number of exams increase as shown in Figure 2d. This can also be seen in Table 6 as the share of school-level compliers are higher in large schools (0.97) relative to small schools (0.93). To further show that this is not caused by the existence of some special defier schools we calculate the average written exam penalty at the school level by pooling all four years in our sample. Then the share of school-level compliers increases to 99 percent in all schools, and to 100 percent for large schools.



(a) Average grades by enrollment

(b) Average exam gains by written exam score

(c) Gains from oral and written exams

(d) School level compliers by enrollment

Figure 2: School-year level average effects of exam form

Note: Binned scatter plots of average school-year level exam scores and internal grades by second grade enrollment in panel a). Binned scatter plots of gains in written and oral exams relative to internal grades in panel b). Scatter plots of average oral and written exam gains at the school-year level in tpanel c) excluding school-years with less than 5 exams in either exam form (3% of students). Binned scatter plots of share of school-year level compliers by enrollment size in panel d), defined as schools where the average gains from written exams exceed the average gains from oral exams in a given year. The binned scatter plots are constructed by dividing enrollment and written exam scores into 20 groups, each containing the same number of students. Enrollment is measured as the number of second grade exams at the school, corresponding to the number of students.