The Final straw: High school dropout for marginal students

Andresen, Martin Eckhoff and Løkken, Sturla Andreas

Statistics Norway

19 December 2020

The final straw: High school dropout for marginal students

Martin Eckhoff Andresen Sturla A. Løkken

Statistics Norway*

Abstract

We investigate the consequences of failing a high-stakes exam in Norwegian high schools. Second-year high school students are randomly assigned to either a locally graded oral exam or a centrally graded written exam. Students assigned to written exams consistently receive lower grades and have a greater probability of failing, particularly in the case of already low-performing students. Because passing the exam is required to obtain a high school diploma, this translates into a reduction in high school graduation rates that remains significant over time, permanently shifting a group of marginal students into dropping out of high school altogether. We show that these marginal students are severely disadvantaged across several dimensions, even more so than dropouts in general. Our analysis of what predicts dropout among these marginal students suggest that effective policies for combating high school dropout should target students exclusively on the basis of poor academic performance, rather than other measures of disadvantage such as socioeconomic status, even though these characteristics are associated with dropout among students in general.

Keywords: Exam type, high school dropout, school performance

JEL codes: I21, I26, J24

*E-mail to mrt@ssb.no and/or sal@ssb.no

1 Introduction[1]

High school dropout is a social problem that has received extensive and sustained attention from practitioners and policymakers across the world (Card and Lemieux, 2001; Heckman and LaFontaine, 2010; Hussar et al., 2020; OECD, 2018). Increasing high school graduation rates has long been a political priority worldwide (e.g. Educate America Act (1994) in the US and Livsoppholdsutvalget[2] (2018) in Norway) and has only become more pressing in recent years as more jobs require specialized skills and job opportunities for unskilled workers are dwindling (Acemoglu and Autor, 2011; Acemoglu and Restrepo, 2018; Autor and Dorn, 2013; Goos et al., 2009). Many policies and interventions have tried to address this problem, often targeting disadvantaged students or groups otherwise thought to be at risk of dropping out (Dynarski et al., 2008; Lamb et al., 2011). The empirical evidence from such interventions has been mixed, suggesting that dropout is a complex problem with no silver bullet policy that works in all cases. This paper investigates how relatively small shocks can cause a group of marginal students to permanently drop out of high school and proceeds to identify key differences between this group of students and dropouts in general and discuss implications for policy design.

Students drop out of high school for a variety of reasons. While some reasons, like physical and mental health problems or a demanding family situation may be unrelated to schooling, poor academic performance remain one of the most important causes of dropout. This group of students typically struggles to keep up with coursework and pass exams. Marginal students, who are close to the margin between dropping out and staying in school, are interesting for several reasons: First, they are relevant targets for policies that aim to effectively reduce dropout rates. Second, they are the relevant students to consider when evaluating the long-term impacts of these policies aimed at reducing dropout. For these marginal students, the value of a high school diploma may be very different than for the average student.

In this paper we investigate how graduation is affected by a small shock to school performance: failing a high-stakes exam in high school. Passing the exam is a prerequisite for being awarded a diploma at the end of high school, and students who fail must organize and pass a re-take exam if they wish to graduate. In addition, re-taking exams might crowd out other coursework and failing the exam might negatively impact the students self-confidence in their academic ability. For these reasons we argue that investigating the consequences of exam failure on dropout behavior enables us to learn about a highly policy-relevant group of marginal students.

We exploit the quasi-random allocation to exam type conditional on course choices in Norwegian high schools. In their second year of high school students on the academic track are given one

[1]We thank seminar participants at the Frisch Centre, NIFU and IWAEE, Lars Kirkebøen, Magne Mogstad, Astrid Sandsør, 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.

[2]Report on Support to Adults With Low Education.

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 oral and written exams are statistically indistinguishable on important outcome determinants such as past performance and family background conditional on course choice. We show that being assigned to a written exam has large negative effects on exam outcomes, lowering grades and more than doubling the probability of failing compared to the alternative oral exam. This, in turn, has important consequences for students, as these exams are mandatory in order for them to finish high school and earn a diploma.

Using exam type as an instrument for failing an exam allows us to estimate causal effects on further performance in school and subsequent graduation. We document that failing an exam has major negative and persistent effects with respect to subsequent graduation. Among the marginal students shifted into a failing grade by our instrument, around half do not obtain a diploma on time and the great majority of them have still not acquired one by age 27.

Identifying factors shared by students who drop out of school has been an influential area of research (Belfield and Levin, 2007; Levin, 1987,8). Poverty, parents with low education, single-parent households, poor academic performance, low aspirations, and psychological disorders are generally recognized as factors associated with educational failure (Kearney and Levine, 2016; Tharmmapornphilas, 2013). Students with one or more such risk factor are often characterized as *at risk* of dropping out. The empirical findings of policies targeting at-risk students are mixed, with some documenting large effects on the outcomes of targeted interventions (Hernæs et al., 2017; Oreopoulos et al., 2017) and others finding no effects (Bernstein et al., 2009; Freeman and Simonsen, 2015; Guryan et al., 2020).

One explanation for such a range of results could be that for a policy to be effective it must target students that are close to the margin between dropping out and graduating. Students too far away from such a margin, either with little chance of dropping out or with little chance of graduating, are unlikely to be responsive to an intervention. This means that it is essential to identify marginal students in order to design effective policies for combating high school dropout, not merely knowing whether a student is at risk. We contribute to this discussion with a comparison between the characteristics of ordinary dropout students and those of marginal students identified by our instrument, and discuss how these groups relate to various risk factors.

In order to understand the large effects we find on high school dropout and reconcile these findings with the existing literature that focuses on disadvantaged students, we characterize the compliers in terms of background characteristics. Results indicate that the marginal student, defined as a student who would pass the exam if assigned to the simpler oral exam but fail when assigned to the written one, are severely disadvantaged. Boys are overrepresented, and so are students with low-educated, low-earning and single parents, as well as students with poor past academic performance. Despite this, we show that school performance is by far the most important predictor of dropout, even more so among the marginal students identified

by our instruments. In the sample of all students, socioeconomic disadvantaged children as measured by immigrant status, single parent household or low parental education and earnings predict dropout behavior, even conditional on past student performance. Among marginal students, however, only student academic performance is predictive of dropout, suggesting that policymakers should target students solely on the basis of ability in order to design efficient policies to combat high school dropout, rather than targeting them on the basis of their family background.

In addition to the literature on at risk students and dropout behavior, our paper relates to three strands of the literature. First, we estimate the effects of passing an exam on future performance in school, which could be thought of as acting either as a signaling effect within schools (Jacob and Lefgren, 2004; Manacorda, 2012) or through student motivation (Diamond and Persson, 2016). Second, the oral exam allows schools to apply local grading practices to the exam scores, in contrast to the centrally set and graded written exams. This relates our paper to a literature on 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) Third, we contribute to a literature 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), all of which are relatively minor shocks that are found to have large impacts on performance. Our paper builds on and adds to these literature strands by isolating a group of students on the verge of dropping out and showing that these students are shifted to dropping out entirely by a relatively small shock to their exam performance.

This paper continues 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 provides a conclusion.

2 Institutional setting

The Norwegian school system

The Norwegian compulsory education system is operated by municipalities and consists of primary school (ages 6-12) 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 upper secondary 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. In the rest of the paper, we will refer to this level of schooling as high school. High

schools are predominantly public[3] and free, and more than 95 percent of students enroll in high school the 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 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 in order to exploit the quasi-random assignment to exam type. Academic-track high school education lasts for three years and qualify students for tertiary (college) education. At the end of the final year, students who have passed all required courses[4] and exams are awarded a primary diploma. Students who fulfill these requirements at a later time are instead awarded an ordinary diploma. For a more in-depth description of the Norwegian school system see Falch et al. (2014).

High school exams

High school exams are set either centrally or locally. Central exams are set by the Norwegian Directorate for Education and Training (UDIR), the executive agency for the Ministry of Education and Research, and held as nationwide written exams for all students at the same time. All central exams are graded by a randomly assigned external examiner.[5] Locally held exams are formally administered by the county administration but are in practice delegated to the individual high schools. Locally held exams are typically oral exams[6] where the students are assessed by one internal (most often the course 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 for the teacher-assessed internal grades, assigned based on performance through the year and set a few weeks before the exam draw. If a student fails the exam or is sick on exam day, they are still permitted to progress to the next year, and have the opportunity to re-take the exam the following semester.

Students risk being selected for examination in all courses that will appear on their diploma. At the end of their third year all students who pass all required courses and exams within the stipulated three years receive a primary diploma.[7] All high school students must take several

[3] 8 percent of students attend private high schools (SSB Statistikkbanken).

[4] In the Norwegian school system, subjects such as mathematics, English language, or biology consists of one or more courses and in some cases multiple variants, such as Biology I and Biology II or Applied maths I and Advanced maths I. Final courses within any subject can be selected for examination.

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

[6] Other forms of locally given exams are either practical or oral-practical, but these are rare in the academic track. 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.

[7] A primary diploma gives students some benefits when applying for colleges straight after high school.

mandatory exams. In their first year, 20 percent of students sit fora final exam in one course. In second year, all students sit for a final exam in one course. In third year, all students take four exams: one written exam in Norwegian language, final written exams in two subjects and a final oral exam in one subject.

Random assignment of exam type

We exploit a feature of the second year exams: exam type is as good as randomly assigned conditional on predetermined course choices. The process of allocating students to exams starts in early spring when the county administration sends lists of centrally set exams to all high schools in the county. After the internal grades are determined, school administrators[8] 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, in terms of both exam course and type. There are no formal incentives for the principals or schools to engage in this behavior. Furthermore, several formal and informal restriction make such manipulation implausible. First, 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. To maximize grades, school administrators need not only to assess the potential outcomes of students in all courses and both exam forms, but also their potential outcomes relative to all other students. It is unlikely that schools have this kind of information, let alone act on it. 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 need to navigate several informal restrictions when assigning students to exams. First, oral exams are costly and there is a fixed cost component to organizing an oral examination. This means that schools want to minimize not only the number of students selected for oral examination, but also the number of courses. Second, school administrators typically start by planning the examination schedule for third year students, who have four exams each. The exam schedules for second year students are then planned later, depending 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 teachers. Finally, in some cases students may have certain course combinations that limit the choice set of the school administrators, in terms of both subject and type of exam.

Students who change courses, re-take exams or who do not finish on time will get a regular diploma.

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

All these formal and informal restrictions make any attempt by the school administration to inflate (maximize) school-level exam grades by manipulating the assignment of students to exam types very difficult. Moreover, an extremely detailed knowledge of students' potential outcomes in various courses and exam types would be needed in order to execute such manipulation effectively. 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.[9] This institutional setting provides us with an empirical strategy that may enable us to estimate the effect of exam type on exam outcomes: conditional on school, cohort and exam course, the examination type should be as good as randomly assigned. In section 4, we provide evidence of this conditional randomization by showing that important determinants of exam performance and later outcomes such as past performance, parents' education and gender are balanced across examination type within exam courses as long as we condition on exam course..

Local grading practices

An important part of being assigned an oral exam is that schools have the opportunity to apply local grading practices. As our first stage estimates reveal, students are systematically awarded better grades on oral than 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 explanation is that students are simply better at oral examinations, another 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 schooling Galloway et al. (2014) find evidence of systematic school-wide differences in grading. They document a high degree of within-school correlation in grading practices. All students apply for high school admission using their grade point average (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. Schools tend nonetheless to use the full range of the grading scale (as documented by Galloway et al. (2014)) , and there might be strong internal pressure towards grading conformity within schools from principal, other teachers and 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 more pronounced grade inflation in schools with low-ability student compared to high-ability students, a fact we document for our sample in Appendix B. Another potential source of grade inflation is competitive pressure between schools in the same county, because school administrators might encourage grading leniency when they are competing with otherwise similar schools for the same students.

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

Teachers might also engage in grade inflation at the individual student level. Teachers develop personal relationships with students, and in some cases the families, which makes it uncomfortable to award low grades. Another possibility is that teachers sometimes grade students in light of 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 in light of positive factors. Hence, the effect on grades will be positive for the oral exam.

Since written exams are randomly and anonymously graded by external examiners using the same assessment guidelines and cross-validated between examiners, these grades are not affected by grade inflation and 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 means an exam score penalty. We provide evidence for how this penalty varies across the ability distribution in Section B in the appendix.

3 Data and sample

Administrative data

Our data come from Norwegian administrative registers for the years 2004 – 2017. All data are hosted at Statistics Norway and contain unique individual identifiers that allow us to link students across registers and connect students to parents' characteristics. From the residency registers, we obtain data on municipality of residence, gender and birth dates. Enrollment data for high school and tertiary education come from the national education registers, while the results of individual courses and exams are reported annually by each school to VIGO, which administers the IT systems and databases for the counties. Graduation data come from the national transcript records (*"nasjonal vitnemålsdatabase"*), covering all issued high school diplomas with results in individual courses. Finally, we measure the students long term labor supply and parents' earnings using annual income data from tax records.

Sample definition

To define our sample, we start out with all 17-year olds resident in Norway on January 1st in the years 2004 to 2007.[10] We focus on those who start their second year of a three-year academic-track high school program in one of these years, ruling out redshirting, delayed students or

[10] An education reform ("Kunnskapsløftet") was introduced in 2007, affecting second year students in 2008, which prohibits us from using later years in our sample.

Table 1: Sample selection

	2004	2005	2006	2007	Sum	%
Norwegian resident at age 17	55,385	56,933	60,276	62,231	234,825	100 %
Enrolled in 2nd year of high school	50,022	51,268	54,688	56,252	212,230	90.4 %
Registered for only one study program	45,968	47,661	50,357	52,940	196,926	83.9 %
Academic-track students	21,780	22,424	23,838	25,429	93,471	39.8 %
Standard student status	21,307	22,137	23,519	24,761	91,724	39.1 %
Full-time student	20,662	21,863	23,143	24,502	90,170	38.4 %
Registered for at least one course with exam	19,547	20,654	21,850	23,807	85,858	36.6 %
Takes at least one exam	18,223	19,496	19,651	22,466	79,836	34.0 %
At most one exam	17,343	18,899	18,976	21,778	76,996	32.8 %
Exclude courses with only one exam type	17,330	18,889	18,964	21,765	76,948	32.8 %

those who start early.[11] We impose a few other sample restrictions to ensure that our students are as comparable as possible and all subject to the quasi-random exam draw, as documented in Table 1.

We then link these students to all registered courses with results, allowing us to compute the GPA for each student that is based on the grades assigned by the teacher prior to the exam, excluding the course in which the student is assigned to an exam. Next, we link these students to the results of the end-of-year exams. For around 6,000 students, or around 7% of the remaining sample, we cannot find any exam results and consequently have to exclude them from the sample. The main reason for this is that our data lack the results from postponed or re-taken exams due to illness, no-show or a failed exam. If this sample selection is endogenous to being assigned to a written exam, our estimates can potentially be biased, but we show in section 4 that the groups assigned to written and oral exams are balanced with respect to pre-assignment performance and other characteristics. We also exclude a few hundred students who are registered with more than one exam. This may happen if students take courses privately, without registering for instruction. These courses are not subject to the standard exam draw.

Summary statistics

This leaves us with a final sample of 76,948 students. Summary statistics for these students are given in Table 2, both for the sample as a whole and for students assigned to oral and written exams separately. Girls are somewhat overrepresented 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 to written exams, however, have a somewhat higher compulsory school GPA and teacher-evaluated grade in the course for which they are assigned to an exam. This indicates

[11] Both early and late school start is uncommon in Norway, as is grade retention in compulsory school.

Table 2: Summary statistics

	Full sample		Written exam		Oral Exam	
Variable	mean	s.d.	mean	s.d.	mean	s.d.
A: Background characteristics						
Female	0.54	(0.50)	0.53	(0.50)	0.54	(0.50)
Single parent	0.14	(0.35)	0.14	(0.35)	0.15	(0.35)
Immigrant background	0.15	(0.35)	0.14	(0.35)	0.16	(0.37)
Father has higher education	0.37	(0.48)	0.37	(0.48)	0.36	(0.48)
Mother has higher education	0.40	(0.49)	0.40	(0.49)	0.39	(0.49)
Father's labor earnings	861.2	(982.5)	860.5	(1020.3)	862.9	(882.5)
Mother's labor earnings	489.4	(343.7)	490.4	(335.1)	486.7	(364.0)
GPA, compulsory school	4.49	(0.56)	4.52	(0.55)	4.42	(0.57)
Number of exam-eligible courses	10.86	(1.39)	10.87	(1.37)	10.84	(1.44)
Grade in exam course	4.05	(0.78)	4.08	(0.78)	3.97	(0.79)
GPA, other courses second year	3.89	(1.09)	3.83	(1.09)	4.04	(1.07)
B: Exam type and outcomes						
Written exam	0.71	(0.45)	1.00	(0.00)	0.00	(0.00)
Failed exam	0.05	(0.21)	0.06	(0.23)	0.02	(0.13)
Exam grade	3.65	(1.29)	3.41	(1.24)	4.22	(1.24)
C: Outcomes						
Second year completed	0.89	(0.32)	0.89	(0.32)	0.89	(0.32)
Third year started on time	0.99	(0.09)	0.99	(0.09)	0.99	(0.10)
Third year completed on time	0.80	(0.40)	0.81	(0.40)	0.79	(0.41)
GPA, third year	4.05	(0.86)	4.08	(0.85)	4.00	(0.87)
Primary diploma obtained on time	0.72	(0.45)	0.73	(0.44)	0.71	(0.46)
Any diploma obtained within allowed time	0.74	(0.44)	0.74	(0.44)	0.72	(0.45)
In tertiary education at age 22	0.70	(0.46)	0.71	(0.45)	0.67	(0.47)
Labor earnings at age 27	512.4	(307.1)	517.5	(308.9)	500.0	(302.2)
N	76,948		54,678		22,270	

Note: The table reports means and standard errors for different individual characteristics and outcomes. All earnings measured in 1,000's of Norwegian kroner (2019). Grades and GPA range from 1 (lowest) to 6 (highest).

that straightforward comparisons of the results of students with written and oral exams will not reflect any causal differences of the exam type. Rather, they reflect a mix of the effect of exam type and the fact that more demanding courses may have different exam types. In practice, the courses chosen will affect both the exam type and likely the results of the exam, as some courses are more difficult than others.

Strikingly, Table 2 reveals large differences in the results of the exam depending on exam type. Students assigned to a written exams score on average 0.8 of a grade lower and have an almost 4 percentage points higher probability of failing the exam than students assigned to an oral exam. Nonetheless, they have a higher probability of obtaining a high school diploma on time, perhaps because they are generally stronger students, as reflected in their choice of courses.

Our student sample takes exams in a total of 42 distinct courses in the sample window. The empirical strategy described in the following section exploits this variation between students who sit exams in the same course, but with different exam types.

4 Empirical strategy

We aim to investigate some of the barriers to graduating from high school faced by the highly policy-relevant group of marginal students. Our strategy is to identify the effect of failing a mandatory exam on important school outcomes, focusing on dropout behavior. Failing an exam places an additional burden on marginal students already struggling with their current course load. First, they need to organize and pass a re-take exam in order to be awarded a diploma. Second, since the re-take exams are typically scheduled at the start of the following semester, the preparation and stress of an additional exam may crowd out time and resources that would otherwise be allocated to other courses. Third, failing the exam could negatively impact the students self-concept of academic ability, which may hamper future academic performance.

Instrumental variable strategy

The most straightforward way to evaluate the effect of failing a high-stakes exam on high school dropout or graduation is to simply compare outcomes for students who fail and for those who pass. However, this ignores the fact that these groups of students are very different, in both observable and unobservable characteristics. In order to identify causal effects we need exam failure to be as good as randomly assigned. As exam results are strongly affected by exam type, and the institutional setting suggests that exam type is as good as random, we use this as an instrument for failing the exam. Although exam type is supposed to be drawn randomly by school administrators, restrictions are imposed by the course choices made by the students. In particular, the exam type may vary systematically with the difficulty of the courses, and

students may sort into harder courses based on unobserved ability. To account for this, it is crucial to control for course choices.

Our basic strategy is to instrument exam failure with exam type conditional on course choice. We do this by controlling flexibly for fixed effects in a two-stage least squares setup that we refer to as the *simple model.*

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

where y_{ikt} is an outcome for student i in cohort t at school k , α_k and δ_k are school fixed effects and π_t and η_t are cohort fixed effects. The fixed effects $\kappa_{c(i)}$ and $\lambda_{c(i)}$ are exam course fixed effects, ensuring that we compare students who are assigned to different exam types in the same course. Finally, F_{ikt} is a dummy for student i failing the exam, and Z_{ikt} is the instrument, a binary variable equal to 1 if student i is assigned to a written exam. The two parameters of interest are γ, reflecting the first stage effect of the exam type on exam failure, and β representing the IV estimate of the effect of failing the exam on the outcome for the compliers.

Instrument validity

As with all IV applications, we rely on relevance, exclusion and monotonicity assumptions for valid estimation of causal effects. The relevance assumption is testable, and we show in the next section that the instrument does indeed affect exam outcomes. The exclusion restriction requires the instrument to affect the outcome y only through the treatment F. In order for β to be interpreted as a causal effect of failing the exam, we need exam type to be as good as randomly assigned given the set of fixed effects. As discussed in Section 2, the institutional setting provides a promising background for such conditional randomization. Nonetheless, given the discretion afforded to local school administrators when assigning exams, we cannot rule out allocation of students to exams on the basis of potential outcomes. We therefore rely on balancing tests to show that the samples of students assigned to different exam types are statistically indistinguishable on important predetermined covariates when course choice is controlled for.

Table 3 presents balancing tests for exam type assignment. In Column 1, we condition only on cohort fixed effects. The two samples are clearly unbalanced with respect to predetermined variables: Students assigned written exams come from families with higher education and themselves have higher GPAs, but have lower scores in the exam course than students assigned to oral exams. This could indicate that courses predominantly using written exams are more difficult than courses using oral exams, but that these courses are more often selected by strong

Table 3: Balancing tests

	(1)		(2)		(3)	
Female	-0.0246***	(0.00527)	-0.0260***	(0.00472)	-0.00106	(0.00200)
Single parent	0.00113	(0.00599)	-0.00249	(0.00555)	0.00397	(0.00249)
Immigrant background	-0.0178**	(0.00857)	0.00766	(0.00588)	-0.00151	(0.00268)
Mother has higher education	-0.00576	(0.00471)	-0.000425	(0.00374)	0.0000664	(0.00163)
Father has higher education	-0.00345	(0.00393)	-0.00675*	(0.00369)	-0.00201	(0.00189)
Mother's earnings	-0.00170	(0.00218)	0.00465***	(0.00148)	0.000271	(0.000649)
Father's earnings	-0.00414	(0.00647)	0.00942*	(0.00561)	-0.00162	(0.00296)
GPA, compulsory school	0.0801***	(0.00870)	0.0666***	(0.00638)	0.00433	(0.00265)
Number of courses	0.0000852	(0.00366)	0.00172	(0.00335)	0.00158	(0.00139)
GPA, other courses	0.0861***	(0.00643)	0.0902***	(0.00563)	-0.00296	(0.00214)
Teacher grade, exam course	-0.0960***	(0.00413)	-0.0865***	(0.00404)	-0.000624	(0.00134)
Fixed effects						
Cohort	✓		✓		✓	
School			✓		✓	
Exam course					✓	
N	76,948		76,948		76,948	
joint F	42.38		41.17		1.197	
p	<0.0001		<0.0001		0.276	

Note: Table reports separate regressions per column of a dummy for written exam on predetermined variables as indicated in row headers and sets of fixed effects as indicated in bottom panel. Earnings are measured in millions of Norwegian Kroner (2019). Grades and GPA range from 1 (lowest) to 6 (highest). For covariates with missing values (a few thousand observations for father's earnings, a few hundred for compulsory school GPA and mother's earnings), these covariates have been set to the mean in the sample and a separate dummy have been included to indicate missing for that covariate. Coefficients for these dummies are not reported, but they are included in the joint F-tests reported at the bottom of the table. Standard errors clustered on school. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

students. These differences across exam type are strongly significant, with a joint F-statistic of around 42.

Moving to Column 2, we add school fixed effects to account for any cross-school differences in grading practices, student ability and teaching that correlate with course choices and/or exam type. This alone does not resolve the issue with balancing, but when exam-course fixed effects are added in Column 3, all coefficients are insignificant and the joint F-statistic drops dramatically. This reduction is largely driven by the point estimates moving closer to zero, not by the standard errors increasing due to the reduction in degrees of freedom. Our interpretation is that this accounts for the effect of individual course choices: Some courses are easier than others and the likelihood of being assigned to a written exam differs from course to course, allowing unobserved ability to affect both the likelihood of being assigned to a written exam and performance on the exam. This suggests that exam type is as good as randomly assigned conditional on cohort, school and exam course.[12] Furthermore, for the IV estimate to be

[12]A final challenge for the exclusion restriction is that the instrument may affect other margins of exam outcomes than just failing. We return to this in Section 5 and Appendix C.

interpretable as a local average treatment effect (LATE), traditional IV analysis relies on a monotonicity (or rather, uniformity) assumption: The instrument(s) cannot affect some people's exam outcomes positively and others negatively. At first glance, this is 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 at oral exams. Remember, however, that oral examination also allows schools and teachers to influence the grades of their students as the teacher is one of the examiners, allowing a teacher who knows the student to 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 graded written exams. Only students who are considerably better at written than at oral exams, so that this effect dominates the local 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 B in the appendix we provide evidence that supports the monotonicity assumption, and show that any group of potential defiers is likely to be very small. Furthermore, de Chaisemartin (2017) show that 2SLS still identifies a LATE when the monotonicity assumption is weakened. As long as there is a sub-group of the compliers with the same treatment effect distribution as the defier group and as long as there are more compliers than defiers, the IV still estimates 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.

Interacted IV-strategy

In settings with strong first stage heterogeneity (i.e. the impact of the instrument varies across subgroups), the pooled estimator in Equation (1) may be inefficient since the instrument is weak in parts of the sample and 2SLS with interacted instruments may be optimal (Angrist and Imbens (1995); see also the discussion in Abadie et al. (2019)). To increase precision, our preferred model interacts the instrument and all fixed effects with the teacher given grade in the exam course, a precise and predetermined measure of student skills in the exam course:

$$\begin{aligned} y_{igkt} &= \alpha_{gk} + \pi_{gt} + \kappa_{c(i)g} + \theta F_{igkt} + \epsilon_{ikt} \\ F_{igkt} &= \delta_{gk} + \eta_{gt} + \lambda_{gc(i)} + \textstyle\sum_{h=2}^{6} \gamma_h \mathbb{1}(h=g) Z_{igkt} + \mu_{igkt} \end{aligned} \tag{2}$$

where g is the teacher grade awarded to student i in the exam course, ranging from 2, the weakest passing grade, to 6, and the fixed effects for cohort, school and exam course are now indexed by g. This allows the instrument to have different effects for weak and strong students and increases precision, while avoiding selection based on the first stage (Abadie et al., 2019), invalidating inference.[13] For this specification, which we refer to as the *interacted model,* the

[13]Reassuringly, our main estimates are relatively similar to the simple specification in Equation (1) for most outcomes, although less precise. Results from this model are presented in Appendix Table A2 for the main outcome of high school graduation.

instrument needs to be as good as random within groups of predetermined ability. Table A1 in the appendix suggests that this is the case: The balancing variables are jointly insignificant for all five ability groups.

The *saturate and weight* theorem (Angrist and Imbens, 1995) implies that the parameter of interest, θ, will be a weighted average of the group-specific IV estimates within groups of teacher grade. These weights are given by the variance of the instrument-induced shifts in treatment, essentially assigning more weights to groups with a stronger first stage.[14] In practice, this places by far the most weight on the weakest students with teacher grade 2 (around 0.82) and 3 (around 0.17) and very little weight on remaining students, where compliers are rare.

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 failing the exam on the outcome for the population of compliers: Students who fail the exam because they are assigned to a written exam but would pass if assigned to an oral exam. This group may be strikingly different from the average student. In particular, we argue that compliers on the pass/fail 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 instrument are a very policy-relevant group: They represent students on the margin of dropping out of high school, who may differ significantly from at-risk students in general. As such, our results may be of particular interest 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 are effective at all. After presenting our main estimates in Section 5, we describe the characteristics of the compliers to our instrument, indicating that disadvantaged students on a range of background characteristics are overrepresented among the complier group.

5 Results

In this section, we first present first stage estimates of the effect of exam type on various exam score margins in Table 4 and discuss the complications caused by the effects of multiple exam score margins. Table (5) then presents IV estimates of the effect of failing the exam on graduating from high school the following year, and Table(6) presents estimates of other schooling outcomes. We provide a description of complier characteristics in Table 7 and finally a comparison of what predicts dropout behavior among marginal students and the overall student population in Table (8).

[14] The model in eq. (2) is not fully saturated in covariates. We verify that a weighted average of the subgroup-specific IV estimates using the weights from the saturate and weight theorem produces a very similar estimate to our baseline estimate in appendix Table C1.

First stage results

Table 4a shows the impact of being assigned to a written exam on all exam outcome margins from the simple model in Equation (1). Starting with the fail margin, our endogenous variable, we see that being assigned to the written exam increases the probability of failing by 3.4 percentage points: This is a large effect, amounting to an increase of 75% over the sample mean or a tripling of the probability of failing compared to the mean among students assigned to the oral exam. This estimate is highly significant, with F-statistic of 82, far above the conventional levels needed for a strong instrument. The instrument does, however, also affect other exam outcome margins: It decreases the average score on the exam by 0.62 of a grade and increases the probability of the student receiving lower grades at all margins as evidenced by the other columns of Table 4a. Moving to Table 4b, we use the interacted specification to see how students are affected across the ability distribution. Unsurprisingly, the effects on the fail margin are concentrated at the lower end, with students with grade 2being 12.1 percentage points more likely to fail when given a written rather than an oral exam. The corresponding number for a grade 3-student is 5.1 percentage points. However, even top students with grade 6 are shifted into failing by the instrument. Moving down the table, the impact for higher-performing students is concentrated at the top of the exam score distribution, reducing the probabilities of obtaining top grades. Taken together using the linear score in Column 6, the aggregate effect of a written exam on the linear score is relatively flat at -0.5 to -0.6 points for students with initial grades from 3 to 6, and somewhat smaller for the weakest students.

The joint F-statistics for these first stages are shown at the bottom of the table. Jointly, our five instruments have an F-statistic of 21.2 for the fail margin, well above the critical value for a maximum of 5% bias from Stock and Yogo (2005), which is 18.4.[15] As with the simple specification, we see that the instruments in this specification also affect other margins of exam outcomes, reducing the exam grades at all margins for students with all backgrounds. We find the uniformity of the effects of the instrument across grade margins and student ability reassuring, providing some support for the monotonicity assumption.

The response to the instrument across margins of exam outcomes, however, may pose a challenge for the exclusion restriction when we instrument for failing the exam. Table 4 indicates that students receive lower grades on written than oral exams, even when they pass, which might violate the exclusion restriction (Angrist and Imbens, 1995) and bias our estimates. This, however, requires these instrument-induced changes to exam outcomes at other margins to affect outcomes (Andresen and Huber, 2020). In our setting it is likely that the fail margin is a considerably more important determinant of long-term outcomes than the exam grade, and

[15]Likewise, we find no evidence of weak instruments using the Montiel-Pfluger robust weak instrument test (Olea and Pflueger, 2013; Pflueger and Wang, 2015) - effective F-statistic 39.3 compared to the critical value of 35.4 for 5% bias. As a robustness exercise, we also show in Appendix C that our main results do not change when estimated by limited information maximum likelihood (LIML) which has favorable properties compared to 2SLS in the presence of weak instruments (Angrist and Pischke, 2008).

Table 4: First stage: all margins

(a) Simple model (eq. (1))

	Exam score <					Linear
	2 (fail)	3	4	5	6	score
Written	0.0342***	0.113***	0.195***	0.188***	0.0890***	-0.620***
exam	(0.00378)	(0.00902)	(0.0122)	(0.0102)	(0.00729)	(0.0342)
N	76,948	76,948	76,948	76,948	76,948	76,948
F	81.98	158.2	255.2	342.7	148.9	328.6
Mean dep.	0.0452	0.202	0.457	0.725	0.926	3.645

(b) Interacted instrument (eq. (2))

Teacher	Exam score <					Linear
grade	2 (fail)	3	4	5	6	score
2	0.121***	0.151***	0.0504***	0.00607	-0.000172	-0.328***
	(0.0198)	(0.0254)	(0.0147)	(0.00537)	(0.000382)	(0.0490)
3	0.0524***	0.206***	0.235***	0.0896***	0.00855***	-0.592***
	(0.00669)	(0.0187)	(0.0213)	(0.0115)	(0.00309)	(0.0442)
4	0.0122***	0.103***	0.270***	0.232***	0.0555***	-0.673***
	(0.00234)	(0.00946)	(0.0177)	(0.0153)	(0.00741)	(0.0375)
5	-0.00173	0.0140***	0.132***	0.322***	0.213***	-0.679***
	(0.00232)	(0.00441)	(0.0111)	(0.0212)	(0.0188)	(0.0447)
6	0.00319**	0.000550	0.0292***	0.132***	0.299***	-0.465***
	(0.00134)	(0.00464)	(0.00922)	(0.0261)	(0.0409)	(0.0646)
N	76,948	76,948	76,948	76,948	76,948	76,948
F	21.19	45.57	65.64	88.14	35.18	90.31

Notes: Each column shows results from separate regressions of a dummy for exam score being equal to or above the score indicated 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, cohort and exam course fixed effects, which have been interacted with dummies for the teacher grade in panel B. Standard errors are clustered by school ($G = 312$) and robust to heteroskedasticity. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

Table 5: IV effects of exam failure on graduation

	Diploma type	
	Primary	Ordinary
Exam failure	-0.555***	-0.469***
	(0.153)	(0.162)
N	76,948	76,948
G (schools)	312	312
First stage F	21.2	21.2

Note: Table shows 2SLS estimates from Equation (2) of the effect of exam failure on receiving a diploma on time, using written exam interacted with teacher grade as the instrument and controlling for school, cohort and exam-course fixed effects interacted with teacher grade. Standard errors in parentheses, clustered at the school level. $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

we argue that these instrument-induced changes in exam scores at other margins are unlikely to affect outcomes. We verify this using an interacted model that estimates the effect both of failing the exam and of the exam score itself in Appendix C, finding a) very similar estimates for the fail margin and b) very small effects of the exam grade at other margins, which suggests that this particular type of violation of the exclusion restriction is unlikely in our setting.

Effects on dropout

One important question we set out to answer is how relatively small hurdles can become insurmountable for marginal students, causing them to drop out of high school. In Table 5 we present the 2SLS estimates of the effect of exam failure on obtaining a diploma on time the following year. The first column shows that students who are shifted into a fail grade by our instrument are 56 percentage points less likely to obtain a primary diploma the following year.

As the possibilities for re-taking exams are limited with respect to obtaining a primary diploma, we expect some of the students who did not manage to obtain a primary diploma to instead obtain an ordinary diploma. Our estimate of the probability of obtaining any high school diploma on time indicates that compliers at the fail margin are 47 percentage points less likely to obtain a diploma if they failed the exam. The estimates are large, indicating that a small group of marginal students may be pushed to drop out entirely by being assigned to the relatively more difficult written exam. These students may simply not have the perseverance or skills to re-take the exam, which is mandatory for receiving a high school diploma. It is also possible that the burden of preparing for a re-take exam on top of the normal course load the following year has spillover effects that are detrimental to academic performance in other courses. Since primary diplomas can only be obtained at the time of graduation, the effect of exam failure for this outcome will not change over time. An ordinary diploma, on the other hand, can be obtained at any time as long as the student has passing grades in an approved combination of

courses and exams. As close to half of the compliers who fail due to exam type are left without any diploma in the year they were supposed to graduate, a natural question is whether they are just temporarily delayed in redoing the courses and exam or whether they have dropped out altogether. To explore 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 estimate the specification from Equation (2) separately by age to see if the effects on graduation fall over time.

These results are plotted in Figure 1. The estimates are remarkably stable over time, decreasing only slightly, and students who fail the exam because of the exam draw are still more than 40 percentage points less likely to have a high school diploma at age 26, 7 years after they were scheduled to graduate. This is a clear indication of persistent long-term effects on graduation rates for a group of marginal students: students who manage to get a diploma on time if they are assigned to the easier oral exam, but who fail and subsequently drop out completely when assigned to 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 type. Although this complier group is unique to the instruments at hand, it can be argued that they are similar to compliers to other policies that aim to boost high school completion.

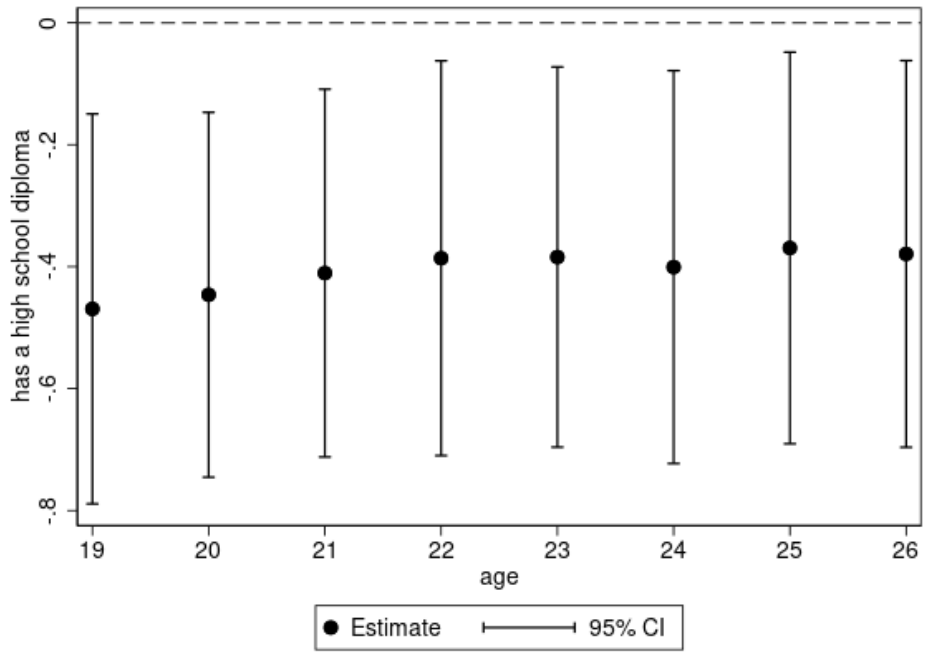


Figure 1: Effect of exam failure on a high school diploma over time

Note: The figure show the results of our interacted IV specification (eq. (2)) of the effect of failing the exam on the probability of having a high school diploma by age. The 2SLS estimates are represented by black dots, while the vertical bars represent the 95% confidence interval. Standard errors are clustered at school level.

Several studies investigate the distinction between temporary and permanent dropouts (Chuang,

1997; Entwisle et al., 2004; Rumberger and Lim, 2008). Initial dropouts who go on to earn a diploma at a later time are systematically different on a wide range of characteristics from those who have dropped out permanently. A common finding is that temporary dropouts have characteristics (family background and individual ability) and outcomes (education, employment, income) that are more similar to the graduating students than to the permanent dropouts.

Looking at the overall group of high school dropouts, all students who did not graduate with a diploma on time, more than 30 percent of these students will have received an ordinary diploma within 3 years of when they were supposed to graduate. This is considerably larger than for the group of marginal students. Figure 1 suggests that a mere 8 percent of the marginal students without a diploma at the time when they should have graduated have earned one after 3 years, and this share does not change much as as more time passes.

This implies that the ordinary dropout group is different than the marginal students in our analysis. Students who changed their study program or are otherwise delayed will constitute a relatively low share of permanent dropouts. Students who did not graduate because of temporary issues related to family or health will also have the opportunity to earn an ordinary diploma at a later time. Furthermore, these results imply that a very large share of marginal students become permanent dropouts. This is especially important as a large share of these students drop out because of random assignment to exam type, and might just as easily have graduated on time with a more favorable draw of exam type. This strong persistence further strengthens our view that this group of students is highly relevant for policymakers aiming to combat high school dropout.

Other outcomes

Failing students need to re-take the exam in order to obtain a diploma, typically at the start of the following semester. This means studying for an extra exam on top of the normal course load, which may have negative spillover effects for academic progression in other courses and places an extra burden on failing students. For some, this means re-optimizing the current course selection, picking less demanding schedules. In Table 6 we present several within-school outcomes ordered by timing from top to bottom. Starting with the first row, failing the exam unsurprisingly has a large negative effect on the probability of successfully finishing second year, as passing the exam is a requirement. As students do have some opportunities to re-take the exam immediately, the coefficient is not –1, but almost 80 percent of students induced to fail the exam by the instruments do not finish second year on time. However, there is no significant effect on the probability of starting third year, as having passed second year is not a prerequisite for starting third year. Row 4 indicates that compliers are less likely to finish high school the following year, but this result is not significant at conventional levels.

We find no significant impact on the number of courses the failing students register for (row 3),

Table 6: IV effects of exam failure on school outcomes

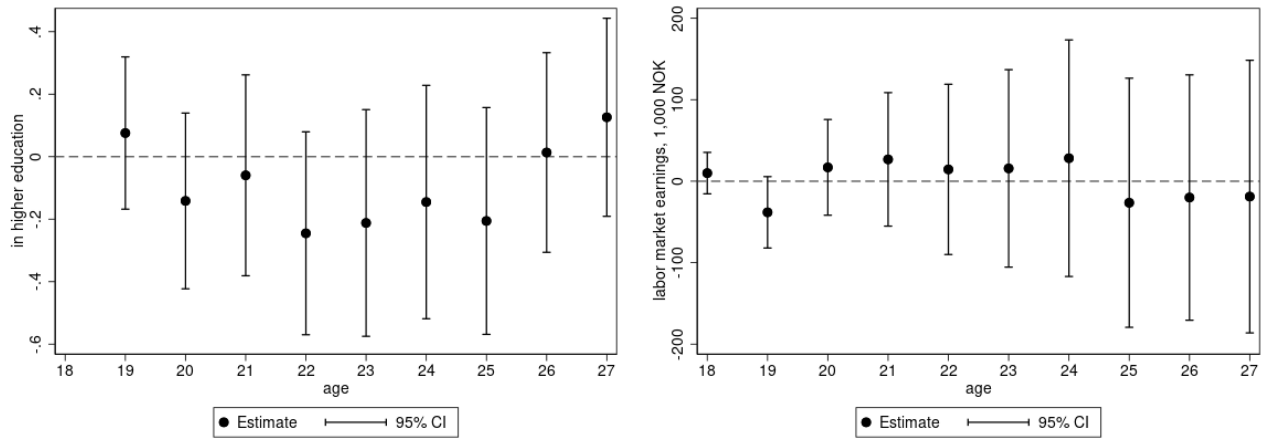
	Outcome	Fails exam		N
(1)	Finishes second year on time	-0.772***	(0.139)	76,948
(2)	Starts third year on time	-0.0894	(0.0692)	76,948
(3)	Number of courses in third year	-1.089	(1.083)	76,948
(4)	Finishes third year on time	-0.192	(0.160)	76,948
(5)	Passed courses in third year	-2.321**	(1.103)	76,948
(6)	GPA in third year	-0.641**	(0.289)	74,199

Notes: IV estimates using written exam interacted with teacher grade as the instrument for failing the exam. Each row contains results from a separate run of the interacted IV model using the outcome in the row header as dependent variable. GPA ranges from 1 (lowest) to 6 (highest). Controls include school, cohort and course fixed effects interacted with teacher grade. Standard errors in parentheses, clustered at school level. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

but they complete 2.3 fewer courses during third year (row 5). Their GPA on third-year courses is 0.6 points lower than passing students, a large effect corresponding to around 75 percent of a standard deviation of this variable. Note, however, that we lose a few thousand students for this outcome as they did not finish any courses from which we calculate their GPA. This sample reduction is probably endogenous to our instrument. Assuming students with written exams are relatively overrepresented among the students without observable GPA, this estimate should be biased towards zero, underestimating the true effect of exam fail on third-year performance for those who continue.

A natural next step is to ask how the lack of a high-school diploma affects these marginal students in terms of later education and earnings. We show this in Figure 2, where we use a dummy for whether a person is enrolled in tertiary education and labor earnings over time for each year from 18(age at start of the year of stipulated graduation) through 27 as the outcome of our IV model. As before, we keep the specification from Equation (2) and run the model separately by age. Unfortunately, precision is relatively low for these outcomes, but there is an 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 earnings is plotted in the right-hand 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 translate into long-term effects on education, as enrolling in tertiary education means the student will enter the labor market later. The coefficients are relatively close to zero, indicating that passing the exam, and perhaps having a high school diploma, do not matter much for the earnings of these marginal compliers, but unfortunately the estimates are far too imprecise to draw firm conclusions.



(a) In higher education

(b) Labor earnings

Figure 2: Long run effects of failing the exam

Note: Figures show IV estimates of the impact of failing the exam on the probability of being in higher education and on labor market earnings over time. All earnings measured in thousands of Norwegian Kroner (2019). Enrollment in tertiary education is an indicator variable equal to one if the individual is registered as enrolled into any tertiary education in that year.

Marginal student characteristics

Our headline estimate is that marginal students who fail a high-stakes exam due to the exam type are almost 50 percentage points less likely to obtain a high school diploma. We interpret this as the effect of a relatively small shock having long-term effects for a set of at-risk students because they are unable or unwilling to re-do the exam in order to obtain a diploma. For policymakers, learning about the characteristics of these marginal students from experiments such as this can be crucial for designing effective school interventions. To further understand who these students are, we therefore estimate the characteristics of the complier population at the fail margin and compare these students to the average student.

To this end, we estimate Abadie's kappa, a weighting scheme for compliers that allows us to estimate any statistical characteristic of compliers (Abadie, 2003). We first estimate the means of central covariates for compliers within each internal grade group, and then use the weights from the saturate and weight theorem on these subgroup-specific means to obtain overall means for our compliers.

These comparisons are provided in Table 7,[16] which for comparison also contains the means for the entire sample (column 1), among graduates (column 2) and dropouts (column3). Focusing on complier means in column (4), we see that disadvantaged students are overrepresented in the

[16]In addition to the means provided in Table 7, Figures C1 and C2 in the Appendix provide plots of PDFs and CDFs of our measures of past academic performance and parents' earnings among compliers and other parts of the sample, indicating that the academic performance and parent income of compliers are stochastically dominated by both the sample altogether and dropouts in particular.

Table 7: Characterizing compliers: Means of covariates across subgroups

	Overall sample			Compliers	
	All	Graduates	Dropouts	means	ratio (3) – (1)
Covariate	(1)	(2)	(3)	(4)	(5)
Female	0.54	0.56	0.47	0.39	0.72
Single parent	0.14	0.13	0.18	0.20	1.43
Immigrant	0.15	0.13	0.18	0.21	1.43
Mother higher educ.	0.37	0.39	0.30	0.22	0.60
Father higher educ.	0.40	0.42	0.34	0.27	0.68
Father's earnings	861.2	885.9	790.6	700.6	0.81
Mother's earnings	489.4	500.9	457.0	429.8	0.88
GPA, comp. school	4.49	4.61	4.18	3.88	0.86
Number of courses	10.86	10.96	10.59	10.57	0.97
GPA, other courses	4.05	4.24	3.50	2.99	0.74

Note: Columns (1)–(3) shows simple means among various subgroups of the sample. Column (4) show the estimated means among the sample of compliers from the interacted IV specification, constructed using Abadie's kappa as a weighting scheme within each grade group and then the weights from the saturate and weight-theorem to aggregate. Column (5) shows the ratios of the mean among the complier from column (4) to the overall mean in the sample for column (1), a measure of the relative overrepresentation of people with a particular statistic among the compliers. All earnings measured in thousands of Norwegian Kroner (2019). GPA ranges from 1 (lowest) to 6 (highest).

complier population. To give some examples, boys are overrepresented among the compliers compared to the full sample, and even compared to the sample of high school dropouts. Children from households where the parents have low education or low income and children from single parent households are likewise overrepresented in the complier group. The same is true for students with a weak academic history. Perhaps more surprisingly, disadvantaged children seem to be more prevalent among compliers to the exam type instrument than among dropouts as a whole, even though almost half the compliers end up with a primary diploma.

A survey article by Doll et al. (2013) aggregates several studies that document the reasons students report for dropping out of high school. They distinguish between schooling-related push factors and pull factors related to circumstances outside the school setting. Push factors include absenteeism, poor grades, or not keeping up with coursework, while pull factors can be work-related, pregnancy or other family-related factors. Similar results have been reported in Norway by Markussen and Seland (2012), which in addition highlight the importance of physical and mental illness.

Arguably, the marginal students in our analysis are most susceptible to push factors, since failing to keep up with the course progression or a poor academic performance will leave them vulnerable to the negative shock of a more difficult exam type. This helps to explain some of the differences we find in our main analysis between the overall dropouts and the marginal dropouts.

A higher share of the ordinary dropout group left school for reasons at least partly unrelated to school performance such as jobs, family situation, psycho-social issues or even starting a different education. If these students at a later age are motivated to go back to school, they might find it easier to obtain a diploma than the group of marginal students who dropped out due to the exam shock. We do, in fact, see that a much larger share of the ordinary dropout group goes on to earn a high school diploma in the years after the stipulated graduation date than the marginal students who drop out (33 pct. vs 8 pct. after 3 years).

From Table 7 we see that the marginal students (Column 4) typically have poorer academic performances than the dropout group (Column 2), which is to be expected if the marginal students are indeed more strongly affected by negative academic shocks. Strikingly, in most characteristics the dropout students resemble the graduates more closely than they resemble the marginal students. However, the most apparent difference between the marginal students and the overall dropout group is in characteristics that are associated with reduced fatherly input such as a low-earning father, low-educated father and single parent household. Figure C1 in the Appendix show the different sample distributions for compulsory school GPA, high school GPA, fathers' earnings and mother' earnings. All distributions for the marginal students are first-order stochastically dominated by all other sample distributions, including ordinary dropouts.

Targeting marginal students

We now return to the question of how best to target policies in order to effectively combat high school dropout. We have argued that it is more effective to target marginal students than at-risk students in general. In order to describe at-risk and marginal students and how a policymaker may target these students, we estimate a simple linear probability model of a dummy for not attaining a diploma on time on various background characteristics, presenting the results in Table 8. In Column (1), we estimate this regression on the entire sample, showing that socioeconomic background characteristics such as immigrant status, single parent household and parents' earnings and education predict dropout, even when conditioning on the obvious covariates that measure student ability. If policymakers want to target at-risk students, one way to do it is to target students with disadvantaged family backgrounds as described in Table 8.

When Abadie's kappa is used as a weighting scheme to estimate the same relationship among our compliers in column (2), this relationship is gone. Socioeconomic status as measured by single parent status, immigrant status and parent earnings and education no longer predicts dropout when conditioning on student performance. For most of the coefficients, the lack of a significant relationship is not only driven by the increased standard errors, but also by the coefficients dropping towards zero in Column (2) compared to Column (1). This suggests that even though disadvantaged socioeconomic background characteristics predict dropout behavior

Table 8: Predicting dropout among all students and among marginal students

Covariate	(1) All	(2) Marginal students
Female	0.000456	-0.00990
	(0.00367)	(0.0113)
Single parent	0.0242***	0.0184
	(0.00511)	(0.0155)
Immigrant	0.0220***	0.00242
	(0.00630)	(0.0142)
Father has higher education	0.00784**	-0.00305
	(0.00371)	(0.0144)
Mother has higher education	0.0185***	0.00214
	(0.00338)	(0.0131)
Father's earnings	-0.00467**	-0.0132
	(0.00198)	(0.00947)
Mother's earnings	-0.00893	-0.00177
	(0.00558)	(0.0209)
GPA, compulsory school	-0.0339***	-0.0616***
	(0.00697)	(0.0126)
Number of courses	-0.0207***	-0.0137***
	(0.00328)	(0.00394)
GPA, other courses second year	-0.215***	-0.245***
	(0.00497)	(0.00976)
Effective N	76,948	34,418

Note: The table shows the results of regressions of a dummy for dropout (not obtaining a diploma on time) on various background characteristics. Column (1) is for the entire sample.Column (3) uses Abadie's kappa weighting scheme separately within groups of internal grade, then computes a weighted average of these five coefficients using the weights from the saturate and weight theorem. This produces an average relationship between each covariate and dropout behavior among marginal students - our compliers. All earnings measured in millions of Norwegian Kroner (2019). GPA ranges from 1 (lowest) to 6 (highest). Standard errors are clustered by school, $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

among students in general, conditioning on student ability, this is not the case for marginal students as identified by our instrument. In order to design effective policies to combat high school dropout, our results suggest that policymakers should focus exclusively on school performance in order to target marginal students who are more likely to benefit from the program.

6 Conclusion

This paper investigates the consequences of failing a high-stakes exam on future performance in school and on graduation. By exploiting the unique quasi-random assignment to exam type in Norwegian high schools we address selection issues and estimate causal effects of passing the high-stakes exam. For struggling students even a relatively small added burden can mean the difference between graduating with a diploma and permanently dropping out of high school.

We show that exam type appears to be as good as randomly assigned conditional on course choice, and exploit this to instrument for failing the exam. This allows us to isolate a group of marginal students who fail the exam when assigned the written exam, but pass when assigned an oral exam. A large share of these marginal students are shifted out of school altogether by failing the exam. This allows us to shed light on the response to an exam shock for a group of students on the brink of dropping out of school and potentially providing information about the effects of other policies aimed at combating high school dropout.

Our results show that assignment to a more difficult written exam does indeed shift otherwise identical students into worse performance on the exam, increasing fail rates by 3.4 percentage points in the entire sample and by as much as 12.1 percentage points in subgroups of students with low predetermined ability. As passing the exam is required in order to eventually graduate, this translates into large and long-term consequences for the compliers. More than 55 percent fail to earn a primary diploma on time and more than 45 percent of them still do not hold any high school diploma by age 27, even though they are allowed to re-take the exam. We also find evidence of dropout already after year two of high school, and some indication of worse performance in the final year of high school.

Importantly, our complier group comprises students who fail the exam when assigned to the written exam, but would pass if they were assigned to the simpler oral exam. We show that this group is severely disadvantaged, as measured by various predetermined characteristics such as parents' earnings and education and their own past academic performance, even compared to other dropouts.

Furthermore, we provide evidence that background characteristics predict dropout in different ways for students in general and the marginal students identified by our instrument. In particular, we show that while important measures of socioeconomic disadvantage predict dropout among students in general, even conditional on past performance, this is not the case among

marginal students. This suggests that policies could improve efficiency by targeting students that share characteristics with our compliers rather than targeting at-risk students in general, who may include students with no real chance of switching from dropout to graduation. As policymakers aiming to effectively combat high school dropout should target marginal students rather than at-risk students in general, our results suggest that targeting on the basis of socioeconomic background characteristics as well as school performance may be less effective than targeting exclusively on the basis of school performance. Of course, different policies might want to target different types of students, but there are almost certainly efficiency improvements to be gained by thinking carefully about who is on the relevant margin and how they can be targeted by policy.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231 – 263.

Abadie, A., Gu, J., and Shen, S. (2019). Instrumental variable estimation with first-stage heterogeneity. Technical report, Working paper.

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In Card, D. and Ashenfelter, O., editors, *Handbook of Labor Economics*, volume 4 of *Handbook of Labor Economics*, chapter Chapter 12, pages 1043 – 1171. Elsevier.

Acemoglu, D. and Restrepo, P. (2018). Low-skill and high-skill automation. *Journal of Human Capital*, 12(2):204–232.

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

Andresen, M. and Huber, M. (2020). Instrument-based estimation with binarized treatments: Issues and tests for the exclusion restriction. Working paper.

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.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

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

Autor, D. H. and Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the us labor market. *American Economic Review*, 103(5):1553–97.

Belfield, C. R. and Levin, H. M. (2007). *The Price We Pay : Economic and Social Consequences of Inadequate Education*. Brookings Institution Press.

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

Bernstein, L., Rappaport, C. D., Olsho, L., Hunt, D., and Levin, M. (2009). Impact evaluation of the us department of education's student mentoring program. final report. ncee 2009-4047. *National Center for Education Evaluation and Regional Assistance*.

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.

Chuang, H.-L. (1997). High school youths' dropout and re-enrollment behavior. *Economics of Education Review*, 16(2):171–186.

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.

Doll, J. J., Eslami, Z., and Walters, L. (2013). Understanding why students drop out of high school, according to their own reports: Are they pushed or pulled, or do they fall out? a comparative analysis of seven nationally representative studies. *Sage Open*, 3(4):2158244013503834.

Dynarski, M., Clarke, L., Cobb, B., Finn, J., Rumberger, R., and Smink, J. (2008). Dropout prevention. ies practice guide. ncee 2008-4025. *National Center for Education Evaluation and Regional Assistance*.

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.

Entwisle, D. R., Alexander, K. L., and Olson, L. S. (2004). Temporary as compared to permanent high school dropout. *Social forces*, 82(3):1181–1205.

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

Freeman, J. and Simonsen, B. (2015). Examining the impact of policy and practice interventions on high school dropout and school completion rates: A systematic review of the literature. *Review of Educational Research*, 85(2):205–248.

Galloway, T. A., Kirkebøen, L. J., and Rønning, M. (2014). Grading practices in norwegian middle schools,. *Statistics Norway Report*, (14).

Goos, M., Manning, A., and Salomons, A. (2009). Job polarization in europe. *American economic review*, 99(2):58–63.

Guryan, J., Christenson, S., Cureton, A., Lai, I., Ludwig, J., Schwarz, C., Shirey, E., and Turner, M. C. (2020). The effect of mentoring on school attendance and academic outcomes: A randomized evaluation of the check & connect program. Working Paper 27661, National Bureau of Economic Research.

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.

Hernæs, Ø., Markussen, S., and Røed, K. (2017). Can welfare conditionality combat high school dropout? *Labour Economics*, 48:144 – 156.

Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., Smith, M., Bullock Mann, F., Barmer, A., and Dilig, R. (2020). The condition of education 2020.

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.

Kearney, M. S. and Levine, P. (2016). Income inequality, social mobility, and the decision to drop out of high school. *Brookings Papers on Economic Activity*, 47(1 (Spring)):333–396.

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.

Levin, H. M. (1987). Accelerated schools for disadvantaged students. *Educational leadership*, 44(6):19–21.

Levin, M. (1986). Educational reform for disadvantaged students: An emerging crisis. *West Haven, CT: NEA Professional Library*.

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

Markussen, E. and Seland, I. (2012). Å redusere bortvalg-bare skolenes ansvar? en undersøkelse av bortvalg ved de videregående skolene i akershus fylkeskommune skoleåret 2010-2011.

Marshall, J. (2016). Coarsening bias: How coarse treatment measurement upwardly biases instrumental variable estimates. *Political Analysis*, 24(2):157–171.

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*.

Olea, J. L. M. and Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, 31(3):358–369.

Oreopoulos, P., Brown, R. S., and Lavecchia, A. M. (2017). Pathways to education: An integrated approach to helping at-risk high school students. *Journal of Political Economy*, 125(4):947–984.

Pflueger, C. E. and Wang, S. (2015). A robust test for weak instruments in stata. *The Stata Journal*, 15(1):216–225.

Rumberger, R. W. and Lim, S. A. (2008). Why students drop out of school: A review of 25 years of research.

Stock, J. H. and Yogo, M. (2005). Testing for weak instruments in linear iv regression. In Andrews, D. W. K. and Stock, J. H., editors, *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. New York: Cambridge University Press.

Tharmmapornphilas, R. (2013). Impact of household factors on youth's school decisions in thailand. *Economics of Education Review*, 37:258 – 272.

Appendix

A Multivalued treatment and alternative models

Table A1 provides balancing tests within teacher grade cells, relevant for the interacted model in eq. 2. This table suggests that exam type is balanced with respect to a range of covariates, even within teacher grade cells: of the 50 coefficients reported, only 4 indicate any statistical significance, 2 of them at the 10% level. Furthermore, the covariates are not jointly significant for explaining exam type in any teacher grade cell. This is important evidence supporting the exclusion restriction. As briefly discussed in the main body of the paper, however, the exclusion restriction might be violated even if exam type is balanced if the instrument affects other margins of treatment than those captured by the endogenous variable (Angrist and Imbens, 1995), causing "coarsening bias" (Marshall, 2016). Andresen and Huber (2020) show how this may happen when a) the instrument affects treatment at other margins than the one used to construct the binary treatment indicator and b) these instrument-induced changes at off-threshold margins of treatment affect outcomes, in addition to providing a testable implication of b). We argued in the main body of the paper that it is likely that the fail margin is far more important than other margins of treatment in our setting, suggesting that a) might hold. This appendix provides results and robustness checks for alternative models designed to address this concern directly, and largely confirms our baseline estimates.

To allow the instrument to affect multiple treatment margins, we exploit the fact that students of different initial ability are

affected at different margins. Specifically, we instrument not only for the fail margin, but also for the exam score itself:

$$\begin{aligned} y_{igkt} &= \alpha_{gk} + \pi_{gt} + \kappa_{c(i)g} + \theta_F F_{igkt} + \theta_s F_{igkt} + \epsilon_{ikt} \\ F_{igkt} &= \delta^F_{gk} + \eta^F_{gt} + \lambda^F_{gc(i)} + \textstyle\sum_{h=2}^{6} \gamma^F_h \mathbb{1}(h=g) Z_{igkt} + \mu^F_{igkt} \\ s_{igkt} &= \delta^s_{gk} + \eta^s_{gt} + \lambda^s_{gc(i)} + \textstyle\sum_{h=2}^{6} \gamma^s_h \mathbb{1}(h=g) Z_{igkt} + \mu^s_{igkt} \end{aligned} \tag{3}$$

where s is the exam score (1 to 6) and F, as before, is a dummy for failing the exam. As in the interacted model with a single endogenous regressor, α and δ are school-by-teacher-grade fixed effects, π and η are cohort-by-teacher-grade fixed effects and κ and λ are exam-course-by-teacher-grade fixed effects. The parameters of interest are θ^F, the effect of failing the exam, and θ^s, the effect of getting a one grade improvement in exam score. This model exploits the fact that higher ability students are much less affected at the fail margin than weak students, and allows us to a) estimate the effects of the score margin itself and b) control for effects at the score margin when investigating effects at the fail margin.

There are now two endogenous variables and two first stage regressions, both of which are

Table A1: Balancing tests: interacted specification

Teacher grade	2	3	4	5	6
Female	-0.00125	-0.00292	0.00104	0.000514	-0.00888
	(0.00516)	(0.00328)	(0.00302)	(0.00340)	(0.00787)
Single parent	-0.00246	0.00422	0.00618	0.00720	0.0103
	(0.00616)	(0.00465)	(0.00463)	(0.00531)	(0.0112)
Immigrant background	-0.00601	-0.00820	-0.000198	0.00379	0.00449
	(0.00711)	(0.00558)	(0.00417)	(0.00453)	(0.0126)
Father has higher education	-0.0108*	0.00596*	-0.00209	0.000464	0.00282
	(0.00646)	(0.00357)	(0.00317)	(0.00350)	(0.00870)
Mother has higher education	-0.00139	-0.00631	-0.000931	0.000990	-0.00349
	(0.00597)	(0.00401)	(0.00304)	(0.00331)	(0.00802)
Father's labor earnings,	-0.00131	0.000814	0.00177	-0.000379	0.00204
(millions of NOK (2019))	(0.00192)	(0.00251)	(0.00185)	(0.00105)	(0.00233)
Mother's labor earnings,	0.00214	-0.00408	-0.00105	-0.00259	-0.00420
(millions of NOK (2019))	(0.00777)	(0.00836)	(0.00459)	(0.00482)	(0.0103)
GPA, compulsory school	0.00863	0.000240	0.0125***	-0.00570	0.0122
	(0.00580)	(0.00439)	(0.00457)	(0.00451)	(0.0124)
Number of courses	-0.000573	0.00206	0.00244	0.00190	0.00321
	(0.00201)	(0.00176)	(0.00152)	(0.00220)	(0.00254)
GPA, other courses	-0.0125***	-0.00125	-0.00550	0.000196	0.00955
	(0.00468)	(0.00355)	(0.00365)	(0.00424)	(0.00986)
N	9,305	18,262	25,196	20,030	4,155
F	1.171	1.329	1.277	0.855	1.347
p	0.300	0.194	0.226	0.602	0.185

Note: The table reports coefficients for balancing variables from a regression of a dummy for written exam on covariates indicated in the row header separately by teacher grade as indicated in the column header. Included and tested in the F-tests, but not reported, are dummies for a few hundred observations with missing variables for mother's earnings and compulsory school GPA, as well as less than 3,000 observations with missing data for father's earnings. The specification includes teacher grade-specific fixed effects for school, cohort and exam course. The specification includes coefficients (not reported) for dummies for missing mother's and father's earnings data and compulsory school GPA, which have been set at sample means. Standard errors are clustered on school. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

shown in table 4 b) in the main paper. As described in the main paper, the instrument affects the fail margin much more strongly for weak students (see column (1)), and the joint F-statistic for all groups is 21.2, well above the critical value for weak instruments of 18.4 from Stock and Yogo (2005). For the linear score in column (6), the instrument has a relatively similar effect across student ability: Students of all ability levels see exam scores dropping around 0.5 to 0.6 grades when assigned the written rather than the oral exam, with the exception of the weakest students, where the scope for reducing grades is naturally limited and the effect on the exam score is around -0.33. Jointly, the instrument is strongly significant with an F-statistic above 90.

IV estimates for this specification are provided in column (3) in table A2. For comparison, the baseline results from eq. (2) is repeated in column (2), and the results from the simple model in (1) are reported in column (1). Two patterns are evident in this table: First of all, there is little evidence of any exam score effects at other margins, with significant estimates only for starting and finishing third grade, and very small and insignificant estimates for our main graduation outcomes in rows (10) and (11). Second, the main effects of failing the exam are very stable to this way of controlling for other exam outcome margins. If anything, the effects of failing the exam on graduation are larger than in the baseline model in column (2) and still highly significant. Table (A2) also reports results from the simple model from eq. (1). In line with expectations, the results are less precise in this specification, but it is reassuring to see that point estimates are relatively similar for most outcomes, in particular our main high school graduation outcomes. This loss of precision illustrates that the interacted model is more efficient in a setting with strong first stage heterogeneity such as ours.

Table A2: Robustness: IV effects of exam fail with alternative models

		(1)		(2)		(3)				N
	Model	Simple		Interacted		Interacted 2				
	Outcome / endogenous variable	Fail exam		Fail exam		Fail exam		Exam score		
(1)	Finishes second grade same year	-0.790***	(0.169)	-0.772***	(0.139)	-0.776***	(0.141)	-0.000558	(0.00834)	76,948
(2)	Starts third grade following year	-0.0142	(0.0516)	-0.0894	(0.0692)	-0.126	(0.0807)	-0.00622**	(0.00300)	76,948
(3)	Number of courses following year	-0.258	(1.564)	-1.089	(1.083)	-1.541	(1.126)	-0.0764	(0.0849)	76,948
(4)	Finishes third grade following year	0.159	(0.235)	-0.192	(0.160)	-0.411**	(0.183)	-0.0371***	(0.0142)	76,948
(5)	Passed courses following year	-1.259	(1.527)	-2.321**	(1.103)	-2.821**	(1.202)	-0.0846	(0.0861)	76,948
(6)	GPA following year	-0.927*	(0.474)	-0.641**	(0.289)	-0.607*	(0.321)	0.00527	(0.0205)	74,199
(7)	Number of similar courses following year	-0.112	(0.898)	0.184	(0.517)	0.120	(0.545)	-0.0108	(0.0530)	76,948
(8)	Passed similar courses following year	-0.485	(0.927)	-0.172	(0.527)	-0.249	(0.559)	-0.0132	(0.0551)	76,948
(9)	GPA in similar courses, following year	-0.567	(0.648)	-0.451	(0.373)	-0.611	(0.404)	-0.0222	(0.0266)	69,124
(10)	Primary diploma on time	-0.428*	(0.233)	-0.555***	(0.153)	-0.621***	(0.167)	-0.0111	(0.0128)	76,948
(11)	Any diploma on time	-0.331	(0.233)	-0.469***	(0.162)	-0.548***	(0.181)	-0.0134	(0.0131)	76,948
(12)	In higher education at age 22	-0.105	(0.206)	-0.236	(0.165)	-0.318	(0.194)	-0.0137	(0.0138)	75,604
(13)	Earnings at age 27, 1,000 NOK (2019)	-83.22	(156.1)	-41.25	(112.3)	-8.625	(129.0)	5.490	(10.47)	75,116
	First stage F (full sample)	81.98		21.19		21.19		90.31		
	Fixed effects									
	Cohort	✓		✓ × G		✓ × G				
	School	✓		✓ × G		✓ × G				
	Exam course	✓		✓ × G		✓ × G				

Notes: IV estimates using alternative models. ✓ indicates inclusion of fixed effect,✓ × G indicates inclusion of fixed effect interacted with teacher grade in exam course. Instrument is written exam in col (1), written exam interacted with teacher grade in row (2) and (3). Each row contains results from a separate run of our IV models using the outcome in the row header as dependent variable. Standard errors in parentheses, clustered at the school level. $^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$

B Grade inflation and instrument monotonicity

The monotonicity assumption means that all who are affected by the instrument are assumed to be affected in the same direction, which in our setting means that no one can get a better exam score by drawing a 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 specific exam score bonus for oral exams that persists across several school characteristics. Second, nearly all schools inflate the internal grades and oral exam scores of their students, and 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 overshadows most individual differences in exam-taking ability.

Table B1 shows the average grades and examination characteristics in our sample at school-year level. In all sub-samples, the average oral exam scores are higher than the average written exam scores, with the average GPA and teacher given exam course grade 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 a written exam than they would have if selected for an 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 B1a shows that internal grades and oral and written exam scores all increase with school enrollment size. Nonetheless, the difference between the exam score 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 B1b 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 is nowhere near the difference in written exam score. A one grade increase in the average written exam scores results in a 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 oral exams, on the

Table B1: School grades

	2004	2005	2006	2007	All schools	Large schools
GPA	4.02	4.05	4.06	4.04	4.05	4.10
Internal grade	3.85	3.89	3.91	3.89	3.89	3.96
Oral exam score	4.12	4.19	4.27	4.18	4.19	4.28
Written exam score	3.51	3.42	3.45	3.32	3.42	3.51
Oral exam gain	0.18	0.21	0.21	0.18	0.19	0.19
Written exam gain	-0.29	-0.43	-0.41	-0.52	-0.42	-0.38
Written exam penalty	-0.47	-0.63	-0.61	-0.69	-0.61	-0.57
School-level compliers	0.88	0.96	0.93	0.96	0.94	0.95
Number of exams	95.9	104.9	104.9	115.8	106.0	147.8
Number of oral exams	28.6	30.6	29.2	33.2	30.5	43.6
Number of written exams	67.3	74.4	75.7	82.6	75.4	104.3
Share oral exams	0.29	0.29	0.28	0.29	0.29	0.30
N schools	274	281	279	281	1.115	272
N students	17.330	18.889	18.864	21.765	79.948	38.074

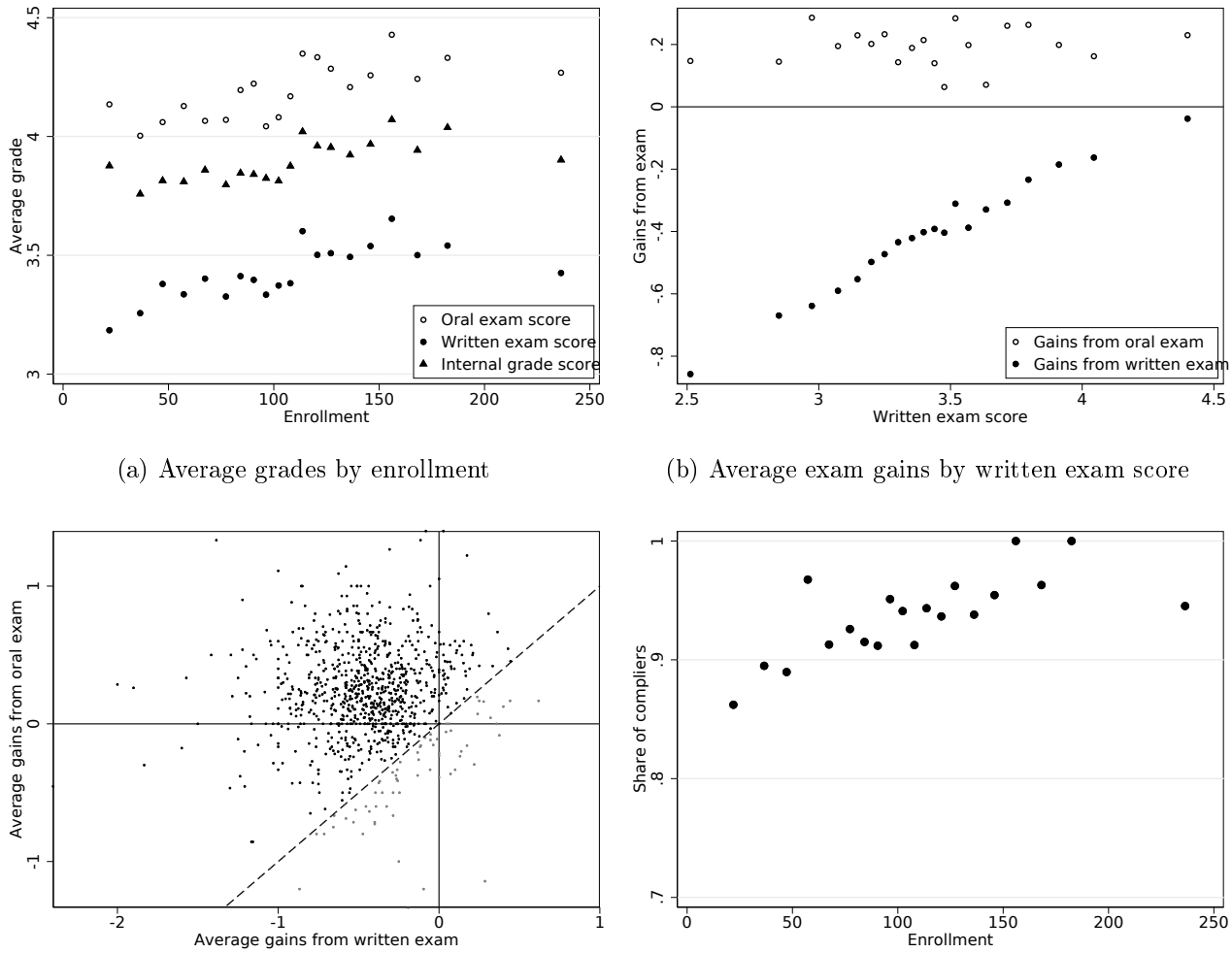
other hand, seem to be stable across schools with different levels of written exam results. This suggests that school- and teacher-level grade inflation constitute a large part of the written exam penalty we observe, but there is also an oral exam-specific benefit. Figure B1c plots the oral exam gains against the written exam gains at school-year level, where compliers are above and to the left of the dashed 45 degree line. Most of the observations are concentrated in the second quadrant, where oral exam gains are positive and written exam gains are negative.

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 gain from written exams does not strictly dominate the benefit the student gains from oral examinations, such as help from the subject teacher and exposure to grade inflation, the potential outcome of an oral exam will still be higher than that of 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 point lower than their own internal grade in the same subject, and 95.8 percent of students attend schools where the average gains attributable to written exams are negative. For oral exams, students are on average graded more than 0.2 grade point higher than their internal grades, and average gains

are higher than written exam gains for 96.3 percent of schools. This translates into an average written exam penalty of -0.6 grade point.

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, especially oral exams, can be quite low in some schools in some years, which increases the probability that random noise from student-exam allocation or exam day conditions will dominate the written exam penalty. This interpretation is supported by the decline in the share of schools with non-negative written exam penalties as the number of exams increases, as shown in Figure B1d. This can also be seen in Table B1, as the share of school-level compliers is higher in large schools (0.95) relative to all schools (0.94). To further show that this is not caused by the existence of some special defier schools, we calculate the average written exam penalty at school level by pooling all four years of our sample. Then the share of school-level compliers increases to 99 percent for 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 B1: 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 written and oral exams gains relative to internal grades in panel b). Scatter plots of average oral and written exam gains at school-year level in panel c) excluding school-years with less than 5 exams of either exam type (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 written exam gains exceed the average oral exam gains 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.

C Other results

C.1 Robustness checks and decomposition

C.2 Complier characteristics

Table C1: Robustness and decompositions

A: Subgroup specific IV estimates

Teacher grade	2	3	4	5	6
First stage	0.121***	0.0524***	0.0122***	-0.00173	0.00319**
	(0.0198)	(0.00669)	(0.00234)	(0.00232)	(0.00135)
F	37.14	61.31	27.11	0.554	5.628
Reduced form	-0.0684***	-0.0158	0.0108	-0.000207	-0.0160
	(0.0220)	(0.0161)	(0.0125)	(0.00993)	(0.0212)
IV estimate	-0.566***	-0.302	0.889	0.120	-5.004
	(0.196)	(0.305)	(1.060)	(5.781)	(6.801)
N	9,305	18,262	25,196	20,030	4,155
weight	0.820	0.169	0.00989	0.000216	0.000744

B: Alternative models

	(1) baseline	(2) saturate and weight	(3) Weighting eq (1)	(4) Weighting eq (2)	(5) LIML
Failing exam	-0.469***	-0.510***	-0.511***	-0.523***	-0.471***
	(0.162)	0.170	(0.166)	(0.172)	(0.165)
N	76,948	76,948	76,948	76,948	76,948

Note: Panel a) shows teacher-grade specific first stage, IV and reduced form estimates, all controlling for school, year and exam course fixed effects. Bottom row shows the relative variance of the residual, instrument-generated variation in exam fail, correspoding to the subgroup weight in the overall IV estimate by the saturate and weight theorem (Angrist and Imbens, 1995). Panel b) shows alternative estimators, where column (2) weights the subgroup specific weights from panel a), column (3) estimates eq. (1) using the weights from panel a) and column (3) estimates equation (2) with a pooled instrument, weighting with the weights from panel a). Finally, column (5) uses Limited Information Maximum Likelihood to estimate eq. (2), an estimator with favorable properties with weak instruments. Standard errors clustered by school, $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

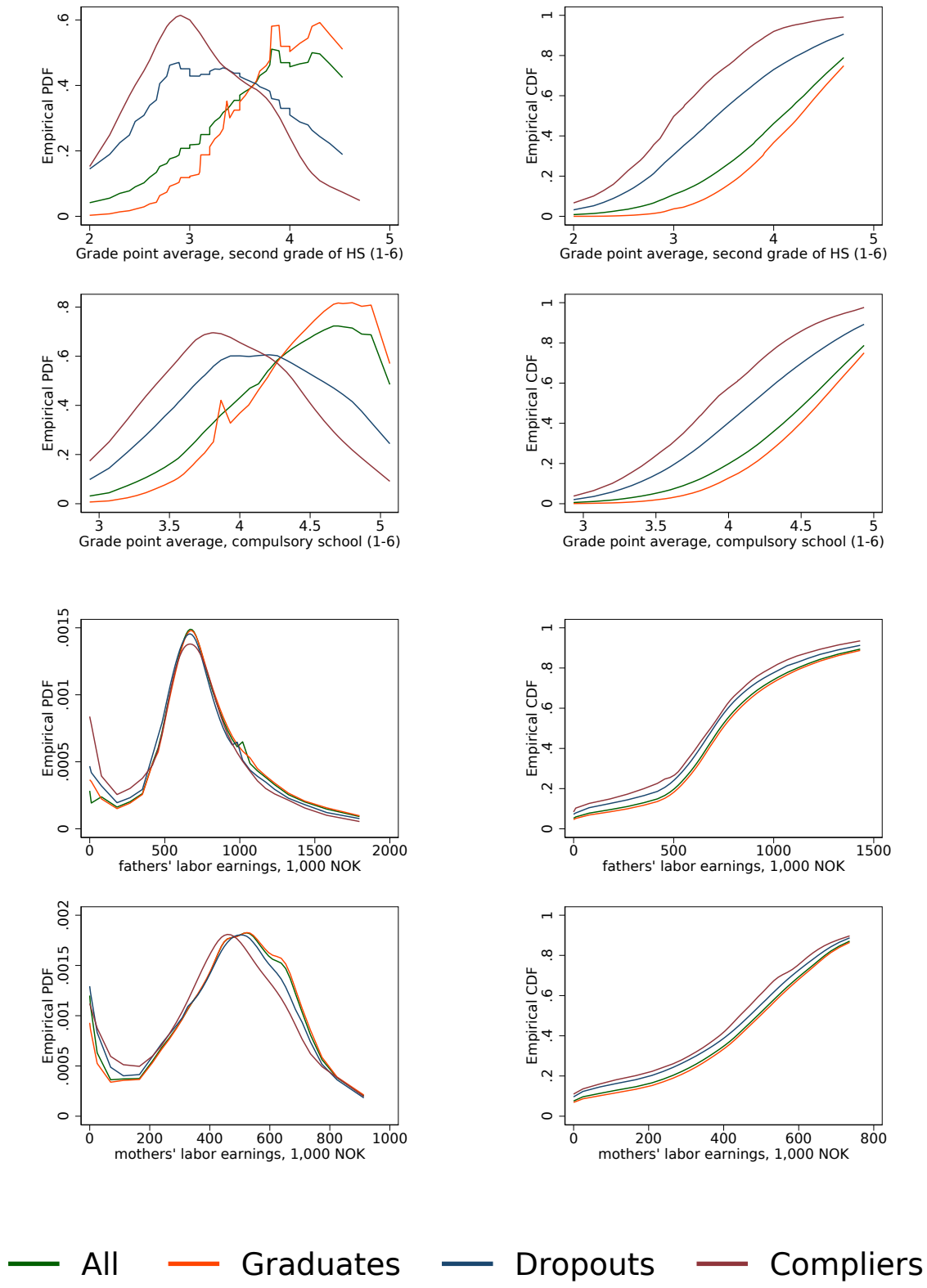


Figure C1: Distribution of complier characteristics

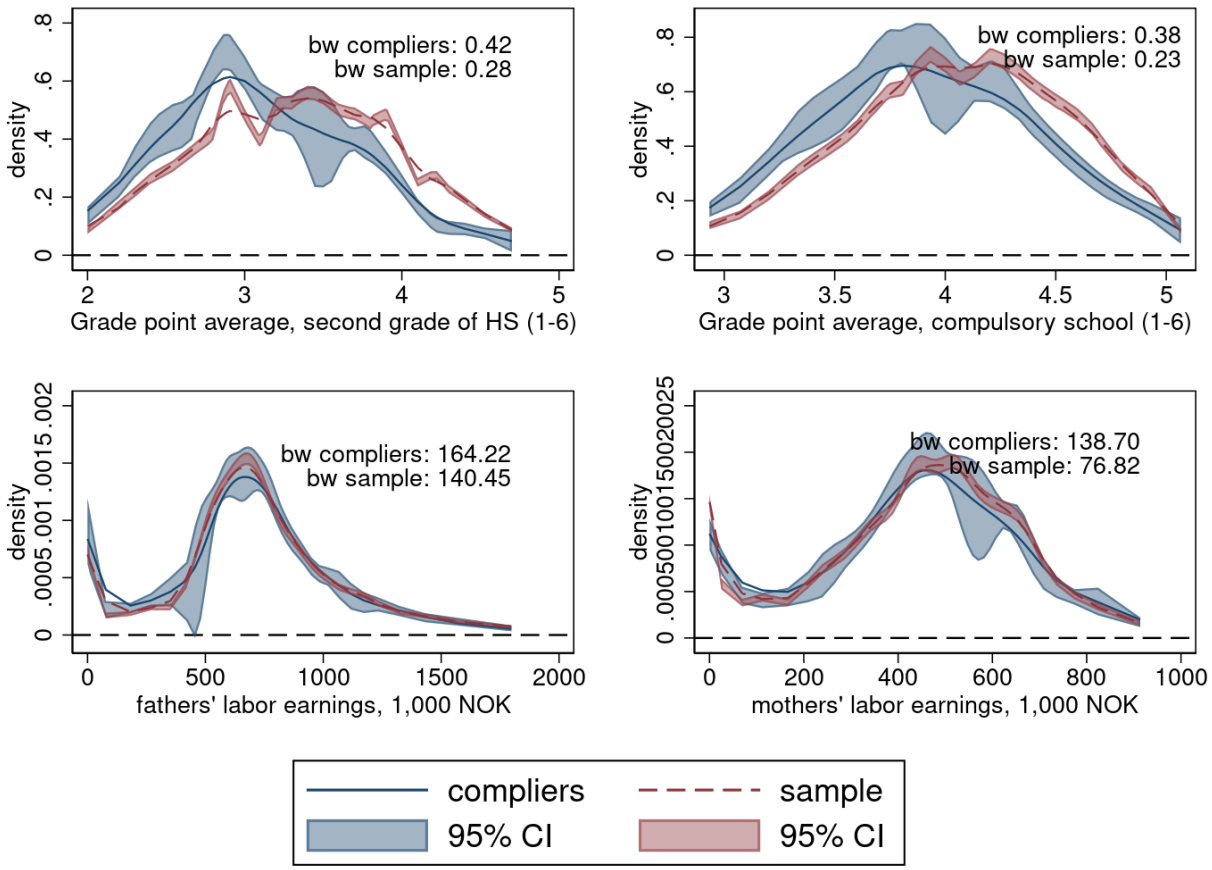


Figure C2: PDFs of compliers and the weighted sample