Scheduled to gain:

Short- and longer-run educational effects of examination scheduling

Simon Søbstad Bensnes*

November 13, 2019
Norwegian University of Science and Technology, and Statistics Norway,
NO-0131 Oslo, Norway
simon.bensnes@ssb.no.

Abstract This article presents findings concerning the effect of examination timing on high-stakes exam scores and longer-run outcomes. It shows that random variations in exam schedules that increase the time students have to prepare have positive effects on test scores. The effect is highly concave, and stronger for females and in quantitative subjects. I trace the effects of preparation time into tertiary education outcomes, finding significant effects for female students on the extensive and intensive margin. The paper shows how easily exam scores and, consequently, longer-run outcomes are affected by a random institutional factor unrelated to student ability.
JEL: I20 I21
Keywords: Upper secondary, exam scores, preparation time, exam schedules.

*I am grateful to three anonymous referees for thorough comments and input from Chris Van Klaveren, Bjarne Strm, Torberg Falch, Magne Mogstad, Sarah Bana, Chang Lee, and participants at various seminars and workshops. I am also grateful for institutional insights from Øyvind Kvanmo. Any remaining errors are my own.

I Introduction

Each spring, millions of students across the world take high-stake examinations at the end of upper secondary schooling. Examples include the General Certificate of Secondary Education in the UK, the Baigrut in Israel, and the SAT and Advanced Placement exam in the US. The scores students receive on these tests are then used to sort students into jobs and tertiary education. When exam scores are utilized in this manner, the underlying assumption is that exam scores are reliable proxies for student ability. However, recent evidence indicates that random disturbances during exams such as ambient air pollution and pollen proliferation on the examination day, have relatively large effects on both exam scores and longer-run outcomes (Bensnes, 2016; Lavy, Ebenstein, & Roth, 2015). Furthermore, institutional factors may also affect exam conditions (e.g. Pope & Fillmore, 2015). These findings show that (i) the accuracy with which exams reflect students' abilities depends in part on how sensitive exam grades are to random variations in exam conditions; and (ii) the more sensitive this relationship is, the less useful exams are as a placement tool. This paper focuses on an institutional factor that is unrelated to cognitive ability but affects exam scores: exam scheduling. The aim of this paper is to first estimate the effect of random variations in exam scheduling on exam scores, and then trace this effect to longer-run outcomes.

Identification in this paper is achieved by using a specific feature of the national exam system for Norwegian upper secondary schools, combined with very detailed administrative data. Each year, on a specific date shortly before their exams, students in the Norwegian upper secondary school system are informed when their written end-of-year exams will be held and which subjects they will be examined in. The time between the date of the announcement and the date of each exam serves as the measure of preparation time. The identification used stems from subject-specific random variation in the exam schedule across years. Results suggest that increasing the preparation time from 5-8 days to 9-12 days increases test scores by 5.3% percent of a standard deviation. Interestingly, the marginal return to preparation time approaches zero when preparation

time is increased beyond this point. Overall, the results indicate that students only study for a limited number of days, regardless of the amount of preparation time they are granted. Furthermore, the effect is stronger in quantitative subjects and for females, who are also more strongly affected in the longer-run: Increasing the share of exams with relatively long preparation periods by one standard deviation increases the probability of a female upper secondary school student enrolling in a tertiary education program by half a percentage point.

While this paper use the direct variation in the conditions associated with exams, the underlying cause of this variation stems from the institutional framework rather than from environmental factors as is the case in Lavy et al. (2015), for example. The effects of institutional factors on students' outcomes per se is not an unstudied topic. Areas which have received some attention include school start times (e.g. Carrell, Maghakian, & West, 2011; Edwards, 2012) and course scheduling (e.g. Dills & Hernandez-Julian, 2008). However, most studies are confined to measuring short-run outcomes. An exception to this restriction is the literature on school starting age. Black, Devereux, and Salvanes (2011) show that starting school later leads to better test scores, but that these effects are driven by age. They also find that starting school later leads to better mental health at age 18, but lower earnings at age 30. Landersø, Nielsen, and Simonsen (2017) find that starting school later causes students to score better at exams and study in fields with higher entry requirements. Fredriksson and Öckert (2014) find that starting school later negatively impacts life cycle earnings[1]. As with school starting age, the institutional framework around exit exams can randomly affect students' performance and later opportunities, regardless of their innate ability or human capital. Identifying and remedying such effects can potentially improve the match quality between individuals and education paths.

The effect of institutionally driven exam timing has previously been studied by Pope and Fillmore (2015). They explore how the time between exams affects the exam scores of American students taking the Advanced Placement exams. The authors have found a linear relationship up to 10 days (the maximum number of days in their data), with

[1]The papers included here are in no way a comprehensive list of previous studies. For further contributions see references in the above papers.

a more pronounced effect for whites, Asians, and females. Their findings indicate that only the score on the second exam is impacted by how many days pass between exams. The authors argue that this effect is likely due to fatigue. However, they admit that the heterogeneity patterns that emerge are more likely to be caused by "cramming", i.e. an intense period of studying just before exams. The current paper differs from Pope and Fillmore (2015) in four key ways. First, the variation in preparation time is approximately twice as large, allowing me to identify a non-linear effect over a longer period than 10 days. Second, Norwegian students take exams in more subjects, allowing me to uncover heterogeneous effects across subjects. Third, students taking the AP exams are generally more academically talented than the average upper secondary school population (Pope & Fillmore, 2015). It is therefore of interest to explore, as I do in this paper, whether the effects Pope and Fillmore (2015) have uncovered hold for a more typical student population. Finally, I can follow students through longer-run outcomes. This is an important contribution as it is not clear ex ante whether increased preparation time only increases test scores, or whether it also impacts longer-run outcomes such as university enrollment.

This paper makes two main contributions. First, I estimate the causal effect of exam scheduling and preparation time on exam performance; this expands our current understanding of how institutionally driven shocks to exam conditions affect exam grades while human capital levels remain very similar. Second, and importantly, I show how these random variations in exam conditions affect longer-run outcomes. From this, we can see how the sensitivity of exam scores to these conditions can significantly impact students future opportunities and educational paths. The more general lesson from these findings is that the institutional framework around exit exams can be significant in shaping students' future matching to further education and the labor market, and even more so than for example school starting age. This might in turn inform on how exit exams should be executed and how much weight to put on student performance in these relative to other performance measures, when evaluating both student performance and the performance of schools and teachers.

While the experimental variation is unique, students in other systems are likely to experience different exam schedules across cohorts or across schools, causing similar variations. Students in some systems might know which exams they will be taking a longer time in advance. I show that the effect of increased preparation time also seem to apply to the mandatory written exam in Norwegian language. As students know they will take this exam when they start upper secondary school, it more closely resembles the exam system in other countries. Furthermore, it is not uncommon for colleges and universities to have a set number of days for students to prepare ahead of exams, for example in the Commonwealth, where a revision week ahead of exams is common. This paper show that students are likely to benefit from such periods, but with sharply declining marginal returns. In addition, students' future schooling outcomes might also be affected.

The rest of the paper is structured as follows: Section 2 explains the institutional setting and the exam system. Section 3 describes the data and the empirical approach. Section 4 presents the results, with heterogeneity analyses between students and subjects in Section 5. In section 6 I explore the longer-run effects on tertiary education outcomes. Section 7 provides concluding remarks.

II Institutional background and exam system

School system

The Norwegian school system consists of ten years of compulsory schooling, which begins the year students turn six, followed by an elective upper secondary education. It is impossible to fail a class during the compulsory component; consequently, grade repetition is practically non-existent so nearly all students complete their compulsory education at age 16. In this study, all students who did not finish mandatory schooling at the age 16 were excluded from the sample. More than 95% of students choose to enroll in elective upper secondary education the fall after graduating from mandatory schooling. Upper secondary schooling is tracked, consisting of 12 tracks that can be grouped into two broad categories: academic and vocational. The three academic tracks consist of the sub-specializations: dance, drama and music; sports; and specialization in general

studies. These span three years, at the end of which graduating students are eligible to apply for higher education.

Of the students choosing to enroll in upper secondary schooling the year they turn 16, roughly 50% opt for an academic track. Only students who have enrolled in the academic tracks are included in the analysis. Students who have enrolled in a vocational track at the age of 16 and then switch to an academic track are therefore excluded.

Each April, before exams, applications for tertiary education programs are submitted through a centralized platform.[2] Students list their ordered preferences for a program by institution (e.g. mechanical engineering at the; University Science and Technology in Trondheim), with up to 15 preferences. Students are allowed to re-arrange their preferences until the end of June. In mid-July, students are given offers of admission based on their application score. The application score is based almost entirely on the average of teacher-assessed grades and exam grades (Kirkeboen, Leuven, & Mogstad, 2016).[3] Higher education institutions cannot consider any application letter or other student attributes (e.g. the student being valedictorian). Therefore, written exam grades account for approximately 15% of the typical student's application score.[4]

Exam system and preparation time

This section outlines the exam system that results in the variation in preparation time. Exams in the academic track may be oral/practical or written, and can be identified as such. Oral/practical exams are excluded from the analysis for several reasons. First, their content is created at the level of the individual school, which also assigns students their exams. Second, the oral/practical exam grade is likely to be influenced by non-academic characteristics.[5] By comparison, written exams are comprehensive, and are held each

[2]There are a few private institutions that are not part of the centralized platform. However, nearly all students enroll in institutions through the platform.

[3]In addition to grades, students' age, gender and subject selection play a limited role, as do military service and folk high school. Specifically, students get extra application points for their age up to a cut-off, and points for one year of military service or one year of optional folk high school. In addition, students get extra points for their gender in some programs, and for science and math subjects in others.

[4]A typical student has around 20 subject grades and 4-5 exam grades, including oral exams. Thus, each exam counts for around 4% of each student's application score.

[5]Oral/practical exams are announced and held after the written exams have been completed, and therefore do not interfere with students' preparation for written exams. This is demonstrated in

school year for each subject, nationwide. Written exams are also anonymized and graded by two teachers from a different school. This ensures that all students taking a written exam in a specific year and subject are given an exam of the same difficulty, and that the results are directly comparable. Exam grades range from one (as the lowest) to six (as the highest) and are distributed in a bell shape with three as the median grade.

Exams in upper secondary schooling are spread across all three years, with most exams taking place in the third year. The first year, 20% of students are randomly selected to take written or oral/practical exams. In the second year, all students must take one exam in a randomly assigned subject. These exams may be written or oral/practical. In the third year, the number and types of exams vary between sub-specializations. However, the majority of students take two written exams and one oral exam, in addition to the mandatory exam in Norwegian.[6]. The details are presented in Table 1

Table 1: Number of exams by specialization

	General	Sports	Music/dance/drama
First year:	20% have exam, written or other		
Second year:	1 exam: written or other		
Third year:	Norwegian language exam		
	+ 2 written exams and 1 oral exam	+ 3 exams, written or other incl. 1 in specialization	+ 2 written or other incl. 1 in specialization.

This table shows the number and types of exams students are required to take during upper secondary schooling, depending on which specialization they follow. The main language exam is in one of the forms of written Norwegian for nearly all students in the sample. Source: Norwegian Directorate for Education and Training.

These are high-stake exams for three reasons. First, if a student fails an exam, she is required to retake it. Second, in order to graduate, students have to pass all exams. Third, students compete for places in higher education on the basis of the average of their subject and exam grades (Kirkeboen et al., 2016). The fact that these are high-stake exams suggests that students will utilize their assigned preparation time to the best of their ability.

A key feature of the exam system in Norwegian upper secondary schools is the assignment of students to specific exams, and the announcement of this assignment. Each year,

Table A.7.

[6]Some Sami students have an exam in their native language rather than in Norwegian

the Norwegian Directorate for Education and Training sets up a schedule for when the written exams in each subject are to be held. This schedule is distributed to schools and announced on the same date at all schools for all subjects. Each of Norway's 19 counties is responsible for assigning students to exams, and they are required to ensure that this assignment is random (Norwegian Directorate for Education and Training, 2009). Students can only be assigned exams in mandatory subjects, or in subjects they elected to take the preceding year. On the announcement date, schools are required to notify students if and when they are to be examined in the various subjects. Note that schools cannot alter the students' preparation time, as they are required to hold the exams on the dates set in the announcement.

The number of days between the announcement date and the exam in each subject is used as the measure of preparation time in the analysis, as mentioned above. Illustrative examples are shown in Figure 1: The two third-year students i and j take the same subjects at the same school in the same year. They are informed on the same date (day zero) which exams they will take, and on which dates these exams will be held. The number of days of preparation time they have for each exam is calculated as the number of days between the announcement date and the exam date. Thus, student j has 9 days to prepare for the exam in German, 17 for the exam in physics, and so on. As is evident from the example, the preparation time is calculated independently of the number of exams a student has in an exam period and the number of exams already taken. Both of these issues are addressed below. In the sample period, preparation time was between 5 and 25 days, with an average of 13.5 days in the final data. The distribution of exam observations across the preparation days is presented in Figure A.6.

After students are notified as to which exams they must take and when, they generally follow normal instruction schedules, although schools and teachers differ with respect to how much time is devoted to preparation for exams during school hours. Some schools provide students with specific study days or extra classes, while others do not. At all schools, however, teachers are available for guidance during the preparation period. If a teacher has students who will be taking an exam in a subject she teaches, she might

offer extra classes or focus instruction on relevant material to help students prepare. The amount of extra instruction offered is teacher-dependent, as there is no national guideline in this regard. Thus, the preparation period generally includes some teacher instruction and some self-directed study.

It should be pointed out that the exam system was not designed specifically to create variation in preparation time; it's objective, rather, is to ensure that students are prepared to take any exam. As students may be randomly assigned to an exam in any of their subjects, they have an incentive to maintain a high level of effort in all subjects throughout the year. To facilitate this, exams are spread out over a long period, resulting in variation in preparation time.

III Data and empirical strategy

Exam scores

The dependent variable in the analysis will be the grade awarded on written exams. The individual grade records of students who enrolled in upper secondary schooling from 2006 through 2009 are collected from register data made available by Statistics Norway covering exams taken in the period 2008 through 2012.[7] The data contains identifiers for the subject for which the grade is awarded. The exam grades are merged with the annual list of exam dates from the Norwegian Directorate for Education and Training.

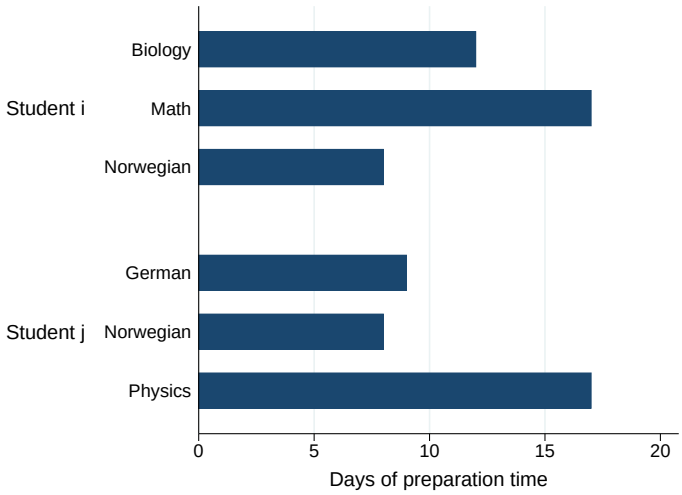
The number of exams each student takes varies somewhat due to attrition, exemptions, sub-specialization, retakes, and the exam assignment system. Students who retake exams to improve their grade are marked as such, and only first attempts are retained in the data. Those who are sick or otherwise unable to meet up for their assigned exam can provide a medical certificate from a physician stating the reason for their absence. They must take a make-up exam in the same subject the following fall semester. The total number of these make-up exams constitutes less than 1% of all exams (Bensnes, 2016).

[7]Prior to 2008, the Norwegian Directorate for Education and Training did not distribute a common announcement date. Rather, the individual schools were required to announce which students were to be examined "at least 48 hours prior to the exam". This rule was interpreted differently across schools (Norwegian Directorate for Education and Training, 2009).

It is important to note that, although re-take exams are identified in the data, make-up exams are not, and the results of make-up exams are thus not distinguishable from ordinary exam results. The final data include students who took two to five written exams, with an average of three and a half written exams per student. This is consistent with the normal number of exams students take in upper secondary school, given student attrition.

Because they choose to take different subjects, students within the same school are generally not all tested on the same days. In any given year, the assignment of students to exams - and hence their preparation time - is random, as it is conditional on the subjects students have chosen as electives

Figure 1: Illustration of preparation time.



Illustrative example of preparation time. The third-year students i and j follow the same subjects at the same school in the same year. On a specific date (day 0) they are informed which written exams they must take, and when these exams will be held. As third-year students, both have to take the mandatory written exam in Norwegian 8 days after the announcement and therefore receive 8 days of preparation time for that exam. Student j is assigned to take exams in German and physics as well, while student i must take exams in math and biology.

Other variables

In addition to exam observations, the data include students' teacher-assessed subject grades and information about several background characteristics: parental education, parental labor market status, and students' gender and immigration status. Descriptive

statistics for selected variables are presented in Table 2[8]. The upper section of the table reports variables at the student level, while the lower part reports variables at the exam level.

Teacher-assessed grades in upper secondary school are slightly higher, on average, than the exam grades. Exam observations that cannot be linked to a teacher-assessed grade are excluded from the sample. If a student does not receive a teacher-assessed grade in a subject, the exam grade is invalid, which in turn gives the student little incentive to perform. The majority of exams are taken in various language subjects. This is largely because all students are required to take a Norwegian language exam. Around a quarter of all exams are taken in subjects classified as mathematics or natural sciences, with the remaining exams taken in "other" subjects.[9]

Table 2: Summary statistics for selected variables. Complete summary statistics in Table A.1.

Variable	Mean	(Std. Dev.)	Min.	Max.	N
Student level					
Average preparation period	13.531	(2.458)	5	24	98,012
Average exam score	3.27	(0.896)	1	6	98,012
Average teacher-assessed grade	3.837	(0.922)	1	6	98,012
Number of exams for students	3.548	(0.756)	2	5	98,012
Dropped out of tertiary education	0.128	(0.334)	0	1	98,012
Enrolled in tertiary education	0.85	(0.357)	0	1	98,012
Enrolled in STEM program	0.16	(0.366)	0	1	98,012
First-generation immigrant	0.029	(0.168)	0	1	98,012
Second-generation immigrant	0.035	(0.183)	0	1	98,012
Female	0.556	(0.497)	0	1	98,012
Exam level					
Preparation period	13.549	(4.871)	5	25	342,187
Exam score	3.289	(1.142)	1	6	342,187
Exam subject					
Norwegian	0.419	(0.493)	0	1	342,187
Natural sciences	0.24	(0.427)	0	1	342,187
Languages, including Norwegian	0.576	(0.494)	0	1	342,187
Other	0.184	(0.387)	0	1	342,187

The preparation period for each examis defined as the number of days between the announcement date and the exam. "Average teacher-assessed grade" is the average for the subjects in which students are examined. "Dropped out of tertiary education" is defined as equal to 1 if the student enrolls in tertiary education, but drops out or changes program by the second year. "Enrolled in STEM program" is equal to 1 if the student enrolls in a science, technology, engineering or math program after completing upper secondary. The subject type "Other" includes the humanities, social sciences and other subjects such as business economics. The three subject types at the end of the table are mutually exclusive. Sources: Statistics Norway, and the Norwegian Directorate for Education and Training.

[8]Extensive summary statistics are reported in Table A.1.

[9]The category "other subjects" includes subjects in the social sciences and humanities, and subjects like business economics and marketing.

Identification strategy for short run effects

The equation to be estimated is presented in Equation (1). The dependent variable is the exam score in subject c for student i taken in year y at school s. The exam score is standardized by subject to have a zero mean and a standard deviation of 1 for each subject.[10] η_c, ψ_y and μ_s are exam -subject, year, and school fixed effects, respectively.[11] Z_{iyc} is a vector of student observables that vary across exams, within students. It includes the number of exams taken by the students in a given year, the number of exams that year, and the standardized teacher-assessed grade in the subject. X_i is a vector of student level controls, including subject-taking controls and cohort dummies. By including subject-taking dummies, I isolate the variation to come from students following the same subjects and therefore eligible to be examined in the same subjects. ϵ_{iyc} is a random idiosyncratic error. In estimations, standard errors will be clustered at the school level. The coefficients of interest are β_1 through β_3, which measure the effect of extra preparation time, measured as the number of days between the announcement of exams and the exams. In the baseline specification, the functional form will be given by three dummies for preparation time in the second to fourth quartiles, with the first quartile as the reference category.[12] This functional form is preferred because it allows for a more flexible relationship to be estimated. This definition is thoroughly challenged in the Appendix, which shows that the results are robust to alternative definitions[13].

$$\begin{aligned} \text{Exam score}_{icys} &= \beta_1(\text{Prep. time 2nd quartile})_{cy} + \beta_2(\text{Prep. time 3rd quartile})_{cy} \\ &+\beta_3(\text{Prep. time 4th quartile})_{cy} + \gamma X_i + \tau Z_{iyc} + \eta_c + \psi_y + \mu_s + \epsilon_{icys} \end{aligned} \quad (1)$$

[10]Standardization is carried out to facilitate interpretation of coefficients. In addition, there are variations in the distribution of grades across subjects prior to standardization. Standardization therefore takes into account that increasing the exam grade in mathematics by one grade point does not necessarily reflect the same effort as increasing the exam grade in Norwegian by one grade point. However, results are not qualitatively sensitive to the transformation of exam scores.

[11]As students take multiple exams, it is also possible to estimate effects conditional on student fixed effects. However, including student fixed effects complicates interpretation as it removes average students' preparation time, as pointed out by an anonymous referee. Results from such a specification provide quite similar results and are available upon request.

[12]The quartiles comprise preparation periods from 5 through 8 days, 9 through 12 days, 13 through 16 days, and 17 through 25 days.

[13]Figures A.3, A.7, Table A.5

The main identifying assumption behind the strategy is that preparation time is in effect random, given the control variables. As both the announcement time and the exam dates are set by the Norwegian Directorate for Education and Training, this seems like a plausible assumption. This assumption is even more likely to hold when estimations include fixed effects in several dimensions. Year fixed effects are included because the exam schedule for each year varies slightly over the period spanned by the data, and small changes in the curriculum might occur. Because the exam year is thus correlated with preparation time, and also potentially with the exam contents, estimates might be biased in the absence of year fixed effects.

Exam-subject fixed effects are necessary to avoid bias. To illustrate this, consider a case without exam-subject fixed effects. If the exam for a specific subject is placed towards the end of the exam period every year, the preparation time for this subject will be longer than for the average subject, for all five years in the data. If exams in subjects that have a higher return to preparation time, also have less preparation time, the estimated effect of increasing preparation time will be biased downward. This would be the case if, for example, the marginal return to preparation time was larger for mathematics and science subjects and these subjects had shorter preparation periods. Both conditions seem to hold: the exams in science and math subjects have 13 days of preparation time on average, whereas subjects that can be classified as "other" have 16 days of preparation time on average. I demonstrate below that the marginal return for science and math subjects is higher than for the other subjects. Thus, including exam-subject fixed effects removes this bias from the estimation.

There might be variation both between and within schools in terms of how much extra instruction time and what kind of instruction students are offered during their preparation period. Moreover, more able students might sort to schools that offer more and better extra instruction during the preparation period. School fixed effects absorb the average differences across schools in this respect and eliminates this source of bias.

To make sure that students with longer preparation periods are as comparable as possible to the students who receive shorter periods, I have also included subject-taking

effects. Subject-taking controls net out variation between students with respect to which subjects they choose to follow and hence which subjects they might be examined in. Including these variables is not crucial to the results, but which exams a student has to take is truly random only conditionally on subjects taken. To make sure there is no remaining variation stemming from the sorting of students into subjects, subject-taking effects must be included[14].

Given these fixed effects, the remaining variation stems from the differences in preparation time within subjects, across years, for students taking the same combination of subjects. Year fixed effects remove variations in preparation time common to all subjects across years, and exam-subject fixed effects remove average differences in preparation time between subjects that is common across years. Subjects for which the preparation time changes, with a different number of days than the average between two years, provide variation that allows identification. Thus, even if all students were assigned to exams in the exact same subjects, it would still be possible to identify the causal effect of preparation time when multiple years of data were observed and changes in preparation time across years were random.

In order to illustrate the variation, I have included some informative figures in the Appendix. Figure A.1 shows the distribution of residual variation in preparation time overall and by subject -group and the residual variation in exam scores, and Figure A.2 plots the residual variation in exam scores against the residual variation in preparation time with a quadratic fit. In sum these figures illustrate that i) there is residual variation in preparation time for each subject group and in exam grades. ii) the residual variation in preparation time exhibits a concave relationship with the residual variation in exam scores. For further details see the text in the Appendix.

For the estimates of Equation (1) to be unbiased, student characteristics must be uncorrelated with preparation time conditional on the fixed effects. Table 3 reports three separate balance tests with preparation time regressed against the background controls[15].

[14]In Table A.4 in the Appendix I have included various estimates of the baseline model with less controls. In Column (5) of the table subject-taking effects are excluded with only small changes to results.
[15]Additional balance tests are reported in Appendix Tables A.2 and A.3

In the first column, exam-subject and subject-taking fixed effects are dropped; in the second column only subject-taking fixed effects are dropped; and the last column includes controls. The last three rows in the table report the p-value for three F-tests for joint significance of (i) the student background characteristics, (ii) the cohort dummies, and (iii) the exam-subject dummies. In the absence of exam-subject fixed effects, the student background controls are highly significant. But both with and without subject-taking controls, the balance tests show that the background effects are jointly uncorrelated with preparation time, supporting the identifying assumption. Note that there is a weak correlation between immigration status and preparation time when subject-taking fixed effects are not included. For this reason, the baseline model will include subject-taking fixed effects. Exam year and cohort effects are significant, which is not surprising as average preparation time shifts from year to year, and cohorts largely determine the years in which students are examined.

IV Effects on exam scores

Results are presented in Table 4. In order to fix ideas, the first column estimates effects using a simple model with exams only from the second year, when students have only one exam. In this specification, results should be interpreted as the marginal effect of increasing the preparation time for a single exam, relative to a period of 5-8 days. Although this specification includes only about a third of observations, it appears that there is a concave effect of preparation time, with no significant improvement in exam scores with 17-25 days, relative to 13-16 days. The point estimate for 9-12 days is insignificant partly because the preparation time quartiles are defined for the sample as a whole, and relatively few exams are held in the second quartile, when the sample is restricted to second year exams. A preparation period of 13-25 days yields a positive return for exam scores of about 6-7.6% of a standard deviation. One benefit of only using second-year students is that they only have one exam - issues related to students preparing for multiple exams, such as in the example provided in Figure 1, are therefore irrelevant. One drawback, however, is that the effect of preparation time on a single

Table 3: Balance test

	(1) No subject FE	(2) Exam-subject FE	(3) Course-taking FE
First-generation immigrant	0.0482	0.0122	0.0300
	(0.0443)	(0.0337)	(0.0503)
Second-generation immigrant	-0.0966**	-0.0626*	0.0240
	(0.0454)	(0.0335)	(0.0444)
Female	-0.114***	0.00803	0.00751
	(0.0153)	(0.0115)	(0.0125)
Mother's education			
Upper sec. school	0.0227	0.0235	0.00720
	(0.0187)	(0.0160)	(0.0156)
Bachelor's degree	0.0168	0.0222	0.0123
	(0.0197)	(0.0174)	(0.0166)
Master's or PhD	0.0279	0.00781	-0.0225
	(0.0277)	(0.0209)	(0.0291)
Father's education			
Upper sec. school	0.0163	0.00215	-0.0164
	(0.0184)	(0.0147)	(0.0158)
Bachelor's degree	0.0432**	-0.000254	-0.0165
	(0.0193)	(0.0158)	(0.0187)
Master's or PhD	0.0489**	-0.0216	-0.0273
	(0.0239)	(0.0183)	(0.0241)
1 parent working	-0.0516	0.00675	-0.0396
	(0.0459)	(0.0402)	(0.0479)
Both parents working	-0.0567	0.0137	-0.0213
	(0.0458)	(0.0393)	(0.0480)
Parental income	1.38e-08**	-9.14e-10	2.95e-08
	(6.90e-09)	(4.09e-09)	(2.01e-08)
Exam in year 2009	4.804***	2.587***	2.463***
	(0.0666)	(0.145)	(0.202)
Exam in year 2010	9.750***	5.500***	5.203***
	(0.0973)	(0.271)	(0.387)
Exam in year 2011	12.59***	5.769***	5.252***
	(0.145)	(0.393)	(0.566)
Exam in year 2012	14.54***	3.530***	2.706***
	(0.192)	(0.517)	(0.747)
GPA lower secondary	0.191***	-0.00943	-0.0135
	(0.0261)	(0.0162)	(0.0198)
Standardized teacher assessed grade	-0.0781***	0.00496	0.00402
	(0.0175)	(0.0107)	(0.0155)
Constant	9.338***	6.010***	6.216***
	(0.154)	(0.0975)	(0.283)
Observations	342,187	342,187	342,187
R^2	0.225	0.563	0.640
Exam year FE	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes
Exam-subject FE	No	Yes	Yes
Course-taking FE	No	No	Yes
p-value joint test background effects	0	.503	.694
p-value joint test cohort effects	0	0	0
p-value joint test exam effects		0	0

The outcome is the number of days of preparation time. Parental income is measured in nominal terms. The third last row reports the p-value for an F-test of joint significance on the background variables. The second from last row and the last row report p-values for similar tests on the cohort and exam-subject fixed effects respectively. Standard errors clustered by school in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Sources: Statistics Norway, and the Norwegian Directorate for Education and Training.

exam has a smaller impact in the longer-run, making it harder to identify effects on the outcomes studied below. Further, students in many school systems have to take multiple exams within a short period. To better understand these cases, it is of interest to assess the importance of preparation time also for the third year students in the sample.

To examine the importance of interaction effects between the number of exams and preparation time, I redefine the sample in Column (2) to include only third-year exams for students with exactly three exams in the same period. This specification also includes a control for the number of exams the student has already taken in the same exam period. This takes into account the fact that students who have multiple exams within a given period must divide their time among more tasks than students who have only one exam in an equivalent period of time. The results show that, conditional on the number of exams being the same, there is a positive return to preparation time than 5-8 days, but the marginal return quickly falls to zero. While the point estimate for 9-12 days is different from that in Column (1), the other coefficients are fairly similar. The similarity between the estimated effects for longer preparation periods could be attributed to students only preparing for a given number of days, regardless of how much time they are given. For example, students who know they face an exam in three weeks might only start to begin preparing about one week before the exam.

When there are multiple exams, a student's performance on an exam after 20 days' preparation time might be influenced by the fact that she had another exam 2 days previously. The results reported here do not take such effects directly into account. However, as the results in Columns (1) and (2) are very similar, such effects are likely to be small compared to the effect of the number of days following the announcement. This should be kept in mind when interpreting the coefficients, nonetheless.[16]

Observing the results in Columns (1) and (2), Column (3) compiles all written exam observations. The model also controls for the total number of exams in the period. The

[16]One referee pointed out that simple methods of controlling for the number of days since the previous exam change what is being estimated from the effect of the number of days since the announcement, to the number of days between the announcement and the previous exam. I therefore do not include controls for the number of days since the previous exam.

effects are very similar to the results in Column (2). This suggests that variations in the number of exams students have do not interact strongly with estimates, when controlling for the number of exams taken and the total number of exams in the exam period. While it appears that the controls take into account the number of other exams, there is also the question of whether there are spillovers in the preparations for relatively similar exams. Table A.12 in the Appendix reports results indicating that there is some spillover when students prepare for similar exams over a short time span.

The fourth column also includes controls for students' subject- selection. This ensures that differences between students in terms of preferences and possible exam combinations are controlled for, thereby increasing the comparability of the students in the sample. The effects found are very similar to the results in Column (3); this is probably due, at least in part, to the exam-subject fixed effects and other controls already included. From this point on, the model estimated in Column (4) will be referred to as the baseline, as the balance test is stronger when subject-taking effects are included.

The baseline estimates include the mandatory exam in Norwegian. However, from the moment they enter upper secondary school, students are aware that they will be taking this exam. Therefore, one could argue that the preparation time for this exam should be irrelevant for exam scores. To check whether students' scores on their Norwegian exam are affected, I re-estimate the baseline model, while excluding all exams on Norwegian. These results are reported in Column (5), and show that the return to preparation time is less concave but quite similar to the results from the baseline specification. This suggests that exam scheduling has an effect on exam scores, even when students know about the exam three years ahead of time. This result is important, as it suggests that the effects estimated in this paper can also be informative for other systems that do not use the same announcement and assignment mechanisms as the Norwegian upper secondary school system.

In the Appendix I report, the results of several specification checks, including changing the grouping of periods (Appendix Figure .3), defining treatment linearly, quadratically, and controlling for holidays and weekends (Appendix Table A.5). None of these checks

change results substantially. For further discussion, see the text in the Appendix. I also change the outcome variable to check which parts of the grade distribution are affected by preparation time. The results of this exercise show that increased preparation time increases the probability of students receiving a grade closer to average than the bottom. See Table A.6 and the related discussion. As a placebo test, I check in Appendix Table A.7 that oral exams held after the written exam period are unaffected by the length of preparation time. This exercise demonstrates that the variation used in the main specifications does indeed measure the preparation time for written exams.

The results thus far raise the question of how students divide their time among exams. Do they front-load their efforts and focus on exams in the order in which they take them, or do they divide their time equally? In the absence of data on how students spend their time, these questions can only be answered indirectly. In Table A.10 in the Appendix, I report additional results aimed at providing some clues, including a model where preparation time is defined as the average number of days per exam for each student, and allowing the effect of preparation time to differ depending on whether the exam is the first one or not. I find no significant effect in terms of average number of days per exam, nor in the return to preparation time depending on whether the exam is the first that year or not. However, I do find that the number of days since the previous exam has a positive effect which peaks at around 12 days, comparable to the effects in the baseline model. In combination, these additional results suggest that students do not front-load, but rather set aside a maximum of approximately 12 days to prepare for any exam. Additional preparation time beyond this point is likely substituted for leisure or activities that do not increase exam scores. Further discussion on mechanisms is allocated to Section 8.

V Heterogeneities

Heterogeneous effects by student characteristics

Taking results so far as causal, one might ask whether there are some underlying heterogeneities with respect to which students benefit most from increased preparation

Table 4: Short run effect of preparation time on exam scores.

	(1) One exam, 2nd year	(2) 3 exams in exam period	(3) All exams	(4) All exams	(5) Excl. mand. Norw. exam
9-12 days of prep.	0.0124	0.0724***	0.0446***	0.0530***	0.0571***
	(0.0198)	(0.0173)	(0.0110)	(0.0124)	(0.0158)
13-16 days of prep.	0.0605**	0.0738***	0.0427***	0.0569***	0.0636***
	(0.0241)	(0.0201)	(0.0113)	(0.0130)	(0.0194)
17-25 days of prep.	0.0762***	0.0836***	0.0451***	0.0656***	0.0808***
	(0.0214)	(0.0231)	(0.0131)	(0.0155)	(0.0190)
Observations	71,033	243,368	342,186	341,735	243,787
Exam period controls	No	Yes	Yes	Yes	Yes
Course-taking FE	No	No	No	Yes	Yes

The outcome in all regressions is the exam grade standardized by course. "Days of prep." refers to the number of days between the announcement and the exam. All specifications control for: teacher-assessed grade in the subject; parental education; income and labor market status; students' immigration status, gender and GPA from lower secondary education, exam year, and exam-subject, school and cohort fixed effects. Standard errors clustered on school in parentheses. The specifications in Columns (3)-(5) also include dummies for the number of exams the student has already taken during the same exam period. Columns (4) and (5) include dummies for the number of exams in the same exam period. Columns (4) and (5) also include fixed effects for all subject combinations students take in school. *** p<0.01, ** p<0.05, * p<0.1. Sources: Statistics Norway, and the Norwegian Directorate for Education and Training.

time. Finding differences in the effects of preparation time across students would be of interest, not just from a policy standpoint, but also from a more general perspective, as it would shed light on how students with various background characteristics prepare for high-stake exams.

In Table 5, I split the sample by gender and previous school performance.[17] Previous school performance is measured as GPA in lower secondary school which is achieved before enrollment in upper-secondary. High-skilled students are defined as students having a GPA above the median. When the sample is split by high and low skilled females in Columns (1) and (2), we can see that girls consistently gain more from increased preparation, with only minor differences between the skill levels. Girls also appear not to exhibit as strong a concave relationship as boys, with marginal returns increasing slightly in time. Boys, on the other hand, have a lower return to preparation time than girls regardless of skill level and exhibit a strong concavity in effects. Interestingly, the

[17] There is no statistical difference between students from different socio-economic backgrounds, measured by parental education or income. These results are reported in Table A.8 in the Appendix.

exam score of high-skilled male students does not increase with preparation time. As high-skilled female students do have a return to preparation time, it appears that there is an interaction between gender and skill level which causes the effect of zero returns for high-skilled males. The heterogeneity pattern here also mirrors the pattern found by Pope and Fillmore (2015), that the number of days between exams has stronger effects for girls. While it is not possible to identify the mechanisms at play here, it is possible that girls have higher aspirations or stronger non -cognitive skills (Fortin, Oreopoulos, & Phipps, 2015; Jacob, 2002). Such factors could make girls inclined to invest more effort in preparing for exams than boys, including setting aside more days to prepare. The gender heterogeneities are further explored with regard to longer-run outcomes in Section 6.

In addition to the gender differences, Table 5 indicates that low-skilled students in general benefit more from preparation time than their higher-skilled peers. This is interesting, because one might expect the latter to be better at taking advantage of preparation time due to their scholastic aptitude. However, the low-skilled students might have higher returns to preparation time as they master a smaller share of the curriculum at the start of the preparation period. The estimates appear to indicate that the latter effect predominates. Additionally, if graders tend to grade on the curve, longer preparation periods will increase the competition for top marks, thus making it harder for students with a strong track record to achieve the best grades. This claim is further substantiated by the estimates in the Appendix Figures A.4 and A.5, where I estimate separate effects for each quintile and decile of previous school performance. Students in the top part of the distribution (i.e. top 10%) have negative but insignificant marginal returns. Eren and Millimet (2008) have previously found that low-skilled students benefit from a longer school year, whereas high-skilled students benefit from shorter school years. The same principle appears to apply to preparation time, suggesting that ceiling effects are quite important.

Heterogeneous effects by subject type

The effects found thus far are not permitted to differ between subjects. It is not a given that the exam score production function is the same across subjects in terms of

Table 5: Heterogeneities across students

	(1)		(2)	(3)		(4)
	Girls			**Boys**		
	High-skilled		Low-skilled	High-skilled		Low-skilled
9-12 days of preparation time	0.0435***		0.0667***	0.0257		0.0552***
	(0.0150)		(0.0176)	(0.0214)		(0.0161)
		[1.32]			[1.83]	
13-16 days of preparation time	0.0428***		0.0854***	0.0133		0.0506***
	(0.0160)		(0.0176)	(0.0232)		(0.0166)
		[2.41]			[2.25]	
17-25 days of preparation time	0.0678***		0.0825***	0.0188		0.0312*
	(0.0192)		(0.0206)	(0.0272)		(0.0180)
		[0.71]			[0.69]	
Observations	111,353		77,851	63,904		88,989

The outcome in all regressions is the exam grade standardized by course. "Days of prep." refers to the number of days between the announcement and the exam. Specifications are similar to the baseline, but with split samples. "High-skilled" refers to students who have an average teacher-assessed course grade in lower secondary school above or equal to the sample median. "Low-skilled' refers to the remainer. t-tests on the difference between coefficients across pairs are reported in square brackets. Both regression pairs are mutually exclusive. *** p<0.01, ** p<0.05, * p<0.1. Standard errors clustered by school in parentheses. Sources: Statistics Norway, and the Norwegian Directorate for Education and Training.

preparation time as an input, because the exam design and concepts differ: the language exams generally consist of essay writing, whereas quantitative exams comprised sets of problems to be solved. One might therefore expect preparation time ahead of quantitative exams to have a greater effect, as it is more difficult to improve writing skills in a short period than it is to learn new concepts in science and math. To explore heterogeneities, I re-estimate the baseline results separately for exams by gender and three subject groups: science and math subjects, language subjects, and "other" subjects.

Results are reported in Table 6. In the first two columns, the sample is confined to exams in natural sciences and math. The effects are generally larger than those in the baseline, but vary more and are generally larger for boys than girls. In part, this is because the distribution of math and science exams across preparation time is not equal to the overall distribution of exams from which the treatment groups are defined. The concavity from the baseline estimates is still present. For languages, a similar pattern emerges for females, with no significant effect for boys. For "other" exams, preparation time has no effect.[18] While it is not possible to identify the exact mechanisms underlying these

[18]In Table A.9, I report results when the effect of preparation time is allowed to vary both across subjects and across students with different performance levels in similar subjects in lower secondary school.

differences, it is possible that they are driven in part by differences in the interests of the genders: boys might be more interested in math and science subjects and devote relatively more time to preparing for exams in these subjects, whereas girls may be more interested in - and therefore focus on - language subjects. I return to subject heterogeneities when analyzing longer-run outcomes in the next section.

Table 6: Heterogeneities across subjects and gender

	(1)	(2)	(3)	(4)	(5)	(6)
	Science and math		**Languages**		**Other**	
	girls	boys	girls	boys	girls	boys
9-12 days of prep.	0.196***	0.181***	0.106***	0.0342	-0.0271	0.0210
	(0.0354)	(0.0293)	(0.0245)	(0.0322)	(0.0336)	(0.0337)
	[0.52]		[2.23]		[1.43]	
13-16 days of prep.	0.0774***	0.120***	0.0869***	0.0214	0.0245	0.0513
	(0.0267)	(0.0301)	(0.0211)	(0.0262)	(0.0428)	(0.0469)
	[1.41]		[2.50]		[0.58]	
17-25 days of prep.	0.0712**	0.180***	0.117***	0.0443	-0.0209	-0.0494
	(0.0351)	(0.0335)	(0.0258)	(0.0288)	(0.0446)	(0.0474)
	[3.25]		[2.52]		[0.60]	
Observations	40,555	41,511	113,126	84,035	35,513	27,401

Specifications are similar to the baseline, but with split samples. "Days of prep." refers to the number of days between the announcement and the exam. Column (1) uses only exams in science and math subjects, Column (2) uses only exams in language subjects, Column (3) uses exams only in other subjects. Columns (1), (3) and (5) only use female students. Columns (2), (4) and (6) use only male students. t-tests on the difference between coefficients across pairs are reported in square brackets. All specifications include the same controls as the baseline specification. *** p<0.01, ** p<0.05, * p<0.1. Standard errors clustered by school in parentheses. Source: Statistics Norway, and the Norwegian Directorate for Education and Training.

VI Longer-run outcomes

The estimated effects thus far show that students' exam performances are sensitive to preparation time. Taking the results as causal, the natural next question is whether the effects are large enough to affect students' outcomes in the longer- run. This is an important question, as it is unclear whether the short-run effects are substantial enough to alter students' longer- run outcomes. To explore this question, I follow students into higher education and analyze how increased preparation time affects outcomes beyond exams in secondary education.

The results of this exercise find little additional heterogeneity for language and other subjects, while there is some heterogeneity across skill levels for math and science exams. In particular, weaker students benefit from longer preparation periods, whereas stronger students benefit from shorter ones. This is in line with the findings in Table 5.

The short run effects are based on multiple observations per student, but for the longer-run effects, I only observe each outcome once for each student - the data must therefore be aggregated to the student level. This is not entirely straightforward, as the short-run estimates showed significant non-linearities. To take the non-linearities into account, I use a method following Bensnes (2016), which uses similar data to answer a different question but faces the same challenges as this paper. The method involves estimating the direct effect of more preparation time across all exams on average exam grades, as well as on higher education outcomes. Specifically, I use the share of exams with a preparation period in the second quartile or above as the independent variable.[19]

I focus on four longer-run outcomes. The first outcome I consider is whether students enroll in higher education. The second outcome is defined as the average application score for other students in the sample starting the same program the same year. This measure is not identical to the application score cut-off, as individuals from older cohorts may apply as well. But it gives some indication as to the general requirements for entering the program[20]. The third outcome is a dummy for students enrolling in math or science programs (STEM).[21] STEM programs generally require higher application scores and are more competitive than most other programs.[22] The final outcome is a dummy for students who drop out of tertiary education or switch to an alternative program before starting their second year. Broadly speaking, the first of these outcomes measures the effect of preparation time on the extensive margin (i.e., the effect on enrollment), while the other three measure the intensive margin (i.e., the effect on where students are accepted and the match quality between the student and the program).

It is crucial to note that students apply for tertiary education programs in April, prior

[19]One alternative might be to use the average preparation time across all exams, but that would not take into account the strong concave relationship between preparation time and performance.

[20]The application scores have to be calculated from the exam score and teacher-assessed grades, and does not include other minor adjustments such as army service.

[21]STEM programs are defined by Statistics Norway, and is also applied by Falch, Nyhus, and Strøm (2014). In addition to pure mathematics and natural science programs, this grouping also includes engineering.

[22]Ideally, I would like to know the exact application score cut-offs for each program, and students' ordered preferences. With such data, I could identify whether more preparation time causes students to enroll in programs that are higher up on their ranked preferences. In the absence of these data, I use STEM-programs as a broad definition of programs with higher requirements, as well as the average application score for other students enrolled in the same program.

to their exams. However, they can re-arrange their preferences until the end of June, and are given enrollment offers in mid-July. Effects on longer-run outcomes may therefore potentially arise from students re-arranging their preferences due to the amount of preparation time and consequent exam performance. However, due to the manner in which the enrollment system is designed (a serial-dictatorship), students have little incentive to re-arrange their stated preferences unless underlying preferences change (Kirkeboen et al., 2016).

Long-run effects can operate through three main channels. The first reason to expect an effect in this model framework is that students who are randomly assigned more preparation time receive higher exam grades, which in turn increases their university application score. The second reason is that students who perform well in certain exams due to increased preparation time might change their underlying preferences for certain programs and therefore change their applications, which may alter which program they are accepted into. Last, students might gain minor increases in human capital due to increased preparation time which might reduce the drop-out probability. When a reduced form effect is estimated, these mechanisms cannot be directly distinguished; however, the reduced form estimates show the policy-relevant measure of how preparation time affects longer-run outcomes.

Before turning to longer-run results, I first present a balance test in Column (1) in Table 7. The outcome in this regression is the share of exams that are in the second to fourth quartiles in the distribution of preparation time. The balance test is intuitively similar to the one reported in Table 3. In addition to reported student characteristics, all regressions in Table 7 include controls for parental education, subject-taking effects, school-by-cohort fixed effects and dummies for the exams to which students are assigned and the years in which exams are taken[23]. Full results are reported in Table A.14 in the Appendix. The results in Column (1) show that, with the exception of parental labor market status, there are no significant correlations between students' background

[23]School-by-cohort fixed effects are preferred over separate cohort and school fixed effects because they eliminate potential differences in school-specific factors in teaching practices ahead of exams to a larger degree, and allow these to vary across years as new teachers and students enter the sample.

characteristics and their preparation time. Jointly, background characteristics are not correlated with preparation time. The fact that parental labor market status is weakly correlated with preparation time might be coincidental, but suggests that the other estimates presented in the table should be interpreted with caution. In Table A.13 in the Appendix, I run an alternative balance test based on Pei, Pischke, and Schwandt (2018): I separately regress each control variable against the preparation time measure and the fixed effects giving a total of 13 regressions. The results from these regressions indicate that there is a small, but significant, positive correlation between first generation immigration status and preparation time, while there is a small negative correlation between second generation immigration status and preparation time. Specifically, first generation immigration status is associated with a 1% higher share of exams with long preparation periods, with the opposite effect for second generation immigrants. While small, the results indicate that for the immigrants in the sample (6%) preparation time is not completely independent of their background. This is worth keeping in mind when interpreting the longer-run results[24].

Considering the gender heterogeneities uncovered in Table 5, the longer-run effects are permitted to differ between the genders. The results in Column (2) show that increasing the share of exams in the second to fourth quartiles from 0 to 1 increases the average exam grade by 0.06, or about 6.6% of a standard deviation for girls. There is no significant effect for boys. This effect is very similar to the estimated effects in the baseline model of increasing the preparation time for a single exam from the first quartile to a higher quartile, strengthening the credibility of the longer-run results. Additionally, in Table A.15 in the Appendix I report results from the longer-run equivalent of Table A.6. The results show that longer preparation periods increase the probability of students scoring above the median on their exam rather than below, and the effect is most pronounced for female students.

The second column reports the reduced form effect on the probability of enrolling in tertiary education. According to the point estimates, an average female student is about

[24]See additional discussion in the appendix.

2%-points more likely to enroll in tertiary education when given the most advantageous amount of preparation time relative to the least advantageous. Again, there is no significant effect for male students.

It also appears that there are effects on the probability of a female student enrolling in a program with more skilled peers. The point estimates in Column (4) suggest that a female student who has all her exams in the second to fourth quartiles rather than the first enrolls in a program where the average application score among her peers is about 0.85 points higher, or about 7% of a standard deviation[25].

In terms of field of study, there is a small effect on the probability of enrolling in the more competitive STEM programs for males only. As the second column show no effect on boys' average exam score, it might be that this effect comes from a shift in boys' preferences due to increased subject-specific confidence or interest rather than an improved application score.

The sixth column estimates the effect of preparation time on the probability of dropping out of tertiary education. Girls are 2.6%-points less likely to drop out if all their exams have long preparation periods rather than short ones. Again, there is no effect for boys. This effect is large as the average drop-out rate in the sample is 14%, and suggests that the female students who score better on their exams due to increased preparation time are matched with a program more closely aligned with their preferences, which in turn reduces the probability of dropping out. However, it is not possible to rule out other explanations. For example, students who receive more preparation time may acquire some skills that are beneficial in tertiary education. Alternatively, a student who does relatively well on an exam might be more motivated to pursue an education closer to the exam subject and re-arrange their applications such that they end up enrolling in a program closer to their new preferences. There is some evidence of this as female students have a relatively higher return to preparation time for language exams while also being more likely to enroll in language programs. Similarly, male students have a relatively high return to preparation time in science and math exams. Male students are also more

[25]The average peers' application score in the sample is 39.3 with a standard deviation of 11.4

likely to enroll in STEM programs when preparation time for math and science exams increase. (Tables 8 and 6).

How large are these effects? The average share of exams in the 2nd to 4th quartiles is 0.825, with a standard deviation of 0.194. Increasing the share by one standard deviation therefore increases the probability of a female student enrolling in university by half a percentage point[26]. In the sample, the college enrollment rate for females is about 94%. The reason for this high number is that data is restricted to students who enrolled the academic track and therefore consists of more talented students who are more likely to seek further education.[27]. Considering the college enrollment rate in the sample, the effect is quite large. However, previous studies from Norway have found a wide span of effect from institutional factors on longer-run outcomes. Black et al. (2011) find no effect of school starting age on educational attainment. Using the random allocation of students to math exams relative to language in lower secondary school in Norway, Falch et al. (2014) find that being allocated to take a math exam increases the probability of enrolling in university by 0.15% points, a third of the effect magnitude shown here. So while effects are relatively large, they are not disproportional in a Norwegian context.

I will now explore the field of study-effect from Table 7 in more depth. Table 6 showed that the effect of preparation time on exam scores is largest for science and mathematics subjects; I now estimate the effect of preparation time prior to science and math exams on the probability of enrolling in STEM programs.[28] Results are reported in Column (1) in Table 8. The results show that, for girls, more preparation time for any exam has little to no effect on the probability of their enrolling in STEM programs. For boys, more preparation time for math and science exams has a strong effect on STEM enrollment: increasing the share of all exams in the second to fourth quartiles by one standard deviation increases the probability of enrollment in a STEM program by about 1.5%-points[29].

[26] $0.194(0.00128 + 0.0191) \approx 0.004$

[27] 50% of students start the academic track in upper secondary school

[28] I report the effect of preparation time for various subjects on tertiary education and STEM enrollment in Tables A.16 and A.18 in the Appendix.

[29] A standard deviation in the overall average of share of exams in the second to fourth quartiles is 0.19, and 0,21 for math and science subjects. 0.2*(-0.0214+0.0960)≈0.015. For girls the equivalent estimate is 0.2*(-0.0214+0.0960+0.0381-0.125)≈-0.002. The independent mean is 0.27 for males and 0.11 for females.

Table 7: Longer-run effects of preparation time

	(1) Balance	(2) Avg. exam score	(3) Enroll in college	(4) Avg. appl. score for enrolees	(5) Enroll in STEM	(6) Drop out
Share of exams						
2nd to 4th quartile		0.0188	0.00128	-0.0291	0.0245*	0.0111
		(0.0307)	(0.00910)	(0.394)	(0.0137)	(0.0142)
Share exams						
2nd to 4th quartile x girls		0.0638***	0.0191**	0.876**	-0.0178	-0.0261**
		(0.0240)	(0.00894)	(0.389)	(0.0118)	(0.0116)
GPA lower secondary	-2.02e-05	0.890***	0.0573***	4.393***	-0.0314***	-0.0384***
	(0.000971)	(0.00758)	(0.00257)	(0.105)	(0.00275)	(0.00294)
1 parent working	-0.00527*	-0.00687	0.0223***	0.814**	0.00402	0.00164
	(0.00307)	(0.0181)	(0.00806)	(0.335)	(0.00767)	(0.00815)
Both parents working	-0.00571*	-0.0117	0.0309***	1.171***	0.00787	-0.00106
	(0.00302)	(0.0174)	(0.00803)	(0.330)	(0.00761)	(0.00770)
Average parental income (NOK)	4.65e-10	-1.51e-09	7.24e-10	5.09e-08	-4.19e-09	2.21e-10
	(3.02e-10)	(1.45e-09)	(6.06e-10)	(4.07e-08)	(2.91e-09)	(9.24e-10)
Female	-0.000498	-0.0909***	0.0201***	0.549*	-0.0444***	-0.0196**
	(0.000964)	(0.0213)	(0.00765)	(0.332)	(0.0102)	(0.00979)
First-generation immigrant	0.00419	-0.0938***	0.0252***	1.379***	0.00205	-0.0385***
	(0.00271)	(0.0207)	(0.00680)	(0.292)	(0.00847)	(0.00794)
Second-generation immigrant	-0.00441	-0.100***	0.0478***	2.249***	-0.0128	-0.0483***
	(0.00271)	(0.0169)	(0.00506)	(0.219)	(0.00871)	(0.00711)
Observations	89,297	89,297	89,297	89,297	89,297	89,297
p-value joint sign. background	0.174					

The outcome in the first column is the share of exams a student has that are in the second to fourth quartiles in the distribution of preparation time as used, in the short-run specifications above. The specification is a balance test. The last row in the first column reports the p-value of a test joint significance of background characteristics. The second to sixth columns report the reduced form effect of the share of exams in the second to fourth quartiles in the distribution of preparation time on various outcomes, as described in the column titles. The outcome "Enroll in college." is a dummy equal to 1 if the student enrolls in tertiary education. "Avg. appl. score for enrolees" is the average application score of students enrolling in the same program the same year as the student. The outcome "Enroll in STEM" is a dummy equal to 1 if the student starts a STEM program in tertiary education. The outcome "Drop out" is a dummy equal to 1 if the student drops out or changes program before his second year. All specifications include cohort-by-school, and course-taking fixed effects, dummies for taking exams in various years, and number of exams taken. Full results, including estimates of coefficients on parental education are reported in Table A.14 in the Appendix. Standard errors clustered by school in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Sources: Statistics Norway, and the Norwegian Directorate for Education and Training.

The difference between boys and girls found in Column (1) could in part reflect the gender difference in Table 6, namely that boys get a higher return to preparation time in math and sciences relative to girls. It could also be a motivational effect where male students assigned to math exams perform well and become more confident or interested in STEM subjects and pursue a STEM program.

If more preparation time for science and math exams increases the probability that male students enroll in STEM -programs, one might wonder where these students would have gone had they had shorter preparation periods. In Table A.19 in the Appendix, I show that more preparation time for science and math exams is associated with a lower probability of male students enrolling in social science and humanities programs, but has no significant effect on language programs. Male students with relatively long preparation periods therefore seem to shift from the social sciences and the humanities

to STEM-programs, displacing other students with shorter preparation periods.

The next two columns of Table 8 report the effect of more preparation time for language exams on the probability of enrolling in a language program, and similarly for social science and humanities programs. The results show that females' enrollment in language programs is increased by more preparation time for language exams. Having more preparation time for exams in subjects related to the program students are eventually accepted for has no direct effect for other subjects. This last result is not very surprising, as these programs are far less competitive than other programs and the effect of preparation time is smaller in these subject groups[30]. The results from this table suggest that students who perform better in certain subject due to increased preparation time might re-arrange their preferences and thereby change the programs for which they are accepted. The results resonate with Falch et al. (2014) who find that exposure to a math exam in lower secondary school can alter students' motivation to pursue STEM in college.

It is also possible to scrutinize the effect of preparation time on enrollment requirements further. In Table 7, I show that increased overall preparation time is associated with the student enrolling in a program with higher entry requirements. In Table A.17 in the Appendix I show the results of re-estimating Table 8 while changing the outcome to be the average application score among enrolling students in the same program. The overall results are quite similar to the results in Table 8: For females, more preparation time is associated with enrolling in programs with higher requirements, and the effect is strongest for more preparation time in language subjects. For males, only more preparation time in science subjects generates an effect. These results support the findings in this paper: females benefit the most from preparation time in general, but males benefit the most from preparation time for math and science.

[30]it is possible to estimate these effects splitting the sample by gender rather than using interactions. However, because of the very demanding model used, with a large number of fixed effects that is used, sample sizes become too small to identify effects.

Table 8: Longer-run effect of preparation time, heterogeneous effects on program enrollment

	(1) Outcome: Start STEM	(2) Outcome: Start language	(3) Outcome: Start other
Share exams 2nd to 4th quartiles	-0.0214	0.00591	-0.0318
	(0.0279)	(0.00628)	(0.0363)
Share exams 2nd to 4th quartiles x female	0.0381*	-0.00716	-0.0181
	(0.0230)	(0.00660)	(0.0302)
Share science exams 2nd to 4th quartiles	0.0960***		
	(0.0304)		
Share science exams 2nd to 4th quartiles x female	-0.125***		
	(0.0214)		
Share language exams 2nd to 4th quartiles		-0.0165	
		(0.0113)	
Share language exams 2nd to 4th quartiles x female		0.0189***	
		(0.00613)	
Share language exams 2nd to 4th quartiles			0.00348
			(0.0443)
Share other exams 2nd to 4th quartiles x female			0.0153
			(0.0360)
Observations	49,484	88,622	46,968

Each column estimates the effect of longer preparation periods for exams in a subject group on the probability of enrolling in a related program in tertiary education. "Share exams 2nd to 4th quartile" is the share of exams the student takes that are in the second to fourth quartile in the preparation time distribtution. The outcome "Start STEM" is a dummy equal to 1 if the student starts in a STEM program in tertiary education. The outcome "Start language" is the equivalent for language programs. The outcome "Start other" is the equivalent for starting in social science, or humanities. All specifications include cohort-by-school, course-taking fixed effects, dummies for taking exams in various years, and number of exams taken. Full results are reported in the Appendix. Standard errors clustered by school in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Sources: Statistics Norway, and the Norwegian Directorate for Education and Training.

VII Conclusions

This paper has shown that the exam performance of Norwegian upper secondary school students is sensitive to exam scheduling. I have found that when students are randomly assigned to a longer exam period, they perform around 5-6% of a standard deviation better than their peers. The size of the effect is comparable to the effect of increasing the length of the school year by about the same number of days (Lavy, 2015). The effect is stronger in natural science and math subjects, and stronger for girls than for boys, and shows a very concave pattern. The strongest hypothesis explaining the effect is that regardless of how much time they are given, most students only use a limited number of days to prepare for exams, amounting to somewhere between one and two weeks, as noted above. When the preparation time they are given is less than this minimum, they score worse on their exams.

I have also shown that variations in exam scores due to exam scheduling are large enough to impact longer-run outcomes. Specifically, I have demonstrated that female students who are randomly assigned to exam schedules that have more exams with a relatively long preparation period are more likely to enroll in university, and less likely to drop out before their second year. Girls also appear more likely to enroll in programs with higher entry requirements, while boys are somewhat more likely to enroll in STEM programs. The fact that girls are more affected overall in the longer-run mirrors the finding that they are more affected in the short-run. There are two likely mechanisms behind the longer-run effects. One is that students who gain higher application scores due to the preparation time given improve their probability of being accepted into university and enrolling in programs that they prefer more. The second mechanism is that students who perform relatively well due to long preparation periods might change their underlying preferences and sort towards programs closer to the subjects in which they did well.

The longer-run effects add to the literature on how institutional factors unrelated to students' abilities, such as school start age, can affect their exam scores at a critical juncture in their lives, potentially shaping future opportunities. Understanding how institutional factors affect students' exam scores and, in turn, longer-run outcomes are important when new designs are implemented into the current school systems. In particular, it implies that putting too much weight on exam performance when sorting students into higher education might cause some mismatch between students and education paths. A potential caveat in the longer-run findings is that there is some evidence that balancing is slightly imperfect, as discussed above. This should be considered when interpreting results.

The strong underlying heterogeneities in terms of gender, student skill level, and subject type uncovered in this paper are worth underlining. In the gender dimension, female students are far more affected than boys, in both the short and the longer-run. In the short-run this is reflected by both high- and low-skilled female students having point estimates up to twice those of low-skilled male students. High-skilled male student are unaffected by preparation time. As pointed out above, the reason behind these hetero-

geneities is unclear, but provides new insights into the literature on the gender gap in school performance.

The heterogeneity between the genders is also evident when subject types are considered separately. Both female and male students improve their performance in sciences and math, while only female students are able to improve their language exam scores with more preparation time. Notably, male students have a higher marginal return relative to females on the longest preparation periods in sciences and math. This could reflect that male and female students decide to prepare for different exams on the basis of interest, or that they perceive the marginal returns differently. However, it is not possible to identify the deeper causes of this result.

An important finding in this paper is the strong concavity in the returns to preparation time. While it is not possible to precisely discern the underlying mechanisms, as I have not observed how students spend their time in the preparation period, there are some clues as to what goes on. One hypothesis is that students are mentally fatigued and therefore unable to improve their performance further (Ackerman & Kanfer, 2009). In the current study, there is little evidence to support this, as the effects are quite similar whether students have multiple exams or only one. It also seems unlikely that fatigue should play such a major role when the exam period is as long as that observed in the data. Moreover, the heterogeneous effects across genders and skill levels are hard to explain unless there are strong differences across these dimensions in how easily students become fatigued. A similar argument was put forward in Pope and Fillmore (2015).

An alternative hypothesis is that students "cram" during their preparation time. However, as the effect of 17-25 days of preparation ahead of a single exam is remarkably similar to the effect of the same number of days ahead of an exam when there are multiple exams in the exam period (Column (4) Table 4), cramming does not appear to be driving the effect.

A final hypothesis is that students only study for a limited number of days ahead of exams. The results from Column (4) Table 4 suggest that this limit is around 9-12 days, or one to two weeks. This hypothesis can also explain the difference found between the

genders, with girls having a positive return to preparation time for the entire period and boys having no marginal return beyond the 9-12 days. Fortin et al. (2015) and Jacob (2002) have both found that girls' relatively strong performance compared to boys is attributable to factors such as aspirations and non-cognitive skills that are not measured by an observed skill. Such differences could indicate that girls spend more time preparing and thereby perform better on exams compared to boys with a similar track record.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*(2), 163.

Bensnes, S. (2016). You sneeze, you lose. The impact of pollen exposure on cognitive performance during high-stakes high school exams. *Journal of Health Economics*, *49*, 1-13.

Black, S. E., Devereux, P. J., & Salvanes, K. G. (2011). Too young to leave the nest? the effects of school starting age. *The Review of Economics and Statistics*, *93*(2), 455–467.

Carrell, S. E., Maghakian, T., & West, J. E. (2011). A's from zzzz's? the causal effect of school start time on the academic achievement of adolescents. *American Economic Journal: Economic Policy*, *3*(3), 62–81.

Dills, A. K., & Hernandez-Julian, R. (2008). Course scheduling and academic performance. *Economics of Education Review*, *27*(6), 646–654.

Edwards, F. (2012). Early to rise? the effect of daily start times on academic performance. *Economics of Education Review*, *31*(6), 970 - 983.

Eren, O., & Millimet, D. L. (2008). Time to learn? The organizational structure of schools and student achievement. In *The economics of education and training* (pp. 47–78). Springer.

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

Fortin, N. M., Oreopoulos, P., & Phipps, S. (2015). Leaving boys behind gender disparities in high academic achievement. *Journal of Human Resources*, *50*(3), 549–579.

Fredriksson, P., & Öckert, B. (2014). Life-cycle effects of age at school start. *The Economic Journal*, *124*(579), 977–1004.

Jacob, B. A. (2002). Where the boys aren't: Non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education review*, *21*(6), 589–598.

Kirkeboen, L. J., Leuven, E., & Mogstad, M. (2016). Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, *131*(3), 1057–1111.

Landersø, R., Nielsen, H. S., & Simonsen, M. (2017). School starting age and the crime-age profile. *The Economic Journal*, *127*(602), 1096–1118.

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal*, *125*(588), F397–F424.

Lavy, V., Ebenstein, A., & Roth, S. (2015). The Long Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution. *American Economic Journal: Applied Economics*, *8*(4), 36–65.

Norwegian Directorate for Education and Training. (2009). *Trekkordning ved eksamen i kunnskapsløftet. The random selection arrangement for examinations under the Knowledge Promotion Reform.*

Pei, Z., Pischke, J.-S., & Schwandt, H. (2018). Poorly measured confounders are more useful on the left than on the right. *Journal of Business & Economic Statistics*(just-accepted), 1–34.

Pope, D. G., & Fillmore, I. (2015). The impact of time between cognitive tasks on performance: Evidence from advanced placement exams. *Economics of Education Review*, *48*, 30–40.