Goulas, Sofoklis; Megalokonomou, Rigissa

Working Paper

Marathon, Hurdling or Sprint? The Effects of Exam Scheduling on Academic Performance

IZA Discussion Papers, No. 11624

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Goulas, Sofoklis; Megalokonomou, Rigissa (2018) : Marathon, Hurdling or Sprint? The Effects of Exam Scheduling on Academic Performance, IZA Discussion Papers, No. 11624, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/185084

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11624

Marathon, Hurdling or Sprint? The Effects of Exam Scheduling on Academic Performance

Sofoklis Goulas
Rigissa Megalokonomou

JUNE 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11624

Marathon, Hurdling or Sprint? The Effects of Exam Scheduling on Academic Performance

Sofoklis Goulas
Stanford University

Rigissa Megalokonomou
The University of Queensland and IZA

JUNE 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Marathon, Hurdling or Sprint? The Effects of Exam Scheduling on Academic Performance

Would you prefer a tighter or a prolonged exam schedule? Would you prefer to take Math before Reading or the other way around? We exploit variation in end-of-course exam schedules across years and grades to identify distinct effects of the number of days between exams, the number of days since the first exam, and the exam order on subsequent performance. We find substantially different scheduling effects between STEM and non-STEM subjects. First, we find a positive relationship between exam performance in STEM subjects and exam order, controlling for other influences of scheduling, suggesting that the later in the schedule an exam is taken the higher the average performance. We call this phenomenon, exam warm-up. Second, we find a negative relationship between the number of days from the very first exam and subsequent exam performance in STEM subjects, suggesting the existence of a fatigue effect. For STEM subjects, the fatigue effect is estimated to be less than half the size of the warm-up effect. For non-STEM subjects, an additional day between exams is significantly associated with lower performance in subsequent exams. Students of lower prior performance have lower fatigue effects and higher warm-up effects in STEM subjects compared to students of higher prior performance. Also, we find that exam productivity in STEM increases faster for boys than it does for girls as they take additional exams due to a higher warm-up effect. Our findings suggest that low-cost changes in the exam schedule may have salient effects on student performance gaps.

JEL Classification: I20, I24

Keywords: exam schedule, cognitive fatigue, exam warm-up, practice, gender gap, STEM subjects

Corresponding author:
Sofoklis Goulas
Hoover Institution
Stanford University
434 Galvez Street
Stanford, California 94305
United States
E-mail: goulas@stanford.edu

1 Introduction

During finals week(s), most high school students take a sequence of exams with only a few days at most in between exams. While some students may feel energized before every exam, many struggle to stay awake and sharp. In fact, survey evidence shows that only 23 percent of college students get eight hours of sleep per night during finals week and a quarter report that sleep deprivation affects their academic achievement (US San Diego College Health Association, 2008).

As families, teachers, and administrators seek ways to improve student academic performance, some question whether the scheduling of exams may hinder the achievement for teenagers. Cognitive psychology research supports this notion, finding that a tight schedule of cognitive tasks is conducive to sleep deprivation (Wolfson and Carskadon, 2003) and that sleep deprivation affects performance in cognitive tasks (Blagrove et al., 1995). In particular, scheduling may affect cognitive fatigue and memory, which consequently influence performance in cognitive tasks. Cognitive or mental fatigue can be defined as a decrease in cognitive resources over time due to sustained cognitive demands, independently of sleepiness and is found to be associated with decreased task performance (Boksem et al., 2005; van der Linden et al., 2003; Lorist et al., 2000; Hockey and Earle, 2006). Cognitive memory as well as meta-memory, which includes both own memory capabilities and the processes of memory self-monitoring, are both affected by the scheduling of tasks (Rohrer, 2009; Rohrer and Taylor, 2007; Taylor and Rohrer, 2010; Rohrer and Taylor, 2006; Bjork et al., 2013).

It is not only in psychology that researchers have studied the performance in cognitive tasks. Understanding the determinants of task productivity is also a central question in economics. In everyday life, individuals are faced with multiple tasks. The time horizon for the completion of several tasks is not endless, but rather limited. Given the scarce time and limited attention, individuals need to decide how to allocate their resources in order to maximize their utility, which is assumed to be positively related to the outcomes of the undertaken tasks. Thus, the different tasks one is faced with compete for their attention and time. In a context with many tasks and limited resources, particularly time, the way the different tasks are scheduled over time has salient effects in task performance (Buser and Peter, 2012). In fact, athletes have been found to benefit when the scheduling of athletic events allows them to have several weeks of recuperation (Chambers et al., 1998). Additionally, in a study of the duration of court case completion by Italian Judges (Coviello et al., 2014), it was found that completing cases simultaneously takes longer time,

on average, than completing tasks sequentially.

The context in which this study explores the effects of scheduling on cognitive task performance is an educational one, and in most educational systems students are required to complete several tasks in a finite time period, including projects and exams. The scheduling of those tasks is an important driver of students' performance. For example, students' performance has been found to be positively associated with the time between cognitive tasks (Pope and Fillmore, 2015). At the same time, postponing an exam for the end of the year has been found to have negative effects on performance (Di Pietro, 2013). Time between tasks is only one aspect of task scheduling. Additional aspects of scheduling include the order in which the tasks are completed as well as the number of tasks to be completed in a given time period. What is the effect of having completed an additional task on later performance? Is there fatigue associated with task completion?

In this study we explore the different channels through which exam scheduling affects performance. A particular type of scheduling, class scheduling, has received some attention both in the education and the economics literature (Carrell et al., 2011a; Dills and Hernndez-Julin, 2008; Edwards, 2012), but the scheduling of exams has been studied very little. Our research question is motivated by practical concerns: how to schedule cognitive tasks optimally. School principals, much like managers, are always looking for innovations that increase exam productivity with little to no increase in resources. History has demonstrated that simple innovations, such as crop rotations, work schedules and other simple managerial practices have been successful at increasing efficiency (Pope, 2016).

Policy initiatives may be interested in understanding the implications of exam scheduling, especially when exam scheduling may affect the performance gap in STEM subjects between males and females. Although the performance gap in STEM subjects, such as mathematics, between males and females has been well-documented in the literature (Fryer Jr and Levitt, 2010; Else-Quest et al., 2010; Dee, 2007; Nosek et al., 2009; Hyde et al., 2008), as have the differences in cognitive learning between males and females (Zimmerman and Martinez-Pons, 1990; Halpern, 2004, 2013; Fennema and Sherman, 1977), but the extent to which the gender gap can be explained by the setup of cognitive learning at school has not be investigated. Mechanisms that influence the performance gap in STEM-related tasks between males and females may be of interest to policy-makers, particularly when those mechanisms can be influenced by low-cost interventions, such as changing the exam schedule. Our setting allows for the investigation of whether males and females have different reactions to exam scheduling in terms of performance. Understanding how aspects such as

exam scheduling impact different genders allows us to understand how different schedules may lessen or widen the gender gap in performance.

We disentangle the timing effects associated with exam scheduling on exam performance by exploiting variation in exam schedules across grades and years in a novel data set on student performance in each exam taken in the 10th and 11th grade in high school. In particular, we obtained data on exam performance in 32 subjects of nine cohorts of students in the 10th and 11th grade in Greece between school year 2001-2002 and 2009-2010. At the end of each school year, between May and June, high school students take exams in every subject taught during the school year. This process lasts between three and four weeks with each student taking exams in more than 13 subjects, on average.

The research question at hand is not easy to answer as task assignment is usually not random and individuals who select into certain tasks may also select into a particular completion schedule for those tasks. At the same time, individuals' preferences or other engagements may also influence their task scheduling. Endogeneity arising from unobservables driving selection into tasks and selection into particular task completion schedules renders the identification of the timing effects of scheduling on task performance challenging.

Our paper contributes to the bodies of research in economics and psychology exploring cognitive fatigue, time between cognitive tasks, cognitive load, and memory recall in several ways. Unlike the previous papers, our approach disentangles three distinct, contemporaneous effects of scheduling on exam performance, allowing us to shed light on and compare the influences of the mechanisms through which scheduling impacts exam performance. Another interesting feature of our paper lies on the orthogonality of scheduling of compulsory courses across grades and years to subject type (STEM or non-STEM) and student characteristics, particularly prior academic performance. The consistent grading structure of every course in the Greek educational system allows for a consistent measure of student achievement; faculty members teaching the same course in each year use an identical syllabus and follow the same examination protocols during a common testing period, allowing for standardized grades within a course-grade combination. The combination of stable institutional characteristics and randomized variation in exam scheduling over time allow us to cleanly identify the causal paths through which scheduling can impact exam performance. Additionally, we identify differential effects of exam scheduling on performance by gender. Moreover, we explore scheduling effects across a broader part of the ability distribution, an endeavor not previously attempted in the literature.

To our knowledge, we are the first ones to identify three channels through which the scheduling of the exams may affect performance. The first channel through which scheduling affects exam performance is the number of days between exams. Time between exams may lend itself to preparation, or recuperation. The effect of the time length between exams on subsequent performance may be a composite effect of both preparation and potential distraction. We call this Scheduling Effect I. Considering a sequence of tasks individuals have to complete in a given time period, the second channel corresponds to the effect of the time distance between the first task completed and the one that they are about to attempt. We call this Scheduling Effect II. The third channel relates to how many exams have been taken before sitting an additional exam. We call this Scheduling Effect III.

We show that the number of days between exams, the number of days since the first exam , and the exam order have distinct marginal influences on exam performance. We find significant scheduling effects on STEM subjects. Our results indicate that exam productivity in STEM courses increases with exam order, suggesting the existence of a learning effect positively associated with taking an additional exam. We find that the exam order affects student achievement, with exams taken later in the schedule being associated with higher performance (practice or warm-up effect), controlling for other influences of exam scheduling such as preparation time between exams and overall fatigue, proxied by the number of days since the first exam taken. Students randomly assigned a later exam order earn a grade in that exam significantly higher compared to students randomly assigned an earlier exam order for the same course.

At the same time, exam performance is found to decrease with the number of days since the first exam, suggesting an additional day in the exam season is associated with exam fatigue in STEM courses. An one-day increase in the day count since the first exam a student took significantly decreases their subsequent performance in STEM-related exams (fatigue effect). For STEM subjects, the fatigue effect is estimated to be less than half the size of the warm-up effect. The number of days between exams is found to be associated with decreasing exam performance in non-STEM courses. One additional day between exams significantly decreases performance in non-STEM-related exams, controlling for other influences.

We also explore differential effects across different levels of prior midterm performance. Students in the top quantile of prior midterm performance enjoy a significantly higher warm-up effect in both STEM and non-STEM courses from additional exams compared to students in the bottom quantile of prior midterm performance. Additionally, students in

the top quantile of prior midterm performance exhibit lower fatigue effect in STEM courses associated with an additional day in the exam season. More days between exams are found to benefit more students in the top quantile of prior midterm performance compared to students in the bottom quantile in terms of STEM-related performance. The results are reversed when performance in non-STEM courses is considered. In particular, students in the bottom quantile of prior midterm performance have a more positive effect in non-STEM courses from an additional day between exams compared to students in the top quantile of prior midterm performance. What is more, we find that exam productivity increases faster for boys than it does for girls as they take additional exams.

This paper is organized as follows. In section one, we discuss the existing evidence to motivate hypotheses that can be tested empirically. In section two and three, we provide information on the institutional setting of exam scheduling and discuss the data that we use in our study, respectively. In section four, we lay out our empirical methodology. We report and discuss our results in section five, and we conclude in section six.

2 Why we would expect exam scheduling to affect performance

In this section we bring together the existing evidence from the economics, sociology, education, and psychology bodies of literature to predict how students' performance would react to different exam scheduling. We consider three potential mechanisms through which exam scheduling may impact performance: time between exams, number of days since the very first exam, and having taken an additional exam in the current exam season. The three channels are annotated as Scheduling Effects I, II, and III, respectively. Exam performance may vary with exam scheduling because scheduling may affect the cognitive conditions under which an exam is taken. Cognitive fatigue, for instance, may emerge when one has to take several exams in narrow time intervals. On the other hand, improvement of meta-cognitive accuracy may appear when one repeats certain cognitive tasks, such as exams. Not every task or exam puts the same stress on every aspect of cognition. For example, tasks than require mathematical calculations may stress cognitive accuracy more than tasks that are more memory-intensive. In this paper, we distinguish between STEM and non-STEM subjects. Exams in STEM subjects are more likely to stress the cognitive capacity that is related to mathematical calculations, logic and decision making, while exams in non-STEM subjects may be more likely to utilize more the parts of cognition

that are associated to memory or critical thinking.

One may argue that if different subjects put different levels of stress in different parts of cognition and one has demonstrated a certain level of achievement in a particular subject in the past, then it is likely their past achievement may reveal the degree to which they possess the cognitive skills that are used intensively in that subject. Consequently, student of different prior performance in a certain type of subjects may possess different levels of the cognitive skills those subjects put more stress on, and subsequent task performance may be explained by prior performance.

At the same time, students of different gender may also possess different cognitive skills or at different levels which would suggest that the scheduling of cognitive tasks may influence heterogeneously males and females. The cognitive differences between males and females has been well established (Halpern, 2013; Hyde et al., 1990; Hyde and Linn, 1988). The existing literature proposes that males may be exhibit higher returns to practice than females in terms of performance in cognitive tasks, although females may be better in self-regulated learning than males (Law et al., 1993; Ablard and Lipschultz, 1998).

We present the three channels in a simple graphic of sequential exam productivity in which a student is subjected to a finite number of exams under different schedules. It is important to stress that the goal is not to test a particular underlying mechanism for the observed effects but rather to disentangle the different ways task scheduling can impact performance.

One potential channel that we consider here is the time between exams. This channel is illustrated in the comparison between exam schedules (a) and (b) in Figure 2. Exam schedules (a) and (b) both include two exams, but schedule (b) allows for more time between the first and the second exam, compared to schedule (a). If we assume that students spend time between exams to recuperate from the last exam and prepare for the subsequent one, we may expect that the longer time students have between exams, the higher their performance in the later exam will be, on average. In this case, average exam performance in the second exam under schedule (b) should be higher than the performance in the second exam under schedule (a) ($p(2nd)_b > p(2nd)_a$). On the other hand, if students do not take advantage of the time between exams to prepare for the next exam but they rather get distracted and abandon studying efforts, we may anticipate a zero effect of the time between exams on subsequent exam performance, and the average performance in the last exam under schedule (b) should be no different than that under schedule (a). If we assume that the potential distraction during a longer time period between exams affects

focus and readiness to complete cognitive tasks in a negative way, then we may expect even a negative effect of the time between exams on subsequent exam performance. In that case, the average performance in the last exam under schedule (b) should be lower than the performance in the last exam under schedule (a). The positive association between time between exams and performance is documented in Pope and Fillmore (2015). The authors propose multiple explanations for the findings. One explanation is based on the cognitive load theory (CLT) and the fact that working memory is limited, thus more time between tasks allows for more recuperation from fatigue. A second explanation is based on last-minute preparation for exams. More time between exams allows for cramming. The third explanation is that when students have very little time between exams, they may focus only on a few. Their findings, however, come from a sample of students that self-select into taking particular exams (e.g. Advanced Placement (AP) exams). If higher achieving students choose to take AP exams, while lower-achieving students do not, the positive estimated effect of the time-between tasks on performance may be associated with the fact that higher-achieving students are also more likely or more capable of cramming at the last minute. On the other hand, lower-achieving students could be less willing or capable of cramming between exams. Thus, the evidence of Pope and Fillmore (2015) cannot be extrapolated to lower-achieving students.

The second channel of influence of exam scheduling on performance that we consider is the days lapsed since the beginning of the exam season, while holding constant the time between exams. This case is illustrated in the comparison between schedules (c) and (d) in Figure 2. Exam schedules (c) and (d) contain the same number of total exams, namely three, and the time between last and the next to last exam is the same in both schedules. The difference between schedules (c) and (d) is that schedule (d) spans a longer number of days than schedule (c). This is depicted as a longer time distance between the first and the second exam. If the time between the first and the second exam can be used to prepare for the third exam, then the performance in the third exam under schedule (d) should be higher than the performance in the third exam under schedule (c), on average ($p(3rd)_d > p(3rd)_c$). On the other hand, one may expect the average performance in the third exam of schedule (d) to be lower than that under schedule (c) if the time between the first and the second exam decreased a student's readiness to take the third exam. One possibility could be that if a student spends a longer time preparing for the second exam under schedule (d) compared to schedule (c), then it may be more difficult for the student to prepare the new material for the third exam, potentially due

to fatigue. The fatigue-based explanation predicts that exam productivity diminishes with additional exams. Cognitive fatigue has been documented in the psychology as well as in the economics literature (Webster et al., 1996; Jensen et al., 2013; Meijman, 1997). For a student to benefit from the time between earlier exams, as shown in the comparison between schedules (c) and (d), they must possess certain metacognitive attributes, such as time management, self-discipline, and multi-tasking skills. High-achieving students may be more likely to have these skills, as students with a history of low achievement have been documented to have problems meeting deadlines and are more prone to passively procrastinate, compared to high achieving students (Kármen et al., 2015; Metcalfe and Finn, 2013; zsoy et al., 2017). High-achieving students have been found to exhibit more self-regulated learning skills and time management (Zimmerman and Martinez-Pons, 1990; Eilam and Aharon, 2003; Nadinloyi et al., 2013). Time management is associated with decreased procrastination, priority setting, and completing more tasks. Time management may allow for studying for exams (Nadinloyi et al., 2013).

The third channel of influence of exam scheduling on exam performance that we explore is related to the order exams are taken. Consider exams schedules (e) and (f) in Figure 2. Exam schedules (e) and (f) have the same lengths and the time between the last and the next to last exam is the same under both schedules. The two exam schedules differ only in that under schedule (c) the last is the third exam taken, while under schedule (d) the last exam is the fourth exam taken. If we assume that, while holding everything else the same, having an additional exam earlier on might be associated with learning related to the subsequent exam, we may expect the performance in the last exam under schedule (f) to be higher than the performance in the last exam under schedule (e), on average $(p(4th)_f > p(3rd)_e)$. If no additional learning can be obtained from taking an additional exam, the performance in the last exam under schedule (f) should be no different from the performance in the last exam under schedule (e). The learning gain associated with taking additional exams may not be strictly related to material tested but also on test-taking strategies or experience in best studying practices as well as the best test-taking strategies. This learning gain has been investigated in the psychology literature as metacognitive accuracy. Metacognitive accuracy (bias scores and Gamma correlations) has been found to improve with practice (Kelemen et al., 2007; Finn and Metcalfe, 2007). Therefore, performance in cognitive tasks, such as exams, may improve as students take additional exams.

2.1 Hypothesis Testing

We now summarize the main hypotheses regarding the predicted sign of each exam scheduling effect on performance, based on the existing evidence for each channel discussed in the previous section.

Hypothesis 1: The more time between exams students have, the higher their performance in the subsequent exam will be on average, ceteris paribus. We hypothesize that as the time between exams is used -to a certain extent- productively in preparation for the subsequent exam and thus it will be positively associated with the student's performance in the next exam, on average.

Hypothesis 2: The higher the number of days since the very first exam, the lower the students' performance will be on average, ceteris paribus. We hypothesize that the more time (in days) a student spends in exam preparation and exam taking the more likely exhaustion is to prevail and decrease subsequent performance.

Hypothesis 3: The higher the number of exams a student has taken at a certain point in time, the higher their performance will be in the next exam on average, ceteris paribus. We posit that each exam may offer experience and knowledge that is positively associated with the student's performance in the next exam.

The hypotheses presented here rely on certain assumptions about how individuals spend their time between cognitive tasks, how prone their are to mental exhaustion, as well as their capacity in learning from practice. These assumptions may be more appropriate for certain types of cognitive tasks (e.g. exams in STEM fields) or individuals with certain characteristics (e.g. higher prior performance in a particular type of cognitive task, such as a language exam). Therefore, we test each of our hypotheses for different types of exams (STEM or non-STEM), for students with different levels of prior performance in each subject, as well as for students of different gender.

3 Exam Scheduling

In this section we describe the institutional setting of exam taking and exam scheduling. At the end of the school year, students take exams an average of 13 subjects in a period of 27 calendar days on average. Exams usually start one week after the last day of classes. End-of-course exams account for fifty percent of the Average Grade in a given grade. The remaining fifty percent of the Average Grade comes from midterm scores. Students need to achieve an Average Grade of at least 9.5 out of 20 to progress to the next grade. In the 11th

grade students must choose one of three Concentrations: Classics, Science, or Information Technology. Each Concentration consists of three compulsory classes. Students in the 10th grade take 14 classes, 11 of which have end-of-course exams. Students in the 11th grade take 17-18 classes, 15-16 of which have end-of-course exams, depending on additional electives, that are always tested after all other exams. The exam schedule for the 11th grade of the 2004-05 school year is shown in Figure 1 as an example.

Exam scheduling in orthogonal to the choice of classes as student choices do not affect scheduling and concentration electives are tested on the same day. By focusing on compulsory subjects only we are not imposing any bias as all compulsory subjects are tested on the same date for students of the same grade in a given year. Additionally, all non-compulsory subjects are tested on the same dates for every students, regardless of their elective.

Selection into concentration electives may be non-random and it may depend on academic strengths as well as other student characteristics (e.g. family background). We focus on nine compulsory subjects across grades.[1] Algebra, Geometry, Physics, Chemistry, Ancient Greek, Modern Greek, Greek Literature, English, History. Algebra, Geometry, Physics, and Chemistry are considered STEM[2] fields, while Ancient Greek, Modern Greek, Greek Literature, English, History are considered Non-STEM fields.

4 Data Description

4.1 School Database

The data set is drawn from a high school in central Greece. We follow students over two grades - 10th and 11th grade- and nine cohorts from 2001-02 to 2009-10. Our data set combines three types of data: enrollment, test scores, and test dates. First, for every student in each year we have student ID number, grade enrolled, classroom assignment, gender, year of birth, and complete course history. Second, for all students we have midterm scores and final exam scores for every subject taken. Third, we have data on the exact dates and times students took any end-of-course exam test.

The school year consists of two semesters: fall and spring. Students are assessed during each semester and receive a score in each subject. We average the two scores each student

[1] Our analysis excludes concentration electives, additional electives, as well as a compulsory course on religion.

[2] STEM is an acronym for fields of study related to Science,Technology Engineering, or Mathematics.

receives from the fall and spring semester in each subject to form a measure of performance in the specific subject prior to the cumulative end-of-the-year exam.

Data from these nine years allow for a comparison of exam performance under different exam schedules. Limiting the analysis to one school reflects a trade-off between homogeneity of scheduling components and more schools. Schools are not required to maintain a record of exam dates and times and retrieving this information for multiple years is challenging. The sampled school has average characteristics similar to the national average, based on data provided by the Hellenic Ministry of Education[3].

The main analysis draws on a nine-year and 1,024-student pooled dataset of 14,258 individual exam scores. Our outcome variable is the final exam score, standardized by subject and grade. Midterm and final exam scores are standardized at the subject and grade level.

4.2 Scheduling Variables Definitions

In this section we provide the definitions of the exam scheduling variables that we construct for our analysis.

Days Between Exams. —The measure of days between exams captures how many days intervene between subsequent exams. For example, the measure of days between exams is equal to one for a specific student and subject if the student took the exam for this subject the day following another exam that he or she took. The measure of days between exams takes the value zero for subjects that were tested first.

Days lapsed since the Exam Season started. —An additional dimension of exam scheduling captured in the data is the number of days since the first exam a student took. For example, the measure of days since the first exam is equal to one for a specific student and subject if the student took the exam for this subject on the day following the first exam in the given year. The measure of days since the first exam takes the value zero for subjects that were tested first.

Exam Order. —The data include detailed information on the timing of each exam, including the order in which each exam was taken for each student. For example, the measure of exam order is equal to one for a specific student and subject if the student took the exam for this subject before any other exam in a given year.

[3]For example, the average GPA in the sample school is 15.2 out of 20 with a standard deviation of 2.9, compared to the national average of 14.2 out of 20 and a standard deviation of 2.8. Similarly, the average percentage of females in our sample is 56 percent, compared to the national average of 57 percent.

Each scheduling variable described in this section captures a district channel through which exam scheduling affects performance. Our empirical approach allows us to disentangle the separate channels of scheduling influence on performance. To illustrate the channel the coefficient of each scheduling variable captures we provide a simple graphic with a comparison of test scheduling patterns for each channel of scheduling effect. The top panel in Figure 2 corresponds to the type I scheduling effect and is estimated by the coefficient of the "days between exams" variable. The middle panel of Figure 2 corresponds to the type II scheduling effect and is estimated by the coefficient of the "days since the exam season started" variable. The bottom panel of Figure 2 corresponds to the type III scheduling effect and is estimated by the coefficient of the "exam order" variable.

The basic statistics of the exam scheduling variables are reported in Table 1. Each student in the 10th grade takes on average 11 exams, while each student in the 11th grade takes on average 15 exams. The exams 10th grades take span 25 days on average, while the exams 11th graders take span 28 days on average. The average time distance between exams for a 10th or 11th grader is approximately 2 days.

4.3 Statistics of Exam Scheduling

The exam scheduling variables vary across years, grades, and subjects. There are two grades (10th and 11th) and nine cohorts. Therefore, each scheduling variable takes 18 values for each subject. We demonstrate the variation of exam scheduling in Table 4, 3, and 2. Table 2 shows how *Days between Exams* varies across subjects. Each entry in Table 2 shows how frequently the subject in that column was tested in the number of days since the previous exam shown in that row. The maximum number of days students have between exams is five days, as shown in the first column of Table 2. For example, Algebra was tested on the same day as the previous exam twice, one day after the previous exam zero times, and so on. History was tested on the same day as the previous exam four times, one day after the previous exam twice, and so on. At the bottom of Table 2 we report the mean and the standard deviation of the number of days lapsed since the previous exam each subject is tested. We observe considerable variation in the average time distance from the previous exam each subject is tested. On average, English and Modern Greek have the shortest average time distance from the previous exam than the other subjects, although we do not see any systematic differences in the testing pattern of STEM and non-STEM subjects in terms of the number of days lapsed since the previous exam.

Table 3 shows how *Days lapsed since the Exam Season started* varies across subjects.

Each entry in Table 3 shows how frequently the subject in that column was tested in the number of days since the beginning of the exam season shown in that row. Students take compulsory exams for a maximum duration of 30 days, therefore the maximum number of days since the first exam is 29 days, as shown in the first column of Table 3. For example, Algebra was tested on the first day twice, two days after the first exam three times, and so on. History was tested on the first day four times, two days after the first exam once, and so on. Physics was never tested more than 24 days since the first exam. Algebra was never tested more than 25 days since the first exam. At the bottom of Table 3 we report the mean and the standard deviation of the number of days lapsed since the first exam each subject is tested. We observe considerable variation in the average time distance from the first exam each subject is tested. On average, Algebra and Physics are tested earlier than the other subjects, although we do not see any systematic differences in the testing pattern of STEM and non-STEM subjects in terms of the number of days lapsed since the first exam.

Table 4 shows how the scheduling variables *Exam Order* varies across subjects. Each entry in Table 4 shows how frequently the subject in that column was tested in the order shown in that row. Students take 16 compulsory exams, therefore there are 16 places in the order of exams, shown in the first column of Table 4. For example, Algebra was tested first twice, second three times, third once, fourth once and so on. History was tested first four times, second once, third once, fourth zero times, and so on. Algebra and Physics were the only subjects that were never tested later than the 11th place in the order of exams. At the bottom of Table 4 we report the mean and the standard deviation of the place in the exam order each subject is tested. We observe considerable variation in the average order each subject is tested. On average, Algebra and Physics are tested earlier than the other subjects, although we do not see any systematic differences in the testing pattern of STEM and non-STEM subjects.

4.4 Statistics of Student Data

Table 5 shows descriptive statistics for the student data across the 2002-2010 cohorts. Across years, we observe 900 distinct students in the 10th and 836 distinct students in the 11th grade. 56 percent of the students are females. Students' average age is 16.41 years. The students have an average GPA of 15.23 out of 20. The average midterm score is 17.25, while the average final score is 13.27 out of 20. Only 2 percent of the students are retained in the same grade. Comparing the descriptive statistics across grades, we see that

there are no substantial differences between 10th and 11th graders characteristics. Detailed descriptive statistics for each cohort are reported in the Appendix in Tables 12, 13, and 14. In those 9 cohorts, we have data for a total of 1,024 students. In Tables 12, 13, and 14, we present average student characteristics for students in each school year. The various measures of student characteristics are substantively similar across cohorts and grades.

5 Empirical Strategy

We calculate the effects of exam scheduling on standardized exam performance in a straightforward manner. We exploit across year and grade variation in exam scheduling to identify three separate channels. We use a panel of nine compulsory subjects in 10th and 11th grade. Our outcome variable is exam score S of student i, in subject s, in grade g, and year t, standardized by subject and grade. We exclude from our analysis the first exam taken by each student. We regress the outcome variable for student i, in subject s, and grade g on the *order the exam was taken*, *days since first exam*, *days between exams*, standardized average midterm score M in subject s, day of the week fixed effect, grade $\times$ subject fixed effects, grade $\times$ year fixed effects, year fixed effects, a linear time-trend retained status, and gender. The average midterm score consists of the average of two scores, one for the fall and one for the spring semester. Controlling for the midterm score in each specific subject allows for precision in capturing a student's differential level of preparedness across subjects. The coefficient of the *days since first exam* can be interpreted as a fatigue effect, while the coefficient of the *exam order* captures a practice or learning effect associated with exam experience. The coefficient of the *days between exams* reflects the effect of recuperation time between tasks. Using variation across years and across subjects, we disentangle the practice effect from the fatigue and the recuperation effects by controlling for the *days since first exam*, the *exam order*, and the *days between exams* simultaneously. Specifically, we run the following regression:

$$\begin{aligned} S_{i,s,g,t} =& \alpha_0 + \alpha_{1c} Exam\ Order_{s,g,t} + \alpha_{2c} Days\ Since\ First_{s,g,t} + \alpha_{3c} Days\ Between_{s,g,t} \\ &+ \alpha_4 M_{i,s,g,t} + \alpha_5 X_{i,s,g,t} + \kappa_{sg} + \lambda_{gt} + \nu_t + \xi_g + \zeta_t + y_t + \eta_{i,s,g,t} \end{aligned} \tag{1}$$

Where $c \in \{stem, non-stem\}$. The two estimated α_1 coefficients, α_{1stem} and $\alpha_{1non-stem}$ reflect the average impact of exam order on performance in STEM and non-STEM subjects. Our specification allows for a comparison of the estimated scheduling effects on exam

performance in STEM and non-STEM subjects. Standards errors are corrected for clustering at the cohort × classroom level to allow for heteroskedasticity and serial correlation at that level, as students that learn in the same room in a given year may share some error patterns. Vector X captures student characteristics such as gender and a binary variable capturing retained status. We also control for age by including a full set of year of birth × cohort dummies in vector X. This provides for slightly higher precision compared to including age dummies or continuous variables for age and age squared. We also control for day of the week fixed effects in vector ζ. Subject × Grade fixed effects are controlled for in vector κ. Grade × Year -specific fixed effects are captured in vector λ. Year-specific fixed effects are reflected in vector ν. A linear time-trend is captured by y.

The identification stems from the randomization of the date of the exam of each subject from one year to the next as well as between 10th and 11th grade. By controlling for grade-by-year fixed effects, we rely on within grade-by-year and across subjects variations in exam timing. Based on this approach, we examine whether subject-to-subject changes in exam scheduling of the same subjects within the same grade and year are systematically associated with subject-to-subject differences in exam performance. The basic idea is to compare the outcomes of students from adjacent cohorts who have similar characteristics and face the same school environment, except for the fact that one cohort has a different exam schedule than the other due to purely random factors. The identification assumption is that performance in STEM subjects tested in a given scheduling pattern would be similar to the performance in STEM subjects tested in the same scheduling pattern in a different year for students of similar characteristics, with an analogous assumption of non-STEM subjects. One limitation of our analysis is that exams in different subjects may provide different practice effects to subsequent exams. If selection into concentrations is driven by student characteristics, then student characteristics may also influence the practice effects. Our estimates reflect the average scheduling effects and ignore differential effects from having taken exams in a different mix of subjects.

We explore non-linear scheduling effects on exam performance using the following specification.

$$\begin{aligned} S_{i,s,g,t} =& \alpha_0 + \alpha_{1c} Exam\ Order_{s,g,t} + \alpha_{2c} Days\ Since\ First_{s,g,t} + \alpha_{3c} Days\ Between_{s,g,t} \\ &+ \alpha_{4c} Exam\ Order^2_{s,g,t} + \alpha_{5c} Days\ Since\ First^2_{s,g,t} + \alpha_{6c} Days\ Between^2_{s,g,t} \\ &+ \alpha_7 M_{i,s,g,t} + \alpha_8 X_{i,s,g,t} + \kappa_{sg} + \lambda_{gt} + \nu_t + \xi_g + \zeta_t + y_t + \eta_{i,s,g,t} \end{aligned} \tag{2}$$

We also employ an alternative approach to investigate non-linear scheduling effects on exam performance. We break down each scheduling effect into binary variables capturing different levels of the underlying scheduling variables. We estimate the following equation.

$$\begin{aligned} S_{i,s,g,t} =& \alpha_0 + \alpha_{1c} Exam\ Order >= 6\ \&\ <= 9_{s,g,t} + \alpha_{2c} Exam\ Order > 9_{s,g,t} \\ &+ \alpha_{3c} Days\ Between\ Exams = 3_{s,g,t} + \alpha_{4c} Days\ Between\ Exams > 3_{s,g,t} \\ &+ \alpha_{5c} Days\ Since\ First >= 10\ \&\ <= 19_{s,g,t} + \alpha_{6c} Days\ Since\ First > 19_{s,g,t} \\ &+ \alpha_7 M_{i,s,g,t} + \alpha_8 X_{i,s,g,t} + \kappa_{sg} + \lambda_{gt} + \nu_t + \xi_g + \zeta_t + y_t + \eta_{i,s,g,t} \end{aligned} \tag{3}$$

The *Exam Order* variable, which is associated with Scheduling Effect III, is broken down to two binary variables; one capturing exams with order between six and nine, and a second one capturing exams with order above nine. A similar break-down to binary variables is applied to the other scheduling variables. The *Days Since First Exam* variable, which is associated with Scheduling Effect II, is broken down to two binary variables; one capturing exams taken between 10 and 19 days since the first exam, and a second one capturing exams taken more than 19 days since the first exam. The *Days Between Exams* variable, which is associated with Scheduling Effect I, is broken down to two binary variables; one capturing exams taken three days (the mode of this variable) since the last exam, and a second one capturing exams taken more than three days since the last exam. The comparison group is exams taken within the first nine days from the first exam, up to the fifth place of exam order, and not later than two days from the previous exam. This condition corresponds to roughly 15 percent of the exams in our data set.

We extend the baseline specification 1 to explore differential effects for males and females and compare scheduling effects by gender × subject type, STEM ad non-STEM.

$$\begin{aligned} S_{i,s,g,t} =& \alpha_0 + \alpha_{1cf} Exam\ Order_{s,g,t} + \alpha_{2cf} Days\ Since\ First_{s,g,t} + \alpha_{3cf} Days\ Between_{s,g,t} \\ &+ \alpha_4 M_{i,s,g,t} + \alpha_5 X_{i,s,g,t} + \kappa_{sg} + \lambda_{gt} + \nu_t + \xi_g + \zeta_t + y_t + \eta_{i,s,g,t} \end{aligned} \tag{4}$$

Where $f \in \{male, female\}$ indicates the student's gender. We use the estimates from the model above to test whether the scheduling effects are statistically different between males and females.

In an extension to these specification, we also consider how the impacts of exam scheduling vary by prior performance, proxied by standardized midterm score in each subject.

Extending equation 1, we interact the exam scheduling variables with quantiles of prior midterm performance:

$$\begin{aligned} S_{i,s,g,t} =& \alpha_0 + \alpha_{1c} Exam\ Order_{s,g,t} + \alpha_{2c} Days\ Since\ First_{s,g,t} + \alpha_{3c} Days\ Between_{s,g,t} \\ &+ \alpha_{4c} Exam\ Order_{s,g,t} \times M_{i,s,g,t} + \alpha_{5c} Days\ Since\ First_{s,g,t} \times M_{i,s,g,t} \\ &+ \alpha_{6c} Days\ Between_{s,g,t} \times M_{i,s,g,t} + \alpha_7 M_{i,s,g,t} + \alpha_8 X_{i,s,g,t} + \kappa_{sg} + \lambda_{gt} + \nu_t + \xi_g \\ &+ \zeta_t + y_t + \eta_{i,s,g,t} \end{aligned} \tag{5}$$

We also explore differential effects by prior student performance using equation 6. The difference between equations 6 and 5 is that equation 6 allows us to estimate each scheduling effect for different quantiles of midterm performance.

$$\begin{aligned} S_{i,s,g,t} =& \alpha_0 + \alpha_{1cp} Exam\ Order_{s,g,t} + \alpha_{2cp} Days\ Since\ First_{s,g,t} + \alpha_{3cp} Days\ Between_{s,g,t} \\ &+ \alpha_4 M_{i,s,g,t} + \alpha_5 X_{i,s,g,t} + \kappa_{sg} + \lambda_{gt} + \nu_t + \xi_g + \zeta_t + y_t + \eta_{i,s,g,t} \end{aligned} \tag{6}$$

Where $p \in \{1, 2, 3, 4\}$ indicates the within grade and year quantile of student i based on his/her midterm score in subject s. Quantile 4 represents students in top 25 percent of midterm performance distribution. For each quantile of midterm performance, the estimated scheduling effects $\alpha_{1p}, \alpha_{2p}, \alpha_{3p}$ reflect the impact of exam order, days since the very first exam, and days between exams, respectively, on exam performance of students in that quantile. We can then explore whether students of different quantiles of midterm performance have different scheduling impacts on their performance in STEM and non-STEM subjects. The analysis by quantile of midterm performance requires a stronger identification assumption than equation 1, that midterm performance is not otherwise associated with exam scheduling patterns, which we discuss earlier along with showing differences in midterm performance by year and subject.

6 Results and Discussion

6.1 Average Effects for STEM and non-STEM subjects

We start the presentation of our findings with the average scheduling effects for exams in STEM and non-STEM subjects.

Scheduling Effect I

Scheduling Effect I is found to be statistically significant only for non-STEM subjects. An additional day between two exams decreases exam performance in the subsequent non-STEM exam by 0.012 of a standard deviation. A student's exam performance in a non-STEM two days after the previous exam, the average gap between any two exams in the data set, is found to be 0.024 of standard deviation lower than the previous exam. The estimated Scheduling Effect I for STEM subjects is not statistically significant. Although our overall results dispute our hypothesis that the average student uses their time between exams productively in order to prepare for the subsequent exams, a potential explanation may be found in the part of the psychology literature that suggests that performance in short-memory-relying tasks decreases as time between tasks increases (Baddeley, 2003).

Scheduling Effect II

Scheduling Effect II is found to be statistically significant, confirming our hypothesis 2 on the potential existence of a fatigue effect, only for STEM subjects. In particular, an additional day since the first exam a student took decreases their performance in the subsequent STEM exam by 0.006 of a standard deviation. For example, a student who takes an exam 24 days since their first exam - the average number of days compulsory exams span - experiences a decrease in the subsequent performance in a STEM subject by 0.14 of a standard deviation. On the contrary, the estimated Scheduling Effect II for non-STEM subjects is not statistically significant, indicating that the underlying mechanism, potentially relating to cognitive fatigue, is relevant only in STEM subjects. Our finding suggests a higher prevalence of cognitive fatigue in analytic reasoning -intensive tasks.

Scheduling Effect III

Similar to Scheduling Effect II, Scheduling Effect III is also statistically significant, confirming our hypothesis 3 on a the existence of a practice or warm-up effect, only for STEM subjects. Specifically, taking an exam one place later in the exam order increases exam performance in STEM subjects by 0.016 of a standard deviation. For example, a student's performance in the 11th exam they take is estimated to be 0.18 of standard deviation higher than their performance in the first exam, ceteris paribus. The estimated Scheduling Effect III for non-STEM subjects is not statistically significant. This indicates that there is a performance gain for STEM subjects associated with taking exams later in the exam schedule. We call this effect, warm-up effect and it may be worth comparing it to the marginal effects of other educational inputs. For example, a movement of one place in the order of exams in the schedule has the equivalent benefit as raising teacher quality by roughly one tenth of a standard deviation (Carrell et al., 2011b). Our finding on the effect of

exam order on performance supports the association of the potential underlying mechanism of improvement of metacognitive accuracy with cognitive practice on the performance in tasks that are intensive in analytic reasoning, such as exams in STEM subjects.

6.2 Non-Linear Effects for STEM and non-STEM subjects

We found economically and statistically substantial non-linear effects of exam scheduling on performance. Column 2 of Table 6 shows the estimated non-linear effects of three distinct channels of exam scheduling on performance. Scheduling Effect III is found to have non-linear effects only on the exam performance of STEM subjects. The positive coefficient on the squared variable associated with Scheduling Effect III reveals the upward curvature of the effect of the underlying mechanism. Exams in STEM subjects taken at a later order in the exam schedule are associated with increasingly higher performance, while controlling for other influences. Scheduling Effect I is found to have non-linear effects only on the exam performance of non-STEM subjects. The positive coefficient on the squared variable associated with Scheduling Effect I reveals the downward curvature of the effect of the underlying mechanism. Exams in non-STEM subjects taken farther in days from the previous exam are associated with decreasingly lower performance, while controlling for other influences. On the contrary to the other Scheduling Effects, Scheduling Effect II is found not to have non-linear effects in either STEM or non-STEM subjects.

Table 7 shows non-linear effects by breaking the variables reflecting the Scheduling Effects into bins. We explore the existence of non-linear scheduling effects in comparison to the exams taken within the first nine days from the first exam, up to the fifth place of exam order, and not later than two days from the previous exam. We find that for non-STEM subjects, the dummy capturing exams that are taken exactly three days after the previous exam are associated with significantly lower performance compared to exam taken less than three days after the previous exam. At the same time, non-STEM exams taken more than three days after the previous exam are associated with performance comparable to that in exams taken less than three days after the previous exam. For STEM subjects, exam taken either exactly three days after the previous exam or more than three days after the previous exam are found to be associated with performance comparable to the exams taken less than three days after the previous exam, suggesting that non-linear effects of the time between exams are not present in STEM exams.

The second panel in Table 7 shows our estimates on non-linear effects of Scheduling Effect II. For STEM subjects, exams taken between 10 and 19 days since the first exam are

associated with performance comparable to that of exams taken less than 10 days since the first exam, but this is not the case for the next category. STEM exams taken more than 19 days since the first exam are associated with performance significantly lower than the performance in exams taken earlier. For non-STEM subjects, exam taken either between 10 and 19 days since the first exam or more than 10 days since the first exam are found to be associated with performance comparable to the exams taken less than 10 days since the first exam, suggesting that non-linear effects of the time since the first exam are not present in non-STEM exams.

The third panel in Table 7 shows our estimates on non-linear effects of Scheduling Effect III. For STEM subjects, exams taken between the 6th and the 9th place of order in the exam schedule are associated with performance comparable to that of exams at a lower place of order, but this is not the case for the next category. STEM exams taken at the 10th or higher place of order in the exam schedule are associated with performance significantly higher than the performance in exams at lower places of order, suggesting the existence of significant non-linear effects in Scheduling Effect III. For non-STEM subjects, exam taken either between the 6th and 9th place of order in the schedule or at the 10th or higher place of order in the schedule are found to be associated with performance comparable to the exams at lower places of order, suggesting that non-linear effects of exam order are not present in non-STEM exams.

6.3 Differential Effects by Student Gender

Table 10 shows heterogeneous effects of exam scheduling on performance by gender. We compare scheduling effects in STEM and non-STEM exams between boys and girls. Because neither males nor females were omitted from our model, the estimated effects are interpreted in comparison to the average effects across both groups. Boys are more responsive to all types of scheduling effects. We find that scheduling effect I in STEM exams is significantly higher for boys than for girls. In particular, one additional day between exams improves the subsequent performance of boys by 0.07 of standard deviation more than for girls. One additional day between exams seems to have a negligible effect on the subsequent exam performance of girls. The estimated Scheduling Effect I for non-STEM exams, although negative and significant for either boys or girls, it is not found to be significantly different across genders. The exam fatigue effect which is captured by our Scheduling effect II does not seem to differ significantly across genders, although girls seem to be less sensitive to that influence of exam scheduling on performance in STEM exams.

Girls experience a roughly 45 percent lower fatigue effects in STEM exams than boys. The estimated Scheduling Effect II is not significant in non-STEM subjects for either boys or girls. The warm-up effect which is reflected in our Scheduling Effect III differs significantly between boys and girls. In particular, boys have roughly 75 percent higher warm-up effect compared to girls. Having taken an additional exam earlier in the exam schedule improves the subsequent performance of boys in STEM exams by 0.023 of a standard deviation. This warm-up effect is 2.5 times bigger than the fatigue effect boys experience in STEM subjects. The estimated Scheduling Effect III is not significant in non-STEM subjects for either boys or girls.

6.4 Differential Effects by Student Prior Performance

We run specification 5 and we are interested in the coefficient of the triple interaction between each scheduling effect, the STEM binary variable, and a continuous variable capturing the standardized prior performance of each student in each subject. The results are shown in Table 8. For Scheduling Effect I the triple interaction is positive and statistically different from zero for non-STEM exam, whereas it is negative and significant for STEM. This indicates that the gap of the effect of an additional day between exams between STEM and non-STEM subjects decreases with prior performance.

The estimated coefficient of the triple interaction for Scheduling Effect II in non-STEM subjects is zero, while the coefficient of the triple interaction for STEM subjects is positive and significantly different from zero. Higher achieving students benefit more from an additional day since their very first exam for STEM subjects, while this is not the case of exam in non-STEM subjects.

The estimated coefficient of the triple interaction for Scheduling Effect III in non-STEM subjects is zero, while the coefficient of the triple interaction for STEM subjects is negative and significantly significant. Higher achieving students benefit less from taking a STEM exam one additional place later in the order of exams in the schedule, whereas for non-STEM subjects, the exam order does not seem to play any important role.

We are interested in comparing the scheduling effects each part of the prior performance distribution experiences. Table 9 shows heterogeneous effects of exam scheduling on performance by student prior performance. It is important to note that no quantile of midterm performance was omitted, allowing for an interpretation of the estimated effects in comparison to average effect.

Table 9 panel: Scheduling Effect I shows differential slopes between the variable cap-

turing Scheduling Effect I and exam performance in STEM and non-STEM subjects across different quantiles of prior performance. The lowest quantile of prior performance (quantile 1), an additional day between exams decreases performance in non-STEM subjects by 0.06 of a standard deviation. We notice that the effect is increasing with prior performance. In particular, an additional day between exams for the second quantile of prior performance harms final exam performance in non-STEM subjects even less, while it has an impact close to zero for the third quantile of prior performance. The effect of an additional day between exams improves the performance of the highest quantile of prior performance (quantile 4) in non-STEM subjects by 0.027 of standard deviation, confirming Hypothesis 1 only for non-STEM subjects.

For STEM subjects we do not find significant scheduling effect I for quantiles of prior performance 1, 2, and 3. However, scheduling effect I is negative and significant for the highest quantile of prior performance in STEM exams. Scheduling effect I is associated with a decrease in the exam performance in STEM subjects for the highest performing student of 0.014 of standard deviation. We plot these marginal scheduling effects III on the top graph of Figure 4.

Scheduling effect II is found to be small and occasionally significant for non-STEM subjects, while scheduling effect II in STEM exams is increasing significantly with prior performance. In particular, an additional day since the beginning of the exam season decreases the performance of quantile 1 (bottom quantile) by 0.021 of standard deviation. Although scheduling effect II for quantile 2 is zero, the effect becomes positive and statistically different from zero for quantiles 3 and 4. Specifically, an additional day since the beginning of the exam season improves quantile 3 and 4's performance in STEM exams by 0.012 and 0.024 of a standard deviation, respectively. The marginal scheduling effects II by quantile of midterm performance are shown in the middle graph of Figure 4.

The pattern of estimated coefficients across quantiles of prior performance for scheduling effect III is very different from that of scheduling effect II. Although the scheduling effect III on non-STEM subjects is small and occasionally significant, the effect on STEM subjects is decreasing with prior performance. An additional place in the order of exams in the schedule increases the performance in STEM exams for quantile 1 of prior performance by 0.052 of a standard deviation. Although the effect on STEM subjects for quantile 2 is zero, taking a STEM-related exam one additional place later in the order of exams in the schedule is associated with a decrease in quantile 3 and 4's performance by 0.03 and 0.06 of a standard deviation, respectively. These marginal effects are graphically depicted in

Figure 4.

6.5 Robustness Checks

We verify the robustness of our estimates to several changes in model specification with results shown in Table 11. All specifications include a full set of individual controls and course by year by grade fixed effects. Column 1 shows our estimates with the inclusion of student fixed effects. Column 2 shows results when including elective courses to address concerns of selection into courses and consequently selection into exam schedules. In column 3, our model excludes exams tested on the same day because of concerns that same-day exams may affect each other differently than exams that are farther apart. The estimates from our robustness specifications are quantitatively similar to those from our main specification, and provide strong evidence that the results are not driven by inconsistencies in the data.

7 Conclusion

Workers and managers are interested in finding ways to improve task productivity. While psychologists have studied the effects of cognitive fatigue and cognitive learning on task performance, the existing literature has not simultaneously measured and compared those effects. Additionally, the literature has not -until now- disentangled the different channels through which task scheduling may affect performance. Researchers have attempted to answer the question of how exam scheduling affects performance; however, to this point unraveling the causal effects of exam scheduling on student achievement has been difficult due to issues related to self-selection and measurement error.

This study identifies the different causal channels of exam scheduling on student academic achievement using data on every exam taken by nine cohorts of high school students to take advantage of the randomized assignment of exam dates to courses. Randomized exam dates, mandatory attendance, stable curriculum and assessment protocols, along with extensive background data on students, allows us to examine how exam scheduling affects student achievement without worrying about confounding factors or self-selection issues that bias existing estimates.

Exploiting variation in exam schedules across grades and years, we disentangle the three channels of scheduling effects on academic performance across STEM and non-STEM subjects, as well as at different parts of the ability distribution. We find that the time between exams (Scheduling Effect I) has, on average, negative and significant effect on

students' productivity in non-STEM subjects, while the effect for STEM subjects is not statistically different from zero. This suggests that exam productivity in more memory-intensive subjects such as non-STEM ones (Language, History etc) are more responsive to having one additional day of gap since the last exam. This effect differs across the ability distribution. In particular, students of high prior performance exhibit positive and significant effects of Scheduling Effect I on their performance in non-STEM subjects. At the same time, the Scheduling Effect I is negative and significant for students of lower prior performance in non-STEM subjects. This suggests that high achieving students, who may possess stronger cognitive meta-memory, may be more capable of cramming and thus more likely for their performance in non-STEM subjects to benefit from an additional day since their last exam. An additional day between exams for low achievers is likely to distract them more from studying efforts.

We also find that time distance since the first exam (Scheduling Effect II) has negative and significant effect on the performance in STEM subjects, while the effect for non-STEM subjects is not significant. During the exam season students are likely to change their studying, sleeping, or eating habits; essentially adopting an exam mode schedule. We hypothesize that the longer a student stays in exam mode, the harder it becomes to maintain focus and discipline to one's exam preparation strategies. We call the progressive loss of focus and relaxation of preparation efforts exam fatigue. Thus, during the exam season, students' performance in exams later in the schedule is likely to be lower due to exam fatigue, all else being equal. STEM subjects may rely less on working memory and more on longer metamemory, rendering performance in those subjects more conducive to fatigue. Scheduling Effect II is found to be positive and significant on STEM subjects for students of higher prior performance, whereas students of lower prior performance exhibit negative and significant Scheduling II effects on their performance in STEM subjects. This could result from lower achieving students being more prone to fatigue, as their studying skills may be less developed than those of the high achieving students. Higher achieving students may have studying routines that allow them to overcome fatigue and take advantage of longer gaps earlier in the exam schedule.

We also find positive and significant effect of the number of previously completed exams (Scheduling Effect III) on subsequent performance in STEM subjects. On the contrary, non-STEM subjects exhibit a Scheduling Effect III that is not statistically different from zero, on average. Having taken additional exams prior to a new one may provide some learning or practice to the students. Learning from exams can pertain to time manage-

ment, stress management, as well as preparation strategies. Thus, in a sequence of exams, students may be more likely to perform better in later exams due to practice or experience, all else being equal. Practice from previous exams improves analytical thinking more than it does cognitive meta-memory, suggesting that the warm-up (Scheduling Effect III) is higher for STEM subjects compared to non-STEM ones. Similarly, to the other scheduling effects, there are heterogeneous effects across prior performance. Low achieving students experience positive and significant scheduling III effects, while the corresponding effect is negative and significant for high achievers. High achievers, whose meta-cognition is already high, may experience lower cognitive returns to practice, compared to low achievers, while additional practice may induce fatigue. Moreover, we find that exam productivity in STEM increases faster for boys than it does for girls as they take additional exams (warm-up).

Our findings have important implications for education policy and task management; administrators aiming to improve student achievement should consider the potential benefits of delaying important exams. A movement of one place in the order of exams in the schedule has the equivalent benefit as raising teacher quality by roughly one tenth of a standard deviation. Hence, later exam dates for important tests may be a cost-effective way to improve test outcomes for adolescents, particularly in STEM fields. Furthermore, manipulating the exam schedule may affect the gender gap in STEM-related performance and potentially enrollment in STEM fields.

References

Ablard, K. E. and R. E. Lipschultz (1998). Self-regulated learning in high-achieving students: Relations to advanced reasoning, achievement goals, and gender. *Journal of Educational Psychology 90*(1), 94.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature reviews neuroscience 4*(10), 829.

Bjork, R. A., J. Dunlosky, and N. Kornell (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology 64*, 417–444.

Blagrove, M., C. Alexander, and J. A. Horne (1995). The effects of chronic sleep reduction on the performance of cognitive tasks sensitive to sleep deprivation. *Applied Cognitive Psychology 9*(1), 21–40.

Boksem, M. A., T. F. Meijman, and M. M. Lorist (2005). Effects of mental fatigue on attention: An erp study. *Cognitive Brain Research 25*(1), 107 – 116.

Buser, T. and N. Peter (2012, Dec). Multitasking. *Experimental Economics 15*(4), 641–655.

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

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

Chambers, C., T. D. Noakes, E. V. Lambert, and M. I. Lambert (1998). Time course of recovery of vertical jump height and heart rate versus running speed after a 90-km foot race. *Journal of Sports Sciences 16*(7), 645–651.

Coviello, D., A. Ichino, and N. Persico (2014, February). Time allocation and task juggling. *American Economic Review 104*(2), 609–23.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources 42*(3), 528–554.

Di Pietro, G. (2013). Exam scheduling and student performance. *Bulletin of Economic Research 65*(1), 65–81.

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

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

Eilam, B. and I. Aharon (2003). Students planning in the process of self-regulated learning. *Contemporary Educational Psychology 28*(3), 304 – 334.

Else-Quest, N. M., J. S. Hyde, and M. C. Linn (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychological bulletin 136*(1), 103.

Fennema, E. and J. Sherman (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. *American educational research journal 14*(1), 51–71.

Finn, B. and J. Metcalfe (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition 33*(1), 238.

Fryer Jr, R. G. and S. D. Levitt (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics 2*(2), 210–40.

Halpern, D. F. (2004). A cognitive-process taxonomy for sex differences in cognitive abilities. *Current directions in psychological science 13*(4), 135–139.

Halpern, D. F. (2013). *Sex differences in cognitive abilities*. Psychology press.

Hockey, G. R. J. and F. Earle (2006). Control over the scheduling of simulated office work reduces the impact of workload on mental fatigue and task performance. *Journal of experimental psychology: applied 12*(1), 50.

Hyde, J. S., E. Fennema, and S. J. Lamon (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological bulletin 107*(2), 139.

Hyde, J. S., S. M. Lindberg, M. C. Linn, A. B. Ellis, and C. C. Williams (2008). Gender similarities characterize math performance. *Science 321*(5888), 494–495.

Hyde, J. S. and M. C. Linn (1988). Gender differences in verbal ability: A meta-analysis. *Psychological bulletin 104*(1), 53.

Jensen, J. L., D. A. Berry, and T. A. Kummer (2013, 08). Investigating the effects of exam length on performance and cognitive fatigue. *PLOS ONE 8*(8), 1–9.

Kármen, D., S. Kinga, M. Edit, F. Susana, K. J. Kinga, and J. Réka (2015). Associations between academic performance, academic attitudes, and procrastination in a sample of undergraduate students attending different educational forms. *Procedia-Social and Behavioral Sciences 187*, 45–49.

Kelemen, W. L., R. G. Winningham, and C. A. W. III (2007). Repeated testing sessions and scholastic aptitude in college students metacognitive accuracy. *European Journal of Cognitive Psychology 19*(4-5), 689–717.

Law, D. J., J. W. Pellegrino, and E. B. Hunt (1993). Comparing the tortoise and the hare: Gender differences and experience in dynamic spatial reasoning tasks. *Psychological Science 4*(1), 35–40.

Lorist, M. M., M. Klein, S. Nieuwenhuis, R. De Jong, G. Mulder, and T. F. Meijman (2000). Mental fatigue and task control: planning and preparation. *Psychophysiology 37*(5), 614–625.

Meijman, T. F. (1997). Mental fatigue and the efficiency of information processing in relation to work times. *International Journal of Industrial Ergonomics 20*(1), 31 – 38.

Metcalfe, J. and B. Finn (2013, Apr). Metacognition and control of study choice in children. *Metacognition and Learning 8*(1), 19–46.

Nadinloyi, K. B., N. Hajloo, N. S. Garamaleki, and H. Sadeghi (2013). The study efficacy of time management training on increase academic time management of students. *Procedia - Social and Behavioral Sciences 84*, 134 – 138. The 3rd World Conference on Psychology, Counseling and Guidance, WCPCG-2012.

Nosek, B. A., F. L. Smyth, N. Sriram, N. M. Lindner, T. Devos, A. Ayala, Y. Bar-Anan, R. Bergh, H. Cai, K. Gonsalkorale, et al. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences 106*(26), 10593–10597.

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

Pope, N. G. (2016). How the time of day affects productivity: Evidence from school schedules. *The Review of Economics and Statistics 98*(1), 1–11.

Rohrer, D. (2009). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education*, 4–17.

Rohrer, D. and K. Taylor (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology 20*(9), 1209–1224.

Rohrer, D. and K. Taylor (2007). The shuffling of mathematics problems improves learning. *Instructional Science 35*(6), 481–498.

Taylor, K. and D. Rohrer (2010). The effects of interleaved practice. *Applied Cognitive Psychology 24*(6), 837–848.

US San Diego College Health Association (2008). How to Prevent Sleep Deprivation During Finals Week.

van der Linden, D., M. Frese, and T. F. Meijman (2003). Mental fatigue and the control of cognitive processes: effects on perseveration and planning. *Acta Psychologica 113*(1), 45 – 65.

Webster, D. M., L. Richter, and A. W. Kruglanski (1996). On leaping to conclusions when feeling tired: Mental fatigue effects on impressional primacy. *Journal of Experimental Social Psychology 32*(2), 181 – 195.

Wolfson, A. R. and M. A. Carskadon (2003, January). Understanding adolescent's sleep patterns and school performance: a critical appraisal. *Sleep Medicine Reviews 7*(6), 491–506.

Zimmerman, B. J. and M. Martinez-Pons (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology 82(1)*, 51–59.

zsoy, G., A. Memi, and T. Temur (2017). Metacognition, study habits and attitudes. *International Electronic Journal of Elementary Education 2*(1), 154–166.

Table 1: Descriptive Statistics of Exam Scheduling Variables

Variable	Obs	Mean	Std. Dev.	Min	Max
Panel A: 10th Graders					
Number of Exams	900	11.39	0.54	11	13
Days since first exam	900	24.91	3.05	21	31
Average Days Between Exams	900	2.18	0.20	1.75	2.58
Panel B: 11th Graders					
Number of Exams	836	15.33	0.47	15	16
Days since first exam	836	28.45	0.77	27	30
Average Days Between Exams	836	1.85	0.07	1.75	2.00
Panel C: All students					
Number of Exams	1736	13.29	2.03	11	16
Days since first exam	1736	26.61	2.87	21	31
Average Days Between Exams	1736	2.02	0.22	1.75	2.58

Table 2: How does Days Between Exams vary across subjects?

Days Between Exams	Ancient Greek	Literature	Modern Greek	History	Algebra	Geometry	Physics	Chemistry	English
0	5	0	2	4	2	0	2	0	0
1	0	9	9	2	0	0	0	6	13
2	6	6	4	5	7	7	8	10	4
3	4	3	3	5	5	9	4	1	1
4	3	0	0	2	2	0	2	1	0
5	0	0	0	0	2	2	2	0	0
Total	18	18	18	18	18	18	18	18	18
Mean	2.00	1.67	1.44	1.94	2.61	2.83	2.56	1.83	1.33
SD	1.46	0.77	0.92	1.35	1.38	0.92	1.38	0.79	0.59

Table 3: How does Days Since First Exam vary across subjects?

Days since Exam Start	Ancient Greek	Literature	Modern Greek	History	Algebra	Geometry	Physics	Chemistry	English
0	4	0	2	4	2	0	2	0	0
2	0	3	0	1	3	0	2	2	1
3	0	1	0	0	0	0	0	0	0
4	1	0	1	0	0	1	1	1	0
5	2	0	0	1	0	1	1	0	0
7	0	0	1	0	2	1	3	1	2
9	4	2	0	0	0	2	2	2	0
10	0	0	0	0	0	0	0	0	0
11	0	0	0	1	0	1	0	1	0
12	0	0	0	0	3	1	0	0	0
13	0	0	0	1	0	0	1	0	0
14	0	0	3	1	3	2	2	1	0
15	0	2	1	0	0	1	0	0	1
16	1	1	0	2	0	1	1	0	3
17	1	1	0	0	0	1	0	0	1
18	0	0	2	1	1	2	0	0	0
19	1	0	1	1	0	0	0	0	1
20	1	0	0	0	1	0	1	0	1
21	0	4	2	1	2	0	1	1	1
22	2	0	1	0	0	0	0	3	2
23	0	1	1	0	0	0	0	2	2
24	0	1	2	1	0	1	1	1	1
25	0	0	1	0	1	0	0	0	0
26	0	0	0	0	0	1	0	1	0
27	0	0	0	1	0	2	0	1	1
28	1	2	0	1	0	0	0	0	1
29	0	0	0	1	0	0	0	1	0
Total	18	18	18	18	18	18	18	18	18
Mean	10.78	15.39	15.72	13.50	11.28	15.17	9.67	16.50	18.06
SD	8.93	8.92	8.08	10.30	7.94	7.26	7.32	9.15	7.00

Table 4: How does Exam Order vary across subjects?

Exam Order	Ancient Greek	Literature	Modern Greek	History	Algebra	Geometry	Physics	Chemistry	English
1	4	0	2	4	2	0	2	0	0
2	0	4	0	1	3	0	2	2	1
3	3	0	1	1	1	3	2	1	0
4	1	0	1	0	1	0	4	2	2
5	3	2	0	2	1	2	3	1	0
6	0	1	1	0	3	4	0	1	0
7	0	0	3	1	1	2	1	1	0
8	1	3	1	2	1	1	2	0	4
9	3	0	1	1	1	2	1	0	1
10	2	2	2	2	2	0	0	1	3
11	0	3	2	1	2	0	1	4	3
12	0	1	1	0	0	1	0	2	2
13	0	0	3	0	0	2	0	0	0
14	0	1	0	0	0	1	0	2	1
15	1	0	0	3	0	0	0	1	1
16	0	1	0	0	0	0	0	0	0
Total	18	18	18	18	18	18	18	18	18
Mean	5.67	7.94	8.11	7.06	5.78	7.50	4.78	8.56	9.28
SD	4.03	4.32	3.94	5.00	3.49	3.52	2.82	4.41	3.39

Table 5: Summary Statistics

	Female	Age	GPA	Midterm Score	Final Exam Score	Retained
Panel A: 10th Graders						
Mean	0.55	15.94	15.40	17.21	13.47	0.01
SD	0.50	0.38	2.72	1.79	3.87	0.10
N	900	900	890	900	900	900
Panel B: 11th Graders						
Mean	0.57	16.92	15.03	17.29	13.06	0.04
SD	0.50	0.51	2.99	1.83	4.10	0.19
N	836	836	804	836	836	836
Panel C: All students						
Mean	0.56	16.41	15.23	17.25	13.27	0.02
SD	0.50	0.66	2.85	1.81	3.99	0.15
N	1736	1736	1694	1736	1736	1736

Table 6: The Effect of Exam Timing on Performance

VARIABLES	(1)	(2)	(3)	(4)
Scheduling Effect III for non-STEM	-0.002		-0.016	
	(0.004)		(0.015)	
Scheduling Effect III for STEM	0.016***		-0.024	
	(0.005)		(0.017)	
Scheduling Effect III for non-STEM^2			0.000	
			(0.001)	
Scheduling Effect III for STEM^2			0.002**	
			(0.001)	
Scheduling Effect II for non-STEM	0.002		0.005	
	(0.002)		(0.006)	
Scheduling Effect II for STEM	-0.006***		0.004	
	(0.002)		(0.006)	
Scheduling Effect II for non-STEM^2			-0.000	
			(0.000)	
Scheduling Effect II for STEM^2			-0.000	
			(0.000)	
Scheduling Effect I for non-STEM	-0.012***		-0.038***	
	(0.004)		(0.011)	
Scheduling Effect I for STEM	0.002		0.016	
	(0.003)		(0.023)	
Scheduling Effect I for non-STEM^2			0.006**	
			(0.002)	
Scheduling Effect I for STEM^2			-0.003	
			(0.003)	
STEM × Scheduling Effect III		0.019***		-0.018
		(0.005)		(0.018)
STEM × Scheduling Effect II		-0.007***		0.002
		(0.002)		(0.007)
STEM × Scheduling Effect I		0.002		0.026
		(0.003)		(0.023)
STEM × Scheduling Effect III^2				0.002*
				(0.001)
STEM × Scheduling Effect II^2				-0.000
				(0.000)
STEM × Scheduling Effect I^2				-0.004
				(0.003)
Observations	14,258	14,258	14,258	14,258
R-squared	0.985	0.985	0.985	0.985

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.10

Note: The dependent variable in each specification is the normalized final exam score at the subject and grade level. Cluster-robust standard errors at the classroom by year level are reported in parentheses. All specifications include: grade fixed effects, year fixed effects, grade by year fixed effects, subject by grade fixed effects, day of the week fixed effects, a full set of birth year by cohort fixed effects, and individual controls. Individual controls include indicators for students who are female, and are retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 7: Non-Linear Effects of Exam Scheduling on Performance

STEM			Non-STEM		
Scheduling Effect I					
3 days	>3 days	Difference	3 days	>3 days	Difference
-0.013	-0.006	0.007	-0.040***	-0.014	0.026**
(0.009)	(0.008)	(0.008)	(0.010)	(0.013)	(0.013)
Scheduling Effect II					
10-19 days	>19 days	Difference	10-19 days	>19 days	Difference
-0.016	-0.045**	-0.029**	0.000	0.009	0.009
(0.013)	(0.021)	(0.013)	(0.015)	(0.026)	(0.014)
Scheduling Effect III					
6-9th place	$>9th$ place	Difference	6-9th place	$>9th$ place	Difference
0.026	0.055***	0.029**	0.014	0.010	-0.005
(0.014)	(0.021)	(0.012)	(0.014)	(0.022)	(0.013)

Note: sample: 14,258 obs. The dependent variable is the normalized final exam score at the subject and grade level. Cluster-robust standard errors at the classroom by year level are reported in parentheses. Specification includes: grade fixed effects, year fixed effects, grade by year fixed effects, subject by grade fixed effects, day of the week fixed effects, a full set of birth year by cohort fixed effects, and individual controls. Individual controls include indicators for students who are female, and are retained. The comparison group is exams taken within the first nine days from the first exam, up to the fifth place of exam order, and not later than two days from the previous exam. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 8: Heterogeneous Effects of Exam Timing on Performance by STEM and Prior Performance

VARIABLES	(1)
Midterm Score	0.470***
	(0.005)
Scheduling Effect I for STEM	-0.012***
	(0.004)
Scheduling Effect II for STEM	0.018***
	(0.003)
Scheduling Effect III for STEM	-0.044***
	(0.007)
Scheduling Effect I for non-STEM	0.025***
	(0.005)
Scheduling Effect II for non-STEM	-0.015***
	(0.004)
Scheduling Effect III for non-STEM	0.038***
	(0.007)
Scheduling Effect I for non-STEM × Midterm Score	0.016***
	(0.002)
Scheduling Effect I for STEM × Midterm Score	-0.004***
	(0.001)
Scheduling Effect II for non-STEM × Midterm Score	-0.000
	(0.002)
Scheduling Effect II for STEM × Midterm Score	0.008***
	(0.001)
Scheduling Effect III for non-STEM × Midterm Score	0.001
	(0.003)
Scheduling Effect III for STEM × Midterm Score	-0.020***
	(0.003)
Observations	14,258
R-squared	0.992

Note: The dependent variable in each specification is the normalized final exam score at the subject and grade level. Cluster-robust standard errors at the classroom by year level are reported in parentheses. All specifications include: grade fixed effects, year fixed effects, grade by year fixed effects, subject by grade fixed effects, day of the week fixed effects, a full set of birth year by cohort fixed effects, and individual controls. Individual controls include indicators for students who are female, and are retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0$

Table 9: Heterogeneous Effects of Exam Timing on Performance by STEM and Prior Performance

	Non-STEM		STEM		Difference	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Scheduling Effect I						
Quantile 1	-0.060***	(0.008)	0.004	(0.004)	0.064***	(0.007)
Quantile 2	-0.028***	(0.004)	-0.003	(0.003)	0.025***	(0.004)
Quantile 3	0.005*	(0.003)	-0.003	(0.003)	-0.008*	(0.004)
Quantile 4	0.027***	(0.005)	-0.014***	(0.004)	-0.041***	(0.006)
Scheduling Effect II						
Quantile 1	0.006	(0.006)	-0.021***	(0.004)	-0.027***	(0.006)
Quantile 2	0.005*	(0.003)	0.000	(0.001)	-0.005*	(0.003)
Quantile 3	0.002*	(0.001)	0.012***	(0.002)	0.009***	(0.002)
Quantile 4	0.001	(0.004)	0.024***	(0.005)	0.022***	(0.006)
Scheduling Effect III						
Quantile 1	-0.014	(0.012)	0.052***	(0.008)	0.066***	(0.013)
Quantile 2	-0.015***	(0.005)	0.001	(0.003)	0.016***	(0.006)
Quantile 3	-0.005*	(0.003)	-0.031***	(0.005)	-0.026***	(0.005)
Quantile 4	-0.001	(0.008)	-0.059***	(0.010)	-0.058***	(0.011)

Note: sample: 14,258 obs. The dependent variable is the normalized final exam score at the subject and grade level. Cluster-robust standard errors at the classroom by year level are reported in parentheses. Specification includes: grade fixed effects, year fixed effects, grade by year fixed effects, subject by grade fixed effects, day of the week fixed effects, a full set of birth year by cohort fixed effects, and individual controls. Individual controls include indicators for students who are female, and are retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 10: Heterogeneous Effects of Exam Timing on Performance by Gender

	Males		Females		Difference	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Scheduling Effect I						
Non-STEM	-0.013***	(0.005)	-0.011**	(0.004)	0.002	(0.003)
STEM	0.006*	(0.004)	-0.001	(0.004)	-0.007**	(0.003)
Scheduling Effect II						
Non-STEM	0.002	(0.003)	0.001	(0.002)	-0.001	(0.002)
STEM	-0.009***	0.003	-0.005*	0.003	0.003	(0.003)
Scheduling Effect III						
Non-STEM	-0.003	(0.005)	-0.002	(0.004)	0.001	(0.005)
STEM	0.023***	(0.006)	0.013**	(0.006)	-0.010*	(0.005)

Note: sample: 14,258 obs. The dependent variable is the normalized final exam score at the subject and grade level. Cluster-robust standard errors at the classroom by year level are reported in parentheses. Specification includes: grade fixed effects, year fixed effects, grade by year fixed effects, subject by grade fixed effects, day of the week fixed effects, a full set of birth year by cohort fixed effects, and individual controls. Individual controls include indicators for students who are female, and are retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 11: Robustness of the Effects of Exam Timing on Performance

VARIABLES	(1) With Student FE	(2) Including Track Electives	(3) Same-day exams excluded
Scheduling Effect III for non-STEM	-0.000	-0.001	-0.000
	(0.005)	(0.005)	(0.005)
Scheduling Effect III for STEM	0.015***	0.014***	0.015***
	(0.005)	(0.004)	(0.005)
Scheduling Effect II for non-STEM	0.001	0.001	0.001
	(0.003)	(0.002)	(0.003)
Scheduling Effect II for STEM	-0.006**	-0.006***	-0.006**
	(0.002)	(0.002)	(0.002)
Scheduling Effect I for non-STEM	-0.013***	-0.013***	-0.013***
	(0.004)	(0.004)	(0.005)
Scheduling Effect I for STEM	0.001	-0.001	0.001
	(0.003)	(0.003)	(0.004)
Observations	14,258	16,747	14,092
R-squared	0.987	0.980	0.987
Student FE	YES	NO	NO

Note: The dependent variable in each specification is the normalized final exam score at the subject and grade level. Cluster-robust standard errors at the classroom by year level are reported in parentheses. All specifications include: grade fixed effects, year fixed effects, grade by year fixed effects, subject by grade fixed effects, day of the week fixed effects, a full set of birth year by cohort fixed effects, and individual controls. Individual controls include indicators for students who are female, and are retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0$

Figure 1: Example of Exam Schedule

ΠΡΟΓΡΑΜΜΑ ΕΞΕΤΑΣΕΩΝ
ΜΑΪΟΥ - ΙΟΥΝΙΟΥ 2005

Β΄ ΤΑΞΗ

Ημ/νια	Ημέρα	Μάθημα	Ώρα Εναρξης
20/5	Παρασκευή	Ιστορία	8 : 15
23/5	Δευτέρα	Λατινικά –Χημεία –Τεχν. Επικοινωνιών (Κατ)	8 : 15
25/5	Τετάρτη	Φυσική Γ. Π.	8 : 15
27/5	Παρασκευή	Εισαγωγή στο Δίκαιο - Γερμανικά	8 : 15
30/5	Δευτέρα	Αρχαία –Μαθηματικά –Μαθηματικά (Κατ)	8 : 15
1/6	Τετάρτη	Θρησκευτικά - Σχέδιο	8 : 15
3/6	Παρασκευή	Άλγεβρα	8 : 15
6/6	Δευτέρα	Αρχαία Γ. Π.	8 : 15
8/6	Τετάρτη	Αγγλικά	8 : 15
10/6	Παρασκευή	Αρχές Φιλοσοφίας-Φυσική-Φυσική (Κατ)	8 : 15
13/6	Δευτέρα	Γεωμετρία	8 : 15
14/6	Τρίτη	Νεοελληνική Γλώσσα	8 : 15
15/6	Τετάρτη	Χημεία	8 : 15
16/6	Πέμπτη	Βιολογία	8 : 15
17/6	Παρασκευή	Νεοελληνική Λογοτεχνία	8 : 15

Ο Διευθυντής

Note: The picture above shows the exam schedule of students in the 11th grade in May-June 2005. The first and second column show the date and day of the week of the exam, respectively. The third column shows the subject tested. The fourth column shows the time the exam starts. Since all concentration electives are tested on the same date, the choice of elective courses does not affect students' exam schedule. For example, on May 23rd, 2015 tests on three concentration elective courses (one for each concentration) were administered for students of the same grade: Latin, chemistry, and communications technology.

Figure 2: Scheduling Effects on Exam Performance

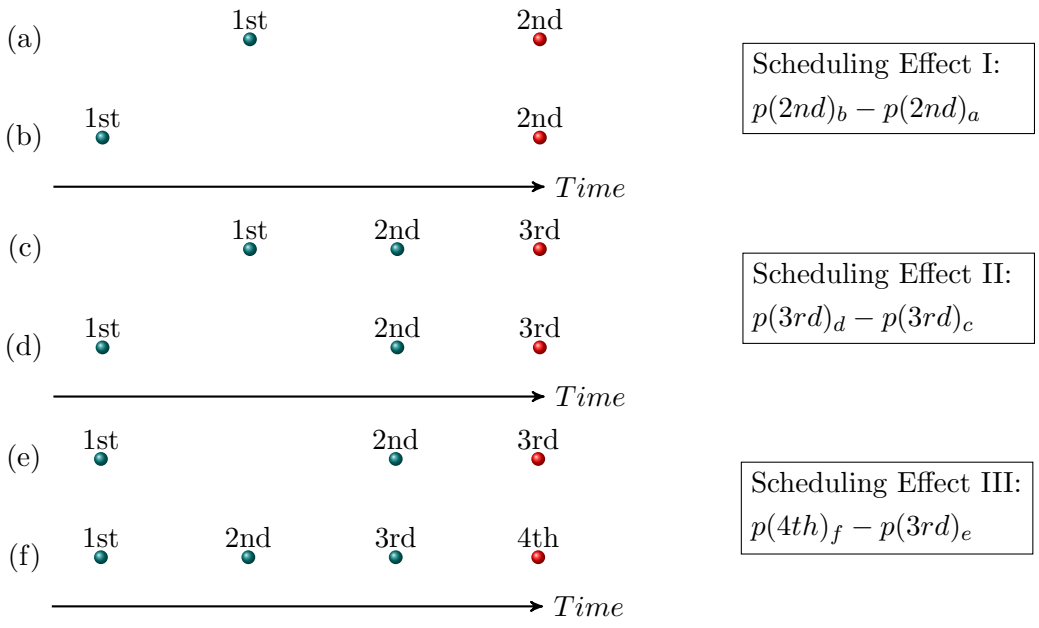


Figure 3: Estimated Scheduling Effects on Performance

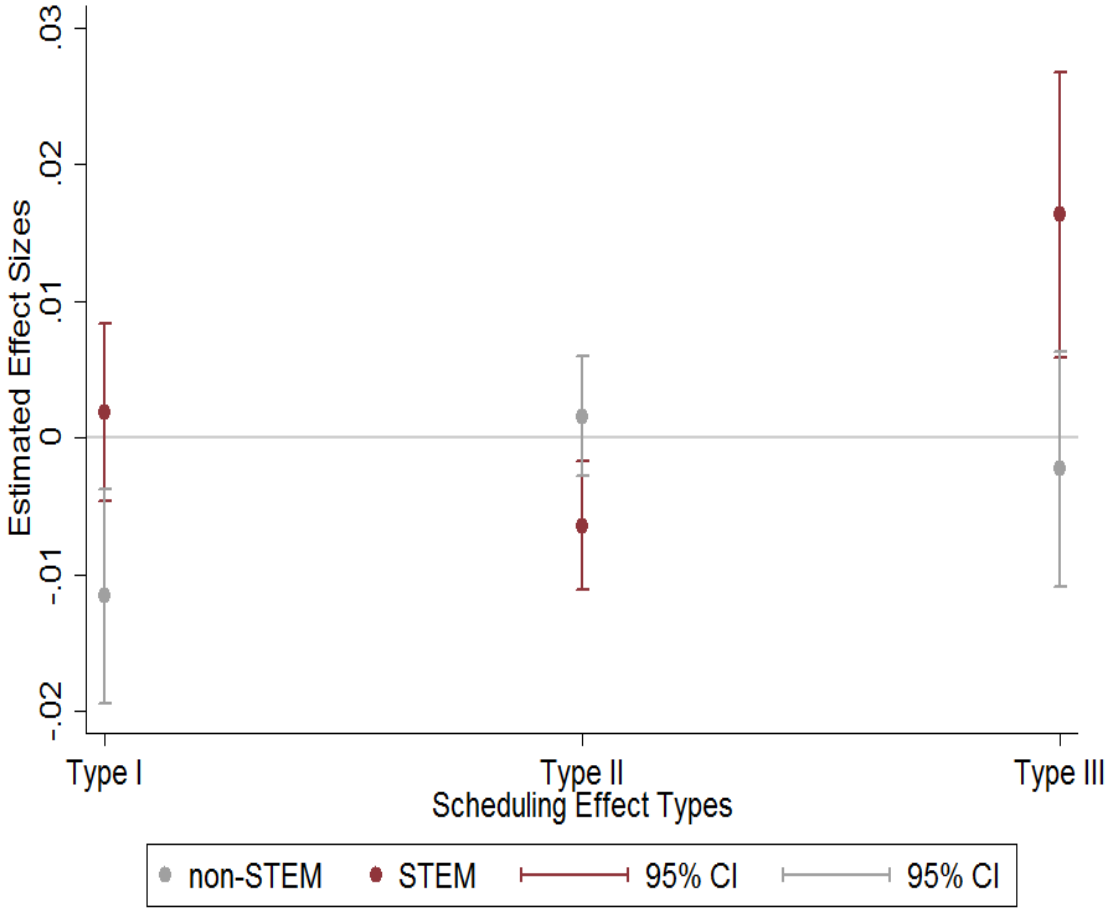
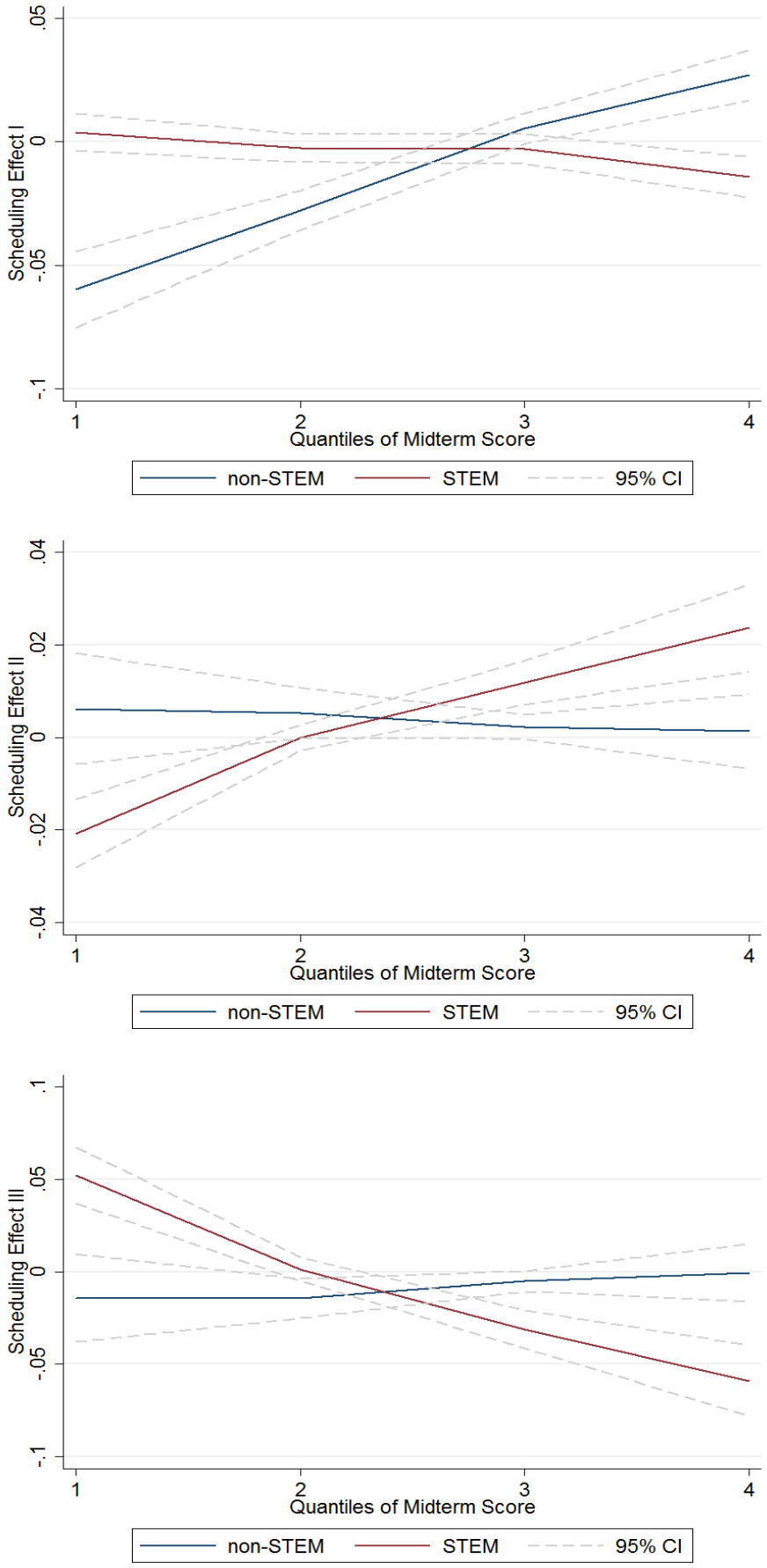


Figure 4: Estimated Scheduling Effects on Performance



8 Appendix

Table 12: Summary Statistics for the 10th Grade

Year		Female	Age	GPA	Midterm Score	Final Exam Score	Retained
2002	Mean	0.60	15.84	14.69	16.76	12.81	0.00
	SD	0.49	0.43	2.81	1.80	3.80	0.00
	N	91	91	91	91	91	91
2003	Mean	0.50	15.76	15.33	17.18	13.50	0.00
	SD	0.50	0.48	2.99	1.79	4.21	0.00
	N	86	86	86	86	86	86
2004	Mean	0.66	15.88	15.88	17.40	14.21	0.01
	SD	0.47	0.55	2.43	1.61	3.55	0.10
	N	101	101	100	101	101	101
2005	Mean	0.48	15.81	14.98	16.68	13.05	0.02
	SD	0.50	0.44	3.04	2.12	4.14	0.14
	N	95	95	93	95	95	95
2006	Mean	0.53	15.95	15.10	17.06	13.14	0.01
	SD	0.50	0.35	2.49	1.62	3.44	0.10
	N	108	108	107	108	108	108
2007	Mean	0.57	16.06	15.30	17.09	12.90	0.04
	SD	0.50	0.36	2.73	1.99	4.35	0.20
	N	116	116	111	116	116	116
2008	Mean	0.55	16.03	15.24	17.32	13.20	0.00
	SD	0.50	0.17	2.98	1.92	4.03	0.00
	N	98	98	98	98	98	98
2009	Mean	0.47	16.02	15.99	17.62	14.38	0.00
	SD	0.50	0.15	2.22	1.44	3.03	0.00
	N	92	92	92	92	92	92
2010	Mean	0.57	16.04	15.99	17.71	14.09	0.01
	SD	0.50	0.19	2.52	1.52	3.84	0.09
	N	113	113	112	113	113	113
Total	Mean	0.55	15.94	15.40	17.21	13.47	0.01
	SD	0.50	0.38	2.72	1.79	3.87	0.10
	N	900	900	890	900	900	900

Table 13: Summary Statistics for the 11th Grade

Year		Female	Age	GPA	Midterm Score	Final Exam Score	Retained
2002	Mean	0.61	16.70	12.05	16.54	10.82	0.12
	SD	0.49	0.46	3.17	1.86	3.65	0.32
	N	102	102	90	102	102	102
2003	Mean	0.63	16.85	14.83	17.18	12.16	0.06
	SD	0.49	0.48	2.80	1.87	4.11	0.24
	N	84	84	79	84	84	84
2004	Mean	0.52	16.78	15.34	17.44	12.59	0.06
	SD	0.50	0.47	3.04	2.04	4.99	0.25
	N	79	79	74	79	79	79
2005	Mean	0.66	16.93	15.13	16.81	13.08	0.05
	SD	0.48	0.82	2.69	1.93	4.18	0.22
	N	99	99	94	99	99	99
2006	Mean	0.50	16.86	15.09	16.88	13.26	0.02
	SD	0.50	0.63	3.12	2.16	4.27	0.15
	N	88	88	86	88	88	88
2007	Mean	0.56	17.01	15.39	17.39	13.46	0.02
	SD	0.50	0.60	2.43	1.52	3.69	0.14
	N	103	103	101	103	103	103
2008	Mean	0.58	17.01	15.79	17.76	14.08	0.00
	SD	0.50	0.10	2.59	1.49	3.62	0.00
	N	103	103	103	103	103	103
2009	Mean	0.59	17.03	15.74	17.72	13.85	0.01
	SD	0.49	0.18	2.84	1.72	4.02	0.11
	N	90	90	89	90	90	90
2010	Mean	0.47	17.05	15.84	17.93	14.23	0.00
	SD	0.50	0.26	2.47	1.33	3.38	0.00
	N	88	88	88	88	88	88
Total	Mean	0.57	16.92	15.03	17.29	13.06	0.04
	SD	0.50	0.51	2.99	1.83	4.10	0.19
	N	836	836	804	836	836	836

Table 14: Summary Statistics for All Students in the Sample

Year		Female	Age	GPA	Midterm Score	Final Exam Score	Retained
2002	Mean	0.61	16.29	13.38	16.65	11.76	0.06
	SD	0.49	0.62	3.26	1.83	3.84	0.24
	N	193	193	181	193	193	193
2003	Mean	0.56	16.29	15.09	17.18	12.84	0.03
	SD	0.50	0.73	2.90	1.83	4.20	0.17
	N	170	170	165	170	170	170
2004	Mean	0.60	16.28	15.65	17.41	13.50	0.03
	SD	0.49	0.69	2.71	1.80	4.30	0.18
	N	180	180	174	180	180	180
2005	Mean	0.57	16.38	15.05	16.75	13.06	0.04
	SD	0.50	0.87	2.86	2.02	4.15	0.19
	N	194	194	187	194	194	194
2006	Mean	0.52	16.36	15.10	16.98	13.19	0.02
	SD	0.50	0.67	2.78	1.88	3.83	0.12
	N	196	196	193	196	196	196
2007	Mean	0.57	16.51	15.34	17.23	13.16	0.03
	SD	0.50	0.68	2.59	1.79	4.05	0.18
	N	219	219	212	219	219	219
2008	Mean	0.57	16.53	15.52	17.55	13.65	0.00
	SD	0.50	0.51	2.79	1.72	3.84	0.00
	N	201	201	201	201	201	201
2009	Mean	0.53	16.52	15.87	17.67	14.12	0.01
	SD	0.50	0.53	2.54	1.58	3.56	0.07
	N	182	182	181	182	182	182
2010	Mean	0.52	16.48	15.93	17.81	14.15	0.00
	SD	0.50	0.55	2.49	1.44	3.64	0.07
	N	201	201	200	201	201	201
Total	Mean	0.56	16.41	15.23	17.25	13.27	0.02
	SD	0.50	0.66	2.85	1.81	3.99	0.15
	N	1736	1736	1694	1736	1736	1736