Economics of Education Review 94 (2023) 102377

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Using high-stakes grades to incentivize learning☆

Andreas Fidjeland *

Nordic Institute for Studies of Innovation, Research and Education, Oslo, Norway & University of Stavanger Business School, Stavanger, Norway

ARTICLE INFO

JEL Codes:
D02
D04
I20
I28

Keywords:
Incentives
High-stakes testing
School choice
Learning

ABSTRACT

I use a natural experiment in Norwegian high school to investigate how high-stakes grades affect students' investment in schooling. By exploiting variation across space and time I compare the performance of students taking the same exit exam in compulsory school, but where the test is high-stakes for only a subset of students. Using a staggered triple-difference framework, I find that exam grades increase in the high-stakes setting if students have a sufficient number of prospective high schools within traveling distance. Results from low-stakes ability assessments suggest actual learning — and not test-taking strategy — could largely explain the effect.

1. Introduction

Investments in human capital can yield great economic returns both for the individual and society. Typical models for the production of human capital posit that both public inputs such as investments in school resources, facilities, and teachers, and on private inputs such as student effort are necessary. However, students often fail to make a sustained effort in school, perhaps because short-term costs are more salient than rewards materializing in adulthood (Levitt et al., 2016). Whereas economic research has provided a number of policy prescriptions for the design of public inputs, it is less clear how policymakers can influence private investments by students. This might be particularly challenging in the case of adolescents, who tend to be less intrinsically motivated than younger students (Eccles & Midgley, 1989; Eccles et al., 1993), instead increasingly seeking external sources of motivation and validation at ages when they transition to middle and high school (Harter, 1981; Midgley et al., 1995).

Economic theory predicts that we are motivated by incentives. We would therefore expect grades to provide students with stronger incentives to learn in cases where they are high-stakes, than if they are low-stakes (Becker & Rosen, 1992; Grove & Wasserman, 2006; S.L. Wise & DeMars, 2005). This is illustrated by recent evidence suggesting that low effort could explain why many developed countries produce subpar performances in cross-country ability assessments, despite an overwhelming advantage in educational expenditure (Gneezy et al., 2019; Zamarro et al., 2019). If proper incentives can motivate students to exert a sustained learn- ing effort, their improved effort should also increase human capital production. However, we have very little knowledge on the extent to which adolescent students respond predictably to non-pecuniary incentives in the school setting (Bach & Fischer, 2020).

One way to raise the stakes of grades — and to move the rewards from investing more effort in school closer to the present — is to adopt merit-based enrollment regimes when allocating students to schools. A key argument for such policies rests on the hypothesis that letting students compete for access will incentivize effort if they prefer to enroll in specific schools, thereby promoting academic achievement (Friedman, 1962; Hoxby, 2003). However, we have little direct evidence in support of such a disciplinary effect, particularly on young students. This paper is therefore relevant for the many cities and countries that have introduced variants of merit-based high-school enrollment (e.g., Paris, Sweden, and the United Kingdom) but lack causal evidence of their effect on academic performance. It might also be informative for high school contexts where teacher-set grades and high-stakes national exams jointly determine access to higher education.

To investigate the incentivizing effect of high-stakes grades on academic achievement, I exploit a natural experiment created by regional

☆ This paper has benefited from helpful guidance on the part of Mari Rege, Ingeborg Solli, and Eric Bettinger, as well as from excellent comments by Tom Dee, Edwin Leuven, Hans H. Sievertsen, Maximiliaan Thijssen, Sandra McNally and the participants in the UiS PhD Workshop in Education Economics. I acknowledge funding from the Norwegian Research Council, Grant No. 270,703/H20. All remaining errors are my own.

* Corresponding author.
E-mail address: andreas.fidjeland@nifu.no.

https://doi.org/10.1016/j.econedurev.2023.102377
Received 28 March 2022; Received in revised form 7 December 2022; Accepted 15 February 2023
Available online 3 March 2023

differences in Norwegian high-school admission regimes. Whilst historically the norm has been for students to enroll in their neighborhood high school, several counties have in recent decades chosen to adopt merit-based enrollment regimes, more colloquially referred to as "school choice" policies. In these counties, over-subscription to schools is solved by ranking students according to their compulsory school grade-point average (GPA), admitting those with the highest average first. Given that school placement is determined by grades in some counties but not in others, economic theory predicts that students exposed to school choice will attain higher grades, provided that school placement is an outcome they care about. Using rich registry data from a sample period covering six dif- ferent school-choice reforms across Norway, I exploit the county–year variation in enrollment regimes in a triple-difference framework. To mitigate concerns that county-specific trends or shocks might influence the decision to introduce such reforms, I leverage the supply of schools within traveling distance from a student's home as the third difference. Specifically, I differentiate in terms of whether a student, in practice, has a real choice of high schools, defined as having at least three schools within traveling distance. If the prospect of being able to choose your high school is a driver of student performance, students should not be induced to invest more effort if they have few geographically realistic options to choose from however well they perform. This means that the triple-difference model not only estimates prereform and postreform trends in the reforming counties as compared with nonreforming counties, but also leverages *de facto* nonchoice students as a within-treatment placebo group.

To ensure that changes in grading practices in response to the reforms are not driving my results, I focus my attention on how students perform on the national, centralized exam that all Norwegian students are required to take at the end of compulsory school. This is the first mandatory national exam faced by Norwegian students and represents their last chance to improve their GPA, as the teacher-awarded grades are finalized before the exam (but not revealed to the students until after it). Qualitative studies indicate that Norwegian teenagers experience high-school choice as a critical stage in their schooling, with far-reaching implications for their educational and labor market prospects, and that earning good grades is therefore vital to them (Bakken et al., 2018; Inchley et al., 2013; Ruud, 2018).

My results suggest that imposing more high-stakes grades has a positive effect on grades earned on the exam. I find robust estimates of a treatment effect of 4–6 percent of a standard deviation for those students who are both exposed to a school-choice reform and have a sufficient number of schools within traveling distance — that is, those for whom the exam might actually be experienced as high-stakes. I find limited evidence that the reforms had any heterogeneous effects on performance across subgroups, with only some suggestive evidence that the effect is stronger for students tested in mathematics.

There are at least two mechanisms that could explain the effect of higher stakes. First, the students' test effort could change if students faced with a high-stakes exam put in more effort ahead of, and during the test itself. This could include adjusting their test-taking strategy (e. g., taking more risks) or making sure to sleep and eat well in the days before the exam. If so, the treatment effect would have limited relevance for human capital development. The second explanation, which has stronger policy implications, is that students facing high-stakes grades will make a sustained learning effort over time in order to acquire the skills required to succeed on the exam. From a policy perspective, the latter explanation suggests that changes to students' incentive structure can be instrumental in increasing their investment into schooling, with potentially long-lasting effects on subsequent educational and labor market outcomes.

Results from low-stakes national assessment tests conducted in the grade prior to the exit exam indicate that average academic ability increased among exposed students in the wake of the reforms relative to the control group. This evidence suggests that the learning-effort hypothesis is important for explaining the main effect. This is also corroborated by a dynamic response in the treatment effect, where larger effect sizes are observed for cohorts further re- moved in time from the reforms. This increasing effect is consistent with the notion that stu- dents will adapt to the new regime over time, with younger cohorts increasingly aware of the importance of making a sustained effort throughout their schooling and not just toward the end of their final year.

My paper contributes to several strands of literature. First, the results are relevant for the literature examining the links between incentives and academic achievement. A rich accountability literature has documented how schools, administrators, and teachers might respond to stricter performance standards and outcome-based funding (see Figlio & Loeb, 2011, and Deming & Figlio, 2016, for surveys). However, the present study considers a setting where incentives change for the students only. In contrast to many other studies on related topics (e.g., Gibbons et al., 2008, and Figlio & Hart, 2014), the compulsory schools are unaffected by the reforms to high-school admission and have no reason to adjust their behavior or effort. When it comes to student-level effects, a separate but related body of work uses direct financial incentives to increase effort and performance in test-taking situations (e. g., Angrist & Lavy, 2009; Behrman et al., 2015; Bettinger, 2012; Burgess et al., 2022; Fryer, 2011; Kremer et al., 2009; Leuven et al., 2010). Several of these experimental studies have successfully demonstrated a causal link between extrinsic incentives, motivation, and effort among students, although their effectiveness in moving outcomes has been modest (Levitt et al., 2016). Pay- ing students for their performance is also costly in the long term and may not be feasible on a national scale, limiting the policy relevance of this body of research. My paper is therefore most closely related to Hvidman and Sievertsen (2021) and Bach and Fischer (2020), which consider how students respond to nonpecuniary incentives. The former work considers a grade re-scaling reform in Danish high schools that led to students' GPA being arbitrarily raised or lowered, finding that those students who experienced a fall in their GPA, which determines postsecondary enrollment, responded by performing better in subsequent years. The authors argue that enhanced study effort is a plausible explanation for this effect. The latter work exploits changes in Germany's tracking system in early primary school. In this case, students face a choice between different ability tracks rather than schools, where some states use bind- ing recommendations from the teachers based on previous performance. The authors find that relaxing the emphasis on the recommendation in favor of more parental choice reduces student achievement, presumably owing to the reduced incentive to perform well.

On a related note, the paper adds to the literature aimed at understanding how competitive behavior implemented through school-choice regimes can influence the efficiency of educational production (e.g., Angrist et al., 2002; Cullen et al., 2006; Figlio & Hart, 2014; Hoxby, 2000; Lavy, 2010). Theoretical studies suggest that allowing parents and students to choose schools freely will improve the quality and productivity on both the supply and the demand side through the disciplinary effect of competition (Becker & Rosen, 1992; Costrell, 1994; Friedman, 1962; Hanushek, 1986). Further, there could also be a positive sorting effect as a result of students (or parents) being allowed to make choices that better fit their needs and preferences, leading to more efficient allocation of students across schools (Epple & Romano, 2003; Hoxby, 2003). However, a weakness of this literature is that outcomes are often mea- sured after the right to choose has been exercised. This makes it difficult to evaluate whether any gains achieved by introducing school choice are indicative of greater learning effort on the part of students or are instead the result of students being in different schools and peer groups. Unlike this literature, I do not study the effect of school choice *per se*, but rather investigate whether the prospect of being *able* to choose, given sufficient academic success, can incentivize students to improve their performance earlier in their education. Hence my results give a clearer indication of the disciplinary effect of high-stakes grades on student behavior, as opposed to school responses to competitive

pressure or the effects of changing peer groups.

Lastly, my paper contributes causal evidence to the interdisciplinary stream of research into the significance of test consequences for performance. The notion that academic tests devoid of consequences will be too low-stakes to make students perform to the best of their abilities is well established in the literature (A. Wise & DeMars, 2005). Although the results in many cases stem from correlational studies, existing empirical work indicates that motivation and effort are associated with test stakes, while the evidence regarding performance is more mixed (Napoli & Raymond, 2004; Wolf & Smith, 1995). A primary challenge in this literature, as highlighted by a recent vein of research (Gneezy et al., 2019; Segal, 2012; Zamarro et al., 2019), is separating effort and ability in test-score outcomes. If policymakers are more interested in the students' ability than in their test scores *per se*, the policy relevance of the association may be undermined by the fact that the correlation between test stakes and performance might simply reflect innate differences in intrinsic motivation and stress resistance (Levitt et al., 2016). My paper provides evidence suggesting that students respond to incentives by exerting effort over time, thereby raising their academic ability. This highlights a channel for policymakers to stimulate private investment into schooling.

2. Background

2.1. Institutional setting

The setting for this study is the universal, publicly funded primary and lower-secondary school (henceforth "compulsory school") in Norway, in which attendance is free and mandatory. Norwegian compulsory school comprises ten grades and ends in graduation in the year when students turn 16.[1] Private options are limited, with the public-school participation rate exceeding 96% in 2016 (Norwegian Directorate of Education & Training, 2017). The allocation of students to individual compulsory schools is decided on the basis of neighborhood catchment areas. Since having inclusive schools with heterogeneous groups of students is a policy objective, formal parental influence on which school their child attends is limited. In the first seven years, no grades are awarded, as competition between students is played down in favor of focusing on individual development. Classroom tests might be given but are typically not scored or ranked and primarily serve as a tool for the teacher to chart the progress of individual students. Grades 8 through 10 represent a separate stage of compulsory school, and students are typically required to change schools after grade 7; this typically also entails being assigned to a new class.[2] Parental influence on assignment to classes or schools remains limited, and nor is there any tracking at this stage. Indeed, The Education Act (Opplæringslova) (1998) specifies that the classes should reflect the aggregate population, without consideration of ability, gender, or ethnicity, effectively advocating as-good-as random assignment of students to classes.

Grade 8 also marks the introduction of teacher-assessment grades. In general, grades 8 through 10 represent a more advanced level of study, where subjects are more academically and theoretically oriented and where students are regularly assessed using graded tests and assignments. Every semester, students are given a transcript consisting of a grade on a scale from 1 to 6 for each subject, set by their teachers. However, only those grades received at the end of year 10 will enter their official school record. The final teacher-assessment grades (in all subjects) along with the grades from the final exit exam make up a student's compulsory-school GPA, with all grades given equal weight. The exit-exam grade is one out of approximately 13 grades on the transcript, meaning that the direct impact of the exam on school placement may be limited for the student population as a whole. Even so, a two-step increase in the grade earned on the exam will by itself move a student roughly five percentiles up in the GPA distribution, which is more than enough to have a real impact for students who are at the margin between two schools. Moreover, whether the incentive represented by the exit exam has a performance- enhancing effect depends not so much on its objective impact as how it is perceived by students. Both Norwegian and cross-country surveys indicate that Norwegian students experience above- average levels of school-related stress toward the end of compulsory school (Bakken et al., 2018; Inchley et al., 2013). Some studies report that students in grade 10 link stress to internal and environmental pressure to perform well, so that they do not spoil their chances of obtaining a good education and having successful careers (Bakken et al., 2018). Anecdotally, some students claim that not getting accepted to their preferred school would mean that "everything is ruined" (Ruud, 2018). The final exam represents the last opportunity to better their chances of admission to their preferred school, and it is therefore likely that many students will experience it as high-stakes.

After graduating from compulsory school, students can apply to enroll in high school. While not mandatory, students have a statutory right to acceptance for upper-secondary education, and very few end their education before or immediately after finishing compulsory school. When applying to high school, students make their first choice of education track, choosing between a variety of vocational and academic programs.[3] Within programs, the allo- cation of students between high schools is left to county-level politicians' discretion and varies from county to county.

2.2. High school enrollment reform

High-school admission in Norway is based on one of two opposing regimes. The neighborhood-catchment (NC) regime follows the principles of compulsory school in requiring students to attend their nearest school, that is, the high school closest to their place of residence that offers their preferred educational program. Proponents of the NC regime emphasize that this allows students to stay close to home, limiting lengthy commutes and keeping youths attached to their local communities. It also serves to promote heterogeneity within the student body, as it constrains students' ability to self-select into specific schools (on parameters other than pro- gram preferences). In contrast the school-choice (SC) regime allows students to apply to any school within their county. This includes the option of applying to the same type of program in several schools, or for several different programs in the same school. In densely populated areas, there will typically be several schools offering the same programs. Where the number of applicants exceeds school capacity, students are ranked by compulsory-school GPA, with the highest scores being prioritized.[4] The cutoff for admission to a particular school is thus equal to the GPA of the last student admitted in that particular year (in the case of ties, admission officials will perform a random draw between those at the cutoff). Cutoffs to specific schools vary substantially with their popularity and perceived quality and will

[1] In the Norwegian educational system, grades 1–7 make up primary school while grades 8–10 make up lower-secondary school, which is roughly equivalent to junior high school in the United States.

[2] In this context, "class" refers to a set group of students within a cohort who share a classroom and attend most subjects together. A class typically stays together for all three years of lower-secondary school.

[3] The vocational track leads to an apprenticeship within a trade. The primary function of academic-track programs is to prepare students for higher education.

[4] For a few programs, such as music and sports, there are additional tests for ability in the domain area. Moreover, in certain instances some counties also take into account a student's travel distance, but this is done on a discretionary case-by-case basis.

also fluctuate from year to year in accordance with application patterns.[5][5] It is therefore hard to predict with certainty which GPA will be required to get accepted to a school in a given year, for example for students at the margin of acceptance compared to previous years' cutoffs. Hence the SC regime places a significant emphasis on the grades attained by students in compulsory school, meaning that the final exit exam involves higher stakes for students in SC counties than for students in NC counties.

In the last decades an increasing number of counties have been adopting the SC regime. In the first year of my sample period, 2002, eight of nineteen counties were already using an SC regime. The variation exploited in this study is provided by six counties that expanded school choice throughout the 2000s, with reforms carried out in 2003, 2005, 2009, 2012 (two counties), and 2014.[6] Thus, in the final year of my sample period, 2015, only five counties still applied an NC regime. The timing of the reforms allows me to observe outcomes both before and after the reforms, but with varying length. The geographical distribution of admission regimes in the first and last years of my sample period is illustrated in Fig. 1. The SC reform decisions followed a timeline similar to that presented in Fig. 2. Thus, students in their final year at the time of the relevant vote had only their last semester to adjust to the new regime.

A county survey of the student population conducted in the wake of one such reform indicated that SC disrupted existing enrollment patterns (Arbeidslaget Analyse, Utgreiing og Dokumentasjon, 2005). In the county of Hordaland, one-quarter of the first cohort affected responded that their preferred high school was not the one they would have been assigned in an NC regime, and 75 percent of those had succeeded in enrolling in their first-choice school. Of the remaining students, who would have preferred to enroll in their geographically closest school, 85 percent were accepted by their first-choice school. In both cases, acceptance rates indicate that enrollment was competitive. However, there is substantial heterogeneity across geography and ability, with the most popular schools being located in city centers. Teacher responses suggest that the primary realignment effect brought about by merit-based enrollment consists in allowing high-ability students in suburban and rural areas to enroll in popular ur- ban schools, displacing low-ability students from the city centers who have to settle for less competitive schools further away. This is consistent with another evaluation of the Hordaland reform, which finds that introducing merit-based enrollment had positive effects on student performance, and links the effect to a substantial increase in school fragmentation suggesting that many students did in fact seek to enroll in schools outside of their neighborhood when given the opportunity (Haraldsvik, 2014).

2.3. Conceptual framework

The plausibility of the causal between high-stakes grades and performance rests on the hypothesis that linking performance to desirable outcomes creates an incentive that motivates students to exert more effort in school. We would expect an increase in school effort if students perceive, first, that such an increase is clearly related to performance in the relevant domain and, second, that the possible outcomes are of sufficient value to them (Wigfield & Eccles, 2000). In economic terms, we will expect students to put effort into schooling if they expect long-term rewards to exceed the short-term costs of the effort (Levitt et al., 2016). Receiving a grade is not enough in itself to elicit such a response if the associated rewards are not of sufficient magnitude (Grant & Green, 2013). One of the goals of the Norwegian school-choice admission regimes is to implicitly provide a reward through merit-based enrollment.

High-stakes grades can be expected to be a more effective incentive for some students than for others. Some studies have suggested that motivation to learn is an innate individual characteristic or trait (Brophy, 1987; Segal, 2012). Students who have a strong motivation to learn (whether innate or not) would be expected to work hard and try their best, even in the absence of any extrinsic incentives that policymakers might offer. Segal (2012) finds that students displaying these traits also perform well on low-stakes assessments, suggesting that they are already properly motivated to capitalize on learning opportunities even when there is no tangible benefit to be gained. Hence high-stakes grades can be expected to provide a more effective incentive for students who do not exhibit those characteristics. Assuming that such students invest strategically in school, effort levels will also vary across individuals as a function of students' relative probability of achieving their desired outcome (Vroom, 1964). Further, it is often assumed that effort and ability are complementary, and that the marginal effect of effort on human capital production increases with ability (Oettinger, 2002). If this is so, high-stakes grades will primarily improve the performance of low-effort, high-ability students.

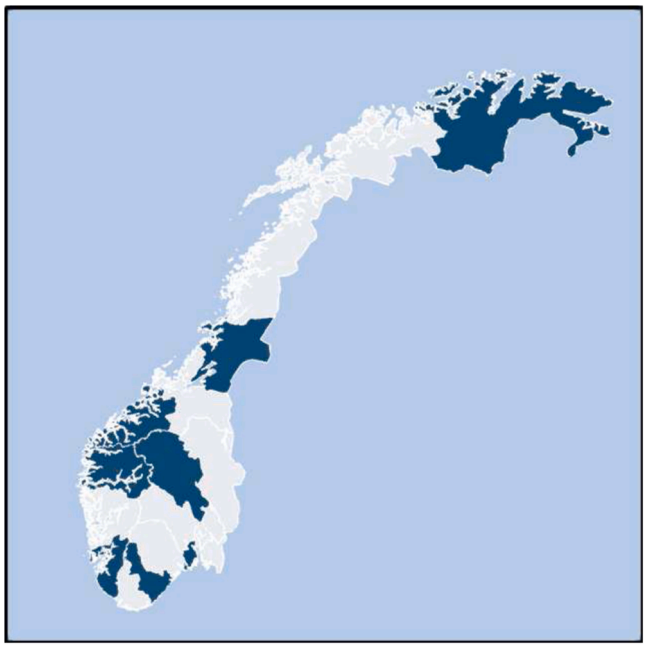
Effort might also be negatively correlated with ability if high-achieving students are able to attain the maximum grade with less effort than average students (Stinebrickner & Stinebrickner, 2008). As we can reasonably assume that motivation at least partly maps to performance through effort, it is also reasonable to assume that many high-achievers will already be sufficiently motivated. We might therefore expect a stronger effect among low-achievers for whom faltering motivation could be the cause of their underperformance. Additionally, some studies have demonstrated that boys respond more than girls to the extrinsic incentives of a competitive environment (Azmat et al., 2016; Hopland & Nyhus, 2016).[7] Provided that boys outnumber girls in the low-achieving segment of the student population, a stronger treatment effect on boys would indicate a stronger effect for low-ability students.

Tying test performance to desirable outcomes might also change the way students approach the test itself. Since exams map a continuous ability distribution to an arbitrary, discrete scale, the expected marginal benefit of performing better is conditional on a given student's latent ability level prior to the exam. If a student is not near the margin between grades, the short- term expected marginal benefit of effort is close to zero, while the marginal costs are positive. Thus, we would primarily expect to see an effect on students whose latent ability level is close to a point where they could earn a higher (or fall to a lower) grade, and therefore have positive expected marginal benefits from investing effort. In line with this theoretical argument, some experimental studies have noted that the effect of introducing extrinsic incentives is greatest for a "marginal group" of students who have success within their reach (Angrist & Lavy, 2009). For example, Burgess et al. (2022) find that while their experimental incentive scheme failed to produce meaningful effects on average, a subset of students (those who scored low at baseline) showed significant gains after being exposed to the intervention. They argue that even though we might not expect large average effects there will often be a group of "right tail" students for which the incentive might be very powerful. In the setting of this paper, we would perhaps not expect to see any substantial effect on the treatment group as a whole. However, for students who perceive themselves to be at the margin between grades, such an incentive might represent a sufficient nudge to make them put in more effort.

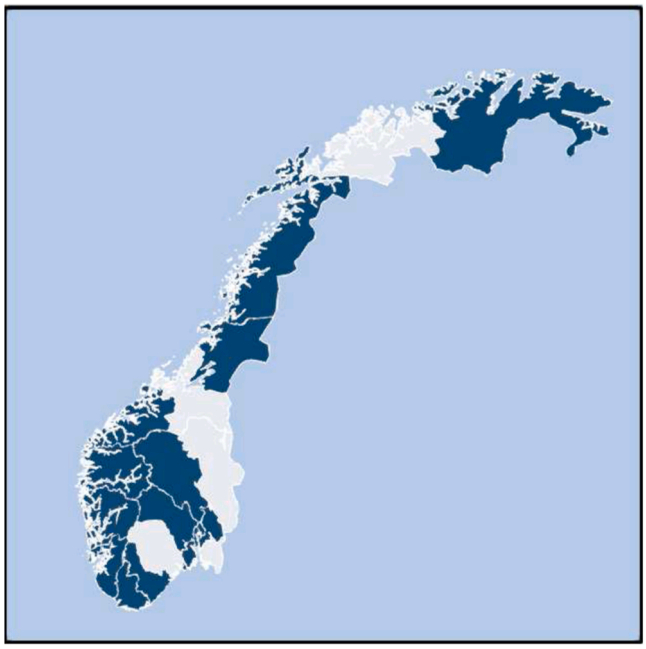
[5] For the least popular schools, admission will typically be uncontested, while it is not uncommon for the cutoff in the most popular urban schools to exceed a 5.0 GPA (out of a possible 6). Information about previous years' cutoffs in specific schools is made available to students.

[6] Specifically, Akershus (2003), Hordaland (2005), Oslo (2009), Vest-Agder (2012), Buskerud (2012) and Nordland (2014).

[7] It should however be noted that this finding is not conclusive. For example, Hvidman and Sievertsen, 2021 find that the incentivizing effect of having your GPA downgraded was strongest for girls, although both genders responded predictably.



(a) 2002



(b) 2015

Fig. 1. Spread of School Choice Regimes in Norway
Note: Illustration of the increase in school-choice regimes in Norwegian counties during the 2002–2015 period. Dark shading of counties indicates some kind of school choice being in effect for students graduating from compulsory school in that particular year.

3. Data and analysis

3.1. Data

The study relies on comprehensive registry data retrieved from the Norwegian National Database of Education, maintained by Statistics Norway. The registry of interest contains compulsory-school outcomes of every student enrolled in a Norwegian school who graduated from grade 10, and it covers the entire student population in the sample period. The sample is limited to 14 adjacent cohorts during the period from 2002 to 2015, which include a total of 856,040 individuals. Central to the analysis are records detailing the final grades attained by each student in all subjects, both through teacher assessments and through written and oral exams. Additionally, the registry contains information about the subject in which a student was tested on the final exit exam as well as about when and where students graduated. Individual identifiers

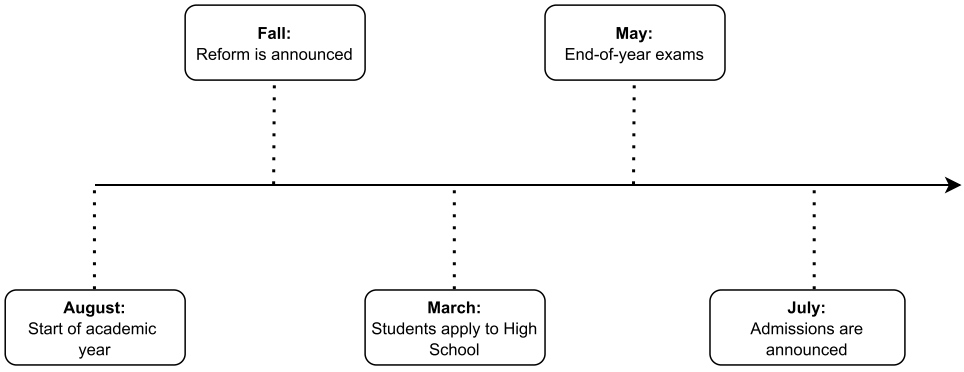


Fig. 2. Timeline for the School Choice Reforms
Note: Overview of the series of events of the school choice reforms in the sample.

allow me to link school outcomes to other registries that provide rich details about demographic characteristics, socioeconomic status (SES), and family origin. These identifiers also allow students to be matched with their parents, producing a rich set of potential covariates that can be controlled for in the estimates.

The focus of the analysis is on students graduating from compulsory school. As each student is only observed once (at the time of graduation), the data are organized as a repeated cross-section, with dummies indicating from which county and in what year a particular student graduated. Graduation takes place in the spring, and most students subsequently enroll in high school the following August. Cohorts are therefore referred to using the year in which they left compulsory school.[8] Similarly, the reforms are deemed to be in effect starting with the first cohort whose members are able to exercise expanded choice in their high-school applications.[9] Details of current high school admission systems are available in each county's regulations (see www.lovdata.no). Some of these also contain notes about significant changes made to the admission regulations, but typically they do not include detailed information about the timing of reforms. To determine when reforms were implemented, I rely on two investigations carried out at the request of members of Parliament that provide additional details on which counties adhered to which systems at the times in question (Dokument 8:41, 2006; Dokument 8:8, 2003). However, as the most recent of those investigations was carried out in 2006, I have supplemented information from public records of county-parliament sessions for later cohorts. In addition, I have cross-checked those records with newspaper articles from local media in the relevant counties to determine the exact timing of the reforms.

3.2. Measures and variables

The key outcome variable is a student's grade on the final exit exam in grade 10, standardized to a mean of 0 and a standard deviation of 1 for ease of interpretation. In the final semester of compulsory school all students are randomly drawn for testing in a centrally administrated written exam in either mathematics, English, or Norwegian.[10] The draw is randomized at the class level, and it is the responsibility of the municipalities in each county to implement the draw in a manner that ensures an even distribution of students across exam subjects, and of exam subjects across schools (Norwegian Directorate for Education & Training, 2018). All students selected for testing in the same subject will take the exact same exam on the same day, and their exam papers will be graded externally by compulsory-school teachers in another part of the country. Both students and teachers remain anonymous throughout the grading process, which uses the same integer scale from 1 (fail) to 6 (top) as teacher-assessment grades and is based on an absolute standard criterion. This anonymity throughout the process and the use of external graders makes the exam grade a more reliable outcome measure than the full GPA, because I cannot rule out whether teachers' grade-setting practices are affected by the reforms. It could for example be the case teachers in school-choice counties are more lenient in an attempt to help their students gain admission to their preferred school. This is clearly less of a concern when the teacher grading an exam does not know who the student is or where they live.[11]

In order to gauge whether students have a real choice of schools, I construct a measure of the number of high schools within traveling distance from the student's home. To determine whether a school belongs to a particular student's choice set, I use the commuting zones (CZ) in which students reside.[12] These represent geographically demarcated areas at a level between county and municipality, that cannot cross county borders. In this regard, the definition of the CZs corresponds with the constraints put on students' high school applications which are also limited to schools in within county borders.[13] Municipalities are defined to be part of the same CZ in part of a sufficient amount of commuting is observed between them. The resulting zones are sub-county regions that thus roughly equates to a geographical within which traveling distances are such that an employee could be expected to commute to work on a daily basis. The variable for the number of schools available to a student thus indicates the number of schools located in his or her commuting zone of residence in the year when he or she graduated from compulsory school. Since there are two main educational tracks to choose from in high school (academic and vocational), I define "real choice" as having at least three high schools within your commuting zone. By doing so, I ensure that at least one of the main tracks will be available in at least two

[8] For example, the cohort enrolled in Grade 10 in the 2002/2003 academic year is referred to as the 2003 cohort.

[9] If students graduating from compulsory school in the spring of 2003, in a reforming county, can exercise school choice the following fall, the reform is defined as being implemented in 2003.

[10] Additionally, students are tested in an oral exam with a similar randomized draw. However, in this case all subjects are eligible for testing, and the exam is carried out locally at each school. The grade from this exam is also added to the student's GPA.

[11] For the curious reader, I include results from using a GPA constructed from all nonexam grades as the dependent variable in Table B.3 in the appendix. The effect sizes in this analysis are largely similar to those estimated in the main analysis.

[12] Definitions and demarcations of these zones are given in an overview provided by Statistics Norway — which refers to them as "economic areas."

[13] There has been a debate in recent years whether cross-county school choice should be allowed, but this was not the case in the sample period.

different schools in that region.[14] A total of 599,885 observations (76 percent) satisfy this condition. However, as most Norwegian high-school students will not be able to obtain a driver's license until their final year (the age limit is 18), the commuting zones probably approximate to the *maximum* traveling distance that a student would consider for a daily commute. Because of their reliance on public transport and other means of transportation, this definition will likely overstate the true choice set that a student would consider, which will bias effect sizes toward zero.

In addition to the commuting zone and the cohort-specific fixed effects necessary to estimate DID and triple-difference models, I control for a rich set of conventional covariates. The Central Population Registry provides details on students' gender, nationality, and year of birth. Records of immigration status are used to construct an indicator of immigrant background, defined as being either a first-generation immigrant or born in Norway but having at least one parent born outside of Norway. Using unique identifiers, I link students with their parents, in order to collect data on parental education and income. Education (the highest level of education completed by each parent) is measured on Statistics Norway's nine-point scale.[15] For income, I use the registered taxable income in Norwegian kroner from official tax records for both parents in the year that the student graduated, with household income being the sum of these incomes rounded to the nearest 1000. Then I divide, for each year separately, households into deciles according to income rank; this is the variable that I include in my analyses. Assuming these covariates are unaffected by the treatment their inclusion in the models should have limited impact on the estimated treatment effects. I primarily include them to reduce noise and increase precision, and show in the main results that my conclusions are not sensitive to whether or not I include them.

3.3. Sample selection

The estimation sample is constructed from the universe of 858,306 individuals having graduated from compulsory school during the years 2002–2015. Of these students, 3057 were exempted from taking the exit exam (e.g., owing to special education needs) and 835 were confirmed sick on the day of testing. A further 1863 students did not show up for the exam without providing a reason for their absence. In accordance with Norwegian guidelines, these were not given a failing grade but rather marked as "Not graded." In the present sample, these cases are coded as missing values. An additional 61,605 observations are missing, mostly due to a large teachers' strike in 2008 that caused exams to be canceled. However, attrition analysis — available in Table C.1 in the appendix — shows that grade missingness is not predicted by treatment status. In total, 68,370 observations without exam grades are excluded from the analysis, leaving an estimation sample of 790,936 unique student-level observations. In cases in which a student is registered with multiple graduation years and outcomes (true for 2556 students, 0.29 percent of the gross sample), I use the earliest observed result. In cases where information is missing for covariates, dummies for missing values are constructed and included accordingly, and the covariates are set to zero.

3.4. Summary statistics

Table 1 details summary statistics for the estimation sample. Column 1 lists mean values and standard deviations for key variables computed for the treated counties (those that implemented reforms to high-school enrollment during the sample period). Column 2 lists corresponding values for the control counties.

Since Norway has a homogeneous population, there are few disparities in the demographic composition of the two groups. One noteworthy exception is the share of immigrants, which is markedly higher in the treated counties. Those counties also have higher levels of average household income than the control counties, despite there being no discernible difference in education level. This is probably due to the fact that some of Norway's largest urban areas, which have a higher frequency of income outliers, are among the reforming counties. This fact is also reflected in the average number of schools available to students as well as in the size of the county cohorts. The average student in the treated counties has thirteen high schools within his or her commuting zone and belongs to an average graduating cohort of 100 students per school. By contrast, students in control counties have an average of six high schools to choose from and the average graduating cohort per school there consists of 89 students.[16]

Students are — by design — evenly distributed between exam subjects. The only discrepancy found with regard to the exam-subject draw is that the sample share of students tested in Norwegian is roughly 10 percent smaller than that for the other subjects. This is due to the aforementioned teachers' strike in 2008, which overlapped with the exam in Norwegian which ended up being canceled. By contrast, exam performance varies substantially. Fig. 3 shows that the likelihood of earning the bottom two grades is markedly higher for those selected to be tested in mathematics, all else being equal. In fact, mathematics exams account for three- quarters of all failing students while over half of the students who obtained the top grade were tested in English. One

Table 1
Summary Statistics.

	Treated		Control	
	Mean	SD	Mean	SD
Background characteristics				
Female	0.488	(0.50)	0.488	(0.50)
Year of birth	1992.6	(4.11)	1992.6	(4.12)
Age at graduation	16.09	(0.95)	16.09	(1.04)
Immigrant	0.125	(0.33)	0.070	(0.26)
Mother's education	13.32	(2.97)	13.14	(2.673)
Father's education	13.47	(2.87)	13.12	(2.59)
Household income	893.2	(1690.8)	790.8	(785.7)
Educational setting				
Number of HS in CZ	13.10	(9.44)	5.74	(4.67)
Share with >2 HS in region	0.83	(0.38)	0.70	(0.46)
Number of students in school	100.8	(53.26)	88.75	(51.79)
Written exam subject				
Math	0.38	(0.48)	0.38	(0.48)
English	0.36	(0.48)	0.36	(0.48)
Norwegian	0.26	(0.44)	0.26	(0.44)
N	350,858		440,078	

Note: Summary statistics for all students in treated counties compared with the control group. Standard deviations in parentheses. The treat- ment group consists of the six counties which implemented high-school enrollment reforms during the 2002–2015 period. All nonreforming counties constitutes the control group. Immigrant is defined as having at least one parent who was born outside of Norway. For the education measure, I convert Statistics Norway's nine-point scale for an individual's highest completed degree to years of education using their own conventions. For reference, completing high school is equal to 13 years of education. Household income is reported in nominal NOK/1000. "HS" = high school, "CZ" = commuting zone.

[14] I assess the sensitivity of the results to this definition in the appendix. Please refer to Section 4.2 for more details

[15] See Statistics Norway (2001) for details.

[16] While differences in observable characteristics do not bias the results in a DID design *per se* (unless underlying trends overlap with the timing of the reforms, which is particularly unlikely in a triple-difference setting), I do control for a rich set of conventional predictors of school achievement, such as parental background and socioeconomic status, in all my estimations in order to increase the precision of the models.

potential concern is that changes in the composition of draws across treatment status and time could threaten the identification strategy. However, considering that the subject draw is randomized within schools across classes, it is unlikely that this would be the case.[17]

3.5. Empirical strategy

3.5.1. The triple difference model

The empirical model of interest in this study is the linear relationship between student performance and high-stakes grades (as proxied by the high-school admission regime), as expressed in Eq. (1).

$$y_i = \mu D_i + \varepsilon_i \tag{1}$$

If students were randomized to admission regimes, the binary variable D_i in (1) would identify an unbiased causal effect on some outcome y_i of exposure to high-stakes grades. However, it is plausible to claim that students are exposed to either regime in a nonrandom fashion. This gives rise to concerns that (1) would falsely attribute mean differences between the student groups to the regime to which they are exposed.

My approach to overcome this identification issue is to exploit the fact that counties implemented merit-based enrollment at different points in time, in a difference-in-differences setup (DID). Using this approach, we can estimate the effect of being exposed to a such a policy change by taking the difference between pretreatment and post treatment periods for both the treatment group and the control group, and then the difference between these two differences. These estimates have a causal interpretation under the assumption that in the absence of an intervention, the trends in outcomes would be equal for treatment and control units, so that any observed deviation from this trend is attributable to the policy change of interest. However, in a setting where the reform is a political decision, this assumption might be problematic, as there could be unobserved trends in outcomes induced particular counties to consider school- choice reform in the first place. Further, these reforms could be the result of changes in political leadership that also led to other changes at the county level around the same time, and those other changes might be correlated with student outcomes. In Fig. A.1 in the appendix I chart the average trend in exam grades for the treatment and control groups centered around the treatment point for the treated units. These trends suggests that the identifying assumption holds only modestly well, and does not allow for a conclusive rejection of the possibility that the treatment group is on a different pretreatment trend than the control group. This raises concerns about the causal nature of DID estimates of the effect of the policy reforms.

To mitigate such concerns, I leverage a third difference that exploits a within-treatment placebo group to construct a triple-difference (DDD) model. Specifically, I consider the supply of schools in a given commuting zone, as detailed in Section 3.1, and make use of those students whom I define as not having a real choice of schools. Those students are in principle treated, because the statutory right to school choice is given to all students in the county, but the minimal supply of feasible options makes them *de facto* non treated. However, they are exposed to the same confounders and investments as the other students within a specific treatment unit. A triple-difference model relaxes the parallel-trends assumption by adding a second control group that is on the same trend as the treatment group because they are both part of the same treatment units, thus taking out the variation in outcomes attributable to the trend rather than to the policy change. The triple-difference model therefore estimates the exam-performance gap between those with and without choice in the treated units, relative to the corresponding gap in the control units — and, moreover, it determines whether this gap changes in posttreatment periods. That is, we identify a treatment effect if the choice/no choice performance gap increases more posttreatment in the treatment units than in the control units. The identifying assumption in this case is therefore that the trend in the choice/no choice gap in exam performance is parallel between treatment and control groups in the pretreatment period. The triple-difference estimate thus accounts not only for changes that occur within the treatment group before and after treatment relative to the control group, but also for changes within the treatment group between students who should and should not be affected by the treatment.

I assess the validity of this assumption in Fig. 4, where I chart the raw difference in grades attained between students defined as having a choice of schools and those defined as having no such choice, separately by time relative to the implementation of school-choice reform and to treatment status.[18] Although there is a slight indication of anticipatory effects in the treatment group in the final pretreatment period (perhaps because students and parents in urban areas are more attuned to ongoing discussions about a possible school-choice reform), the trends in the treatment and control groups prior to the reforms are reasonably parallel — clearly more so than in the double-difference case. It is evident that the difference in performance between students living in commuting zones with a large versus small supply of schools is stable over the sample period in the nonreforming counties (the control group). By contrast, the corresponding gap increases sharply in posttreatment periods in the treatment group, which would suggest a treatment effect.

I estimate the treatment effect more formally by estimating the following model using ordinary least squares:

$$y_{izct} = \alpha_c + \lambda_t + \mu D^{Choice}_{c,t,z} + D_{c,t} + \theta_z \cdot \alpha_c + \theta_z \cdot \lambda_t + \theta_z + \phi_i + \upsilon_{izct} \tag{2}$$

The dependent variable is the (standardized) grade attained in the written exit exam in compul- sory school by student i in commuting zone z in county c, observed in year t, and α_c and λ_t are vectors of unit and time indicators. The binary indicator $D_{c,t}$ takes the value 1 for students graduating in a treated county after a school-choice reform took effect. The third difference is represented by the indicator variable θ_z, which takes the value 1 for students going to school in commuting zone z if and only if that zone has more than two high schools. The variable of interest is thus $D_{c,t,r,}$which is an interaction between $D_{c,t}$ and θ_z where the parameter $\hat{\mu}$ captures the DDD estimate of the effect of imposing high-stakes grades. The triple-difference estimator is essentially a three-way interaction between α_c, λ_t and θ_z. The interaction $\theta_z \cdot \alpha_c$ controls for county-specific differences in outcomes between students living in a commuting zone with real school choice and those not living in such an area, while $\theta_z \cdot \lambda_t$ controls for the possibility that students with real choice have a different linear time trend from those without choice. To control for other predictors of academic achievement, I also add a vector of student-level covariates, represented by ϕ_i, to most models. This includes gender, year of birth, immigrant status, parental education, parents' age when the student was born, and household income. In most specifications, I also control for being tested in mathematics as well as for subject-specific time trends.

3.5.2. Event study analysis

My primary mode of analysis will involve decomposing the aggregate results obtained with the framework outlined above using an event-study type design. There are two reasons for this approach. First, estimating treatment effects for individual periods leading up to or following the treatment point allows a more formal investigation of the validity of the parallel-trends assumption than merely inspecting descriptive trends in outcomes. The presence of statistically significant treatment effects in the prereform periods would suggest that other confounding variables could be correlated with either treatment or

[17] Based on results not included here, I find that neither treatment status nor covariates are predictive of being tested in mathematics rather than a language.

[18] In Figures A.2 and A.3 in the appendix I display more descriptive trends across choice status, treatment status and exam subjects. Overall, I find little evidence that would suggest that the parallel trend assumption is severely violated.

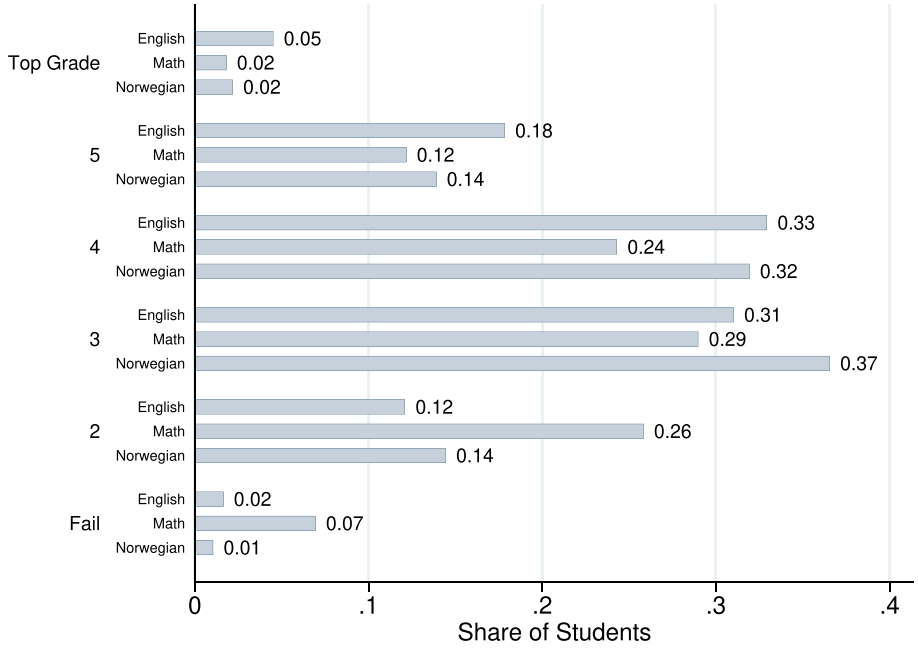


Fig. 3. Distribution of Exam Grades by Subject
Note: Grade distribution for each exam subject, measured as the fraction of students tested in that subject attaining a specific grade. The exams are graded on a six point integer scale, where 6 is the top grade, and 1 is a fail.

choice status and thus bias the results.

Second, recent studies have highlighted that DID designs where the timing and length of treatment exposure vary between units, estimates of aggregate treatment effects represent a weighted average of all the possible two-by-two DID estimators in the sample, which can yield biased results that are intuitively hard to interpret (Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021). For instance, the implicit weights assigned to each estimator are given by relative unit sizes and by the variance of the treatment indicator, that is, the timing of the treatment relative to the sample period. These weights can be unreasonable; for example, they might have negative values (de Chaisemartin & D'Haultfaeuille, 2020). In such cases, an event study or "stacked" DID design might be a more appropriate approach (Goodman-Bacon, 2021). The potential bias inherent in DID and DDD designs with variation in treatment timing can be particularly problematic if the treatment effect is not homogeneous across units and/or not static over the posttreatment period (Borusyak & Jaravel, 2018; Sun & Abraham, 2021). However, in such cases, even event-study designs can suffer from biased estimates as a result of an unreasonable implicit weighting of the estimators.

To overcome this issue, I follow the procedure introduced by Sun and Abraham (2021) to estimate an interaction-weighted (IW) triple difference model. A conventional event-study design decomposes a binary treatment indicator into a set of leads and lags, each of which is interacted with the treatment to achieve period-specific average treatment effects at various points in the window around the treatment occurrence, such as in the following equation.

$$\begin{aligned} y_{izct} \;=\; & \alpha_c + \lambda_t + \sum_{l=-4}^{-2} \mu_l D_{c,t,z}^{l,Choice} + \sum_{l=0}^{L} \mu_l D_{c,t,z}^{l,Choice} + \sum_{l=-4}^{-2} \mu_l D_{c,t}^{l} \\ & + \sum_{l=0}^{L} \mu_l D_{c,t}^{l} + \theta_z{\cdot}\alpha_c + \theta_z{\cdot}\lambda_t + \theta_z + \varphi_i + v_{izct} \end{aligned} \tag{3}$$

In Eq. (3), the four sets of variants of $\sum_l^L \mu_l D_{c,t,z}^l$ are the binary indicators taking the value 1 if the focal student in commuting zone z in county c in time t graduates l periods from the implementation point of the reform (with *Choice* denoting whether or not commuting zone z has more than two high schools). Such a specification relaxes the assumption that the treatment effect is static posttreatment, allowing estimates to take a nonparametric functional form across periods. However, note that when we estimate a model such as (3), we also assume that the treatment effect is homogeneous across treatment units for a given l, meaning that the period-specific estimates for all units follow the same dynamic path for $l \geq 0$. If the treatment units are in fact heterogeneous in terms of baseline characteristics, this assumption quickly becomes unreasonable. Sun and Abraham (2021) propose an alternative procedure that allows the treatment effect to vary both across time and across treatment units. Instead of a model specification like (3), they suggest estimating the cohort-specific average treatment effect on the treated, $CATT_{e,l}$. To do so I group the treated units $e = 1,\ldots, 6$ into cohorts according to their treatment point (in calendar time).[19] In my setting, four units are treated at a different time than the others (2003, 2005, 2009, and 2014 respectively) and therefore constitute their own cohort. Two counties were treated in 2012 and is then grouped together as one cohort. Essentially, the IW approach is then to estimate separate event-study models for each cohort, thereby allowing treatment effects to be heterogeneous across group-time combinations. I then take the weighted average of the group-time treatment estimates for a given time period, with the weights determined by the sample share of each unit. The resulting $CATT_{e,l}$ can be interpreted as the average difference in outcomes at time l relative to never being treated (Sun & Abraham, 2021).

More formally, rather than estimating the indicators $\sum_l^L \mu_l D_{c,t,z}^l$, the IW approach suggest to estimate the set of $CATT_{e,l}$ (that is the period-specific average treatment effect on the treated for a given cohort consisting of units e in period l) given by $\sum_e \sum_{l \neq -1} \delta_{e,l} (1\{E_C = e\}) D_{c,t,z}^l$ (and,

[19] In this study, the treated units e are the subsample of counties C that implemented school-choice reform.

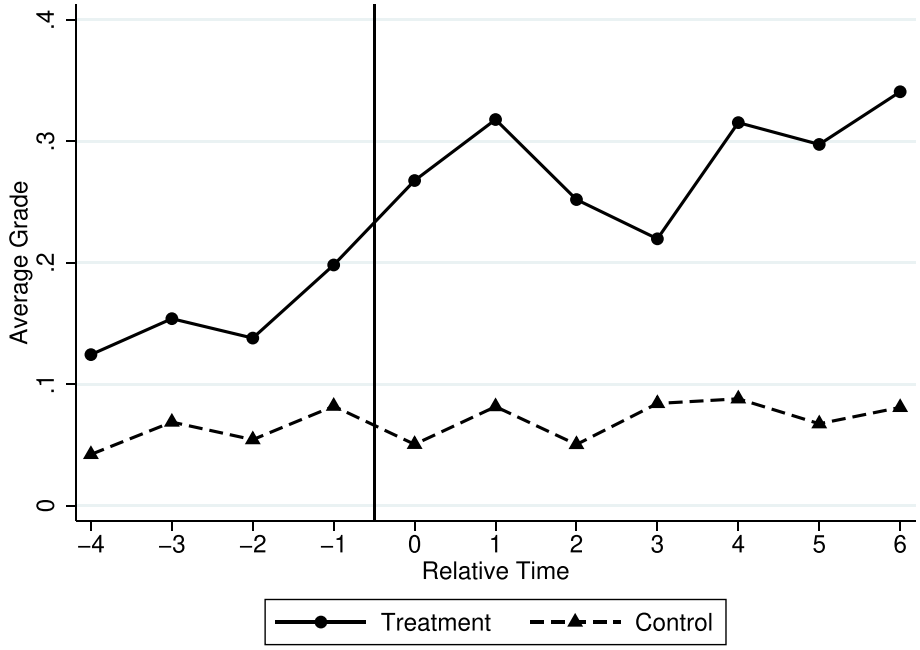


Fig. 4. Trend in Choice/No Choice Differential in Average Exam Grades
Note: The figure charts the difference across *choice* status in average grade attained on the written final exit exam, by cohort and treatment status. Circles (triangles) represent averages for students (not) exposed to a school-choice reform in that particular relative time point. Higher values on the *y*-axis indicate a larger gap in favor of students in *choice* commuting zones.

correspondingly, by $\sum_{e}\sum_{l\neq -1}\delta_{e,l}(1\{E_C = e\})D^{l,Choice}_{c,t,z}$) in (3), where the resulting coefficient $\delta_{e,l}$ is the estimated $CATT_{e,l}$ for unit e in period l. The IW-equivalent specification of Eq. (3) is formulated as

$$y_{izct} = \alpha_c + \lambda_t + \sum_{e}\sum_{l\neq -1}\delta_{e,l}\left(1\{E_C = e\}\cdot D^{l,Choice}_{c,t,z}\right) + \sum_{e}\sum_{l\neq -1}\delta_{e,l}\left(1\{E_C = e\}\cdot D^{l}_{c,t}\right) + \theta_z\cdot\alpha_c + \theta_z\cdot\lambda_t + \theta_z + \varphi_i + v_{izct} \quad (4)$$

For all l, I then take the sample-share-weighted average across the relevant e to get the IW

DDD estimate $\hat{v_l}$ for the lth period relative to the treatment timing. $\hat{v_l}$ thus corresponds to the lead and lag estimates you would produce in a conventional event-study specification (i.e., μ_l in Eq. (3)). To get an aggregate estimate of the average treatment effect for the post- reform period as a whole I follow the approach suggested in Callaway and Sant'Anna (2021) by averaging the group–time specific effects for each treated unit (i.e., averaging over the post- reform treatment effect estimates $\hat{\delta_{e,l}}$ for each e) before averaging across units using the sample share weights derived in the event-study analysis. The resulting parameter is the average effect of being exposed to the reforms experienced by all units that were ever exposed (Callaway & Sant'Anna, 2021, p. 12).

The control group in these specifications consists of all nonreforming counties. However, in this setting this includes both counties that already had school-choice systems in place at the start of the sample period and counties that applied a neighborhood-catchment regime for the duration of that period. The inclusion of counties that had already implemented similar reforms prior to the start of my sample period (the "always-treated") in the control group could potentially bias the results (Goodman-Bacon, 2021). For example, if similar reforms were implemented shortly before the start of my sample period and thus be on a similar dynamic trend. I assess to sensitivity of my results with regards to the composition of the control group in my presentation of the main results.

4. Results

4.1. Event study analysis

I begin my discussion of results by presenting the estimates from the event-study model de- scribed in the previous section. First, the results from the IW event-study model formulated in (4) are depicted in Fig. 5. I report the specific coefficients and standard errors from both this and the conventional event-study model in Table 2. .[20] As per convention, I set the period immediately prior to treatment, $l = -1$, as the reference category. Depicted in the figure is the period-specific estimates produced by (4), $\hat{v_l}$, that is the treatment effect of taking your exit exam in the lth period relative to the implementation of high-stakes grades while in a commuting zone with more than two schools. This corresponds to the three-way interaction between the treatment indicator, period indicators, and choice indicator. Two things are evident from this figure. First, I find little evidence of any anticipatory effects. In particular, the estimates for $l = -4$ and $l = -2$ are very close to zero. The point estimate for $l = -3$ is negative and slightly larger in magnitude, but nonetheless it is not statistically significant. In contrast, I find a moderately sized point estimate of 3.9 percent of a standard deviation (0.039σ), significant at the 10% level, for $l = -3$ when using the traditional event-study specification. This suggests that one of the treated units for which the parallel-trends assumption holds less well is overemphasized in the model. However, application of the sample-size re-weighting approach offered by IW DDD makes this anticipatory effect disappear in the aggregate. It is worth noting, how- ever, that the absence of statistical significance in such pre-trend testing does not in itself prove that the parallel trend assumption holds. Rambachan and Roth (2022) note that researchers might pass such tests due to low power in pre-periods, even when the true pre-trends differ be- tween the treatment in control. Caution is therefore warranted in the extent to which the point estimates for the periods $l \leq 0$ can be interpreted as evidence against potential violations of the identifying assumption in this setting, and should be

[20] Full results, including all δ, are available in Table D.1 in the appendix.

viewed in conjunction with the descriptive evidence provided.[21]

Second, there is a clear dynamic response to the implementation of high-stakes grades: first a sharp immediate response, which then fades, but is followed by continually increasing point estimates as we move further away from $l = 0$. The immediate effect is substantial, with a significant estimate of 0.07σ. However, the period-specific estimates peak for the cohorts graduating five years after the reforms, for which I estimate a treatment effect of 0.10σ. Such an increasing effect size suggests that younger cohorts of students adapt to the new incentive over time, perhaps as the culture and focus within schools change as well.[22] The sharp increase in point estimates is in fact apparent only once the fully treated cohorts — that is, those that through grades 8–10 under the new regime — enter the sample. On the other hand, the quickly dissipating immediate effect might suggest that the reforms and their potential effects were highly salient for the first affected cohort (owing to media attention, uncertainty about how it would affect school enrollment in the short term, etc.) but less so for the second and third cohorts.

Despite the concerns outlined in Section 3.5, the coefficients reported in Table 2 do not indicate that the difference between the IW and a conventional event-study approach is large. In the third column, I report *p*-values from tests of whether the estimates from these different approaches are significantly different. I find that this is the case only for $l = -3$. For all other l, I find broadly similar estimates, suggesting that the conventional event-study approach would be a reasonable approach for this context. Nevertheless, the IW DDD remains my preferred event-study approach throughout the paper, because of its more beneficial properties and assumptions.

4.2. Aggregate results

In this section, I present aggregate estimates of the average treatment effect of implementing high-stakes grades for the posttreatment period as a whole. Results from estimating the triple-difference model (2) using ordinary least squares are presented in Table 3. For ease of exposition, I report only the estimated coefficients for the three key parameters — the indicator for school-choice reform (in essence the *Treat* × *Post* interaction), the indicator for choice, and the triple interaction. First, in Columns 1 and 2 I present results from estimating (2) with and without additional control variables. In both specifications I find statistically significant effects on the triple-difference parameter. In column 2, my preferred specification where I control for student characteristics, parental background, and socioeconomic status (as described in Section 3.2), I estimate an average increase in the exam grade attained of 0.053σ (without additional controls the point estimate is 0.043σ). For an intuitive comparison of the effect size, 0.053σ is about half the estimated performance gap between native and immigrant students using this specification. In line with the identifying assumption of my triple-difference model, I can- not reject the null hypothesis of no effect for the school-choice-reform indicator alone. These results imply that imposing high-stakes grades through school-choice reforms is effective in improving student performance if combined with sufficient levels of choice so that the grades are actually perceived as consequential.[23]

In Column 3, I re-estimate the model, adding an indicator of whether a student was tested in mathematics as well as a subject-specific time trend. Using this specification, I estimate a treatment effect of 0.048σ — somewhat smaller, but substantively similar to the result in Column 2.

As a robustness check, in Column 4 I further examine if the treatment counties were on a differential trend before the reforms were implemented by controlling for a treatment-specific linear trend. In doing so, I relax the parallel-trends assumption to see if such differences are driving the results. As is evident from the estimate, controlling for such a trend increases the key point estimates by 0.014σ relative to the preferred specification, while the other parameters remain virtually unchanged. This substantiates the notion that the effect estimated in fact stems from the school-choice reforms, and not from some other underlying trend specific to the treatment counties. If anything, such (unidentified) underlying trends would appear to depress the initial estimate of the treatment effects.

To check if the results are sensitive to the control-group specification, Columns 5 and 6 exclude always-treated and never-treated counties, respectively. Hence in Column 5 the outcomes for students in the six treated counties are considered only in relation to students in those counties that never implemented high-stakes grades. Conversely, Column 6 estimates the same model using only those counties that were already "treated" prior to the start of my sample period. As evident from the results in Table 3, neither approach changes the substance of the results: while point estimates in both cases are smaller, they remain very close to, and are not significantly different from, those of the main model. When the never-treated counties are excluded, the *p*-value of the estimate does fall below the conventional 5 percent level, but only just. In Column 7, I further consider whether the result is robust to dropping all observations from 2008 from the sample (in that year, a large teachers' strike caused about one-third of exit exams to be canceled, primarily those in Norwegian; see Section 3 for details). It turns out that dropping the observations from that year does not impact the estimates in any meaningful way.[24]

Finally, in Column 8 I aggregate the event-study treatment effects derived from the IW DDD approach. As was the case in the event study, using the IW approach does not move the estimates in any way that would cause the conclusions to change.

Overall, the results presented in Table 3 are consistent with the hypothesis that students are incentivized by the prospect of being able to choose high schools given adequate academic performance. They are also consistent with the notion that, since a prerequisite for this

[21] Rambachan and Roth (2022) also note that the emphasis on such pre-period tests might exarcebate bias if researchers condition an analysis on passing such tests. That is, if researchers discard research projects that fail the pre-period test, it increases the likelihood that the studies that do, do so due to noisy estimates or low power in pre-treatment periods, when the true trend differs between treatment groups. They propose that researchers also report studies where the pre-tests fail, and propose procedures and assumptions that allow researchers to recover causal estimates in such settings.

[22] An alternative explanation for this pattern of effects could be that the composition of the treatment group changes toward the end of the sample window, as not all treated units are observed in all relative time periods. If the units with the strongest response are also those observed in later relative periods, this could potentially give a false impression of this increasing treatment effect. To assess the validity of this concern, I re-ran the analysis using different compositions of the treatment group; the results are reported in Table B.1 in the appendix. Specifically, I re-estimated the model separately using only the first three cases (the "early adopters") and the last three cases (the "late adopters") in the treatment group, respectively. I also ran a model where I used the middle four cases for which I could create a balanced sample window where all treated units are observed in all relative time periods. The results from these exercises indicate that, although it is apparent that the early and middle adopters are driving the observed effects, they themselves display this dynamic increase in effect sizes. Hence the shape of the event-study model does not seem to be an artifact of a changing composition of the treatment group, but rather a reflection of the dynamics within the units most strongly affected by the reforms.

[23] An alternative to this approach would be to estimate a more conventional double-difference model, and to subsample on the choice condition. Doing so yields broadly similar results, with a DID estimate of the effect of the reforms of 0.042σ, significant at the 5% level, for the *choice* subsample, and a nonsignificant estimate of -0.011 for the *no choice* subsample.

[24] I report results from additional robustness checks in Appendixes B and C. For example, I consider alternative approaches to computing the standard errors, such as clustering at the county level (and performing few-clusters corrections) and using randomization inference rather than conventional t-tests. The results are robust to these alternative approaches.

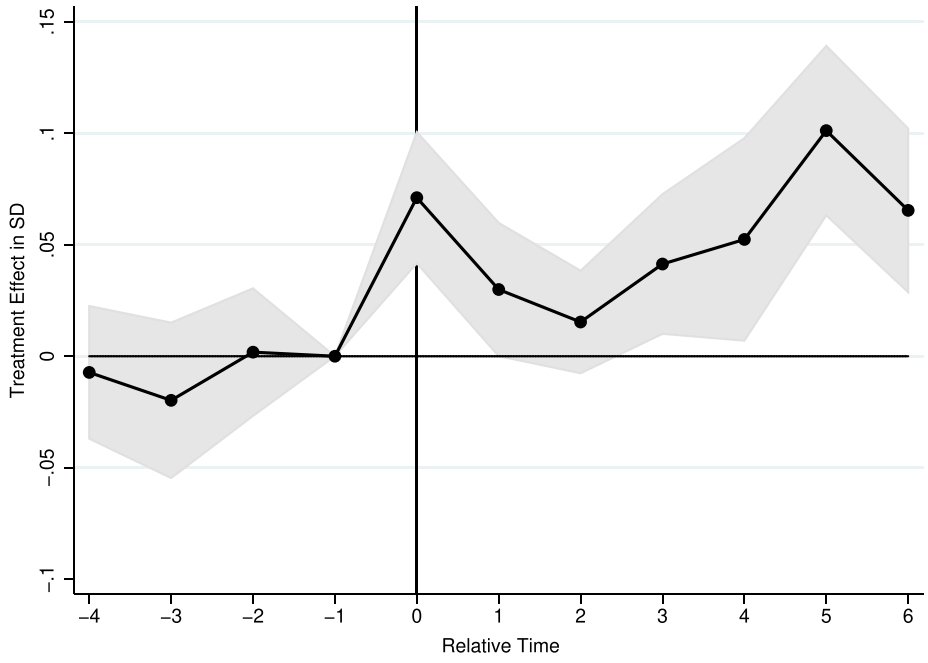


Fig. 5. Event Study Estimates of the Effects of School Choice Reforms on Exam Grades
Note: This figure presents the results from estimating the IW event-study type model formulated in Eq. (4), which decomposes the dynamics of the treatment effect over periods leading up to, and following implementation of the reforms. Reported are the coefficients estimated for the three-way interaction that takes the value 1 if the student lives in a commuting zone with at least two schools and takes her exam in a treated county after a school-choice reform l periods removed from the treatment, where $l \in \{-4, 6\}$. The model is saturated in period indicators so that the indicator for the first and last periods takes the value 1 for all preceding/subsequent periods, respectively. $l = -1$ is omitted as the reference category. The shaded area represents 95% confidence intervals . I report full results in Table D.1 in the appendix.

Table 2
Event Study Analysis.

Relative time	DDD estimates $\hat{\mu_l}$	IW DDD estimates $\hat{\nu_l}$	Difference p-value
−4	−0.022	−0.007	0.671
	(0.039)	(0.015)	
−3	0.039*	−0.020	0.001
	(0.020)	(0.018)	
−2	0.013	0.002	0.613
	(0.026)	(0.015)	
−1	Omitted	Omitted	
0	0.065**	0.071***	0.816
	(0.032)	(0.015)	
1	0.046	0.030*	0.580
	(0.036)	(0.015)	
2	0.056	0.015	0.178
	(0.035)	(0.012)	
3	0.034	0.041**	0.811
	(0.036)	(0.016)	
4	0.071*	0.052**	0.574
	(0.037)	(0.023)	
5	0.092	0.101***	0.873
	(0.057)	(0.019)	
6	0.070*	0.065***	0.903
	(0.037)	(0.019)	
N	790,905	790,905	
Adj. R^2	0.214	0.215	

Note: Estimation of the timing of treatment effects using a conventional event-study design and the Sun and Abraham (2021) IW event study approach. For this estimation, treatment status is replaced with an indicator equal to one in that particular year only, except $l = -4$ and $l = 6$, which are one for all preceding/subsequent years. The year prior to implementation is omitted for reference. In the *Difference* column I report p-values from tests of whether $\hat{\mu_l}$ and $\hat{\nu_l}$ are significantly different. Errors clustered at the commuting-zone level in parentheses.
* $p<0.01$,
** $p<0.05$,
*** $p<0.01$.

mechanism to be effective is having several options within a reasonable commuting distance, students in treated counties but in commuting zones with few choices are viable as a control group. The non-significance of the point estimates for the school-choice-reform indicator supports this conjecture. Similarly, having many schools within traveling distance does not in and of itself seem to have an effect on performance. It is only when a sufficiently large supply of schools is combined with school-choice reform that grades are actually perceived

Table 3
Aggregate Results.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
School choice reform ×	0.043**	0.053**	0.048**	0.067**	0.050**	0.043*	0.054**	0.051***
Choice	(0.021)	(0.024)	(0.025)	(0.030)	(0.023)	(0.025)	(0.026)	(0.13)
School choice reform	−0.021	−0.010	−0.010	0.003	−0.016	−0.008	−0.007	−0.029**
	(0.015)	(0.014)	(0.016)	(0.020)	(0.012)	(0.016)	(0.015)	(0.011)
Choice	−0.111***	0.013	0.014	0.005	0.158*	−0.011	0.019	0.071
	(0.021)	(0.059)	(0.068)	(0.064)	(0.090)	(0.039)	(0.061)	(0.056)
N	790,936	790,905	790,905	790,905	526,303	615,454	750,264	790,905
Adj.R^2	0.017	0.173	0.221	0.221	0.181	0.172	0.174	0.215
Covariates		✓	✓	✓	✓	✓	✓	✓
Subject FE + trend			✓	✓				✓
Linear trend				✓				
IW DDD								✓
Excluding:								
Always treated					✓			
Never treated						✓		
Year = 2008							✓	

Note: The table presents estimates of the average treatment effect on exam grade of imposing high-stakes grades through merit-based school- choice admission schemes. The outcome variable is standardized to have a mean of 0 and a standard deviation of 1. Panel A reports results from estimating the DDD model specified in (2). The coefficient of interest is the three-way interaction School choice reform × Choice in the top row, which gives the average treatment effect of being a student graduating from a treated county, in a labor market region with more than two high schools, after the treatment has been implemented. Conversely, the School choice reform variable controls for the conventional two-way fixed effects difference-in-differences estimator of graduating from a treated county in a posttreatment year. Choice is a dummy equal to one for students who have more than two high schools within traveling distance from their home. The triple difference model in practice interacts the DID estimator with this dummy. The models in Column 5 and 6 exclude all observations from always-treated and never-treated counties, respectively. In Column 7 I exclude all observations from the year 2008 from the regression. In Column 8 I aggregate the IW DDD event study results following the procedures suggested by Callaway and Sant'Anna (2021). Cluster-robust standard errors clustered at the commuting-zone level in parenthesis.
* $p<0.1$,.
** $p<0.05$,.
*** $p<0.01$.

as consequential, which boosts students' performance.[25]

I stress, however, that my analysis cannot conclusively rule out the possibility that changes in either the teachers' or the schools' behavior are contributing to, or driving the results entirely. However, I argue that there are several reasons why this is likely not the case. First, the incentives facing teachers and schools do not change in response to the reforms. For example, whether students get into their preferred high school does not come with any professional con- sequences for either group. Even if parents viewed the compulsory schools as accountable for their children's high school placement, they would not be able to punish low-performing schools as their enrollment is based on strict residential catchment areas. Moreover, the triple-difference approach implies that teachers in treated areas with a large number of schools would have to respond differently than other "treated" teachers for them to be driving the results. Lastly, if the true effect of the reforms were on the teachers it would likely be harder to detect any effect on objective student outcomes as it would require a longer causal chain. The reforms would have to make the teachers exert more effort (even if it would come with little to no benefit to themselves) and that effort had to succeed in increasing learning among their students. The students would then have to able to convert that learning into better performance on the exit exam.

[25] What is more, these results do not appear to be sensitive to the specific choice of school-supply threshold. Figure B.1 in the appendix indicates that the effects are similar — if anything larger — when the choice threshold is set higher. In sum, this exercise suggests that the result is not an artifact of my definition of what constitutes real choice. Rather, it reinforces the notion that the choice set of schools must be sufficiently large to create a competitive market that incentivizes students, suggesting that this effect may increase with the supply of schools. Setting the threshold at three thus represents a conservative constraint.

5. Mechanisms

5.1. Learning vs test effort

The results reported in Section 4 suggest that there is a mechanism by which test scores are influenced by the imposition of higher stakes. From a policy perspective, however, our main interest lies not in test scores *per se* but in students' accumulation of human capital. Indeed, one of the main purposes of testing is to measure the extent to which students have learned the skills they are supposed to learn. However, several papers have pointed out that scores on tests involving low stakes will reflect not only students' ability but also their motivation and effort (Gneezy et al., 2019; Heissel et al., 2021; Segal, 2012; Zamarro et al., 2019). One potential explanation for the difference observed in the present study between treated and nontreated students could therefore be that those students do not really differ in human capital but that what distinguishes them is that the treated ones have a stronger incentive than the nontreated ones to put effort into the exit exam and hence are likely to obtain better grades.

To explore whether the results reflect a sustained learning effort or mere test effort, I exploit the fact that, for the past decade, the Norwegian Ministry of Education has required all students to take a national standardized assessment test in grades 5, 8, and 9, the latter test being specifically implemented to measure students' improvement over the first year of the second stage of compulsory school. These tests are meant to provide a comprehensive assessment of a student's ability level at that point in time, providing school managers and policymakers with a tool enabling them to determine where resources and measures should be directed in order to improve student outcomes. For the students, however, there are no formal consequences associated with the tests. Their test scores do not factor into their grades, do not appear on any transcript, and are available only to their teacher and to their parents. Hence

these tests are low-stakes in nature for the student.[26] According to economic theory, the rational decision for a student, assuming that effort is costly, is therefore to devote less effort to such tests than to high-stakes test such as the exit exam. This, in turn, would imply that scores on these assessment tests may not adequately reflect students' true ability. Importantly, this does not change as a result of school-choice reforms. Consequently, if turning the final exit exam in grade 10 into a high-stakes test affects the effort students make to learn throughout the second stage (grades 8—10) and not just their effort ahead of and during that exam, this should be observable in the development of scores on the national assessment test. In other words, if students subjected to high-stakes grades put in more effort to learn, at least from the start of grade 8, they should have improved their ability level between grades 8 and 9 more than other students. If this is so, this would imply that the incentives provided in order to increase effort have actually worked by placing those students on a higher learning trajectory than they would otherwise be on. I test the hypothesis outlined above by estimating triple-difference models similar to those used in the main analysis as described in Section 4, with scores on the national assessment test in grade 9, that is, in the year prior to the year of graduation, as the outcome of interest. As these tests were introduced for ninth-graders in 2010, the analysis is restricted to the 2010–2015 cohorts. I match grade 9 observations to the same students' scores in grade 8, so that I can control for previous performance. I include students missing tests from eighth grade by constructing an indicator equal to one if the subject score is missing and setting the score to zero. Within the sample period, three counties implemented school-choice reforms (in 2012 and 2014, respectively). This provides a staggered DDD framework similar to that previously used. All students are tested in both mathematics and Norwegian language/reading in both grade 8 and grade 9.[27] To construct my outcome measure, I standardize the scores on each test, average them across the tests, and standardize the resulting average score once more. This composite score is thus a measure of a student's general skill level in the subjects covered by the final exit exam.

I present the results from this analysis in Table 4. That table includes estimates from event studies similar to those described in Section 4.1, decomposing the triple-difference results into leads and lags using both the Sun and Abraham (2021) interaction-weighted design and the conventional event-study specification. As previously, I set $l = -1$ as the reference category. For these grade 9 assessment tests, the period-specific estimates display a similar dynamic evolution in terms of effect size as was observed for the exit-exam grades (Fig. 5). This is inconsistent with the idea that the improvements in test scores result only from changes in the amount of effort spent on the assessment tests themselves, as such an effect should be observable immediately upon implementation and then remain stable. In fact, I find no effect on the scores of those students who took the assessment tests immediately after the implementation of high-stakes grades. On the other hand, for the cohort of students who where in grade 8 when the reforms were implemented, meaning that they had ample time to adjust their effort levels to the new regime, I find a substantial increase in the composite-score measure. Strong effects are also evident for subsequent cohorts, amounting to approximately 0.070σ (unfortunately, the sample period does not allow me to extend the analysis further into the posttreatment period). The fact that these effect sizes appear with a similar dynamic rhythm as the increases in effect sizes in the main analysis lends support to the claim that the main treatment effect observed in scores on the final exit exam is not solely attributable to test effort, but is also explained by an increase in what students have actually learned — that is, in their ability level.[28]

Fig. 6

Table 4
National Assessment Test Event Study Results.

Relative time	DDD estimates $\hat{\mu_l}$	IW DDD estimates $\hat{\nu_l}$	Difference p-value
−2	0.023	0.039	0.571
	(0.033)	(0.033)	
−1	Omitted	Omitted	
0	−0.011	0.010	0.376
	(0.027)	(0.023)	
1	0.030	0.053***	0.290
	(0.024)	(0.019)	
2	0.053	0.071*	0.362
	(0.035)	(0.042)	
3	0.055	0.070**	0.685
	(0.038)	(0.027)	
N	249,602	249,602	
Adj. $R2$	0.767	0.753	

Note: The table presents results from a triple difference event-study analysis using performance on the standardized national assessment tests in mathematics and reading in 9th-grade as the outcome. The event study decomposes the results over the years leading up to, and following, the implementation of the reforms using both the conventional, and the Sun and Abraham (2021) IW event-study approach. I standardize the score of each test, take the mean, and standardize the resulting composite score. The outcome is thus a representation of the general skill level of the student in subjects applicable for the final exam. For these estimations, treatment status is replaced with an indicator equal to one in that particular year only. The year prior to implementation is omitted for reference. In the *Difference* column I report p-values from tests of whether $\hat{\mu_l}$ and $\hat{\nu_l}$ are significantly different. Cluster-robust standard errors clustered at the commuting-zone level in parentheses.

* $p<0.1$,.
** $p<0.05$,.
*** $p<0.01$.

5.2. Interactions analysis

The channels through which the effect of this incentive might work could also be illuminated by its differential effects across subsamples. For example, a widely accepted notion is that a more competitive environment in schools will benefit boys, who tend to thrive more than girls under such conditions (Almås et al., 2016; Azmat et al., 2016; Hopland & Nyhus, 2016). Certain other subsamples are also of particular policy interest, including students from a low socioeconomic background. Socioeconomic status (SES) is a major predictor of educational achievement, and there is a large body of research into interventions at the compulsory-school level aimed at improving the performance of students from low-SES households (Dietrichson et al., 2017). Evidence that such typically at-risk students respond positively to high-stakes grades — learning more in the process — would therefore have obvious policy implications.

Moreover, Almås et al. (2016) demonstrate that there is a strong socioeconomic gradient in terms of competition preferences. In

[26] One could plausibly argue that these are not low-stakes tests for the teachers if they believe that their class' or school's performance reflect on them professionally. However, there is little reason to expect that this perception, and therefore the teachers' behavior, should change in response to the reforms. Nevertheless, I cannot rule out that any effect on these ability assessments are due to changes in teacher practices overlapping with the school choice reforms rather than student effects, and some caution is therefore warranted in interpreting the results.

[27] Students are also tested in English in grade 8, and I include those scores as well in the controls.

[28] In the appendix I also report results from a similar analysis using the test scores in grade 8 as the outcome. In this case, there is no clear pattern to the results — if anything students appear to do somewhat worse after reform implementation, suggesting that the change in behavior starts upon entry to lower-secondary school, not in earlier grades. This is consistent with the notion that lower-secondary school marks a new stage in the students' trajectory, where grades and future academic paths are more strongly emphasized.

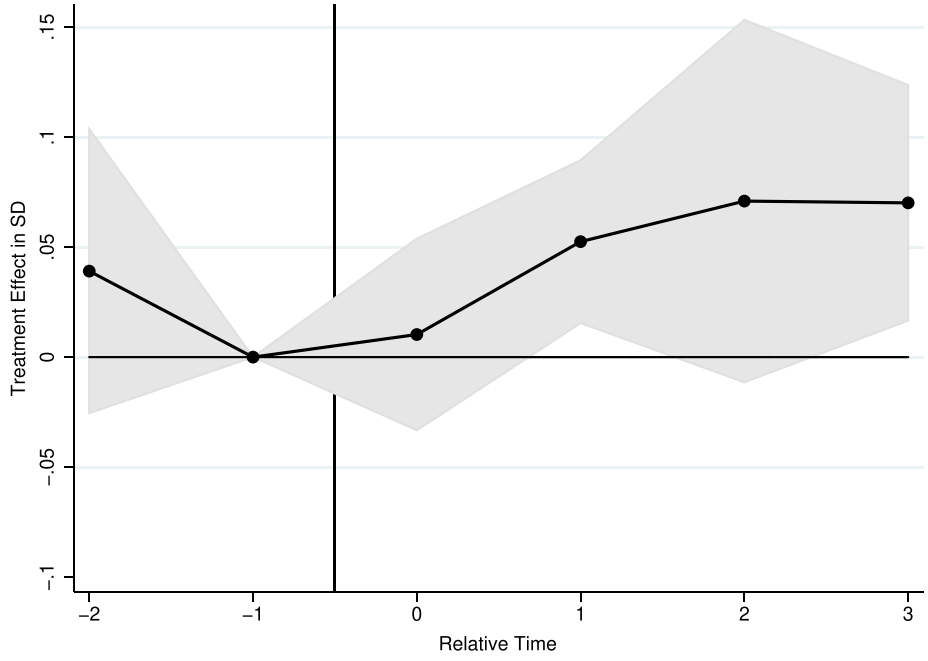


Fig. 6. Event Study Results for Assessment Test Scores in 9th-grade
Note: This figure presents the results from estimating an event-study type model decomposing the dynamics of the treatment effect of introducing high-stakes grades in 10th grade on low-stakes assessment tests conducted in 9th grade. Reported are the coefficients estimated for indicators for being l periods removed from implementation, where $l \in \{-2, 3\}$. The model is saturated in period indicators as the sample period is constrained to the 6-year window in question. $l = -1$ is omitted as the reference category. The shaded area represents 95% confidence intervals. Full results are available in Table D.2 in the Appendix.

particular, boys from lower-SES households are less willing to compete than boys from higher-SES backgrounds. If we believe that the competitive pressure created by high-stakes grades is the driving mechanism behind the observed increase in performance, that increase could therefore also reflect an adverse segregational effect across parental background in that boys from richer homes may benefit to a particularly large extent.

In the following analyses I also consider whether students who were tested in mathemat- ics at the exit exam are more impacted by the treatment than others. As students take only one exam, the subject they are allocated can greatly influence their performance, all else being equal. Generally, students tested in mathematics perform far worse than those tested in a language subject, as illustrated in Fig. 3. In this particular case, it is plausible that mathematical skills can be improved more by high-effort behavior such as cramming and repetition and may thus be more responsive to high-stakes grades. Conversely, language skills may be harder to improve through effort alone, in that they require a longer-term maturation process. This hypothesis takes into account evidence suggesting that students' vocabulary and language skills are strongly tied to their parental background (Buckingham et al., 2013; Dustmann, 1997), and that scores on language tests often appear to be less receptive to interventions than scores on mathematics tests (Bettinger, 2012).

For this purpose, I extend (2) to incorporate either gender or SES as a fourth dimension, to estimate a quadruple-type model of the form

$$\begin{aligned} y_{igzct} = {} & \mu_1 D^{Choice}_{c,t,z,g} + D_{c,t,z} + D_{c,t\cdot\gamma g} + \alpha_c\cdot\gamma_g\cdot\lambda_t \\ & + \theta_z\cdot\gamma_g\cdot\lambda_t + D_{c,t} + \alpha_c\cdot\gamma_g + \alpha_c\cdot\theta_z + \lambda_t\cdot\gamma_g \\ & + \lambda_t\cdot\theta_z + \gamma_g\cdot\theta_z + \alpha_c + \lambda_t + \gamma_g + \theta_z + \phi_i + \upsilon_{izct} \end{aligned} \tag{5}$$

$D_{c,t,r,g}$ takes the value 1 if student i of gender (SES) g in commuting zone z in county c in cohort t takes her exam in a treated county after a school-choice reform has been implemented there, and her commuting zone has more than two high schools. mu_1 captures the DDDD treatment effect estimate. In the model I control for all possible interactions among the four main variables (α_c, λ_t, θ_z, γ_g), and for student-level characteristics ϕ_i. The interaction between α_c and λ_t is, as before, represented by the binary indicator $D_{c,t}$, which takes the value one if the student is in a reforming county and takes her exam after the reform. Similarly, $D_{c,t,z}$ is the interaction terms between $D_{c,t}$ and θ_z, taking the value 1 if the student in addition lives in a commuting zone with more than two schools. I estimate the model separately for the full sample and for the subsamples tested in mathematics and language, respectively.

Table 5 presents the results from estimations of the quadruple-difference models. Panel A reports results for the gender specifications. Evidently, the estimates do not indicate any gender-specific differential effects of the admission reforms. While I find large and significant point estimates for the effect of the reform in general, the estimates for the differential effect on girls are small and statistically insignificant. This is the case both for the overall sample and across exam subjects. As the top row reports the marginal effect of being a treated girl, the coefficients for *School choice reform* × *Choice* give the average treatment effect for treated boys. Columns 1 and 2 indicate that boys randomly drawn to be tested in mathematics respond more strongly to the reforms than those tested in language, but these estimates are imprecise and not significantly different from each other.

Panel B considers low-SES students, defined as having a mother whose highest completed level of education is at most compulsory school (which is true for 22.7% of the sample). Following Almås et al. (2016), we would expect these students to respond less strongly to a com- petitive incentive and hence to manifest smaller treatment effects. However, as with gender, I find limited evidence of such a differential effect using the quadruple-difference model. As reported in Table 5, I find small positive coefficients for both the total sample and the language subsample. Although neither is close to being statistically significant, in both cases the direc- tion of the estimate is the opposite of what the literature would have us expect. This is also the case for the mathematics subsample, for which I find a moderately sized point estimate of 0.054σ. At face value, such an estimate suggests that treated low-SES students who were tested in mathematics increased their performance more than treated students with other socioeconomic backgrounds who were also tested in mathematics. While this estimate is also imprecisely estimated, it provides a suggestive piece of evidence that, if anything, the reforms served to *reduce* the SES gap in performance on the mathematics exam.

Overall, however, the conclusion to be drawn from the analysis presented in this section is that I find limited evidence of differential

Table 5
Interactions Analysis.

	All (1)	Math (2)	Language (3)
Panel A: Gender			
School choice reform × Choice × Female	−0.028	−0.015	−0.017
	(0.022)	(0.040)	(0.022)
School choice reform × Choice	0.080***	0.099*	0.064**
	(0.027)	(0.056)	(0.028)
School choice reform	−0.037**	−0.024	−0.048**
	(0.016)	(0.037)	(0.024)
School choice reform × Female	0.005	−0.021	0.006
	(0.019)	(0.038)	(0.018)
Choice × Female	0.042	0.045	−0.021
	(0.030)	(0.039)	(0.028)
Female	0.396***	0.180***	0.494***
	(0.022)	(0.036)	(0.024)
N	790,905	297,414	493,491
Adj.R^2	0.174	0.214	0.181
Panel B: Socioeconomic Status			
School choice reform × Choice × Low SES	0.013	0.054	0.014
	(0.028)	(0.033)	(0.036)
School choice reform × Choice	0.056**	0.075	0.043
	(0.028)	(0.056)	(0.029)
School choice reform	−0.015	0.004	−0.034
	(0.014)	(0.033)	(0.022)
School choice reform × Low SES	−0.020	−0.087***	0.010
	(0.021)	(0.031)	(0.024)
Choice × Low SES	0.001	0.029	−0.025
	(0.023)	(0.037)	(0.031)
Low SES	−0.802***	−1.012***	−0.735***
	(0.035)	(0.045)	(0.040)
N	771,445	289,554	481,891
Adj.R^2	0.163	0.208	0.168

Note: This table reports results from subsample analyses of differential treatment effects across gender and socioeconomic status. Column 1 estimates effects for the full sample, while columns 2 and 3 estimate identical models for those tested in mathematics or a language separately, using the preferred specification from Table 3. In panel A I consider differential effects between boys and girls. In Panel B I consider whether the effects interact with socioeconomic background. Here I use the mother's education to determine socioeconomic status, where low SES indicates that her highest level of completed education is at most compulsory school (10 years). Errors clustered at the commuting-zone level in parentheses.
* $p<0.1$,.
** $p<0.05$,.
*** $p<0.01$.

treatment effects across important subsamples. Instead, the positive effect of the admission reforms on student performance seems to be rather uniform across the subsamples considered here, with some evidence that the effect is stronger for students tested in mathematics, in particular for those with a low-SES background. It would appear that these results should at least mitigate some of our concern regarding the possibility of strong segregational effects of school-choice policies such as those studied in this paper.

6. Concluding remarks

In this paper, I investigate the incentivizing effect of high-stakes grades on student learning. I exploit a natural experiment created by regional differences in Norwegian high-school admission regimes to compare scores on the final exit exam of compulsory school, which is a high-stakes exam for some students but not for others. I use the supply of schools within stu- dents' traveling distance as a third source of variation, to distinguish students who have a real choice of schools from those who have such a choice only in theory. In line with theory-based predictions, my triple-difference model reveals that tying the final exit exam of compulsory school to salient outcomes improves the grades attained, with an effect size of 4–6 percent of a standard deviation. The effect size is moderate, but it is still economically meaningful, especially considering the fact that students are already heavily incentivized to invest in schooling by the large, inherent returns to education (Burgess et al., 2022). For example, the magnitude is equal to about 20% of the unconditional gender gap in exam performance, and to 10% of the SES gap. While several papers have demonstrated a causal link between test stakes and performance, either through smaller field experiments or by using financial incentives, this paper provides evidence for the viability of exploiting such a mechanism to stimulate students' investment of effort in school at the policy level. Indeed, the results indicate that the change to a merit-based enrollment regime in high school in and of itself improves performance in younger students. That is, performance improves at a stage where no tracking or sorting of any kind is conducted. However, a crucial prerequisite is that the supply of schools must be sufficient to create a sense of real choice. Introducing school choice has little impact if students have only one or two schools within a reasonable traveling distance. Further, my analysis does not find any significant heterogeneity in treatment effect across exam subject, socioeconomic status or gen- der —– a result that contrasts with the results of earlier studies suggesting that school-choice enrollment regimes might have adverse segregational effects (Altonji et al., 2015; Hsieh & Urqiuola, 2006; Lindbom, 2010)

Building on a growing body of work exploring the relationship between effort and performance in low-stakes assessments (Gneezy et al., 2019; Segal, 2012; Zamarro et al., 2019), I assess the extent to which my results can be explained by a sustained learning effort, as op- posed to a more punctual test-taking effort, on the part of students. By contrasting performance on the final exit exam with scores on comprehensive ability assessments conducted in earlier grades, I demonstrate that students exposed to a school-choice enrollment regime appear to be on a higher learning trajectory than students in the control group. These results imply that the main treatment effect is not only a result of increased test effort but is also indicative of a higher, sustained learning effort throughout the final years of compulsory school. Evidence of students making a long-term investment in their schooling should increase the relevance of this study for policymakers. Many countries employ similar systems, where a combination of teacher grades and national exams are used to determine placement to schools, either at the high school or post-secondary level. My results suggest that such systems can incentivize students to exert more effort at school. The effect sizes are nontrivial, but nevertheless moderate, which suggests that some students respond more to these incentives than others. While identifying those students lies beyond the scope of the present study, policymakers can be expected to be interested in finding out who they are, in order to thoroughly assess the distributional effects of implementing high-stakes grades.

Lastly, I also stress that the results presented in this paper does not necessarily imply that merit-based enrollment only produce beneficial effects, nor that any beneficial effects outweigh the costs. Increasing competitive pressure could also have negative effects on students that are not captured in the current analysis, e.g., worsened health outcomes due to stress. Optimal incentive structures must also balance increasing the students' efforts towards the focal task with inducing a desirable allocation of effort across *all* tasks which the students have to accomplish (Holmstrom & Milgrom, 1991). The fact that the students' GPA comprise of both the exam grade as well as teacher grades in all subjects should at least mitigate concerns that students would solely focus on the exam subjects. However, increased emphasis on test scores and grades in enrollment regimes could divert students' attention away from activities that are productive and valuable in their own right, but does not directly translate to an improved GPA. In doing so, merit-based systems might make student effort and motivation more instrumental, perhaps to the detriment of their long-term intrinsic motivation. Assessing how the

incentives associated with merit-based enrollment affect student behavior in other domains as well will be important to determine the desirability of such systems.

CRediT authorship contribution statement

Andreas Fidjeland: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization.

Data availability

The authors do not have permission to share data.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.econedurev.2023.102377.

References

Almås, I., Cappelen, A. W., Salvanes, K. G., Sørensen, E.Ø., & Tungodden, B. (2016). Willingness to Compete: Family Matters. *Management Science, 62*(8), 2149–2162.

Altonji, J. G., Huang, C.-. I., & Taber, C. R. (2015). Estimating the Cream Skimming Effect of School Choice. *Journal of Political Economy, 123*(2), 266–324.

Angrist, J. D., & Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review, 99*(4), 1384–1414.

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural Experiment. *American Economic Review, 92*(5), 1535–1558.

Arbeidslaget Analyse, Utgreiing og Dokumentasjon. (2005). Evaluering av fritt skolevalg. *Hordaland Fylkeskommune*.

Azmat, G., Calsamiglia, C., & Iriberri, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association, 14*(6), 1372–1400.

Bach, M., & Fischer, M. (2020). Understanding the response to high stakes incentives in primary education. *IZA Discussion Paper Series, 13845*.

Bakken, A., Sletten, M.A., & Eriksen, I.M. (2018). Generasjon prestasjon? Ungdoms opplevelse av press og stress. Tidsskrift for ungdomsforskning, (2).

Becker, W. E., & Rosen, S. (1992). The learning effect of assesment and evaluation in high school. *Economics of Education Review, 11*(2), 107–118.

Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: Results From a social experiment in mexican high schools. *Journal of Political Economy, 123*(2), 325–364.

Bettinger, E. P. (2012). Paying to learn: The effect of financial incentives on elementary school test scores. *The Review of Economics and Statistics, 94*(3), 686–698.

Borusyak, K., & Jaravel, X. (2018). Revisiting event study designs, with an application to the estimation of the marginal propensity to consume. *Working Paper*.

Brophy, J. (1987). Synthesis of research on strategies for motivating students to learn. *Edu- cational Leadership, 45*(2), 40–48.

Buckingham, J., Wheldall, K., & Beaman-Wheldall, R. (2013). Why poor children are more likely to become poor readers: The school adulthood. *Australian Journal of Educa- tion, 57*(3), 190–213.

Burgess, S., Greaves, E., & Murphy, R. (2022). Deregulating teacher labor markets. *Economics of Education Review, 88*, Article 102253.

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics, 225*(2), 200–230.

Costrell, R. M. (1994). A simple model of educational standards. *American Economic Review, 84*(4), 956–971.

Cullen, J. B., Jacob, B. A., & Levitt, S. (2006). The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica : journal of the Econometric Society, 74*(5), 1191–1230.

de Chaisemartin, C., & D'Haultfaeuille, X (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review, 110*(9), 2964–2996.

Deming, D. J., & Figlio, D. (2016). Accountability in US education: Applying lessons from K-12 experience to higher education. *Journal of Economic Perspectives, 30*(3), 33–56.

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-. M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research, 87*(2), 243–282.

Dokument 8:41. (2006). *Forslag fra stortingsrepresentantene anders anundsen, jon jæger gåsvatn og Åse M. schmidt om å innføre fritt skolevalg for alle elever i videregående skole.* Oslo, Norway: Stortingets Utredningsseksjon.

Dokument 8:8. (2003). *Forslag fra stortingsrepresentantene ulf erik knudsen, arne sortevik og karin S. woldseth om innføring av elevers rett til fritt skolevalg på videregående skole.* Oslo, Norway: Stortingets Utredningsseksjon.

Dustmann, C. (1997). The effects of education, parental background and ethnic concentration on language. *The Quarterly Review of Economics and Finance, 37*, 245–262.

Eccles, J. S., & Midgley, C. (1989). Stage-environment fit: Developmentally appropriate classrooms for young adolescents. In *Research on motivation in education, 3* pp. 139–186). New York, NY: Academic Press.

Eccles, J. S., Wigfield, A., Midgley, C., Reuman, D., Iver, D. M., & Feldlaufer, H. (1993). Negative effects of traditional middle schools on students' motivation. *The Elementary School Journal, 93*(5), 553–574.

Epple, D., & Romano, R.E. (2003). Neighborhood schools, choice, and the distribution of educational benefits. In C. M. Hoxby (Ed.), The economics of school choice (Chap. 7, pp. 227–286).

Figlio, D., & Hart, C. M. D. (2014). Competitive effects of means-tested school vouchers. *American Economic Journal: Applied Economics, 6*(1), 133–156.

Figlio, D., & Loeb, S. (2011). School accountability. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Chap. 8, Vol. 3, pp. 383–421).

Friedman, M. (1962). *Capitalism and freedom*. Chicago, IL: University of Chicago Press.

Fryer, R. (2011). Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics, 126*(4), 1755–1798.

Gibbons, S., Machin, S., & Silva, O. (2008). Choice, competition and pupil achievment. *Journal of the European Economic Association, 6*(4), 912–947.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights, 1*(3), 291–308.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics, 225*(2), 254–277.

Grant, D., & Green, W. B. (2013). Grades as incentives. *Empirical Economics, 44*, 1563–1592.

Grove, W. A., & Wasserman, T. (2006). Incentives and student learning: A natural experiment with economics problem sets. *American Economic Review, 96*(2), 447–452.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24*(3), 1141–1177.

Haraldsvik, M. (2014). Does performance-based admission incentivize students? *Unpublished manuscript*.

Harter, S. (1981). A new self-report scale of intrinsic versus extrinsic orientation in the classroom: Motivational and informational components. *Developmental Psychology, 17*(3), 300–312.

Heissel, J. A., Adam, E. K., Doleac, J. L., Figlio, D. N., & Meer, J. (2021). Testing, stress and performance: How students respond physiologically to high-stakes testing. *Education Finance and Policy, 16*(2), 183–208.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, & Organization, 7*, 24.

Hopland, A. O., & Nyhus, O. H. (2016). Gender differences in competitiveness: evidence from educational admission reforms. *The B.E. Journal of Economic Analysis & Policy, 16*(1).

Hoxby C.M. (2003). School choice and school productivity: could school choice be a tide that lifts all boats? In C. M. Hoxby (Ed.), The economics of school choice (Chap. 8, pp. 287–342).

Hoxby, C. (2000). Does competition among public schools benefit students and taxpayers? *American Economic Review, 90*(5), 1209–1238.

Hsieh, C.-. T., & Urqiuola, M. (2006). The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program. *Journal of Public Eco- nomics, 90*, 1477–1503.

Hvidman, U., & Sievertsen, H. H. (2021). High-stakes grades and student behavior. *Journal of Human Resources, 56*(3), 821–849.

Inchley, J., Currie, D., Young, T., Samdal, O., Torsheim, T., Augustson, L., . . . Barnekow, V. (Eds.). (2013). *Growing up unequal: Gender and socioeconomic differences in young people's health and well-being*. WHO regional office for Europe: World health organization.

Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics, 91*(3), 437–456.

Lavy, V. (2010). Effects of free choice among public schools. *The Review of Economic Studies, 77*(3), 1164–1191.

Leuven, E., Oosterbeek, H., & van der Klaauw, B. (2010). The effect of financial rewards on students' achievement: Evidence from a randomized experiment. *Journal of the European Economic Association, 8*(6), 1243–1265.

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy, 8*(4), 183–219.

Lindbom, A. (2010). School choice in Sweden; effects on student performance, school costs, and segregation. *Scandinavian Journal of Educational Research, 54*(6), 615–630.

Midgley, C., Anderman, E., & Hicks, L. (1995). Differences between elementary and middle school teachers and students: A goal theory approach. *The Journal of Early Adoles- cence, 15*(1), 90–113.

Napoli, A. R., & Raymond, L. A. (2004). How reliable are our assessment data?: A comparison of the reliability of data produced in graded and un-graded conditions. *Research in Higher Education, 45*(8), 921–929.

Norwegian Directorate for Education and Training. (2018). Trekkordning ved eksamen for grunnskole og videregående opplæring. [Online]. Accessed: 2018-09-07 Available from: Https://www.udir.no/regelverk-og-tilsyn/finn-regelverk/etter-tema/eksamen/trekkordning-ved-eksamen-for-grunnskole-og-videregaende-opplaring-udir-2-2018/.

Norwegian Directorate of Education and Training. (2017). *The education mirror 2016*. Oslo, Norway: Utdanningsdirektoratet.

Oettinger, G. S. (2002). The effect of nonlinear incentives on performance: Evidence from "ECON 101". *The Review of Economics and Statistics, 84*(3), 509–517.

Rambachan, A., & Roth, J. (2022). A more credible approach to parallel trends. Working Paper.
Ruud, M. (2018). Skolepress og stress øker, særlig blant jenter. Retrieved from https://www.utdanningsnytt.no/helse-psykisk-helse-skolemiljo/skolepress-og-stress-oker-saerlig-blant-jenter/152221.
Segal, C. (2012). Working when no one is watching: motivation, test scores, and economic success. *Management Science, 58*(8), 1438–1457.
Statistics Norway. (2001). Norwegian Standard classification of education. *Statistics Norway*.
Stinebrickner, R., & Stinebrickner, T. (2008). The causal effect of studying on academic performance. *The B.E. Journal of Economic Analysis & Policy, 8*(1).
Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics, 225*(2), 175–199.
The Education Act (Opplæringslova). (1998). *Lov om grunnskolen og den videregående opplæringa (LOV-1998-07-17-61)*.
Vroom, V. H. (1964). *Work and motivation*. New York, NY: Wiley.
Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievment motivation. *Contemporary Educational Psychology, 25*, 68–81.
Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17.
Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*(3), 227–242.
Zamarro, G., Hitt, C., & Mendez, I. (2019). When students don't care: Reexamining international differences in achievement and student effort. *Journal of Human Capital, 13*(4), 519–552.