Preparation time, exam scores, and tertiary education

How preparation time affects high-stakes exam scores, and higher education outcomes

WORKING PAPER

Simon Søbstad Bensnes [*]

Department of Economics, Norwegian University of Science and Technology

October 12, 2016

Abstract

In many cases students are allowed some time in preparation for high-stakes tests, and the length and use of this time is likely to affect actual test scores in a similar way as school year length. However, today no empirical evidence exist on the effect of preparation time. This paper adds to the literature by using what is in effect random variation in students' preparation time prior to high-stakes exams. Explicitly, all Norwegian high school students are notified which exams each student will take at a precise date and time. Because students are randomly assigned to take exams in several different subjects, there is a random within-student variation in the length of preparation time varying between 5 and 25 days in the data. Using this randomization and administrative student level data, the study finds that 5 extra days of preparation time increases exams scores between 5.7 and 6.7% of a standard deviation. The effect differs somewhat between the genders, and also materializes strongly in longer-run outcomes indicating increased human capital. Finally, the paper uses the variation in preparation time to obtain an IV estimate of the effect of exam scores on longer-run outcomes.

JEL classification: I20 I21

Keywords: High school, test scores, preparation time.

[*] I am grateful for comments on earlier drafts from Chris Van Klaveren, Bjarne Strøm, Torberg Falch, Ole Henning Nyhus, Magne Mogstad, Øyvind Kvanmo, and seminar participants at the workshop "Education, Skills, and Labor Market Outcomes" at the Norwegian University of Science and Technology June 2016, and "Evidence Based Economics Summer Meeting 2016" in Freising. All remaining errors are my own. Correspondence: simon.bensnes@ntnu.no.

1 Introduction

Economists have long scrutinized the role of various inputs in the education system. A so far overlooked factor that has a strong effect on test scores is the time students have to prepare prior to examinations. This paper provides the first evidence regarding preparation time, further extending the literature to a new area. The research presented here is not only relevant for increasing the understanding of how students are affected by preparation time, but it also has important policy implications. High-stakes examinations are commonly applied in many countries and school systems for student placement, including the Scholastic Aptitude Test (SAT) in the US. This paper shows that the time students have to prepare ahead of these tests is an important factor for the exam outcome. Further, a popular opinion is that private prep courses ahead of these exams give an unfair advantage to the privileged students with access to them. This paper also shed some light on the potential effectiveness of such courses, and can therefore help inform the debate.

Additionally, the findings presented herein show that students who are identical in terms of background and skills, can achieve different test scores merely because they are given a different number of days to prepare for their exam. These effects are shown to be large enough that the future opportunities of these students are significantly altered, reducing the matching quality between individuals and education, and potentially jobs. The paper also show that the educational outcomes in relatively old students van be strongly affected by a relatively mild intervention. Thus, policy makers can use carefully applied variations in preparation time to reduce gaps in the performance between students from advantaged and disadvantaged backgrounds even after early childhood.

Past research has shown that school year length and instruction time both have strong effects on students' test scores. While one might expect certain similarities in the effects of preparation time and instruction time, empirical evidence is needed as there are at least two reasons to expect some difference in the effects. First, instruction time is accumulated over a longer time horizon than preparation time and might therefore differ in its effects depending on the education production function. Second, general instruction from a teacher might have a smaller or larger direct effect on test scores than targeted self-study. The policy implications are also not identical. A significant causal effect of preparation time on exam scores is a strong argument for preparation periods to be consistent across students: In the absence of homogenous preparation periods, high-stakes tests can be more informative of the preparation time rather than students' abilities in the more extreme cases[1].

Identification is achieved by using the unique national exam system in the Norwegian high school system combined with highly detailed administrative data. Specifically, students in the Norwegian high school system are required to take several exams and are notified when exams will be held and which exams they will take on a specific date each year. Because exams in different subjects are held on different dates and because students take multiple exams, it is possible to use random within-student variation in preparation time to assess the effect of increased preparation time. The results suggests that increasing preparation time from 5-8 days to 9-12 days increases test scores by 5.9% of a standard deviation. The effect is relatively large considering that the increase reflects about 4 days extra time spent in preparation for an exam. Considering the relatively large effects uncovered in this paper, I also consider heterogenous effects across students, and show how the effect depends on students characteristics.

The estimated effects on exam scores also persist in longer-run outcomes. To assess the impact of preparation time on students' achievements in the longer-run I follow them into tertiary education, where I find that preparation time not only affects enrollment, but also students' persistence in higher education: A longer average preparation period ahead of exams reduces the probability that students drop out within the second year.

In addition to evaluate the direct effect of preparation time on exam scores and longer-run outcomes, this paper also provides IV-estimates of the effect of exam scores on longer-run outcomes. This is made possible by using the random variations in preparation time as an instrument for exam scores. The IV-estimates are highly significant and large compared to the estimates when exam scores are not instrumented.

During the preparation period studied herein, students mostly follow normal instruction. Some additional guidance can be provided from teachers by their discretion. This implies that students are generally in charge of managing their own preparation for their exams. This study

[1] Preparation time and preparation period is used interchangeably here.

therefore also sheds some light on how students cope with time management prior to large tests.

The paper is organized as follows. Section 2 introduces related literature. The institutional setting is presented in Section 3, while the empirical strategy is presented in Section 4 with the data. Main results are presented in Section 5 with robustness checks in Section 6. Section 7 shows how results vary across students and subjects. Longer-run results are presented in Section 8, while Section 9 show IV-estimates of the effect of exam scores on longer-run outcomes. Last, Section 10 concludes.

2 Related literature

While there are no previous contributions on the effects of preparation time ahead of exams known to the author, this paper is related to the growing literature on instruction time. Generally one might expect that students's performance is increasing in instruction time. Consequently, several school districts in the U.S. and authorities in other countries have been increasing instruction time in order to increase students' academic skills (Rivkin & Schiman, 2015). At the same time as policy makers have increased instruction time, there has been a growing amount of economic research aiming at determining the causal effect of instruction time on academic performance.

In an early contribution Card and Krueger (1992) find that students who grew up in American states with longer term lengths tended to have a higher return to education when comparing across states. When they use wihtin-state variation effects vanishes. However, the identification strategy in Card and Krueger (1992) is unable to take into account a series of potential endogeneities. As the authors point out, it could be the case that states term length might be correlated with other unobservable measures of school quality such as teacher quality.

The contributions to the literature have arrived with increased frequency the last 15 years. Wößmann (2003) uses TIMSS data and compare several cross-country institutional differences, including instruction, time across countries, and finds that an increase in the school year length of 1 standard deviation, or around 6 days, increases test scores in math by around 0.025 standard deviations[2]. Pischke (2007) uses a national reform in West-Germany that decreased school

[2] TIMSS is an acronym for Trends in Mathematics and Science Study. My calculation based on numbers reported

year length and finds that it lead to increased grade repetition and reduced enrollment in upper secondary schools, although there are no identifiable longer-run effects on wages nor employment.

In the past decade since Pischke (2007) researchers have successfully utilized several exogenous variations in school year and instruction time. Marcotte and Hemelt (2008) employs school closing days resulting from extreme weather conditions as a source of variation in school year length in Maryland. They find that 5 extra days of unscheduled school closings increased the number of third grades with satisfactory test scores by 3% with smaller effects for older children. Using a similar approach Hansen (2011) finds that 5 extra days of schooling yields an increase of around 0.05-0.15 standard deviations in test scores among students in Maryland. The results are similar when the source of variation is the moving of the examination day across school years, and similar to the effects found by Wößmann (2003).

Another approach employed in two recent studies is to use student panel data with multiple observations of both test scores and instruction time. This approach was first introduced by Lavy (2015) who finds that one hour extra instruction time per week increases student test scores by 0.06 of a standard deviations on average using within-student variation the PISA data. This first study was followed up by Rivkin and Schiman (2015) who employed the same method to a newer set of PISA data, finding relatively similar results. As Rivkin and Schiman (2015) points out, one might not necessarily expect that increasing instruction time increases test score by much as the effect is likely to depend on the quality of the instruction. Rivkin and Schiman (2015) shows that the effect of instruction time appears to interact with a classroom quality index which is based on measures of the quality of student-teacher interactions and the social environment in the classroom.

It is clearly of interest to understand how the time students spend in school interact or depend on the quality of teaching. As such this paper offers a new contribution that shed more light on this interplay. Specifically, students in Norwegian high schools spend some of their time preparing for exams without direct supervision from a teacher. The marginal return to preparation time estimated here is therefore a weighted average of the marginal return to self-study and the marginal return to instruction time prior to the exam. If the effect of preparation

by Wößmann (2003).

time identified in this paper resembles the effect of instruction time, students are about as well off doing self-study as spending time in the classroom when preparing for an exam, suggesting that in the specific case of preparing for high-stakes exams, the marginal return to self-study is similar to the marginal return to instruction time.

Another finding in Rivkin and Schiman (2015) is that the marginal return to instruction time is diminishing. A finding that is supported by the findings herein. The policy implication of this seemingly robust finding is that prior to increasing instruction time it is important to take into account the existing levels of instruction time. In cases where instruction time or preparation time is already at relatively high levels, it might be more cost effective to increase the quality of teaching.

Interestingly, previous research has shown that there appears to be some heterogeneity in which students benefit from extra instruction time. Eren and Millimet (2008) concludes that students in the upper part of the test score distribution benefit from a shorter school year, while students in the bottom half benefit from a longer school year using US data. The fact that previous studies have found heterogeneities in the effects of school year length motivates an additional exploration in regards to the effects of preparation time. The rich data available to this study is ideal for this purpose and will be executed in section 7.

3 Institutional background and exam system

3.1 School system

This section builds on Bensnes (2016). The Norwegian school system consists of ten years of compulsory schooling, starting the year students turn six, and an elective high school education. It is not possible to fail a class in mandatory schooling, implying that grade repetition is practically non-existent and that nearly all students graduate from mandatory schooling at age 16. In this study, all students who did not finish mandatory schooling at the normal age are dropped from the sample. More than 95% of students choose to enroll in elective high school education the fall after graduating from mandatory schooling. High school is tracked and consists of 12 tracks that can be grouped into two broad categories: the academic tracks and the vocational tracks. Exams in the vocational tracks often have a practical part and it is not possible to identify which exams include this partitioning. For this reason, only students in the academic tracks are included in the analysis. The three academic tracks are: Dance, drama and music; sports; and specialization in general studies. The academic tracks lasts three years and grants graduating students eligibility to apply to higher education regardless of which academic track the student completes.

Of the students choosing to enroll in high school the year they turn 16, roughly 50% opt for an academic track. In the empirical analysis, only students who enroll in the academic tracks are included. Students who enroll in a vocational track at 16 then switch to an academic track are therefore dropped. In the academic tracks, exams are either oral or written and can be identified as such. For comparability to other studies, and because oral examinations are likely to be influenced by non-academic characteristics for the student, only written exams are used in the analysis. Additionally, the contents of oral examinations are determined at the school level. Thus, the difficulty of these exams are not the same across schools and could adjusted to the abilities of the students. The curriculum in each subject is comprehensive and standardized at the national level, and written exams are created for each subject nationally for each school year. This ensures that all students who take a an exam in a specific year and subject are given an exam of the same difficulty and results are directly comparable.

3.2 Exam system

Students in the academic track in the Norwegian high school system are all required to take a written exam in standard Norwegian language. Besides this exam students are further required to take randomly drawn written exams in their second and third years. In the second year all students are required to take one exam that is either written, oral, or practical. In the third year the number and forms of exams students are required to take varies between sub-specializations in the academic tracks. A detailed description of this system is provided in the Appendix. The majority of students are required to take two randomly drawn written exams and one oral exam, while the remaining students are drawn to take two to three exams that are either written or oral. Finally, around 20% of students are also randomly drawn to take an additional exam in the first year of high school, which can be in any of the three forms. Exam scores 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 are anonymized and are graded for all students in the subject-class by two external evaluators who teach the same subject at a different high school.

Each year the The Norwegian Directorate for Education and Training set up a schedule for when written exams in each subject are to be held[3]. This schedule is distributed to schools along with an announcement date. The counties are responsible for assigning students to exams, and are required to ensure that the assignment is random (The Norwegian Directorate for Education and Training, 2009). At the announcement date schools are to notify students if and when they are to be examined in the various subjects. Note that schools are unable to alter the preparation time for students as they are subject to the hold exams at the specified dates and the announcement date is set. In the empirical period the announcement date is between 5 and 25 days prior to the examinations with an average of 13.5 days. The distribution of exam observations across the preparation days are presented in Figure 1. In a few subjects there are only a handful of students assigned to take an exam across the five years observed in the data. This is particularly common in subjects aimed at language training in the mother tongue of minority students. Exams in courses with fewer than 5 students are excluded from the sample. This excludes 11 subjects and 20 observations from the sample.

After students are notified which exams they will have to take and at which dates exams will be held, they generally follow normal instruction, although schools and teachers differ in how much time is devoted to exam preparation during school hours. Some schools allow students specific study days, or extra classes, while others do not. At all schools teachers are available for guidance during the preparation period. If a teacher has students who are taking an exam in a subject she teaches, she might offer extra classes or focus teaching on exam relevant material to help students prepare. The amount of extra instruction the teacher offers will generally depend on the teacher as there is no national guideline in this regard. As such, the preparation period also includes some instruction from a teacher and some self-study. This will be more closely addressed in the following sections.

[3] The dates and content for oral and practical exams are set at the local level.

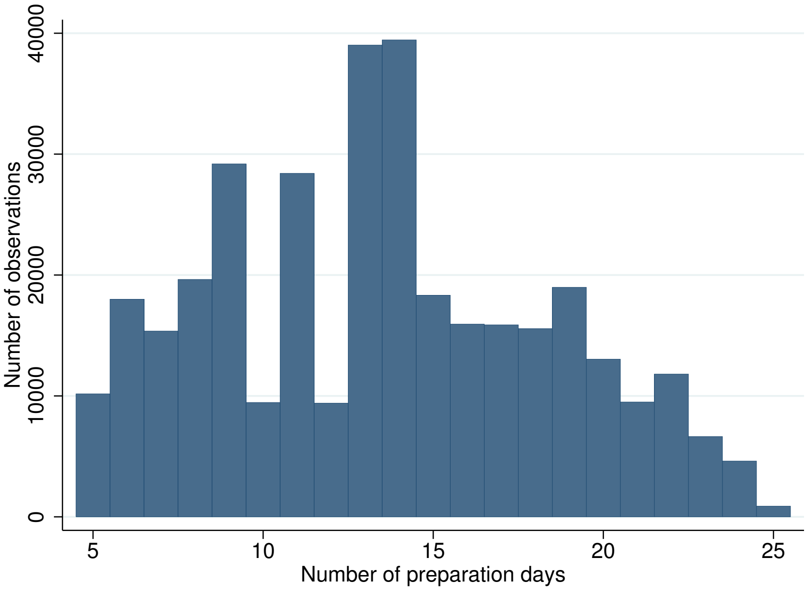

Figure 1: Distribution of exam observations across preparation time

The exams in the Norwegian high school system are high-stake for three reasons. First, if a student fails an exam they are required to re-take the exam if it is in Norwegian languages. If it is a randomly drawn exam they will be redrawn to take a new exam in either the same or a different subject the next semester. Second, in order to graduate students have to pass all exams. Third, students compete for placement in higher education based on the average of their subject grades and their exam grades (Kirkebøen, Leuven, & Mogstad, 2016)[4]. The fact that tests are high stake suggests that students will utilize their assigned preparation time to the best of their abilities.

4 Data and empirical strategy

4.1 Dependent variable

The dependent variable in the analysis will be the grade awarded on written exams for students enrolled in the academic track. Individual grade records are collected from register data made

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

available by Statistics Norway, covering the period including 2008 through 2012[5]. The data contains identifiers of the subject for which the grade is awarded. The exam grades are merged to the yearly list of exam dates from the Ministry of Education.

In the data, the number of exams each student takes in high school varies somewhat due to five factors. First, some students are drawn to take more written exams than others as explained above. Second, some students in the sample drop out of high school and therefore do not take the number of exams required to graduate. Third, the number of exams taken is not the same for all students who graduate due to exemptions. Fourth, a minority of students are randomly assigned to take a written exam the first year in high school. Finally, some students retake exams to achieve a passing grade or a higher score. Students who retake exams in order to improve their grade are marked as such; results from second attempts on any exam are also dropped from the sample. Students who are sick or otherwise unable to show up for the examination in a subject they are picked to take, and can provide a medical certificate from a medical doctor stx2ating the reason for their absence, are required to take a make-up exam the following fall semester. The total number of make-up exams constitute less than 1% of all exams (Bensnes, 2016). It is important to note that although retake exams are identified in the data, make-up exams are not and results from make-up exams are thus not distinguishable from results from ordinary exams. The final dataset includes students who took two to five exams, with the median student taking three written exams. This is consistent with the normal number of exams students take in high school given student attrition.

Because students follow different subjects, all students within a school are generally not tested on the same days. The assignment of students to exams, and hence preparation time, is random given the subject composition of each student. That is, given the the subjects the student has chosen as elective subjects, the expected preparation time will be random.

[5] Prior to 2008 The Norwegian Directorate for Education and Training did not distribute a common announcement date. Rather, schools were to announce which students were to be examined “At least 48 hours prior to the exam”. A rule that was interpreted differently across schools.

4.2 Other variables

In addition to exam observations, the data also includes information on students' teacher assessed subject grades and several background characteristics. In the main analysis the student background variables will not be used as they are absorbed in the student fixed effects, but they allow for some heterogeneity analysis and are consequently introduced here. The student background characteristics include information on parental education, parental labor market status, and the students gender and immigration status. Descriptive statistics for all variables are presented in Table 1. The upper part of the table report variables on the student level, while the lower part report variables measured on the exam level. As can be seen from the Table 1 a little more than half of the students are female, which is due to the fact that more boys than girls follow the vocational track in high school. Further, the teacher assessed grade is a little higher than the exam grades on average. Exam observations that cannot be linked to a teacher assessed grade are dropped from the sample. This is because if a student does not receive a teacher assessed grade in a subject, for example due to a high absence rate, the exam grade is invalid, which in turn gives little incentive to perform. In terms of ethnicity about 7% of the students are either first or second generation immigrants. Regarding subject composition the majority of the exams are taken in various language subjects. This is largely due to the requirement that all students in the academic track are required to take an exam in Norwegian. Around a quarter of all exams are taken in subjects that are characterized as mathematics or natural sciences. The remaining exams are taken in social sciences other subjects[6]. The parental income reported in the table comprise all pensionable income earned by the students' parents reported to the tax authorities the year students are 15 years old, i.e. prior to high school enrollment.

[6] The category "other subjects" includes subjects like accounting, marketing, and history

Table 1: Summary statistics

Variable	Mean	(Std. Dev.)	Min.	Max.	N
Student level					
Preparation time	13.34	(2.709)	5	25	105,024
Average exam score (1-6)	3.247	(0.929)	1	6	105,024
Average teacher assessed grade (1-6)	3.812	(0.956)	1	6	105,024
Number of exams for students	3.424	(0.918)	1	5	105,024
Graduated high school within the normal time	0.867	(0.34)	0	1	105,024
Enroll in tertiary education	0.831	(0.375)	0	1	105,024
Drop out of tertiary education	0.127	(0.333)	0	1	105,024
Enroll in STEM program	0.154	(0.361)	0	1	105,024
First generation immigrant	0.032	(0.176)	0	1	105,024
Second generation immigrant	0.035	(0.184)	0	1	105,024
Girl	0.556	(0.497)	0	1	105,024
GPA lower secondary	4.534	(0.569)	1.786	6	105,024
Father's education					
Mandatory schooling	0.147	(0.354)	0	1	105,024
High school	0.431	(0.495)	0	1	105,024
Bachelor degree	0.269	(0.443)	0	1	105,024
Master or PhD	0.153	(0.36)	0	1	105,024
Mother's education					
Mandatory schooling	0.148	(0.355)	0	1	105,024
High school	0.371	(0.483)	0	1	105,024
Bachelor degree	0.401	(0.49)	0	1 105,024	
Master or PhD	0.081	(0.273)	0	1	105,024
Parental labor market status					
Both parents working at student age 15	0.794	(0.404)	0	1	105,024
One parent working at student age 15	0.182	(0.386)	0	1	105,024
Average parental income at student age 15 (NOK)	445,266.62	(615,882.696)	700	168,725,952	105,024
Exam level					
Preparation period	13.491	(4.876)	5	25	349,203
Exam score	3.282	(1.145)	1	6	349,203
Teacher assessed grade	3.851	(1.105)	1	6	349,203
Subject type					
Norwegian	0.418	(0.493)	0	1	349,203
Natural sciences and mathematics	0.241	(0.428)	0	1	349,203
Languages, including Norwegian	0.578	(0.494)	0	1	349,203
Social sciences	0.181	(0.385)	0	1	349,203

All variables measured at the student level. Preparation period is the number of days between the announcement date and the examination. Parental education is measured as the highest attained degree of either parent. Income is measured in nominal terms. The subject type "Social sciences" includes the humanities and other subjects such as business economics. The three subject types in the end of the table are mutually exclusive. Sources: Statistics Norway, and The Norwegian Directorate for Education and Training.

4.3 Identification strategy

The identification strategy can be presented as in equation (1). The dependent variable is the exam score in subject c for student i taken in the year y. The exam score is standardized by subject to have a zero mean and a standard deviation of 1[7]. C is a constant, and η_i, γ_c and θ_y

[7] Standardization is done to ease interpretation of coefficients. Results are not sensitive to the transformation of exam scores.

are student, subject and year fixed effects respectively. X_{iyc} is a vector of student observables that vary between exams. It includes the number of exams the students take the same year and the number of the exam within the year in chronological order. ϵ_{iyc} is a random idiosyncratic error. The coefficients of interest are β_1 through β_3, which measures the effect of extra exam preparation time. In the main specification functional form will be given by three dummies for the preparation time being in the second to forth quartile, with the first quartile as the reference category[8]. A non-linear functional form is preferred because it allows for a more flexible relationship to be estimated. This definition will be challenged in Section 6 which shows that the results are robust to alternative definitions.

$$\begin{aligned}\text{Exam score}_{iyc} &= C + \beta_1 \text{Preparation time second quartile}_{iyc} + \beta_2 \text{Preparation time third quartile}_{iyc} \\ &\quad + \beta_3 \text{Preparation time forth quartile}_{iyc} + \eta_i + \gamma_c + \theta_y + X_{iyc} + \epsilon_{iyc} \qquad (1)\end{aligned}$$

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

Subject fixed effects are necessary to avoid bias. To see this consider a case without subject fixed effects. If the exam in a subject is systematically placed towards the end of the exam period every year, the preparation time for this subject will be longer than for the average subject all of the five years in the data. If exams in subjects with a higher return to preparation time also have longer preparation time, the estimated effect of preparation time will be upward biased. Thus, including subject fixed effects removes this potential bias from the estimation.

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

The student fixed effects employed here enhances the credibility of identification strategy for several reasons. First, students are likely sort into the three academic tracks based on unobservable characteristics. The subsequent possible subject choices and the number and types of exams are partially dependent on the choice of track. Further, which subjects a student choose within a given track is also likely to be correlated with unobserved student characteristics. These characteristics include general cognitive and non-cognitive abilities, and other time invariant characteristics of the students. Because preparation time will depend on the subjects a student choose, and this will partially depend on the unobserved student characteristics, estimates might be biased in the absence of student fixed effects.

A second, related issue is that exams in some types of subjects on average have shorter preparation period than other subjects. For example, exams subjects that can be classified as either science or math subjects have on average 13 days of preparation time, whereas subjects which can be classified as being in the social science and other category have on average 16 days of preparation time. Because students sort into subjects in part based on unobservables, estimates might suffer from bias when student fixed effects are not employed, as student fixed effects absorb students' choice of subjects and thereby which courses they might be assigned to be examined in.

Third, there might be variation both between and within schools in terms of how much extra instruction time students are offered in their preparation time. Further, more able students might sort to schools which offer more and better extra instruction during the preparation time. School fixed effects would absorb the average differences across schools in this respect, but not differences within schools stemming from students being assigned to exams in different subjects taught by different teachers. Student fixed effects removes average differences in extra instruction time both between students within a school, and between students across schools.

A further benefit of the student fixed effects approach is that students might differ in how able they are in taking advantage of preparation time general. In the absence of student fixed effects this would not cause a bias, but could make estimates less precise. When student fixed effects are included, I only take advantage of within-student variation in preparation time across subjects, and remove all time-invariant differences in the marginal return to preparation time across students.

With fixed effects in three dimensions the remaining identifying variation stems from two sources: (i) Differences in preparation time within subjects across years given differences between students taking the exam in different years. While year fixed effects remove variations in preparation time that is common for all subjects between years and subject fixed effects removes average differences in preparation time between subjects that are common across years, subjects for which preparation time changes with a different number of days than the average between two years provide variation which allow for identification. Thus, even if all students were assigned to take exams in the exact same subjects, it would still be possible to identify the causal effect of preparation time when multiple years of data are observed and changes in preparation time across years are random. This variation remains also after student fixed effects are included. (ii) Students having multiple exams within a single year. Because students have multiple exams in their last year, it is possible to compare how they perform on exams with different amounts of preparation time given the average difference between subjects which are controlled for with subject fixed effects. However, to utilize this variation in the presence of course fixed effects it is necessary that the data spans more than one exam year. If the data only spanned one year it would not be possible to identify both course fixed effects and the effect of preparation time. Both of these sources of variation ultimately stems from the random assignment of students to exams and remain also after the inclusion of student fixed effects. Because all the remaining variation in preparation time stems from a random component of the exam system and allow for student fixed effects it is very well suited for a credible study of the effect of preparation time on exam scores.

Although student fixed effects remove an important source of potential bias, it is, unfortunately, not possible to test the assumption that preparation time is uncorrelated with student characteristics conditional on student fixed effects. However, it is possible to test whether preparation time is correlated with student characteristics conditional on subject fixed effects. To do this I regress preparation time against the student characteristics presented in section 4.3 and dummies for the exam year and the cohort of the student. The results are presented in Table 2. The first column in Table 2 does not include subject fixed effects. The three last rows in the table report the p-value for three F-test for joint significance of (i) the student background characateristics, (ii) the cohort dummies, (iii) the subject dummies. As evident from Column (1) the test clearly rejects the null hypothesis that the student background characteristics have no joint explanatory power. This underlines the need to include subject fixed effects in the main specification according to the arguments above. The test whether the cohort dummies have no explanatory power rejects the null hypothesis. This is not surprising as the cohort largely determines which years students will be examined in. Because the average preparation time changes across years, it also changes across cohorts. When subject fixed effects are introduced in Column (2) there is no strong evidence that preparation time is correlated with student background characteristics. Immigration status has some explanatory power, but the F-test clearly reject joint explanatory power and one would expect that at least 1 of the 12 background variables would be significant at the 10% level merely by chance. Note that both the year in which the exam is held and the cohort of a given student are strongly correlated with preparation time. However, this is not worrisome as these correlations are mechanical and easily controlled for in the main regressions.

As described above, one might be concerned that highly skilled students receive more preparation time because the subjects they choose are persistently placed towards the end of the examination period, giving them more preparation time. To attenuate this concern the estimates in Column (3) includes students' average teacher assessed subject grades from lower secondary education (GPA). Introducing this additional control does not drastically alter the p-value. As for the coefficient on GPA, it insignificant. This suggests that students who on average are of higher ability does not receive any longer or shorter preparation time than other students once their subject selection is controlled for. Although the results in Table 2 suggests that subject fixed effects are necessary to control for student sorting into subjects the baseline model will include student fixed effects as well in order to more accurately estimate the average effect of preparation time given student ability, and to control for unobserved student characteristics as explained above. In summary, the balance tests show support to the identifying assumption that student characteristics are uncorrelated with preparation time.

Before moving on to results, the handling of standard errors should be addressed. In the

Table 2: Testing the randomness of assigned preparation time (# days)

	(1)	(2)	(3)
Mother's highest educational level is high school	0.0284	0.0246	0.0252
	(0.0195)	(0.0160)	(0.0160)
Mother's highest educational level is bachelor degree	0.0264	0.0226	0.0238
	(0.0202)	(0.0171)	(0.0171)
Mother's highest educational level is master or PhD	0.0384	0.0111	0.0125
	(0.0285)	(0.0210)	(0.0208)
Father's highest educational level is high school	0.0186	0.00501	0.00539
	(0.0179)	(0.0147)	(0.0146)
Father's highest educational level is bachelor degree	0.0520***	0.00803	0.00890
	(0.0188)	(0.0158)	(0.0157)
Father's highest educational level is master or PhD	0.0567**	-0.00568	-0.00449
	(0.0239)	(0.0179)	(0.0179)
First generation immigrant	0.0491	0.0330	0.0318
	(0.0438)	(0.0330)	(0.0328)
Second generation immigrant	-0.0891*	-0.0682*	-0.0688*
	(0.0454)	(0.0353)	(0.0353)
Female	-0.0894***	0.00456	0.00617
	(0.0149)	(0.0112)	(0.0112)
Exactly 1 parent working	-0.0400	-0.00203	-0.00156
	(0.0440)	(0.0387)	(0.0385)
Both parents working	-0.0471	0.00285	0.00357
	(0.0439)	(0.0384)	(0.0381)
Parental income	1.31e-08*	1.67e-09	1.70e-09
	(6.89e-09)	(3.47e-09)	(3.46e-09)
Exam in year 2009	4.794***	2.719***	2.718***
	(0.0659)	(0.116)	(0.116)
Exam in year 2010	9.724***	5.739***	5.738***
	(0.0968)	(0.207)	(0.207)
Exam in year 2011	12.58***	6.167***	6.166***
	(0.148)	(0.299)	(0.299)
Exam in year 2012	14.53***	4.055***	4.053***
	(0.195)	(0.390)	(0.391)
GPA lower secondary			-0.00646
			(0.0117)
Constant	8.061***	5.390***	5.420***
	(0.123)	(0.0847)	(0.104)
Observations	349,203	349,203	349,203
Cohort FE	Yes	Yes	Yes
p-value joint test background effects	0	.546	.513
p-value joint test cohort effects	0	0	0
p-value joint test course effects	-	0	0

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 and third last rows report p-values for similar tests on the cohort and course fixed effects respectively. Standard errors clustered by high school in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Source: Statistics Norway, and The Norwegian Directorate for Education and Training.

main results standard errors are clustered on the school level. Clustering errors at the school level allows for any correlation structure within schools. In other words, this approach acknowledge that each exam-student observation within a given school adds less information on the effect of preparation time than a new exam-student observation from a new school. Reasons why this could be the case include, but are not limited to, differences across school in the sorting of students, subjects and tracks offered, teacher quality, peer-effects, and the amount of teacher assistance during the preparation time.

An alternative approach would be to cluster standard errors at the subject level. This approach suggests that more observations within a given subject adds less new information than more observations from a subject not previously included in the data. One might expect this to be the case if students sort into various subjects partially based on common unobservable characteristics. However, subject fixed effects removes at least some of this concern (Cameron & Miller, 2015). Yet, it might be considered desirable to cluster the standard errors on subjects. A problem with this approach in the current setting is the number of subjects observed in the data. In the main specifications there are 82 different subjects, while there are more than 350 schools. Cameron and Miller (2015) argues that in the case of with clusters of equal sizes one needs around 50 clusters, and when clusters are of unequal sizes the number of clusters needed is even higher. The course clusters are of very unequal sizes, and therefore it is unclear whether it is desirable to cluster on the subject level. A final argument to cluster on the school level rather than the subject level is that students are randomly assigned to take exams in different subjects. Through this randomization the intra-cluster correlation in subjects will fall, while the intra-cluster correlation in schools remain unaffected. Considering the issues raised here in combination I choose to cluster standard errors on schools here, and report results when errors are clustered by subject and two-way clustering with schools and subjects the appendix table A1.

5 Main results

Table 3 shows the main results when estimating equation (1) with various number of controls. All estimations include year fixed effects and standard errors are clustered on the school level. Column (1) shows the most parsimonious estimation when only the number of days of preparation time and year fixed effects are included. The estimates indicates that there is a positive return to preparation time, but when preparation time is in the fourth quartile the estimated effect is much smaller than in the preceding quartiles. However, the model estimated in Column (1) does not take into account the potential problems with some exams sorting towards the beginning or the end of the preparation period. In particular, Table 2 shows that a model without subject fixed effects is likely to be biased.

Column (2) expands the model to include subject fixed effects. The estimated results show that increasing preparation time beyond the second quartile has litte if any effect on exam scores. The results actually indicate that students who have more preparation time than the second quartile perform worse than students in the second quartile. This effect is somewhat puzzling, but could arise if students with relatively high ability sort into subjects with on average shorter preparation periods. This does not seem unlikely as science and mathematics exams on average have shorter preparation periods.

Column (3) expands the model further by including the number of exams a student takes the same year in total and the number of taken so far the same year. These additional variables take into account that students must devise their time in preparing for several exams during the same exam period. Because the alternative cost of increasing the time spent preparing for one exam is reducing the time spent preparing for another exam it is relevant to include measures of how many exams the student will have to prepare for. Results are only slightly affected by including these controls, suggesting that students are adept in devising their preparation time between exams.

Moving on, Column (4) includes student fixed effects, only utilizing within-student variation in preparation time. The estimates are very similar, but more credible as unobserved student characteristics are absorbed. Because the model estimated in Column (4) nets out more unobservables than the model applied in Column (3), it will be referred to as the baseline model from here on. As can be seen from the table the estimated effect appears to be non-linear. Increasing the preparation time from 5 to 8 days to 9 to 12 days, that is the first to the second quartile in the distribution of preparation time, increases the exam score by 5.9% of a standard deviation, which is practically the same as moving from the first to the third quartile. Even the difference between the second and the fourth quartile is tiny and statistically indistinguishable from one another. An effect like this could arise if there is a limit to how much students are able to prepare for their exams. In other words, if the marginal return to self-study decreases sharply after around 10 days, one will one expect to see any sizable effect of increased preparation time beyond that point.

Comparing the results in Column (4) to previous research the estimates are remarkably similar. For example, Hansen (2011) reports that 5 extra days of schooling yields a test score increase of between 0.05 and 0.15 standard deviations, while Lavy (2015) finds that increasing instruction time by one hour per week increased test score by about 0.06% of a standard deviation. With a 40 week school year this is equivalent of 5 days of 8 hours study. Thus, the results reported here falls in line with previous studies on the effect of instruction time, suggesting that 5 days of preparation time has about the same impact as one hour extra instruction throughout a year, or increasing school year length by 5 days. However, the marginal return appears to be sharply decreasing. It is also worth noting that the comparison to Lavy (2015) assumes that students have 8 hours of instruction every day in the subject they are to be examined in. As students generally follow normal instruction, they only have a few hours of instruction per week in each subject. Consequently, the estimates reported in Table 3 suggests that (i) students spend a significant amount of time in self-study after normal school hours; or (ii) they might receive very effective extra instruction hours; or (iii) the productivity of the normal instruction rises sharply because students are more motivated and exert more effort closer to the examination; or a combination of these factors. Regardless, the estimates show that preparation time significantly affects exam results.

Column (5) takes the number of controls one step further and includes the teacher assessed subject grade as a control. Because the model estimated here includes student fixed effects, the subject grade is a measure of subject specific skills that are not absorbed by the student fixed effects. The teacher assessed grade has a very high predictive power, but does not alter the point estimates of preparation time. If anything the effect is more precise when the teacher assessed grade is included as a control. Unfortunately, it is not entirely clear if the teacher assessed grade is set prior to the date of the exams. If the teacher assessed grade is set after the exam is held, but still prior to the publication of exam scores, students who perceive that they performed poorly might influence the teacher's assessment and thereby causing a two-way

Table 3: Main results

	(1)	(2)	(3)	(4)	(5)	(6)
9-12 days of preparation time	0.0842***	0.0695***	0.0681***	0.0585***	0.0570***	0.0534***
	(0.0101)	(0.0117)	(0.0122)	(0.0136)	(0.0134)	(0.0196)
13-16 days of preparation time	0.0786***	0.0535***	0.0524***	0.0567***	0.0617***	0.0624***
	(0.0105)	(0.0110)	(0.0134)	(0.0144)	(0.0142)	(0.0230)
17-25 days of preparation time	0.0391***	0.0493***	0.0580***	0.0667***	0.0719***	0.0748***
	(0.00992)	(0.0112)	(0.0159)	(0.0178)	(0.0171)	(0.0229)
Standardized teacher assessed grade					0.314***	
					(0.00456)	
Observations	349,203	349,203	349,203	349,203	349,203	251,861
# Students	105,023	105,023	105,023	105,023	105,023	102,575
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Course FE	No	Yes	Yes	Yes	Yes	Yes
Total number of exams same year	No	No	Yes	Yes	Yes	Yes
Number of exams same taken year	No	No	Yes	Yes	Yes	Yes
Student FE	No	No	No	Yes	Yes	Yes

The outcome in all regressions is the exam grade standardized by course. Specifications in Columns (3)-(6) include dummies for the number of exams the student have the same exam period and the number of exams the student have already taken the same exam period. Specifications in Columns (4)-(6) include student fixed effects. The specification in Column (6) excludes all observations in the mandatory written exam in Norwegian. Standard errors clustered on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Sources: Statistics Norway, and The Norwegian Directorate for Education and Training.

causality. Because of this potential problem the teacher assessed grade will not be included in specifications in the baseline model.

In the Norwegian high school system students are randomly assigned to subjects to be examined in with one exception. This exception is written standard Norwegian (bokmål)[9]. The written Norwegian exam is mandatory and consists of students writing an essay. Although the exam is mandatory, students do not know which date they will be examined any earlier for Norwegian than for other subjects. For this reason the exams in standard Norwegian are included. Yet, one might argue that because students know they will have to take the exam in written Norwegian they start preparing for the exam even before the announcement date. Thus, Column (6) shows estimates when the mandatory exams in written Norwegian are removed from the estimation sample. Standard errors increase somewhat as the sample size falls by about one third, but point estimates remain very similar and highly significant. This suggests that the effect of preparation found in herein also applies to exams students know they will

[9] Students who have chosen the alternative non-standard form of written Norwegian as their main language have to be examined in that form. These students represent a clear minority. The estimation in Column (6) removes exams in the mandatory language regardless of which form it is.

have to take for up to three years prior to the examination (because students enroll three years prior to their exit exam in Norwegian), which further increases the policy relevance of the findings to systems where students are not randomly assigned to exams with a relatively short preparation period. Further, the results show that including the observations of mandatory exams in standard Norwegian does not appear to bias estimates, only increasing precision.

6 Robustness and further results

The main results presented in the previous section groups preparation time in intervals. Following the results, an immediate question is whether the definition of these variables are driving the results. In particular this is important as the results in Table 3 hints at a non-linear relationship between preparation time and exam scores. This section addresses this concern by presenting results where preparation time is included with alternative functional forms. Table 4 replicates Table 3 with the exceptions that Panel a use a linear functional form, and Panel b use a quadratic functional form. The results in Panel a suggests that the effect of preparation time is positive but insignificant in most specifications. When teacher grades are included, increasing preparation time by one day is estimated to increase test scores by almost 3% of a standard deviation.

The effects are much more precise in Panel b. In all specifications there is a clear concave relationship between preparation time and exam scores. This finding is in line with Rivkin and Schiman (2015), who finds diminishing returns to instruction time. The results in Column (4) shows that increasing the number of days preparation time from the average of the first quartile (6.5 days) to the average in the second quartile (10.5 days) increases the exam score by about 3.6% of a standard deviation. This effect is relatively similar to the estimated effect in Table 3. However, the results in Column (4) in Table 4 indicate that the marginal return to preparation time is negative after 17 days. This seems implausible and suggests that this model is less suited to estimate the causal effect of preparation time than the model used in Table 3.

It should be noted that others find no evidence of diminishing returns. Carlsson, Dahl, Öckert, and Rooth (2015) use random variation in the amount of schooling prior to cognitive tests on Swedish male conscripts. Although there is a robust effect on test scores from the amount of schooling, they find no evidence of non-linearities. There are, however, some differences in between the current study and Carlsson et al. (2015). First, they only have data on males, who appear to respond differently to more preparation time than females, as will be shown in section 7.1. Second, this study uses variation in an intense preparation period, whereas Carlsson et al. (2015) use variation in the amount of normal schooling prior to a test which is no directly related to school curricula. Considering these differences, it is not very surprising that findings differ somewhat.

Preparation time is partitioned into quartiles rather than a more fine tuned grouping because it is easier to present in a table format. Partitioning the data in quintiles or deciles give qualitatively the same results. These are presented as figures in Figure 2. The overall results are very similar to the baseline results presented in Column (4) in Table 3, I therefore conclude that the results are not sensitive to an alternative construction of the preparation time variable.

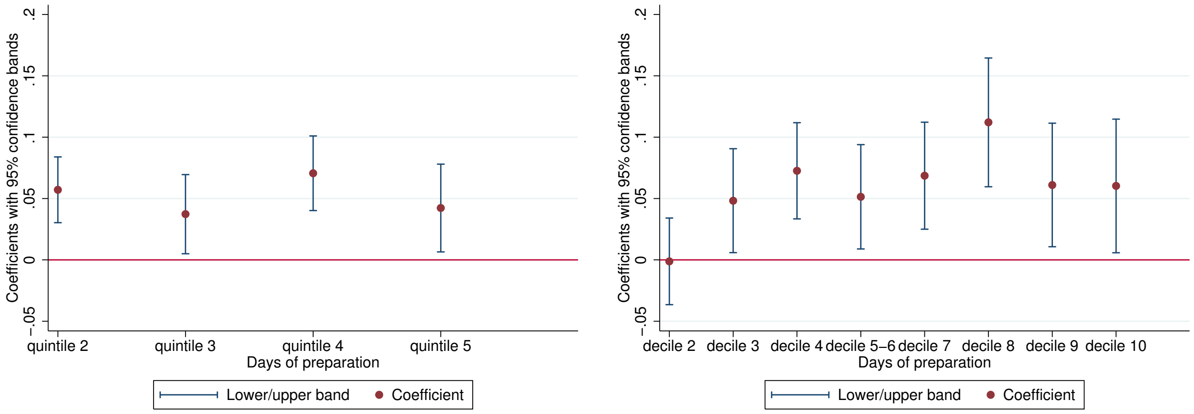

Figure 2: Coefficients with alternative grouping of preparation time. Sources: Statistics Norway, and The Norwegian Directorate for Education and Training.

6.1 Days since previous exam and non-school days

In the Norwegian high school system students most of their exams in their last school year. Thus, when students are informed which dates and subjects they are to be examined in, they will also have to divide their preparation time between the exams. A student who has three exams might

Table 4: Alternative functional forms

Panel a	(1)	(2)	(3)	(4)	(5)	(6)
Days of preparation time	-0.000844	0.00125	0.00193	0.00256*	0.00274**	0.00158
	(0.000682)	(0.000762)	(0.00119)	(0.00141)	(0.00136)	(0.00179)
Standardized teacher assessed grade					0.314***	
					(0.00456)	
Observations	349,203	349,203	349,203	349,203	349,203	251,861
# Students	105,023	105,023	105,023	105,023	105,023	102,575
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Course FE	No	Yes	Yes	Yes	Yes	Yes
Total number of exams same year	No	No	Yes	Yes	Yes	Yes
Number of exams taken same year	No	No	Yes	Yes	Yes	Yes
Student FE	No	No	No	Yes	Yes	Yes

Panel b	(1)	(2)	(3)	(4)	(5)	(6)
Days of preparation time	0.0312***	0.0206***	0.0187***	0.0172***	0.0185***	0.0174***
	(0.00362)	(0.00383)	(0.00432)	(0.00466)	(0.00463)	(0.00648)
Days of preparation time, squared	-0.00111***	-0.000670***	-0.000575***	-0.000494***	-0.000533***	-0.000507**
	(0.000125)	(0.000130)	(0.000143)	(0.000150)	(0.000150)	(0.000198)
Standardized teacher assessed grade					0.314***	
					(0.00456)	
Observations	349,203	349,203	349,203	349,203	349,203	251,861
# Students	105,023	105,023	105,023	105,023	105,023	102,575
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Course FE	No	Yes	Yes	Yes	Yes	Yes
Total number of exams same year	No	No	Yes	Yes	Yes	Yes
Number of exams taken same year	No	No	Yes	Yes	Yes	Yes
Student FE	No	No	No	Yes	Yes	Yes

The outcome in all regressions is the exam grade standardized by course. Specifications in Columns (3)-(6) include dummies for the number of exams the student have the same exam period and the number of exams the student have already taken the same exam period. Specifications in Columns (4)-(6) include student fixed effects. The specification in Column (6) excludes all observations in the mandatory written exam in Norwegian. Standard errors clustered on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Sources: Statistics Norway, and The Norwegian Directorate for Education and Training.

therefore choose to focus on the exam that comes first in the beginning of the examination period, and leave work on the remaining exams for the later part of the examination period. If this is the case, one could argue that the relevant measure of preparation time is the number of days since the previous exam rather than the total number of days between the exam and the announcement date. To address this potential problem I include a linear term for the number of days since the previous exam in Columns (1) of Table 5[10]. The estimation also include a dummy for whether the exam is the first one in the school year. The point estimates changes only slightly relative to the baseline estimates reported in Column (4) in Table 3, but remain statistically indistinguishable. This suggests that the model reported in Table 3 includes the relevant measure of preparation time, and that the students are relatively adept at dividing their preparation time between the exams.

Table 5: Robustness: Days since previous exam and non-school days

	(1)	(2)	(3)	(4)	(5)
9-12 days of preparation time	0.0564***	0.0638***	0.0538***	0.0569***	0.0586***
	(0.0152)	(0.0147)	(0.0140)	(0.0150)	(0.0136)
13-16 days of preparation time	0.0535***	0.0652***	0.0350*	0.0537***	0.0573***
	(0.0196)	(0.0184)	(0.0178)	(0.0195)	(0.0144)
17-25 days of preparation time	0.0630**	0.0801***	0.0442**	0.0624**	0.0680***
	(0.0260)	(0.0263)	(0.0205)	(0.0261)	(0.0186)
Observations	349,203	349,203	349,203	349,203	349,203
# Students	105,023	105,023	105,023	105,023	105,023
Year FE	Yes	Yes	Yes	Yes	Yes
Course FE	Yes	Yes	Yes	Yes	Yes
Total number of exams same year	Yes	Yes	Yes	Yes	Yes
Number of exams taken same year	Yes	Yes	Yes	Yes	Yes
Student FE	Yes	Yes	Yes	Yes	Yes
Control for number of days since previous exam	Yes	No	No	No	Yes
Control for number of weekends in prep. period	No	Yes	No	Yes	Yes
Control for number of holidays in prep. period	No	No	Yes	Yes	Yes

The outcome in all regressions is the exam grade standardized by course. Column (1) is the same as Column (3) in Table 3 with the exception that it controls linearly for the number of days since the previous exam the student had with a dummy if it is the first exam in the school year. Column (2) controls linearly for the number of weekends between the announcement date and the examination. Column (3) controls linearly for the number of national holidays between the announcement date and the examination. Column (4) controls linearly for the number of weekends and number of holidays between the announcement date and the examination. Standard errors clustered on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Sources: Statistics Norway, and The Norwegian Directorate for Education and Training.

The examination period, and hence the preparation period, in the Norwegian high school

[10] Using dummies yields qualitatively the same result.

system coincides with some national holidays[11]. On these holidays schools are closed and students cannot receive any direct instruction from their teachers, as is the case with weekends. The longer the preparation period is, the larger is the probability that it includes one or more of these holidays, and it will include more weekends. As the number of holidays in the preparation period is positively correlated with the length of the preparation period, estimates would be positively biased if holidays offer a particularly good opportunity for students to study. On the other hand, if students require instruction from a teacher to progress in their exam preparation, estimates would be negatively biased when the number of weekends and holidays are not controlled for. Following these arguments, Table 5 show results when linear controls for the number of weekends and holidays are included in the model: Columns (2) controls for the number of weekends; Column (3) controls for the number of holidays; Column (4) controls for both weekends and holidays; Column (5) controls for weekends, holidays, and number of days since the previous exam. When the number of weekends are controlled for estimates are slightly higher than baseline results in Column (4) of Table 3, and slightly lower when the number of holidays are controlled for. When they are jointly controlled for estimates are very similar to the baseline, which suggests that students are about as able to take advantage of self-study time at school as on off-days. Finally, Column (5) show that even when controlling for both holidays and the days since the previous exam results are almost identical to the baseline estimates. Overall, the results in Table 5 are supportive of the baseline model.

7 Heterogeneity

7.1 Heterogeneity by student characteristics

Taking the results so far as causal, one might ask if there is an underlying heterogeneity in which students benefit from increased preparation time. The results above show that otherwise similar students will receive different exam scores if their preparation time differ. However, the estimations might conceal an underlying heterogeneity as the effect is forced to be the same for all students. The effect of preparation time might therefore differ substantially between students

[11] In particular these are the constitution day (May 17), Ascension Day, the first and second day of Pentecost

with various backgrounds. Finding such differences would be of interest from a policy perspective as it would further substantiate the importance of homogenous preparation time for students. It can also shed light on how students with varying characteristics are able to take advantage of preparation time in a more general perspective. It should also be noted that previous studies exploring heterogenous effects of schooling on test scores are inconclusive. While Carlsson et al. (2015) find no differences in the effect of schooling across test takes with different backgrounds, Eren and Millimet (2008) find that the effect of schooling on test scores is dependent on initial skills. Further exploring the interaction between background characteristics and more time spent learning is therefore called for.

Table 6 reports results when the data is split in four different ways. The first two columns report estimation results when the sample is split according to the average teacher assessed grade in lower secondary schooling, the first column is for students who have a GPA that is higher than, or equal to, the median, and the second column shows estimates for the rest of the sample. Splitting the sample in this way allows for heterogenous effects across all the independent variables, which is a benefit relative to a model with interaction terms for the variables of interest. The reason the sample is split according to GPA is that one might expect that students who are on average performing better in school will be better able to further increase their relevant skills. One potential mechanism in play here could be that high performing students have some general skills that makes them better at managing their study time. On the other hand, well performing students might have a more limited opportunity to increase their skills: if they are already performing on the upper part of the scale it might be harder to further increase their skills as marginal returns to studying could be decreasing. Further, there is an upper limit in the grading system potentially causing a ceiling effect. The results show that the students who are performing in the upper half of all students, on average, have a lower return to extra preparation time. The difference between the two groups in the marginal return to preparation time diminishes when preparation time increases to the third and fourth quartiles, suggesting that students reach a level for which they have little opportunity to further increase their subject specific skills. Below each coefficient pair, t-values for a test of equality are reported in brackets. Although none of the coefficient pairs are significantly different at the 5% level, the overall pattern is interesting because it suggests that the production function differs across students with different skill levels, which again implies that how resources are spent in the education sector should partially depend the composition of students: Less skilled students benefit more form extra preparation time than more skilled students. The findings are also supportive of Eren and Millimet (2008) who finds that students in the lower part of the skill distribution benefit more from a longer school year, while students in the upper part of the distribution benefit from a shorter school year.

The claim is further substantiated when the sample is more finely partitioned. Figures 3 and 4 in the Appendix plot coefficients for the various subsamples when the sample is split according to GPA into quintiles and deciles, respectively. The overall results show that the best students have no return to preparation time, while the students around, and slightly above, the middle receive the most. In fact, point estimates for the top 10% best students are all negative, but insignificant. The absence of any significant effect for the strongest students could be due to a very low marginal return. If the strongest students work hard throughout the year, they might already master the subjects they are tested in regardless of preparation time. Although a low marginal return can explain the absence of any significant effect, the negative point estimates cannot be explained through this mechanism. A more likely candidate is the grading system. The exams in each subject in each class are graded by two external sensors. When preparation time is short, the best students might receive the best grade or close to the best grade because they have worked on the subject throughout the year. When preparation time increases, more students are able to deliver exams which qualify for the top grade. If the external sensors grade on the curve, and only give the best grade to a fixed number of students. this will increase competition for the best grade and thereby reduce the probability that student with a high GPA from lower secondary earns the top grade on their high school exams.

Moving on, Columns (3) and (4) estimates the baseline model for girls and boys separately. Often girls are found to outperform boys in school, this is also the case in the Norwegian school system. Yet, it is not clear a priori that girls should benefit more from extra preparation time than boys. However, the results suggest that girls benefit somewhat more than boys when the preparation time increases. Girls also have have a positive marginal return for all levels of preparation time, while for boys, the marginal return is either zero or negative when preparation time increases. Pairwise, the coefficients are also statistically different for when preparation time is in the third and fourth quartiles. The implication of this finding is that preparation time, on average, benefits girls more than boys. As girls already outperform boys, preparation time could further expand the gender gap.

The four last columns divide the sample by socio-economic background. The reason for this part of the analysis is that students who come from a more resourceful background might receive more guidance at home in preparing for their exams. A longer exam preparation period might thus allow for even more assistance from parents. Columns (5) and (6) estimates the effect separately for students whose parents have an average income above and below the median the year students are 15, before they start high school. Columns (7) and (8) divide the sample by students for whom at least one parent has a college degree, and students for whom neither parent has any degree beyond a high school diploma. In neither of the sample splits there is any clear pattern of a different effect for the two groups for any amount of preparation time. This suggests that exam preparation time has the same positive effect on exam scores for students regardless of their socio economic background.

Summing up, there appears to be some heterogeneity in the effect of preparation time between high and low performing students and between the genders. There is no traceable heterogeneity between students from different socioeconomic backgrounds. Thus, the heterogeneity analysis suggests that increasing preparation time further strengthens the differences in performance between the genders, and potentially between the high and low performing students.

Another dimension in which students differ is graduation status. One might argue that students who eventually drop out of school have less motivation to prepare for exams, and thus reduce the point estimates. While this will not cause a bias in the estimates, it is of interest whether the effect of preparation time is stronger for students who are more motivated. To explore this relationship I re-run the main estimations reported in Columns (4) and (5) in Table 3, while dividing the sample between graduates and non-graduates. The results of these estimations are reported in Table 7. Below each coefficient pair I report the t-values for a test of coefficients being equal. Although the majority of the sample graduates in the normal time,

Table 6: Heterogenous effects across student characteristics

	(1) High skill	(2) Low skill	(3) Girls	(4) Boys	(5) High income	(6) Low income	(7) High education	(8) Low education
9-12 days of preparation time	0.0483***	0.0697***	0.0639***	0.0474***	0.0514***	0.0594***	0.0554***	0.0616***
	(0.0170)	(0.0168)	(0.0162)	(0.0170)	(0.0162)	(0.0168)	(0.0164)	(0.0173)
	[1.27]		[0.97]		[0.47]		[0.35]	
13-16 days of preparation time	0.0422**	0.0719***	0.0721***	0.0325*	0.0602***	0.0494***	0.0534***	0.0601***
	(0.0177)	(0.0178)	(0.0167)	(0.0188)	(0.0176)	(0.0179)	(0.0172)	(0.0194)
	[1.67]		[2.11]		[0.60]		[0.34]	
17-25 days of preparation time	0.0616***	0.0736***	0.0893***	0.0303	0.0764***	0.0590***	0.0675***	0.0629***
	(0.0228)	(0.0207)	(0.0214)	(0.0211)	(0.0220)	(0.0205)	(0.0219)	(0.0224)
	[0.58]		[2.80]		[0.85]		[0.21]	
Observations	178,097	171,106	193,153	156,050	174,669	174,534	214,405	134,798
# Students	52,267	52,755	58,387	46,635	51,683	53,339	64,095	40,927
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Course FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Total number of exams same year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of exams taken same year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

"High skill" refers to students who have an average teacher assessed course grade in lower secondary school above or equal to the sample median. "High parental income" refers to students whose parents had an average income in the years 2007-2008 that was equal to or higher than the sample median. "High parental education" refers to students who have at least one parent with a bachelor's degree or higher. All regression pairs are mutually exclusive. Standard errors clustered by high school in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Source: Statistics Norway, and The Norwegian Directorate for Education and Training.

around one tenth of the students do not. Considering the estimates for graduates first, the coefficients are close to those reported in Table 3. For the non-graduates on the other hand, point estimates are marginally lower when preparation time is relatively short, and are only significant when the preparation period is in the fourth quartile. Even though none of the coefficient pairs are statistically different, the estimates suggests that students who end up not graduating are less able or willing to take advantage of the preparation period and have a less diminishing marginal return. Thus, they are only able to catch up with the rest of the students when the preparation period is very long and the marginal return to more self-study for the rest of the students have diminished. However, estimates are too noisy to find strong statistical support for this claim.

Table 7: Results by graduation status

	(1)	(2)	(3)	(4)
9-12 days of preparation time	0.0562***	0.0482	0.0554***	0.0500
	(0.0140)	(0.0343)	(0.0138)	(0.0330)
		[0.23]		[0.16]
13-16 days of preparation time	0.0544***	0.0538	0.0611***	0.0516
	(0.0148)	(0.0379)	(0.0145)	(0.0361)
		[0.02]		[0.26]
17-25 days of preparation time	0.0572***	0.120***	0.0647***	0.114***
	(0.0184)	(0.0431)	(0.0178)	(0.0410)
		[1.46]		[1.21]
Standardized teacher assessed grade			0.307***	0.303***
			(0.00458)	(0.0112)
Observations	310,641	38,562	310,641	38,562
# Students	91,028	13,994	91,028	13,994
Year FE	Yes	Yes	Yes	Yes
Course FE	Yes	Yes	Yes	Yes
Total number of exams same year	Yes	Yes	Yes	Yes
Number of exams taken same year	Yes	Yes	Yes	Yes
Student FE	Yes	Yes	Yes	Yes
Sample limited to	Graduates	Non-graduates	Graduates	Non-graduates

The outcome in all regressions is the exam grade standardized by course. The table replicates the results in columns (4) and (5) in Table 3 with the exceptions that Columns (1) and (3) only include students who graduated high school, and Columns (2) and (4) only include students who did not graduate high school. Standard errors clustered on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Sources: Statistics Norway, and The Norwegian Directorate for Education and Training.

7.2 Heterogeneity by subject type

Although the results so far establishes a robust connection between preparation time and exam scores, the results have not addressed potential differential effects by subject type. A general finding in the economics of education literature is that interventions tend to affect mathematics and science results more than results on reading and languages. Consequently, one might expect the effect of preparation time to be dependent on which subject students are preparing for. One hypothesis could be that students are able to improve their skills in more quantitative subjects with more ease compared to less quantitative subjects. To explore this hypothesis I follow a similar approach as Lavy (2015) applies in part of his analysis. First, I divide the exams in three main groups: natural sciences and math exams, language exams, social science and other exams. Then, I re-run the main specification three times, excluding all exams from one group each time. Following this approach it is possible to estimate the the effect of preparation time separately for exams that belong to each category. An alternative approach could be to estimate the effect of preparation time for each category separately. However, such an approach would only use variation from students who have more than one exam within each category. For example, to estimate the effect of increased preparation time in natural sciences and math, only students who have been examined in at least two math and natural science subjects will contribute to identification. Given that students take relatively few exams this alternative approach suffers more from few observation.

Table 8 reports results from the preferred approach. The first column reproduces the main results for comparison. Column (2) show the results when all exams in language subjects are excluded. The coefficients are generally larger for all amounts of preparation time. A similar pattern is found in Column (3) where social science and other exams are excluded. The only specification where effects are smaller than the baseline is when science and math exams are excluded in Column (4). These finding suggests that students have a higher marginal return to preparation time when studying for quantitative subjects, and that preparation time has little or no effect for less quantitative subjects. It should however be pointed out that the differences between the coefficients in Columns (2) and (3) are not statistically different from the baseline estimates.

While Lavy (2015) does not find any statistically significant support for the productivity to differ across subject types, there appears to be support for that here. One reason for this could be that while Lavy (2015) used variation in instruction time, this paper use variation in preparation time. In effect the difference is then that Lavy (2015) tests if the productivity in instruction varies across subjects, whereas I test if the productivity in self-study ahead of exams differ across subjects. Thus, contrasting the finding herein with Lavy (2015), it appears that the relative productivity difference between self-study and instruction is larger for science and mathematics courses than other courses.

Table 8: Subject heterogeneity

	(1)	(2)	(3)	(4)
9-12 days of preparation time	0.0585***	0.110***	0.0643***	0.0450**
	(0.0136)	(0.0328)	(0.0170)	(0.0190)
13-16 days of preparation time	0.0567***	0.116***	0.0764***	0.0162
	(0.0144)	(0.0351)	(0.0166)	(0.0210)
17-25 days of preparation time	0.0667***	0.0882**	0.112***	0.0191
	(0.0178)	(0.0405)	(0.0204)	(0.0238)
Observations	349,203	147,385	285,962	265,059
# Students	105,023	89,055	104,546	102,308
Year FE	Yes	Yes	Yes	Yes
Course FE	Yes	Yes	Yes	Yes
Total number of exams same year	Yes	Yes	Yes	Yes
Number of exams taken same year	Yes	Yes	Yes	Yes
Student FE	Yes	Yes	Yes	Yes
Exclude subjects	None	Language	Social science	Mathematics and science

The outcome in all regressions is the exam grade standardized by course. All columns report results from a specification identical to the baseline specification in Column (4) in Table (3) with the following exceptions: The second column excludes all observations from courses that are classified as language courses; the third column excludes all observations in social science and other courses; the fourth column excludes all observations in science and mathematics courses. Standard errors clustered on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Sources: Statistics Norway, and The Norwegian Directorate for Education and Training.

8 Longer-run

The estimated effects so far suggests that students' performance on exams is sensitive to preparation time. Taking the results as causal, the next natural question is whether the effects are large enough to affect students' outcomes in the longer-run. This is a pressing question because it is not clear whether the short-run effects are substantial enough to alter students' longer-run outcomes. In the case that students are unaffected in the longer-run, the estimated effects above are more of a peculiarity than an important factor in the education production function. To explore this question I follow students into higher education. A core problem in estimating a longer-run effect of increased preparation time is that preparation time varies within the individual, while the outcomes in higher education are only observed once. Therefore, the student fixed effects approach employed above is not feasible. Instead, two alternative methods are employed.

Both methods here follows Bensnes (2016), which is exposed to the same challenges as this paper using the same data to answer a different question. The first method is to estimate the direct effect of average preparation time across all exams on average exam grade across all exams, and outcomes in higher education. One reason to expect an effect in this model framework is that students who are randomly assigned more preparation time, earn higher exam grades which in turn increase their university application score. A second reason to expect a longer-run effect is that students might increase not only their exam scores in the subjects they are tested, but also their general human capital or develop other skills which are beneficial in the longer run, such as the ability to concentrate on a task over a longer time period. When estimating a reduced form effect these mechanisms cannot be directly distinguished, however, the reduced form estimates show the policy relevant measure of how preparation time affect longer-run outcomes. A drawback with this method is that is is unable to take into account the non-linear relationship identified above. Yet, as the reduced form estimates are relevant for policy they are included here.

The second method is based on a two-step procedure. In the first step I use the data on outcomes in higher education to estimate a descriptive relationship between the average exam grade and these outcomes. Because it is likely that there is some sorting of students between applying to tertiary education and not depending on background characteristics and unobservables the estimated relationship does not reflect a causal relationship, but it does give an impression of the connection between exam grades and enrollment. In the second step I combine the estimated effect of increased preparation time from the baseline model in the main results with the descriptive relationship estimated in this section. The combination of these estimates can then be used to predict the longer-run effects of increased preparation time. This approach is not ideal as the relationship between grades and enrollment is merely descriptive, because some graduates who have high enough grades to enroll chose not to apply. However, opposed to the first method, this method is better suited to take into account the non-linearities found in the main analysis.

The outcomes used in both methods are: enrolling in higher education, enrolling in a math or science program rather than any other program (STEM)[12], and dropping out of higher education or changing program before the second year of university. In the Norwegian education system enrollment in higher education is based on an application score comprised almost exclusively by the average teacher assessed grades and exam grades[13]. Higher education institutions are obliged to admit students based on their application score, and cannot regard any application letter or other attributes such as the student being valedictorian. Therefore, written exam grades will comprise about 15% of the application score for the typical student. As such, one should expect to find effects on enrollment in higher education from preparation time. Further, the results bear interest for other school systems where exit exams play a even more significant role.

In the absence of student fixed effects, all estimations in this section includes school-by-cohort fixed effects to net out any mean differences between cohorts and schools, including differences in average preparation time due to differences in subjects offered, the composition of students and other factors that are fixed within the school-cohort level. In other words, it could be the case that students in a specific cohort at a specific school elects subjects for the average preparation time is longer than at another school or for another cohort. If these students also receive better instruction the estimated effect of preparation time would be biased upwards. Another way to think about the school-by-cohort fixed effects, is that they make it possible to only compare students exposed to the same learning environment, but randomly assigned to take exams in different subjects. In this section only student who graduate high school are included, as a high school diploma is required to eligible to apply to higher education. Standard errors are clustered at the school levels as above.

Before moving on to the longer-run results, Column (1) in Table 9 shows estimates for the relationship between average preparation time and the average exam grade when the data is collapsed on the student level. The results show that an increase in the average preparation time of one standard deviation increases the average exam score by 2.7% standard deviations[14]. In

[12] This definition is set 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.

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

[14] The average preparation time in the longer-run sample is 13.46 days with a standard deviation of 2.63. The student level average exam score is 3.37 with a standard deviation of 0.86.

the baseline model the effect of increasing preparation time by around twice as much increased exam scores by about 6% of a standard deviations. The effect is therefore about the same size as in the baseline model. This is a reassuring result which increases the credibility of the longer-run reduced form estimates.

Columns (2)-(4) report results when the dependent variable is various outcomes in higher education. In Column (2) the outcome is a dummy equal to one if the student is ever registered in a higher education program. In Column (3) the outcome is a dummy which equals one if the student ever enrolls in a STEM program conditional on enrolling in higher education. In Column (4) the outcome is a dummy for whether the student drops out of higher education within the start of the second year. This variable also includes students who change from one program to another. The results in Columns (2) and (3) show that, as expected, more preparation time increases the probability that students enroll in higher education and in programs with higher requirements: A one standard deviation increase in the average preparation time increases the probability that a student enrolls in higher education by 0.9%-points, and the probability that a student enrolls in a STEM program by 2%-points. The effects can be considered to be fairly large. In the longer-run sample (i.e. student who have graduated high school) 92.7% start a higher education program and 17.2% start a science or mathematics program. Thus, the estimates suggests that a one standard deviation longer preparation time increases the probability that a student enrolls in a STEM program rather than another program by about 10%.

The estimated results in Column (4) suggests that students who receive a longer average preparation time are less likely to drop out once enrolled. This suggests that students who by chance have more preparation time, not only improve their exam scores and thereby their application scores, but also do better once enrolled in higher education. A decreased probability that students drop out reflects either (i) that students are matched to programs which they are motived to work on; or (ii) are have improved their skills in a way which enhances their ability to study in tertiary education, or both.

Moving on to Columns (5)-(7) similar results are found. Better performing students, measured as their average exam score, are more likely to enroll in higher education and in the more STEM programs. They are also less likely to drop out of higher education once enrolled. When these results are combined with the baseline estimates they suggest that increasing preparation time from 5 to 8 days to 9 to 12 days increases the probability that a student enrolls in higher education by .2%-points, enrolls in science and math programs by .1%-points, and are .2%-points less likely to drop out. Although the estimated size of the effects estimated differ between the methods, they both indicate an effect of preparation time on longer-run outcomes.

Summing up the results from Table 9 the short-run effects on exams scores found in Table 3 appear to follow students in the longer-run, affecting both enrollment in higher education, and persistence.

9 Assessing the effect of grades on longer-run outcomes

This paper has shown that preparation time has a direct effect on exam scores. Because preparation time is randomly assigned, it is possible to use preparation time as an instrument to ask a general and important question: what is the effect of exam scores on later outcomes? One can easily argue that an OLS estimation with a longer-run outcome as the dependent variable and exam grades as the independent variable will be biased because exam grades will be correlated with unobserved characteristics with the student which could also directly affect the outcomes, such as the aspirations of the student or her parents, innate ability, and more. Using preparation time as an instrument for exam scores circumvents these potentially biasing factors, and makes it possible to give a credible estimate on the effect of exam grades on longer-run outcomes.

The estimation procedure is straightforward and is methodologically similar to the one applied in Lavy, Ebenstein, and Roth (2015). The first stage is parallell to the estimated model in Column (1) in Table 9. The second stage is equivalent of the models estimated in Column (5)-(7) in Table 9 with the exception that the exam score is instrumented with preparation time.

For the estimates to be credible, the instrument must fulfill two criteria: the exclusion criteria and the relevance criteria. The exclusion criteria in this setting demands that the the only way preparation time affects longer-run outcomes is through exam scores. Thus, the exclusion criteria would be violated if longer average preparation time influenced longer-run outcomes in a subtle way. One example of such an effect would be if a longer-preparation

Table 9: Longer-run: Reduced form and descriptive relationship

	(1) Outcome: Exam grades	(2) Outcome: Enroll higher edu.	(3) Outcome: Enroll STEM	(4) Outcome: Drop out	(5) Outcome: Enroll higher edu.	(6) Outcome: Enroll STEM	(7) Outcome: Drop out
Average preparation period	0.00875***	0.00313***	0.00727***	-0.00270***			
	(0.00154)	(0.000459)	(0.000844)	(0.000644)			
Average exam score					0.0331***	0.0161***	-0.0354***
					(0.00138)	(0.00234)	(0.00198)
Exactly 1 parent working	-0.00494	0.0209***	-0.00704	-0.00676	0.0209***	-0.00719	-0.00760
	(0.0187)	(0.00802)	(0.00993)	(0.00973)	(0.00794)	(0.00997)	(0.00965)
Both parents working	-0.00801	0.0318***	-0.00247	-0.0136	0.0320***	-0.00226	-0.0147
	(0.0183)	(0.00788)	(0.00970)	(0.00937)	(0.00781)	(0.00974)	(0.00930)
Average parental income (NOK)	2.43e-10	1.71e-09	-1.82e-09*	-9.86e-10	1.74e-09	-1.74e-09*	-1.03e-09
	(2.39e-09)	(1.38e-09)	(1.03e-09)	(1.03e-09)	(1.35e-09)	(1.01e-09)	(1.01e-09)
Parents highest educational level is high school	-0.0157	0.0221***	0.00403	0.000645	0.0227***	0.00445	-0.000179
	(0.0120)	(0.00567)	(0.00620)	(0.00636)	(0.00565)	(0.00621)	(0.00635)
Parents highest educational level is bachelor degree	0.0585***	0.0474***	0.0222***	0.00703	0.0454***	0.0214***	0.00866
	(0.0122)	(0.00569)	(0.00626)	(0.00629)	(0.00566)	(0.00626)	(0.00630)
Parents highest educational level is master or PhD	0.164***	0.0527***	0.0801***	0.00848	0.0472***	0.0775***	0.0138**
	(0.0130)	(0.00585)	(0.00696)	(0.00665)	(0.00583)	(0.00697)	(0.00669)
Female	-0.0322***	0.0260***	-0.193***	-0.0306***	0.0267***	-0.193***	-0.0318***
	(0.00558)	(0.00192)	(0.00329)	(0.00275)	(0.00192)	(0.00331)	(0.00275)
First generation immigrant	-0.121***	0.0315***	0.0591***	-0.0652***	0.0354***	0.0611***	-0.0699***
	(0.0167)	(0.00565)	(0.00858)	(0.00749)	(0.00567)	(0.00858)	(0.00752)
Second generation immigrant	-0.147***	0.0529***	0.0456***	-0.0828***	0.0574***	0.0474***	-0.0880***
	(0.0162)	(0.00471)	(0.00857)	(0.00779)	(0.00473)	(0.00859)	(0.00784)
GPA lower secondary	0.926***	0.0810***	0.100***	-0.0860***	0.0507***	0.0862***	-0.0535***
	(0.00593)	(0.00217)	(0.00297)	(0.00285)	(0.00240)	(0.00362)	(0.00317)
Observations	94,152	94,152	87,265	87,265	94,152	87,265	87,265
School-by-cohort groups	1218	1218	1205	1205	1218	1205	1205
School-by-cohort FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

The outcome variables are (i) the standardized average written exam score (ii) enrolling in higher education, (iii) enrolling in a natural science, mathematics, or engineering program conditional on enrolling in higher education, (iiii) dropping out of higher education or changing program before the second year. In addition to reported variables each regression also controls for the number of exams the students take and school-by-cohort fixed effects. The parental education variables refer to the highest attained degree by either parent. Parental income is measured in nominal terms. Standard errors clustered by high school in parentheses. * p<0.01, ** p<0.05, *** p<0.1. Source: Statistics Norway, and The Norwegian Directorate for Education and Training.

increases students' enjoyment of studying, for example by an increased sense of mastery. This could in turn increase students' appetite for further studies in tertiary education or change their preference ranking of programs. Although it is not possible to rule out such a mechanism by mean of econometric tools, it seems implausible that this effect would be any more than a small fraction of the effect of increased grades observed in the first stage.

The relevance criteria, that average preparation time affects average exam score, holds. As can be seen from the table the first stages are fairly strong, however, the F-statistic from the Kleinbergen-Paap tests reported in each column in Table 10 clearly exceed the critical values from Stock and Yogo (2005). Specifically, they report the critical value for weak identification under the conditions used here as 16.38 for a 10% maximal IV bias.

Before moving on to results, it is worthwhile to briefly discuss the IV approach in terms of which students are affected by the instrument. In some cases IV-results suffer from weak external validity because the local average treatment effect is based on a very selected group of treated observations. The sub-sample analysis in section 7 showed that all students had some response to increased preparation time. Considering these previous results, the IV-estimates are based on a broad group of treated individuals, covering all or most of the students to various degrees. Thus the IV estimates should be interpreted as a credible estimation of the link between exam scores and longer-run outcomes.

The IV-estimates are all significant. Starting with Column (1) increasing the average exam score by 0.1 standard deviations increases the probability that a student enrolls in higher education by 3.1%-points. The OLS estimate in Column (5) in Table 9 is only about 10% of the IV-estimates. Insofar the instrument affects most students the large difference in effects suggests that OLS estimates should be considered a lower bound. The IV-estimates might be considered too large to be credible, however at least one previous study find effects in the same ballpark. Lavy et al. (2015) use random variation in the levels of pollution on examination days in Israel as an instrument for Bagrut composite scores, and find that increasing the composite score by 1 standard deviation increases the probability that a student enrolls in a post-secondary education by 45%-points[15]. An effect 50% larger than the effect I find here. Although institutions

[15] My calculation based on descriptive statistics and estimates presented in their paper.

differ, the consistent finding of large IV-estimates dampens worries regarding the credibility of the results.

Moving on to the second outcome, the probability of enrolling in a STEM program relative to other programs, 10% of a standard deviation increase in exam scores increases the relative probability that a student enrolls in a STEM program by 7.3%-points[16]. The results in Columns (1) and (2) therefore show an expected pattern. Higher exam scores result in better outcomes in the longer-run with a higher probability to enroll in tertiary education and in more competitive programs.

The last outcome is the probability that a student drops out of tertiary education in the first year. Again there is a large effect of exam scores. Increasing the average exam score by 10% of a standard deviation decreases the probability that a student drops out prior to the second year of university by 2.8%-points.

The key take-away from this section is that the OLS estimates are much smaller than the more credible IV-estimates, which suggests that the OLS-estimates severely underestimates the effect of exam grades on longer-run outcomes. This is seemingly counterintuitive, one might expect that some obvious candidates for causing omitted variable bias, such as innate ability, are positively correlated with both longer-run outcomes and average exam grades. However, exam grades are basically a measure of skills with measurement error. If omitted variables only cause a relatively moderate upward bias, the OLS estimate might be downward biased due to measurement error.

10 Concluding remarks.

This paper has offered the first evidence on the effect of preparation time on students' achievements on high-stakes tests. Using a unique institutional setting and administrative data, the effects are relatively large and significant, and suggest longer-run effects. The contribution is unique in the sense that it is the first paper to directly assess the impact of preparation time as opposed to instruction time. While the identification strategy relies on within-student vari-

[16] The student level average exam score for students enrolling in tertiary education is 3.41 with a standard deviation of 0.85.

Table 10: Longer-run: IV estimates

	(1)	(2)	(3)
First Stage:	Exam score	Exam score	Exam score
Average preparation period	0.00875***	0.00845***	0.00845***
	(0.00154)	(0.00159)	(0.00159)
Second Stage:	Enroll in higher edu.	Enroll in STEM	Drop out
Average exam score	0.357***	0.860***	-0.320***
	(0.0784)	(0.194)	(0.0932)
Observations	94,152	87,265	87,265
R^2 First stage	0.345	0.334	0.334
Kleinbergen-Paap Wald F statistic	32.46	28.30	28.30

The outcome variable in all first stage estimations is the averaged written exam score for each student. The first stage estimation reported in Column (1) is equivalent to Column (1) in Table 10. The second stage outcomes are (i) enrolling in higher education, (ii) enrolling in a natural science, mathematics, or engineering program conditional on enrolling in higher education, (iii) dropping out of higher education or changing program before the second year. In addition to reported variables each regression also controls for the number of exams the students take and school-by-cohort fixed effects. Standard errors clustered by high school in parentheses. * p<0.01, ** p<0.05, *** p<0.1. Source: Statistics Norway, and The Norwegian Directorate for Education and Training.

ation and a specific institutional setting, the results are also indicative on how students utilize self-study time prior to important tests. In particular the results are non-linear with differences between male and female students, being more pronounced when the preparation period is relatively long. In terms of the longer-run effects established here, effects are likely to be significantly larger in systems where high-stake exams play a larger role in students' later placement in tertiary education.

Interestingly, the effects of extra preparation time identified here are remarkably similar to the effects of increasing instruction time by the comparable amounts. Although the effects are very similar it is not possible to discern why because it is unknown how much of their preparation period students spend in self-study and how many hours of extra instruction they receive from their teachers. In one extreme case students could receive no extra instruction from their teacher, but spend many hours every day in motivated self-study. On the other extreme, students could receive intense extra instruction, but spend no time in self-study. A further issue is that instruction might be much more productive when students are close to a high-stakes examination. Regardless, the paper demonstrates how preparation time have strong effects, both for immediate performance on tests, but also human capital accumulation and students' performance in higher education.

In terms of policy, the results have two main implications. First, differences in preparation time between otherwise identical students can distinct and relatively large effects on both test scores and later outcomes. This in turn suggests that homogenous preparation periods across students is important when exit exams are used as placement tools. As the exams in the Norwegian educational system is likely to have a smaller impact on the application score than in some other systems, this argument may be even stronger in other countries. Second, preparation time can be adjusted to affect test score results even among relatively old students. Because preparation time relatively easily adjusted with low costs it can be used as an effective and inexpensive policy tool. However, the marginal return to preparation time is sharply decreasing, creating an upper limit on the effect of policy potential interventions.

In addition to finding exploring the effects of preparation time on both immediate and longer-run outcomes, this paper also estimates the effect of exam grades on long-run outcomes using preparation time as an instrument. Compared to OLS estimates the IV-estimates are up to 10 times larger. This suggests that OLS-estimates severely underestimates the effect of test scores on longer-run outcomes.

References

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.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372.

Card, D., & Krueger, A. B. (1992). Does school quality matter? returns to education and the characteristics of public schools in the united states. *The Journal of Political Economy*, *100*(1), 1–40.

Carlsson, M., Dahl, G. B., Öckert, B., & Rooth, D.-O. (2015). The effect of schooling on cognitive skills. *Review of Economics and Statistics*, *97*(3), 533–547.

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.

Hansen, B. (2011). School year length and student performance: Quasi-experimental evidence. *Available at SSRN 2269846*.

Kirkebøen, L., Leuven, E., & Mogstad, M. (2016). Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, qjw019.

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, forthcoming*.

Marcotte, D. E., & Hemelt, S. W. (2008). Unscheduled school closings and student performance. *Education*, *3*(3), 316–338.

Pischke, J.-S. (2007). The impact of length of the school year on student performance and earnings: Evidence from the german short school years*. *The Economic Journal*, *117*(523), 1216–1242.

Rivkin, S. G., & Schiman, J. C. (2015). Instruction time, classroom quality, and academic achievement. *The Economic Journal*, *125*(588), F425–F448.

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear iv regression. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*.

The Norwegian Directorate for Education and Training. (2009). *Trekkordning ved eksamen i kunnskapsløftet*.

Wößmann, L. (2003). Schooling resources, educational institutions and student performance: the international evidence. *Oxford bulletin of economics and statistics*, *65*(2), 117–170.

11 Appendix

The Norwegian high school system is divided in 12 tracks. 9 of these are vocational, upon completing any of the 3 remaining academic tracks students are qualified to apply for tertiary education. These 3 tracks are: specializations in general studies; music, dance and theater; sports. The types of exams students take through high school depends partly on which track they are enrolled in. Students enrolled in specialization in general studies are required to take two randomly drawn written exams in the third year and one randomly drawn exam that is either oral, practical, or oral-practical. Students enrolled in the music, dance and theater track have to take to randomly drawn exams that can be either written, oral or oral-practical. Further they have to take one exam that is oral-practical and related to their track specialization. Lastly, students in the track for sports have to take three exams that can be either written, oral or oral-practical, at least one of which have to be related to their track specialization. In addition to these exams all students enrolled in any academic track have a 20% probability of having to take one exam in the first year of high school. This exam can be written, oral or oral-practical. In addition, all students have to take one exam in their second year in high school, which also can be written, oral or oral-practical. Lastly, all students have to take a written exam in Norwegian languages in their third year of high school. For students in the academic track it is possible to identify which exams are oral or oral-practical. Such exams are dropped from the sample.

Table A1: Alternative clustering

	(1) Subject clustering	(2) Subject clustering	(3) Two-way clustering	(4) Two-way clustering
9-12 days of preparation time	0.0588**	0.0572**	0.0588**	0.0572**
	(0.0270)	(0.0265)	(0.0237)	(0.0232)
13-16 days of preparation time	0.0570**	0.0620**	0.0570**	0.0620***
	(0.0251)	(0.0242)	(0.0220)	(0.0212)
17-25 days of preparation time	0.0670**	0.0721**	0.0670**	0.0721***
	(0.0311)	(0.0289)	(0.0277)	(0.0256)
Standardized teacher assessed grade		0.314***		0.314***
		(0.0261)		(0.0222)
Observations	341,536	341,536	341,536	341,536
R^2	0.620	0.647	0.620	0.647
Year FE	Yes	Yes	Yes	Yes
Course FE	Yes	Yes	Yes	Yes
Tot al number of exams same year	Yes	Yes	Yes	Yes
Number of exams taken same year	Yes	Yes	Yes	Yes
Student FE	Yes	Yes	Yes	Yes
# Students	97,284	97,284	97,284	97,284

The two first columns are the same as Columns (4) and (5) in Table 3 with the exception that standard errors are clustered by subject rather than school. Columns (3) and (4) are also the same as Columns (4) and (5) in Table 3 with the exception that standard errors are clustered in both subject and school.

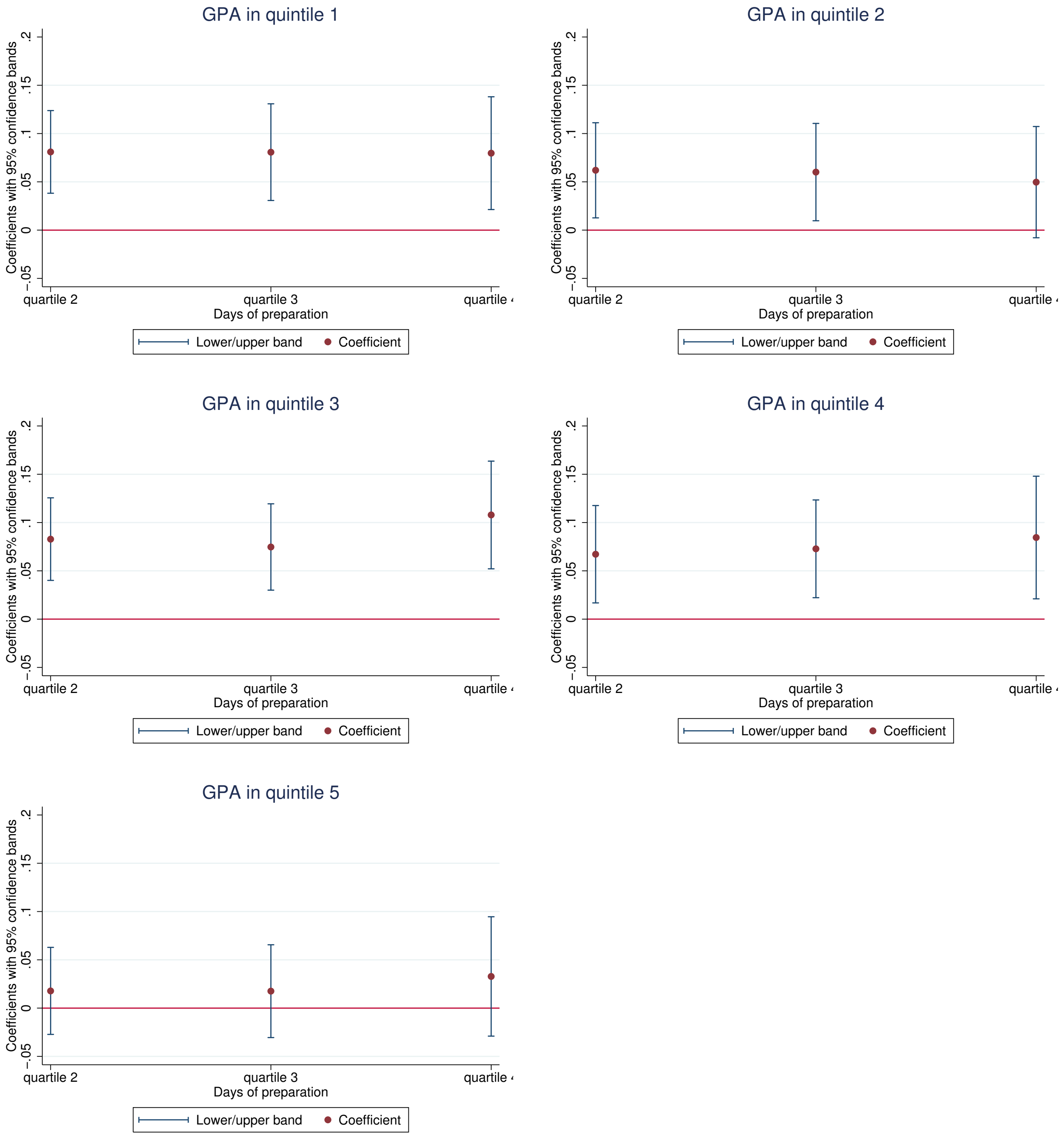

Figure 3: Heterogeneity: Sample split in quintiles by GPA from lower secondary.

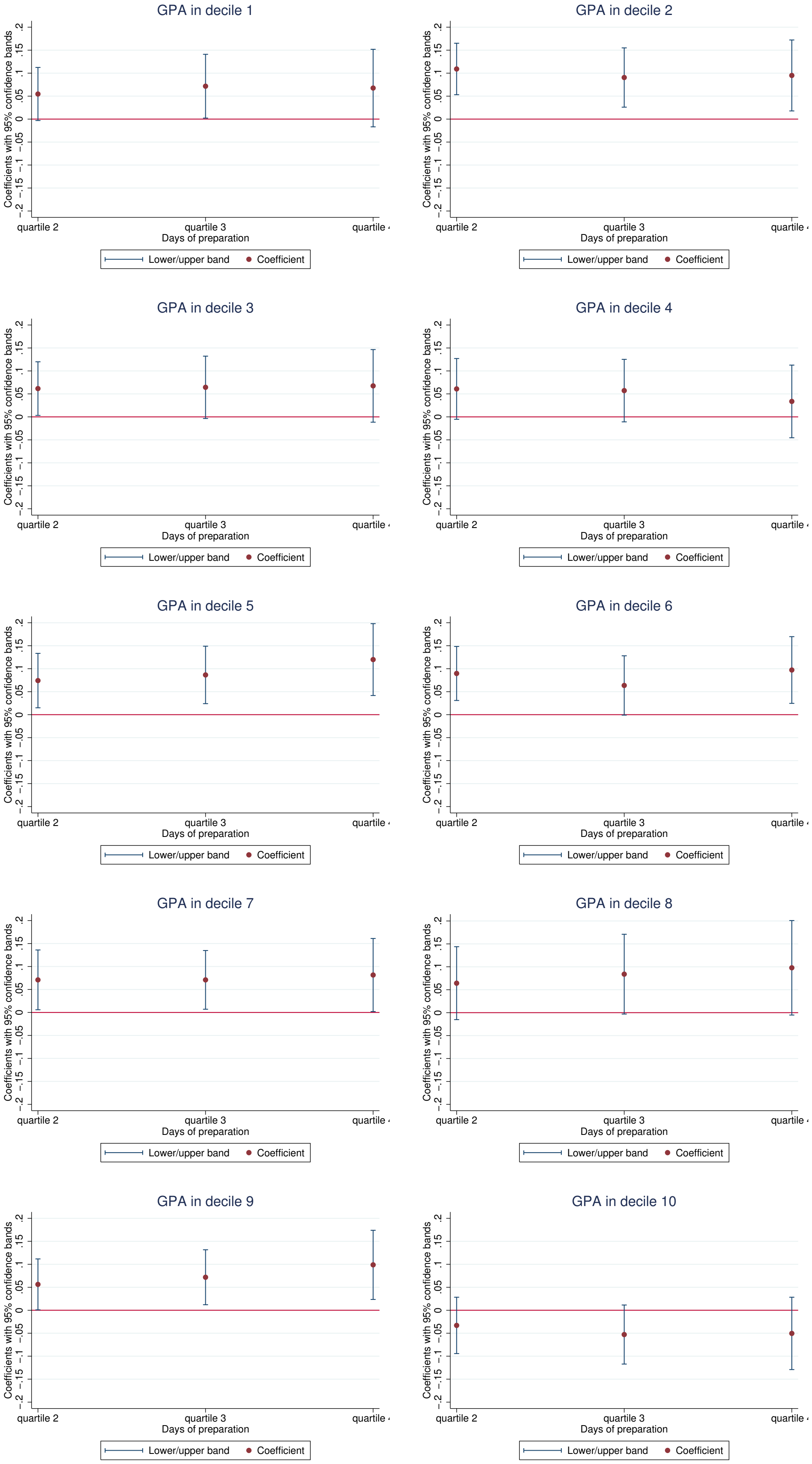

Figure 4: Heterogeneity: Sample split in deciles by GPA from lower secondary.