**Sofoklis Goulas**[1] / **Rigissa Megalokonomou**[2]

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

[1] Hoover Institution, Stanford University, 434 Galvez Street, Stanford, CA, USA, E-mail: goulas@stanford.edu. https://orcid.org/0000-0001-7100-0647.
[2] Department of Economics, University of Queensland, Brisbane, Queensland, Australia, E-mail: r.megalokonomou@uq.edu.au

**Abstract:**
Would you prefer a tighter or a more prolonged exam schedule? Would you prefer to take an important exam first or last? We exploit quasi-random variation in exam schedules across cohorts, grades and subjects from a lottery to identify distinct effects of the number of days between exams, the number of days since the first exam, and the exam order on performance. Scheduling effects are more pronounced for STEM exams. We find a positive and a negative relationship between STEM scores and exam order (warm-up) and number of days since the first exam (fatigue), respectively. In STEM, warm-up is estimated to outweigh fatigue. Marginal exam productivity in STEM increases faster for boys than for girls. Higher-performing students exhibit higher warm-up and lower fatigue effects in STEM than lower-performing students. Optimizing the exam schedule can improve overall performance by as much as 0.02 standard deviations.

**Keywords:** exam schedule, cognitive fatigue, exam warm-up, practice, scaffolding, gender gap, STEM, student performance, lottery

**JEL classification:** I20, I24

**DOI:** 10.1515/bejeap-2019-0177

## 1 Introduction

The question of what determines performance on cognitive tasks has long been of concern to policy makers, social scientists, and academics. Research on cognitive task performance has not been limited to psychology[1]; understanding what affects task productivity is also a central question in economics (Coviello, Ichino, and Persico 2014; Pope and Fillmore 2015). In everyday life, individuals face multiple tasks. The time horizon for the completion of several tasks is typically somewhat limited. Given the scarce time and limited attention, individuals must decide how to allocate their resources in order to maximize their utility, which is assumed to be positively related to the outcomes of the tasks undertaken. To put it differently, the diverse tasks one faces compete for their attention and time. In a context with many tasks and limited resources, particularly time, how different tasks are scheduled over time has salient effects on task performance (Buser and Peter 2012). For instance, athletes have been found to benefit when the scheduling of athletic events allows them to have several weeks of recuperation in between (Chambers et al. 1998). Additionally, in a study of the duration of court case completion by Italian judges (Coviello, Ichino, and Persico 2014), it was found that completing cases simultaneously takes more time, on average, than completing tasks sequentially.

The effects of scheduling on cognitive task performance are of particular interest in education. School principals, much like managers, are often looking for innovations that increase educational outcomes with little to no increase in resources. Even though exam scheduling is more likely to influence measured performance instead of actual learning during the school year, the impact of exam scheduling remains of interest to policy makers who wish to minimize error in measured performance. In most educational systems, students are required to complete several tasks in a finite time period, including projects and exams. In particular, in many countries, all high school students take a sequence of exams during finals week within a short period of time and with only a few days at most between exams. Performance on these exams is paramount for students' progression to the next grade, and exam scheduling may be an important driver of students' performance. For example, students' performance has been found to be positively associated with the time between exams (Pope and Fillmore 2015); however, this is only one aspect of task scheduling. Other aspects may include the order in which exams are completed, and the number of exams to be completed in a given time period. This paper asks: What is the effect

**Sofoklis Goulas** is the corresponding author.

of having completed an additional exam on later performance? Is fatigue associated with exam completion? Our research question is motivated by a practical concern: how exams can be optimally scheduled.

Although the importance of exam schedules is obvious, it has received little attention in the literature. In particular, there is lack of causal evidence on how scheduling affects student achievement, since in most settings course assignment is usually not random, and individuals who select into certain courses may also select into a particular exam schedule for those courses. At the same time, individuals' preferences or other engagements may also influence their exam scheduling. Endogeneity arising from unobservables that drive selection into courses and selection into particular exam schedules renders the identification of the scheduling effects on exam performance challenging.

In this paper, we explore the different channels through which exam scheduling affects performance. We also explore how scheduling effects vary by prior performance and gender; this has not previously been attempted in the literature. We exploit quasi-random variation in exam timing across cohorts, subjects and grades to identify the 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 compare exam performance across subjects, grades and cohorts of students who have similar characteristics and face the same school environment, except for the fact that each subject has different exam timing across grades and cohorts, due to purely random factors. To address the issue of endogeneity, we exploit a novel dataset on exam scheduling and student performance on each exam taken in the 10th and 11th grades in a public high school in Greece that used a lottery to generate exam schedules. Our dataset covers exam performance in all subjects for nine cohorts of students in the 10th and 11th grades between school years 2001–2002 and 2009–2010. At the end of each school year, between May and June, high school students take compulsory written exams in every subject taught during the school year. Exam schedules are announced after the last day of classes in each year. The exam season lasts between 3 and 4 weeks, with each student taking exams in more than 13 subjects on average.

To our knowledge, this is the first paper to use quasi-random variation in exam timing across cohorts, grades and subjects from a lottery to identify *three contemporaneous channels* through which exam scheduling may affect performance. The randomness in exam scheduling achieved through the lottery guarantees the orthogonality of subject type (i. e. STEM[2] or non-STEM) and student characteristics (i. e. prior performance) on the exam scheduling of compulsory subjects across cohorts and grades. A simulation experiment provides evidence that the exam timing of each subject in our empirical data is indeed random, and has average characteristics similar to the exam timing in schedules generated by a truly random schedule-generating process. In addition, balancing tests provide further evidence that the variation in exam scheduling is not associated with students' background characteristics, which allows us to obtain more precise and unbiased estimates of scheduling effects. The large number of exams taken permits us to disentangle three distinct contemporaneous effects of scheduling on exam 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, and the effect of the length of time between exams on subsequent performance may be a composite effect of preparation and potential distraction. We call this Scheduling Effect I. Considering a sequence of exams individuals must take in a given time period, the second channel corresponds to the effect of the time distance between the first exam and the exam that students are about to take. We call this Scheduling Effect II and it may capture fatigue from exam completion. The third channel relates to how many exams have been taken before sitting an additional exam. We call this Scheduling Effect III, and it may reflect warm-up or practice effects from exam completion. 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, which allows for comparable grades within a course-grade configuration.

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, particularly in STEM subjects. Our results indicate that exam productivity in STEM courses increases with exam order, suggesting the existence of a learning effect that is positively associated with taking an additional exam. We find that the exam order affects student achievement: Exams taken later in the schedule being are 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. Students randomly assigned to a later place in the exam order earn a grade on that exam that is significantly higher than the grades of students randomly assigned an earlier place in the 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 that an additional day in the exam season is associated with exam fatigue, particularly in STEM courses. A 1-day increase in the day count since the first exam decreases significantly the student's subsequent performance on STEM-related exams (fatigue effect). For STEM subjects, the estimated warm-up effect is larger

than the estimated fatigue effect. The number of days between exams is found to be associated with decreasing exam performance only in non-STEM courses, controlling for other influences.

We also explore differential effects across different levels of prior performance, proxied by midterm scores. Students in the top quantile of prior 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. In addition, 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. Additional days between exams are found to improve STEM-related performance more for students in the top quantile of prior midterm performance than for students in the bottom quantile. 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 (recuperation effect) compared to students in the top quantile of prior midterm performance.

Policy makers 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 and Levitt 2010; Else-Quest, Hyde, and Linn 2010; Dee 2007; Nosek et al. 2009; Hyde et al. 2008), as have differences in cognitive learning between males and females (Zimmerman and Martinez-Pons 1990; Halpern 2004, 2013; Fennema and Sherman 1977), the extent to which the gender gap can be influenced by exam scheduling in schools has not been investigated. This paper explores differential scheduling effects on exam performance by gender. We find that exam productivity increases faster for boys than it does for girls as they take additional exams. In addition, boys' exam performance benefits more from additional time between exams than girls' performance.

As families, teachers, and administrators seek ways to improve student academic performance, some ask what the optimal exam schedule is. We conduct a series of simulation experiments to show that the higher the number of exams taken the higher the potential benefit from optimizing exams scheduling. Our simulations show that optimizing the exam schedule can improve overall performance by as much as 0.02 standard deviations.

The remainder of the paper is organized as follows. In Section 2, we discuss the evidence to motivate hypotheses that can be tested empirically. In Section 3 and 4, we provide information on the institutional setting of exam scheduling and discuss the data that we use in our study, respectively. In Section 5, we lay out our empirical methodology. We report and discuss our results in Section 6, and simulate the optimal exam schedule in Section 7. Section 8 concludes.

## 2 Why We Would Expect Exam Scheduling to Affect Performance

In this section, we combine evidence from the literatures of economics, sociology, education, and psychology to predict how students' performance would respond to different exam scheduling. We consider three potential channels through which exam scheduling may impact performance: time between exams, number of days since the first exam, and having taken an additional exam during the current exam season. The three channels are Scheduling Effects I, II, and III, respectively. 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.

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 takes several exams in a narrow time interval. On the other hand, improvement of meta-cognitive accuracy may occur when one repeats certain cognitive tasks, such as taking exams. If there is learning associated with exam taking, then as students take more exams, certain knowledge becomes "automatic," freeing up mental resources (a reduction in cognitive load) that can be used in other cognitive tasks. The reduction in cognitive load for repeating cognitive tasks is called "scaffolding" (Askell-Williams, Lawson, and Skrzypiec 2012; Ambrose et al. 2010). Not every task or exam puts the same stress on every aspect of cognition. For example, tasks that 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 employ cognition processes associated with memory or critical thinking.

One may argue that if different subjects put different levels of stress on different aspects of cognition and one has demonstrated a certain level of achievement in a particular subject in the past, it is likely their past achievement will reveal the degree to which they possess the cognitive skills that are used intensively in that subject. Consequently, students of different prior performance in a certain type of subjects may possess differ-

ent 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 males and females heterogeneously. Cognitive differences between males and females have been well established (Halpern 2013; Hyde, Fennema, and Lamon 1990; Hyde and Linn 1988). The literature proposes that males may exhibit higher returns to practice than females in terms of performance on cognitive tasks, although females may be better in self-regulated learning than males (Vygotskiĭ 2012; Ablard and Lipschultz 1998).

We present the three channels in a simple schematic of sequential exam productivity in which a student is subjected to a finite number of exams under different schedules. One potential channel we consider here is the time between exams. This channel is illustrated in the comparison between exam schedules (a) and (b) in Figure 1. Exam schedules (a) and (b) both include two exams, but schedule (b) allows for more time between the first and second exams compared to schedule (a). If we assume that students spend time between exams recuperating from the last exam and preparing for the next one, we may expect that the more time students have between exams, the higher their performance on the later exam will be, on average. In this case, average performance on the second exam under schedule (b) should be higher than performance on 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 rather become distracted and stop studying, we may anticipate a zero effect of the time between exams on subsequent exam performance, and average performance on the last exam under schedule (b) should be no different than that under schedule (a). If we assume that potential distraction during a longer interval between exams affects negatively focus and readiness to undertake cognitive tasks, we may even expect a negative effect of the time between exams on subsequent exam performance. Time between tasks might affect negatively subsequent performance on tasks that rely more on short-term memory (Baddeley 2003). In that case, average performance on the last exam under schedule (b) should be lower than performance on the last exam under schedule (a).

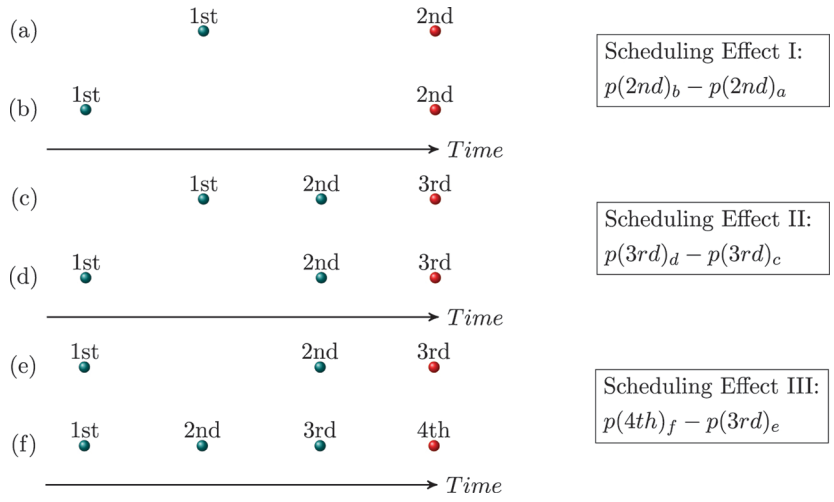


**Figure 1:** Scheduling effects on exam performance.

The positive association between time between exams and performance is documented by Pope and Fillmore (2015), who propose multiple explanations for their findings. One explanation is based on 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 on only a few. Pope and Fillmore's findings, however, come from a sample of students that self-select into taking particular exams (i. e. 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 Pope and Fillmore's evidence cannot be extrapolated to lower-achieving students.

The second channel of exam scheduling's influence on performance we consider is the days lapsed since the beginning of the exam season while holding constant other aspects of scheduling, such as time between exams. This case is illustrated in the comparison between schedules (c) and (d) in Figure 1. Exam schedules (c) and (d) contain the same number of total exams – 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 interval between the first and second exam. If the time between the first and second exam can be used to prepare for the third exam, then performance on the third exam under schedule (d) should be higher than performance on the third exam under schedule (c), on average ($p(3rd)_d > p(3rd)_c$). On the other hand, one may expect average performance on the third exam of schedule (d) to be lower than that under schedule (c) if the time between the first and 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 study 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 literature of both psychology and economics (Webster, Richter, and Kruglanski 1996; Jensen, Berry, and Kummer 2013; Meijman 1997). In order 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 possess these skills. Students with a history of low achievement have been documented to have problems meeting deadlines and are more prone to procrastinate than high-achieving students (Kármen et al. 2015; Metcalfe and Finn 2013; Özsoy, Memiş, and Temur 2017). Also, high-achieving students have been found to exhibit more self-regulated learning skills and time management skills (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 skills may also be beneficial for studying (Nadinloyi et al. 2013).

The third channel of exam scheduling's influence on exam performance we explore is related to the order in which exams are taken. Consider exam schedules (e) and (f) in Figure 1. The two schedules have the same length, 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 (e), the last exam is the third exam taken, while under schedule (f) the last exam is the fourth exam taken. If we assume that, holding everything else the same, having an additional exam earlier on might be associated with learning related to the subsequent exam, we may expect performance on 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, performance on the last exam under schedule (f) should be no different from performance on the last exam under schedule (e). On the other hand, if additional exams are associated with lower performance in subsequent exams, the performance on the last exam under schedule (f) should be lower than the performance on the last exam under schedule (e). The potential learning gain associated with having taken additional exams may not only be strictly related to the material tested, but also to the best strategies for studying or test-taking. This learning gain has been investigated in the psychology literature as metacognitive accuracy (bias scores and gamma correlations), which has been found to improve with practice (Kelemen, Winningham, and Weaver 2007; Finn and Metcalfe 2007). In addition, as discussed earlier, this leads to "scaffolding" – the concept that as a student learns, he/she builds up knowledge that allows them to create new knowledge. As they repeat the "scaffolding" process, certain knowledge becomes automatic as they encounter the material again (Tversky and Kahneman 1974; Vygotskiĭ 2012; Askell-Williams, Lawson, and Skrzypiec 2012). When students re-encounter some learned material, they experience a reduction in their cognitive load, allowing them to acquire new knowledge.[3] Therefore, performance on cognitive tasks, such as exams, may improve as students take additional exams.

We now summarize the main hypotheses regarding the predicted sign of each scheduling effect on exam performance, based on the evidence for each channel discussed so far.

Hypothesis 1 (recuperation effect): The more time between exams students have, the higher their performance on the subsequent exam will be on average, ceteris paribus. We hypothesize that the time between exams is used – to a certain extent – to recuperate from the last exam and to prepare for the next exam, and thus it will be positively associated with the student's performance in the next exam, on average.

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

Hypothesis 3 (warm-up effect): The higher the number of exams a student has taken at a certain point in time, the higher their performance will be on 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.

Our hypotheses rely on certain assumptions about how individuals spend their time between cognitive tasks, how prone they are to mental exhaustion, and their capacity for learning from practice. These assump-

tions may be more appropriate for certain types of cognitive tasks (e. g. exams in STEM fields) or individuals with certain characteristics (e. g. individuals with higher prior performance on 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, and for students of different genders.

## 3 Institutional Setting and Exam Scheduling

In this section, we describe the institutional setting of exam taking and exam scheduling in high school. High school in Greece starts at 10th grade and ends at 12th grade. At the end of the school year, students take exams on an average of 13 subjects[4] within 27 calendar days, on average. Exams usually start 1 week after the last day of classes. End-of-course exams account for 50 % of the course grade, and midterm scores for the other 50 %. Final exams are important for students, as failing to achieve a passing course grade in every subject leads to repeating the grade. Students must achieve a course grade of at least 9.5 out of 20 in every subject to progress to the next grade.

Teaching faculty in each subject and grade collectively construct the final exam and they split the grading load.[5] The principal has to read and approve the exam questions and marked exam papers, and document the marks in the school log (and computer). The principal is responsible for the adherence of teachers to the grading guidelines provided by the Ministry of Education.[6]

In the 11th grade, students choose one of three Tracks or Concentrations: Classics, Science, or Information Technology. Each track consists of three compulsory classes.[7] Students in the 10th grade take 14 classes, 11 of which have compulsory end-of-course exams. Students in the 11th grade take 17–18 classes, 15–16 of which have end-of-course exams, depending on add-on electives, that are always tested after all other exams. The exam schedule for the 11th grade of the 2004–2005 school year is shown in Figure 7 as an example.

Exam scheduling is orthogonal to the choice of courses and student ability, as student choices do not affect scheduling and track electives are tested on the same dates (e. g. one exam date is reserved for one track elective test, regardless of track. A second exam date is reserved for another track elective test, regardless of track, etc.). We focus on compulsory subjects, as all compulsory subjects are tested on the same date for students in the same grade in a given year. This enables the clean separation of scheduling effects from influences from student characteristics. Additional exam dates at the end of the exam schedule are reserved for additional electives, and students of different electives take their respective exams on the same date.

Our empirical investigation focuses on nine compulsory subjects across grades:[8] algebra, geometry, physics, and chemistry are considered STEM fields, while ancient Greek, modern Greek, Greek literature, English, and history are considered non-STEM fields.

Each school's teachers and principal are free to collectively design their exam schedules using specific exam dates set by the Ministry of Education. Our dataset is drawn from a high school in central Greece that chose to employ a lottery to design their exam schedule for the years 2002–2010. The lottery process was described on the school's website. Every year, the Ministry of Education set the dates on which an exam would be given, roughly 1 week after the last day of classes. The proctoring load was split evenly among teachers. In the sampled school, a lottery process without replacement was used to decide on the exam order. For the first exam date, a teacher drew a piece of paper with a subject name written on it from a raffle, then a second paper for the second exam date, etc. The lottery process was conducted during a scheduled meeting of teachers and the principal, who supervised the process. The meeting took place shortly after the last day of classes of the school year. This process was repeated every year separately for 10th and 11th graders. The lottery process used by the sampled school allows for quasi-randomization in exam timing across cohorts, grades, and subjects.

No data are available on the process other schools followed to design their exam schedules. Limiting the analysis to one school reflects a trade-off between institutional stability regarding the schedule-generating process and a larger sample size. In addition, schools are not required to maintain a record of exam dates and times, and retrieving this information for multiple years is challenging.

To understand how the sampled school compares to the schools in its district and the country, we compare the over-time averages on school size, demographics, neighborhood income, STEM-course participation, student performance on national exams, and university admission rates in Table 1. Compared to the average school in the district or the country, the sampled school has higher enrollment, higher share of students in Science and Information Technology track over Classics track, higher university admission rate, and higher average student performance on national exams (university admission score). The differences between the characteristics of the sampled school and those of schools in the district and the country are within one standard deviation of the district or country characteristics, respectively. In addition, we observe that the average income for the sam-

pled school's neighborhood is roughly one standard deviation below the average neighborhood income in the district or the country. The sampled school is a traditional public school located in an urban area: 65 % and 73 % of schools in the sampled school's district and the country are in urban areas, respectively, and 76 % and 85 % of schools in the sampled school's district and the country are traditional public schools,[9] respectively. Overall, the average characteristics of the sampled school are not substantially different from the characteristics of other schools in the same district or the country.

**Table 1:** How does the sampled school compare to the district and the country?

| **Representativeness of sampled school** | | | | | |
|---|---|---|---|---|---|
| | **Sampled school** | **District** | **Country** | **Difference** | **Difference** |
| | **(1) Mean** | **(2) Mean/ (sd)** | **(3) Mean/ (sd)** | **(1)–(2) Difference** | **(1)–(3) Difference** |
| School size† | 183.778 | | 143.876 | | 39.902 |
| | | | (60.445) | | |
| Prop. of female students† | 0.558 | | 0.545 | | 0.013 |
| | | | (0.054) | | |
| Student age† | 16.487 | | 16.636 | | –0.148 |
| | | | (0.915) | | |
| Prop. in classics track† | 0.310 | | 0.367 | | –0.057 |
| | | | (0.094) | | |
| Prop. in science track† | 0.285 | | 0.214 | | 0.071 |
| | | | (0.083) | | |
| Prop. in IT track‡ | 0.407 | | 0.364 | | 0.042 |
| | | | (0.110) | | |
| University admission score‡ | 13.088 | 12.844 | 12.256 | 0.244 | 0.831 |
| | | (0.982) | (1.520) | | |
| Prop. of admitted students‡ | 0.814 | 0.804 | 0.760 | 0.010 | 0.054 |
| | | (0.068) | (0.123) | | |
| Neighborhood income‡ | 14,704.730 | 16,785.246 | 19,345.909 | –2,080.516 | –4,641.175 |
| | | (1,813.745) | (5,567.154) | | |
| Urban‡ | 1 | 0.846 | 0.759 | 0.154 | 0.241 |
| | | (0.376) | (0.428) | | |

Notes: The table reports means of the listed variables for the sampled school across years (column 1), the district in which the sampled school is located (column 2), and the country (column 3). We also report standard deviations in parentheses below the means in columns (2) and (3). In columns (4) and (5) we report the differences between the means of the sampled school (column 1) and the district (column 2), and the differences between the means of the sampled school (column 1) and the country (column 3), respectively. †: Statistics on school size, proportion of female students, average age for 10th and 11th graders, as well as the proportion of 11th graders in each track come from a country-representative dataset of 124 traditional public schools between 2003 and 2010 used in Goulas and Megalokonomou (2015). ‡: Statistics on university admission score, prop. of admitted students, neighborhood income and urban come from a dataset on university admission in Greece between 2003 and 2011 used by Goulas, Megalokonomou, and Zhang (2018). Neighborhood Income is measured in Euro in 2009. Urban is a binary variable that takes the value one if the school is in an urbanized area. University admission score represents students' performance on national exams. This is the only criterion for university admission, and ranges between 0 and 20.

# 4 Data Description

## 4.1 School Database

We follow students over two grades – 10th and 11th – and nine cohorts from 2001–2002 to 2009–2010. Our data set combines three types of data: course enrollment, test scores, and test dates. First, for every student in each year we have a student ID number, grade enrolled, classroom assignment, gender, year of birth, and complete course history. Second, for all students we have midterm 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. Data from these 9 years allow for a comparison of exam performance under different exam schedules.

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 from the fall and spring semester in each subject to

form a measure of performance for each student in the specific subject prior to the cumulative end-of-the-year exam.

The main analysis draws on a 9-year and 1,024-student pooled dataset of 14,258 individual exam scores. Midterm score and final exam score, our outcome variable, are standardized at the subject and grade level.

Using student-level data to investigate the effects of exam scheduling on performance is advantageous because it allows us to control for predetermined characteristics, such as prior performance and gender. In addition, student-level data permit us to explore differential scheduling effects by gender or prior performance. As we demonstrate in Section 6, scheduling effects differ markedly by ability and gender.

## 4.2 Statistics for Student Data

Table 2 shows descriptive statistics for student data across the 2002–2010 cohorts. Across years, we observe 900 distinct students in the 10th grade and 836 students in the 11th grade. Fifty-six percent of the students are female, and students' average age is 16.41 years. Students have an average GPA of 15.23 out of 20. The average midterm score is 17.25 and the average final score is 13.27 out of 20. Two percent of students are retained in the same grade. Comparing 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 Appendix B in Table 11, Table 12, and Table 13. Over a period of nine cohorts, we have data for a total of 1,024 students, 706 of which are observed in both the 10th and 11th grades. In Table 11, Table 12, and Table 13, we present average student characteristics for students in each school year. The various measures of student characteristics are similar across cohorts and grades.

**Table 2:** Summary statistics for student data.

| | 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 |

Notes: This table reports summary statistics for student characteristics and outcomes for 10th and 11th graders, and overall. The variables shown here are gender (female=1); age; GPA (between 0 and 20); midterm score (between 0 and 20); final exam score (between 0 and 20); and a dummy variable for retained status (if grade retained=1). Panel A shows 10th graders, Panel B shows 11th graders, and Panel C shows statistics across 10th and 11th graders.

## 4.3 Exam Scheduling Variables

We define the exam scheduling variables we construct for our analysis as follows:

*Days Between Exams*. – The measure of days between exams captures how many days intervene between one exam and the next. 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 after taking their immediately previous exam. The measure of days between exams is set to missing 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 after the first exam in the given year. The measure of days since the first exam is set to missing 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 by each student. For example, the measure of exam order is equal to two for a specific student and subject if the student took the exam for this subject second in a given year, and so on.

Each scheduling variable described in this section captures a distinct channel through which exam scheduling affects performance, as shown in Figure 1. The top panel in Figure 1 corresponds to type I scheduling effect, which is captured by the "days between exams" variable. The middle panel of Figure 1 corresponds to the type II scheduling effect, which is captured by the "days since the exam season started" variable. The bottom panel of Figure 1 corresponds to the type III scheduling effect, which is captured by the "exam order" variable.

Basic statistics for the exam scheduling variables are reported in Table 3. 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 graders take span 25 days on average, while the exams 11th graders take span 28 days on average. The average time between exams for 10th or 11th graders is approximately 2 days.

**Table 3:** Descriptive statistics for exam scheduling variables.

| Variables | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| Panel A: 10th Graders | | | | | |
| Number of exams | 900 | 11.39 | 0.54 | 11 | 13 |
| Duration of exam season | 900 | 24.91 | 3.05 | 21 | 31 |
| Days between exams | 900 | 2.39 | 0.22 | 1.91 | 2.82 |
| Panel B: 11th Graders | | | | | |
| Number of exams | 836 | 15.33 | 0.47 | 15 | 16 |
| Duration of exam season | 836 | 28.45 | 0.77 | 27 | 30 |
| Days between exams | 836 | 1.98 | 0.08 | 1.87 | 2.14 |
| Panel C: All students | | | | | |
| Number of exams | 1,736 | 13.29 | 2.03 | 11 | 16 |
| Duration of exam season | 1,736 | 26.61 | 2.87 | 21 | 31 |
| Days between exams | 1,736 | 2.20 | 0.26 | 1.87 | 2.82 |

Notes: The table presents summary statistics for exam taking. The number of exams shows how many exams each student took in a grade-year configuration, on average. Duration of exam season shows how many days the exam season lasted, on average, in a grade-year configuration. Days between exams shows the average time distance between consecutive pairs of exam for students in each grade-year configuration. Panel A refers to 10th graders, Panel B refers to 11th graders, and Panel C combines 10th and 11th graders.

#### 4.3.1 Randomness of Exam Schedules

One may worry that students characteristics may vary along the same dimensions as exam scheduling. In our research design, exam scheduling varies across years and grade. To investigate whether student characteristics used in the analysis (standardized midterm score, gender, age) are balanced across years and grades, we perform a balancing regression test for each student characteristic on a full set of year by grade dummy variables (see Table 15 in Appendix C). Table 15 provides evidence that the variation in exam scheduling is not associated with students' background characteristics, which allows us to obtain unbiased estimates of scheduling effects.

Another potential concern is whether exam scheduling is consistent with a random process and orthogonal to subject type. To address this, we provide evidence that the average exam timing, over time, does not differ significantly across subjects. In Figure 2, we present box plots for each of the three scheduling variables for each subject. The top panel refers to exam order, the second to the number of days since the first exam, and the bottom to the number of days between exams. The middle bar of each box plot shows the median. The lower hinge of the whisker box corresponds to the 25th percentile and the upper hinge of the whisker box to the 75th percentile. The upper adjacent line corresponds to the 95th percentile and the lower adjacent line to the 5th percentile. We do not observe the upper adjacent line of any subject being below the lower adjacent line of any of the other subjects for any of the scheduling variables. Simultaneously, we do not observe the lower adjacent line of any subject being above the upper adjacent line of any of the other subjects for any of the scheduling variables. Thus, average exam timing does not differ across subjects in a statistically significant way.

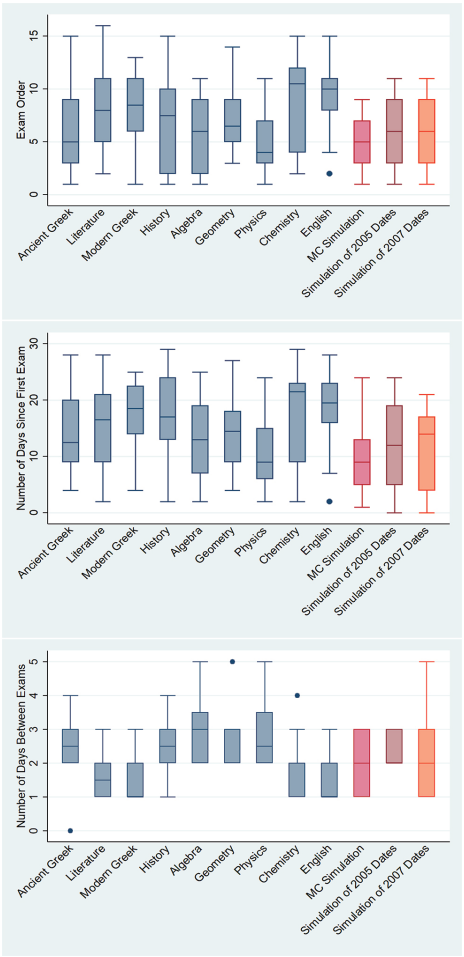


**Figure 2:** Randomness of exam schedules and Monte Carlo simulation.
Note: These figures present box plots for each of the three scheduling variables for each subject. The top panel refers to exam order, the second to the number of days since the first exam, and the bottom to the number of days between exams. We compare the central tendency and dispersion of the exam scheduling variables for each subject to the simulated ones from truly random schedules generated by a Monte Carlo procedure. We also include two box plots of the scheduling variables of 9,000 simulated exams using random draw without replacement from the exam dates of the 2004–2005 and the 2006–2007 school years. The middle bar of each box plot shows the median. The lower whisker corresponds to the 25th percentile and the upper whisker to the 75th percentile. The upper adjacent line corresponds to the 95th percentile and the lower adjacent line to the 5th percentile.

In addition, we perform a simulation exercise using two approaches to show that the average exam timing of each subject is not different from the scheduling of an exam generated by a truly random schedule-generating process. In the first approach we use a random permutation process to generate simulated exam schedules for nine subjects (i. e. the number of subjects we examine) that are indeed random. We use a random number generator to randomly choose number of days between exams, with equal probabilities on each value in the set {1, 2, 3}.[10] For each exam we add the number of days between that exam and all previous exams to calculate the number of days since the first exam. We obtain statistics for the exam scheduling variables for each subject across nearly 30 million simulated exam schedules. In the second approach we obtain statistics for the exam scheduling variables from 9,000 simulated exams using random draw without replacement from the exam dates of the 2004–2005 and the 2006–2007 school years.[11]

There are three box plots of each scheduling variable obtained from simulated exam schedules shown on the right of each panel in Figure 2, as a point of reference. One box plot corresponds to the Monte Carlo simulation of exam schedules, a second box plot corresponds to the simulation of exam schedules using exam dates from the 2004–2005 school year, and a third box plot shows the scheduling variable statistics associates with simulated exam schedules using exam dates from the 2006–2007 school year. We compare the central tendency and dispersion of each scheduling variable for each subject in our empirical data to the corresponding moments in the simulated schedules. Even though the median scheduling variables vary across subjects in Figure 2, we find that for each subject the basic parameters that describe the distribution of exam scheduling are consistent with those obtained from a random schedule-generating processes. In addition, we find that the lower (upper) adjacent line of the box plot for the simulated data is not above (below) the upper (lower) adjacent line of

that of any of the subjects in the empirical data, for any of the scheduling variables. Our findings support the randomness of exam scheduling in our empirical data.

## 4.4 Correlation of Exam Scheduling Variables

There is a mechanical relationship between two of the exam scheduling variables that we construct: the number of days since the first exam and exam order. Exams taken at a higher exam order are also the ones taken at a larger number of days since the first exam. The mechanical relationship between exam order and the number of days since the first exam contributes to a high and positive correlation between those two variables as shown in Table 4. The number of days between exams exhibits low correlation with the other two exam scheduling variables. To understand better how the exam scheduling variables co-move, we plot the average value for each exam scheduling variable across all exams by the number of days since the first exam in Figure 3. Figure 3 shows two facts. First, the average exam order increases with the number of days since the first exam but at a slower rate (average slope = 1/2.2 = 0.45). Second, the average number of days between exams does not appear to systematically increase or decrease with the number of days since the first exam. Even though exam order co-moves with the number of days since the first exam, the two are not a linear combination of each other, allowing for the separate OLS estimators to be uniquely defined from the first order conditions. Collinearity between exam order and the number of days since the first exam could potentially be a threat to the separate identification of the different exam scheduling effects and could lead to insignificant estimates.

**Table 4:** Correlation matrix of exam scheduling variables.

| | Exam order | Days since first exam | Days between exams |
|---|---|---|---|
| Exam order | 1.000 | | |
| Days since first exam | 0.982 | 1.000 | |
| Days between exams | –0.089 | 0.059 | 1.000 |

Note: This table shows the correlation coefficient between every pair of exam scheduling variables.

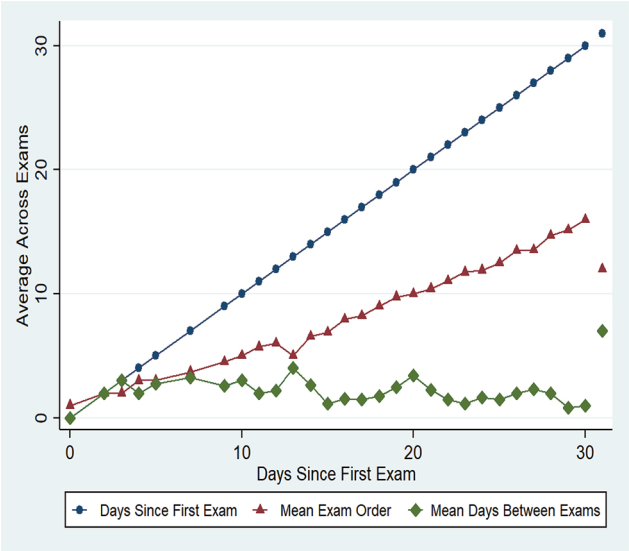


**Figure 3:** Co-movement of exam scheduling variables.
Note: This figure presents the mean value of each exam scheduling variable for each day since the first exam across all exams. Only one out of 323 exams in the dataset was ever administered on the 31st day after the first exam. The mean exam scheduling variables associated with that outlier exam are depicted as disconnected dots.

# 5 Empirical Strategy

## 5.1 Main Scheduling Effects

We calculate the effects of exam scheduling on standardized exam performance in a straightforward manner. We exploit across-cohort, -grade, and -subject variation in exam scheduling to identify three separate channels. We use a panel of nine compulsory subjects in the 10th and 11th grade. Table 16, Table 17, and Table 18 in Appendix D provide evidence that there is significant variation in exam timing across years, grades, and subjects. 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 regress the outcome variable for student $i$, in subject $s$, and grade $g$ on the *order the exam was taken*, the *number of days since the first exam*, the *number of days between exams*, standardized average (fall and spring semesters) midterm score $M$ in subject $s$, day of the week fixed effects, grade × subject fixed effects, grade × year fixed effects, a dummy variable for having been retained in the previous year, and gender. Controlling for the average midterm score in each specific subject allows for precision in capturing a student's differential level of preparedness across subjects. The coefficient of the *number of days between exams* reflects the type I scheduling effect (potential recuperation effect). The coefficient of the *number of days since the first exam* can be interpreted as a type II scheduling effect (potential fatigue effect), and the coefficient of the *exam order* captures type III scheduling effect (potential practice or warm-up effect). Using variation across cohorts, grades and subjects, we disentangle the practice effect from the fatigue and recuperation effects by controlling for *number of days between exams*, *number of days since the first exam*, and *exam order* simultaneously. Specifically, we run the following regression:

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

where $c \in \{stem, non\text{-}stem\}$. The influence of each scheduling variable is estimated separately for STEM and non-STEM subjects. For example, the two estimated $\alpha_3$ coefficients, $\alpha_{3stem}$ and $\alpha_{3non-stem}$, reflect the average impact of exam order on performance in STEM and non-STEM subjects, respectively.[12] Our specification allows for comparison of the estimated scheduling effects on exam performance in STEM and non-STEM subjects.[13] Standard errors are corrected for clustering at the cohort × classroom level to allow for heteroskedasticity and serial correlation at that level, as students who learn in the same room in a given year may share some error patterns.[14] Matrix X captures student characteristics, such as gender and a binary variable that captures having been retained in the previous year. We also control for age by including a full set of year of birth × cohort dummies in matrix X. This provides for slightly higher precision compared to including age dummies or continuous variables for age and age squared. Subject × Grade fixed effects are controlled for in matrix $\kappa$. Grade × Year-specific fixed effects are captured in matrix $\lambda$. We also control for day of the week fixed effects in matrix $\zeta$.[15]

The identification stems from the randomization of the date of the exam across subjects, from 1 year to the next, as well as between 10th and 11th grades. By controlling for grade-by-year fixed effects, we rely on within-grade-by-year and across-subjects variation in exam timing. Based on this approach, we examine whether differences in the exam scheduling of the same subject across combinations of grade and year are systematically associated with differences in exam performance. The basic idea is to compare the outcomes of students from different cohorts who have similar characteristics and face the same school environment, except for the fact that one cohort's exam schedule differs from 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 performance in STEM subjects tested in the same scheduling pattern in a different year for students with similar characteristics, with an analogous assumption for non-STEM subjects. One limitation of our analysis is that exams in different subjects may have different practice effects on subsequent exams, resulting in different scheduling effects.[16] If selection into concentrations is driven by student characteristics, then student characteristics may also influence type II (practice) scheduling effects. Our estimates reflect average scheduling effects and ignore differential effects from having taken exams in a different combination of subjects.

One may worry that the scheduling effects on exam scores may not reflect student performance in different exams spread out in the exam schedule but other subject-by-grade-specific influences, such as marking. If teachers mark exam papers in a time pattern similar to the pattern students take exams in, the lack of within subject-by-grade variation in the research design would render it impossible to tell whether the effect of exam schedules on scores is attributable to different student performance or different marking patterns. We provide evidence that time-specific patterns of marking do not pose a threat to identification in our research design.

For marking to be responsible for the scheduling effects on scores, the arrival pattern of exam papers to each teacher for marking must mirror the exam-taking pattern of the students. In a given exam season, a teacher

receives exam papers for marking as many times as the number of subject by grade configurations he/she is teaching. In the sampled school, each teacher teaches in 2.06 (SD 0.92) subject by grade configurations each year, on average. In other words, each teacher receives exam papers for marking only 2.06 times during an exam season. The exam-paper arrival pattern of the average teacher of 2.06 times is much shorter than the exam-taking pattern of the average student, who takes exams 13.29 times in each exam season (Table 3). The time pattern of exam-paper arrival to the teacher is not sufficient to identify the three distinct exam scheduling effects in our research design. Thus, it is not possible for marking to be a primary mechanism behind the exam scheduling effects on scores.

### 5.2 Nonlinear Scheduling Effects

Our main specification assumes that the scheduling effects are linear. One may expect though that exam scheduling effects may be less likely to kick in in the first two or three exams when student stamina may remain high. In other words, exam scheduling may matter more further down the exam season than earlier on.

To address this, we use natural breaks in the distribution of the exam scheduling variables to break each scheduling variable into binary variables that capture different levels of the underlying scheduling variables. Our specification for the estimation of nonlinear scheduling effects is shown below.

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

The *Days Between Exams* variable, which is associated with Scheduling Effect I, is broken into three binary variables: one captures exams taken up to and including 2 days after the last exam, one captures exams taken 3 days (the mode of this variable) after the last exam, and one captures exams taken more than 3 days after the last exam. The first group (i. e. exams taken up to and including 2 days after the last exam) are omitted as a reference group. A similar breakdown into binary variables is applied to the other scheduling variables. The *Days Since First Exam* variable, which is associated with Scheduling Effect II, is broken into three binary variables: one captures exams taken up to and including 9 days after the first exam, one captures exams taken between 10 and 19 days after the first exam, and one captures exams taken more than 19 days after the first exam. The first group (i. e. exams taken up to and including 9 days after the first exam) is omitted as a point of comparison. The *Exam Order* variable, which is associated with Scheduling Effect III, is broken into three binary variables: one captures exams with order up to five, one captures exams with order between six and nine, and one captures exams with order above nine. The first group (i. e. exams with place of order up to five) is omitted, as a point of comparison. The omitted group in specification 2 is exams taken within the first 9 days from the first exam, up to the fifth place of exam order, and not later than 2 days from the previous exam. This condition corresponds to roughly 15 % of the exams in our dataset.

## 6 Results and Discussion

### 6.1 Overall Scheduling Effects for STEM and Non-STEM Subjects

We begin the discussion of the scheduling effects with a presentation of restricted regression models where one or more exam scheduling variables is omitted. Because of the mechanical relationship between the exam scheduling variables, restricted models show whether some exam scheduling variables capture the same variation in the data generating process, potentially rendering them insignificant in the full model but significant in the restricted model. At the same time, restricted models inform us about the size and direction of omitted variable bias when one or more exam scheduling variables are excluded. Table 5 shows the estimated scheduling effects for STEM and non-STEM subjects using restricted models.[17] The results of the regressions on subsets of the scheduling variables show that there is no subset of the scheduling variables that would give estimates close to the ones obtained when all three scheduling variables are included, suggesting the presence of omitted variable bias.

**Table 5:** Subsets of the effects of exam timing on performance by STEM.

| Variables | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Scheduling Effect III for non-STEM | 0.001* | | | 0.000 | | 0.001 | –0.002 |
| | (0.001) | | | (0.004) | | (0.001) | (0.004) |
| Scheduling Effect III for STEM | 0.003*** | | | 0.017*** | | 0.002*** | 0.016*** |
| | (0.001) | | | (0.005) | | (0.001) | (0.005) |
| Scheduling Effect II for non-STEM | | 0.001* | | 0.001 | 0.001 | | 0.002 |
| | | (0.000) | | (0.002) | (0.000) | | (0.002) |
| Scheduling Effect II for STEM | | 0.001** | | –0.006*** | 0.001** | | –0.006*** |
| | | (0.000) | | (0.002) | (0.000) | | (0.002) |
| Scheduling Effect I for non-STEM | | | –0.014*** | | –0.012*** | –0.012*** | –0.012*** |
| | | | (0.004) | | (0.004) | (0.004) | (0.004) |
| Scheduling Effect I for STEM | | | –0.004 | | –0.005 | –0.005 | 0.002 |
| | | | (0.003) | | (0.003) | (0.003) | (0.003) |
| Observations | 14,258 | 14,258 | 14,258 | 14,258 | 14,258 | 14,258 | 14,258 |
| R-squared | 0.985 | 0.985 | 0.985 | 0.985 | 0.985 | 0.985 | 0.985 |

Notes: The dependent variable is the standardized 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 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' gender, and a dummy that indicates whether a student is retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

When considering all three channels of exam scheduling influence on performance, we are able to estimate the distinct impact of each exam scheduling variable while keeping fixed every other aspect of exam scheduling. Our results for the average contemporaneous scheduling effects for exams in STEM and non-STEM subjects are shown in column 7 of Table 5. We provide here a visual representation of the scheduling effects for STEM and non-STEM subjects in Figure 4. Figure 4 also shows the scheduling effects across all subjects from Table 19 as a reference.

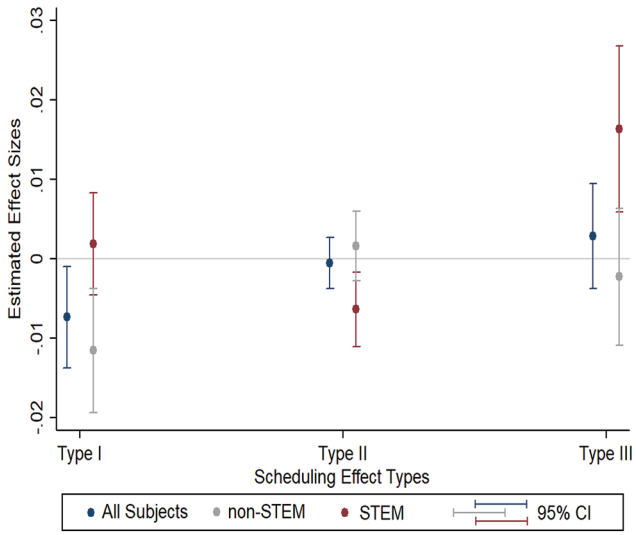


**Figure 4:** Estimated scheduling effects on performance.

*Scheduling Effect I*

Scheduling Effect I is found to be statistically significant only for non-STEM subjects, as shown in Figure 4, as well as in column 1 of Table 20 in Appendix F. An additional day between two exams decreases exam performance in the subsequent non-STEM exam by 0.012 of a standard deviation. A student who takes a non-STEM exam 2 days after the previous exam – the average gap between any two exams in the dataset – experiences a decrease in their performance equal to 0.024 of a standard deviation. The estimated Scheduling Effect I for STEM subjects is not statistically significant. The overall Scheduling Effect I disputes the hypothesis that time between exams is used productively in order to prepare for the subsequent exams, on average.

*Scheduling Effect II*

Scheduling Effect II is found to be statistically significant, as shown in Figure 4 (also in column 1 in Table 20), which confirms our Hypothesis 2 on the potential existence of a fatigue effect, but only for STEM subjects. In particular, an additional day since the first exam a student took decreases their performance on the subsequent STEM exam by 0.006 of a standard deviation. For example, a student who takes an exam 24 days after 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. In contrast, the estimated Scheduling Effect II for non-STEM subjects is not statistically significant, indicating that the underlying mechanism, which may potentially relate to cognitive fatigue, is relevant only for 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, as shown in Figure 4 and column 1 in Table 20, which confirm our Hypothesis 3 on the existence of a practice or warm-up effect, but 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 he/she takes is estimated to be 0.16 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 only for STEM subjects associated with taking exams later in the exam schedule. We call this warm-up effect, and it is found to have the largest impact on performance of the three scheduling effects. 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 or scaffolding on the performance in tasks that are intensive in analytic reasoning, such as exams in STEM subjects.

Even though exam scheduling merely influences test performance and not student learning during the year, it is worth comparing the estimated scheduling effects to the marginal effects of other educational interventions. As examples, we consider the effects of class size, teacher quality, classroom instructional time, and school attendance on test scores. Krueger (1999) shows that a one-standard-deviation reduction in class size (approximately eight students) increases language test scores by up to 0.3 of a standard deviation – sufficient to move a student from the 25th percentile to the median of the score distribution. Several studies report that a one-standard-deviation increase in teacher quality improves student test scores by approximately one-tenth of a standard deviation (Rockoff 2004; Rivkin, Hanushek, and Kain 2005; Aaronson, Barrow, and Sander 2007; Kane, Rockoff, and Staiger 2008). Carrell and West (2010) revise downwards the estimated teacher effects after correcting for the bias in value-added estimates of teacher quality pointed out by Rothstein (2010). In particular, Carrell and West (2010) find that a one-standard-deviation change in professor quality results in a 0.05-standard-deviation change in student performance. Lavy (2015) finds that, on average, a 1-hour increase per week in math, science, or language instruction raises test scores in those subjects by 0.015 standard deviations. Hanushek, Peterson, and Woessmann (2012) estimate a scaling factor of 0.25 standard deviations per year of school attendance (180 days) for all grades and subjects.

In the context of the estimated effects of various education inputs on test performance, we view our estimated scheduling effects as sizable. To understand better how much of other education inputs are required to achieve an impact on performance similar to that of exam scheduling in absolute terms, we summarize our comparisons in Table 6. Table 6 shows the units of change in various factors in the education production function necessary to change student performance by the same amount in absolute terms as a one-unit change in each of the three scheduling variables.[18] An additional day between two exams (Scheduling Effect I) has the same average effect on student performance as an increase in class size of 0.04 standard deviations (less than a student), a decrease in teacher quality of 0.12 standard deviations, a decrease in instructional time of a little less than 1 hour per week, or a decrease in school attendance of approximately 9 days. An additional day since the first exam (Scheduling Effect II) decreases performance by the same amount as an increase in class size of 0.02 of a standard deviation (less than a student), a decrease in teacher quality of 0.06 standard deviations, a decrease in instructional time of almost half hour per week, or a decrease in school attendance of roughly 4 days. A movement of one place in the order of exams in the schedule (Scheduling Effect III) has a benefit equivalent to reducing the class size by 0.05 of a standard deviation (less than a student), raising teacher quality by roughly one-third of a standard deviation, increasing instructional time by roughly 1 hour per week or attending school for additional 12 days.

**Table 6:** Comparison of scheduling effect sizes to education inputs.

| Treatment effect | Class size (Krueger) | Teacher quality (Carrel & West) | Instructional time (Lavy) | Attendance (Hanushek) |
|---|---|---|---|---|
| Scheduling Effect I | 0.04 SD | 0.12 SD | 0.80 hpw | 8.64 days |
| Scheduling Effect II | 0.02 SD | 0.06 SD | 0.40 hpw | 4.32 days |
| Scheduling Effect III | 0.05 SD | 0.32 SD | 1.07 hpw | 11.52 days |

Notes: The table shows the change in various education interventions studied in the literature that would be as effective in changing student performance in absolute terms as a one-unit change in the variables associated with Scheduling Effect I, Scheduling Effect II, and Scheduling Effect III. SD stands for standard deviations; hpw stands for additional hours of instructional time per week; days stands for additional days of school attendance.

The channels through which exam scheduling influences performance are rather interwoven and a policy design is unlikely to manage to exploit only the scheduling aspects with the most beneficial marginal effects. Optimizing the exam schedule requires taking into account as many channels of exam scheduling influence on test scores as possible. We discuss the net benefit of optimizing exam scheduling in Section 7.

## 6.2 Nonlinear Scheduling Effects for STEM and Non-STEM Subjects

We find economically and statistically substantial nonlinear effects of exam scheduling on performance. Table 7 shows nonlinear effects by breaking the variables that reflect the scheduling effects into bins (using specification 2).[19] Small changes in the cutoffs used to define the binary variables leave the estimated non-linearities qualitatively identical. The omitted category (and point of reference) for the each estimated nonlinear effect consists of exams taken within the first 9 days after the first exam, up to the fifth place of exam order, and not later than 2 days after the previous exam. The first panel in Table 7 shows our nonlinear estimates of scheduling effect I. We find that for non-STEM subjects, exams that are taken exactly 3 days after the previous exam are associated with significantly lower performance compared to exams taken less than 3 days after the previous exam. Taking a non-STEM exam 3 days after the last exam decreases performance by 0.04 standard deviations, compared to taking the exam less than 3 days after the last one. At the same time, non-STEM exams taken more than 3 days after the previous exam are associated with performance comparable to that on exams taken less than 3 days after the previous exam. For STEM subjects, exams taken either exactly 3 days after the previous exam or more than 3 days after the previous exam are found to be associated with performance comparable to exams taken less than 3 days after the previous exam, suggesting that nonlinear effects of the time between exams are not present in STEM exams.

**Table 7:** Nonlinear 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) |

Notes: The dependent variable is the standardized final exam score at the subject and grade level. The sample includes 14,258 observations. Cluster-robust standard errors at the classroom by year level are reported in parentheses. Specification includes 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' gender and a dummy that indicates whether a student is retained. The comparison group is exams taken within the first 9 days from the first exam, up to the fifth place of exam order, and not later than 2 days from the previous exam. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

The second panel in Table 7 shows our estimates on nonlinearities of Scheduling Effect II. For STEM subjects, exams taken between 10 and 19 days after the first exam are associated with performance comparable to that of exams taken less than 10 days after the first exam, but this is not the case for the next category. STEM exams taken more than 19 days after the first exam are associated with performance significantly lower than the performance on exams taken earlier. Taking a STEM exam more than 19 days after the first exam decreases performance by 0.045 standard deviations, compared to taking the exam less than 10 days after the first exam. For non-STEM subjects, exams taken either between 10 and 19 days after the first exam or more than 19 days after the first exam are found to be associated with performance comparable to the exams taken less than 10 days after the first exam, suggesting that nonlinear 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 nonlinearities 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 performance on exams at lower places of order, suggesting the existence of significant nonlinearities in Scheduling Effect III. Taking a STEM exam in the 10th place of order or later increases performance by 0.055 standard deviations, compared to taking the exam in the 6th or lower place of order, controlling for other influences. For non-STEM subjects, exams 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 exams at lower places of order, suggesting that no-linear effects of exam order are not present in non-STEM exams.

## 6.3 Differential Scheduling Effects by Student Gender

Although the two genders do not seem to have largely different midterm or final exam scores – particularly in STEM subjects (see Table 14 in Appendix B), we find that males and females experience different effects with respect to exam scheduling. We estimate separate scheduling effects for males and females by allowing scheduling effects coefficients $\alpha_{1c}$, $\alpha_{2c}$ and $\alpha_{3c}$ in specification 1 to vary by gender in the same model. We also compare each scheduling effect across genders by interacting each scheduling variable in specification 1 with a binary variable indicating being female. Table 8 shows the heterogeneous effects of exam scheduling on performance by gender. We compare scheduling effects in STEM and non-STEM exams between boys and girls. Boys are more responsive to all types of scheduling effects than females. We find that Scheduling Effect I (days between exams) on STEM exams is significantly higher for boys than for girls. In particular, one additional day between exams improves the subsequent performance of boys by 0.007 of a 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 both genders, is not found to be significantly different across genders.

**Table 8:** Differential 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) |

Notes: The dependent variable is the standardized final exam score at the subject and grade level. The sample includes 14,258 observations. Cluster-robust standard errors at the classroom by year level are reported in parentheses. Specification includes 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' gender, and a dummy that indicates whether a student is retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

The exam fatigue effect,which is captured by our Scheduling Effect II (days since the first exam) does not seem to differ significantly across genders, although girls seem to be less sensitive to that influence of exam scheduling on performance on STEM exams. Girls experience a roughly 45 % 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 (exam order) differs significantly between boys and girls in STEM exams. In particular, boys have roughly 75 % higher warm-up effect in STEM exams compared to girls. Having taken an additional exam earlier in the exam schedule improves the subsequent performance of boys on 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 Scheduling Effects by Student Prior Performance

We compare the scheduling effects in each part of the prior performance distribution. Figure 5 (and Table 22 in Appendix G) shows heterogeneous effects of exam scheduling on performance by four quantiles of student prior performance.[20] The top graph in Figure 5 shows differential Scheduling Effect I in STEM and non-STEM subjects across different quantiles of prior performance. For 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 a standard deviation, confirming Hypothesis 1 (recuperation effect) only for non-STEM subjects.

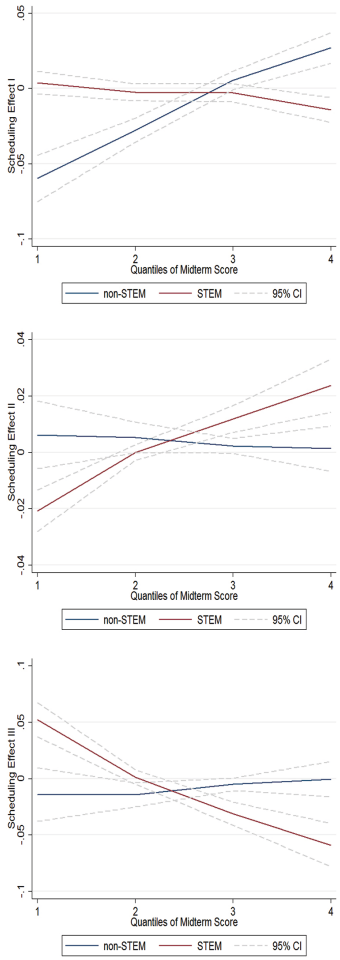


**Figure 5:** Differential scheduling effects by prior performance.
These figures present differential scheduling effects by quantiles of prior performance, proxied by midterm score in each subject. The top panel refers to Scheduling Effect I, captured by the days between exams variable. The middle panel refers to Scheduling Effect II, captured by the days since the first exam variable. The bottom panel refers to Scheduling Effect III, captured by the exam order variable.

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 on STEM exams. Specifically, Scheduling Effect I is associated with a decrease in exam performance in STEM subjects for the highest performing student of 0.014 standard deviations.

Scheduling Effect II (days since the first exam), in the middle graph in Figure 5, is found to be small and occasionally significant for non-STEM subjects, while Scheduling Effect II on STEM exams is increasing significantly with prior performance. In particular, an additional day since the beginning of the exam season decreases performance on STEM exams of quantile 1 (bottom quantile) by 0.021 of a 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's and 4's performance on STEM exams by 0.012 and 0.024 of a standard deviation, respectively.

The pattern of estimated coefficients across quantiles of prior performance for Scheduling Effect III is very different from that of Scheduling Effect II, as shown in the bottom graph of Figure 5. Although the effect of 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 performance on 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's and 4's performance by 0.03 and 0.06 of a standard deviation, respectively.

## 6.5 Robustness Checks

We verify the robustness of our main estimates to several changes in model specification with results shown in Table 9. All specifications include a full set of individual controls and the same fixed effects as specifications (1)–(6). Column 1 presents our main estimated effects from column 7 in Table 5, as a point of reference. Column 2 shows our estimates with the inclusion of student fixed effects. Column 3 shows results when we include elective courses. As mentioned earlier, electives are tested on the same dates (i. e. on specific dates every student takes an exam is some elective). Taking exams in different type of electives (e. g. in chemistry or in history) may impact a student's performance on other exams differently. In column 4, our model excludes exams administered on the same day because of concerns that same-day exams may affect each other differently than exams that are farther apart.[21] The model in column 5 includes an additional scheduling variable: the number days until the next exam. Controlling for the number days until the next exam allows us to account for potential bunching of exams following a longer break between exams. We find no economically or statistically significant effect of the number of days until the next exam, while the coefficients on other exam scheduling variables remained qualitatively unchanged.[22] Overall, the estimates from our robustness specifications are very similar to those from our main specification, and provide strong evidence that the results are not driven by inconsistencies in the data or omitted information.

**Table 9:** Robustness of the effects of exam timing on performance.

| Variables | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Baseline results | With student FE | Including electives | Same-day exams excluded | Control for bunching |
| Scheduling Effect III for non-STEM | –0.002 | –0.000 | –0.001 | –0.000 | –0.003 |
| | (0.004) | (0.005) | (0.005) | (0.005) | (0.005) |
| Scheduling Effect III for STEM | 0.016*** | 0.015*** | 0.014*** | 0.015*** | 0.018*** |
| | (0.005) | (0.005) | (0.004) | (0.005) | (0.006) |
| Scheduling Effect II for non-STEM | 0.002 | 0.001 | 0.001 | 0.001 | 0.002 |
| | (0.002) | (0.003) | (0.002) | (0.003) | (0.003) |
| Scheduling Effect II for STEM | –0.006*** | –0.006** | –0.006*** | –0.006** | –0.008*** |
| | (0.002) | (0.002) | (0.002) | (0.002) | (0.003) |
| Scheduling Effect I for non-STEM | –0.012*** | –0.013*** | –0.013*** | –0.013*** | –0.012*** |
| | (0.004) | (0.004) | (0.004) | (0.005) | (0.005) |
| Scheduling Effect I for STEM | 0.002 | 0.001 | –0.000 | 0.001 | 0.003 |
| | (0.003) | (0.003) | (0.003) | (0.004) | (0.003) |
| Observations | 14,258 | 14,258 | 16,763 | 14,092 | 13,169 |
| *R*-squared | 0.985 | 0.987 | 0.980 | 0.987 | 0.985 |
| Student FE | No | Yes | No | No | No |

Notes: Column 1 presents the main results from Table 20, as the baseline for our robustness checks. Column 2 is based on the same regression model as in column (1), with the addition that we include student FE. In column (3) we include in the regression all electives subjects. Column 4 is based on the same regression as columns (1) and (2), while we exclude same-day exams. Column (5) shows the the regression model of column (1), with an additional control for the number of days until the next exam. The dependent variable in each specification is the standardized 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 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. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

## 7 What Is the Optimal Exam Schedule?

Our empirical investigation allows us to reveal the marginal contributions of three distinct aspects of exam scheduling. Given our findings on the effects of exam scheduling – i. e. as a result of (i) the order of exams, (ii) the number of days between exams, and (iii) the number of days since the first exam – on students' grades, researchers, administrators, and policy makers may ask: What is the optimal exam schedule in terms of overall performance?

To identify the optimal exam schedule, we conduct three simulation experiments. In each experiment, we simulate a large number of exam schedules for a given number of exams, half of which are in STEM and half in non-STEM subjects. For each exam schedule we calculate the average grade across all exams, taken in random order with a random draw of days between them in the range [1,3],[23] and thus a random number of days since the first exam.

We follow a straightforward simulation approach. First, we create random permutations of a given number of exams. We show results using four, six, and eight exams. Second, we use a random number generator to randomly choose a number of days between exams with equal probabilities on each value in the set {1,2,3}. For each exam we add the number of days between that exam and all previous exams to calculate the number of days since the first exam. The score in the first exam is set at zero and every subsequent exam is assigned a standardized score equal to the combined estimated marginal effects presented in column 7 of Table 5, our preferred specification. For simplicity, we ignore nonlinear scheduling effects in this exercise. Finally, we average the scores across subjects for each simulated exam schedule.

In Figure 6 we plot the distribution of average scores for each simulation experiment. The distributions of the simulated exam schedules reveal significant variation in average scores due to exam scheduling. We find that the higher the number of exams taken, the higher the potential benefit from optimizing exam scheduling, since additional exams are associated with positive net marginal effects on performance. For each number of exams taken, we compare exam schedules with low average scores and schedules with high average scores. Table 10 summarizes the range of average performance due to differences in exam scheduling of a given set of exams: four, six, and eight exams.

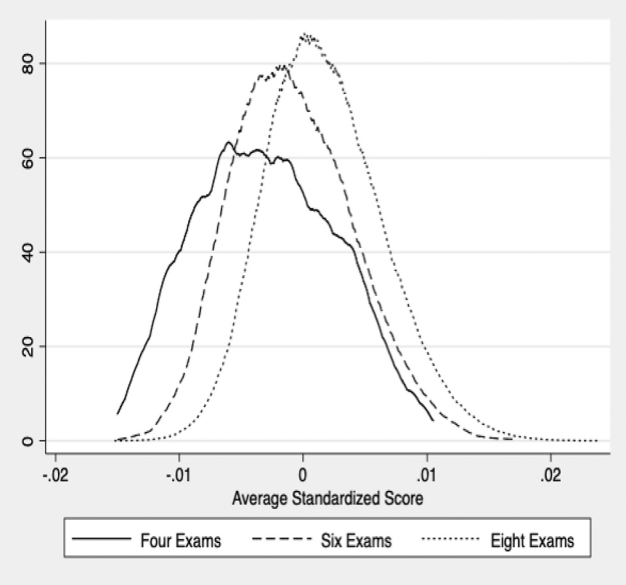


**Figure 6:** Simulated average scores from different exam schedules.
Note: This figure shows the distributions of overall performance across schedules from three simulation experiments for a different number of exams (four, six, and eight exams).

**Table 10:** Range of average scores from simulated exam schedules.

| Simulation | *N* | Mean | SD | Min | Max | Range | 5 % | 95 % | 95 % Range |
|---|---|---|---|---|---|---|---|---|---|
| Four exams | 2,400 | – 0.003 | 0.006 | – 0.015 | 0.011 | 0.026 | – 0.012 | 0.006 | 0.018 |
| Six exams | 72,000 | – 0.001 | 0.005 | – 0.015 | 0.017 | 0.032 | – 0.008 | 0.008 | 0.016 |
| Eight exams | 4,032,000 | 0.002 | 0.005 | – 0.015 | 0.024 | 0.039 | – 0.006 | 0.010 | 0.016 |

Notes: The table shows summary statistics for overall performance from three simulation experiments of different exam schedules for four, six, and eight subjects. In each simulation experiment, we show the variation in overall performance attributable to exam scheduling. For example, when we simulated 2,400 different schedules for exams in four subjects – two STEM subjects and two non-STEM subjects – we found that the range of overall performance due to exam scheduling is 0.026 standard deviations, while the 95 % interval is 0.018 standard deviations.

We find that exam schedules with the highest average grade schedule every non-STEM exam to be taken before every STEM exam. Conversely, exam schedules with the lowest average grade have every STEM exam scheduled to be taken before every non-STEM exam. Pushing STEM exams after non-STEM exams means that both exam order and the number of days since the first exam increase for STEM exams. The positive exam order effect offsets the negative effect of the days since the first exam with the net effect for STEM exams being a small positive, suggesting that pushing STEM exam after non-STEM exams may be optimal. The difference between exam schedules with the lowest average scores and schedules with the highest average scores ranges roughly between 0.03 and 0.04 standard deviations, depending on the number of exams. When we focus on the 95 % interval of average scores, we see that the difference in performance between schedules in the fifth percentile and schedules in the 95th percentile is approximately 0.02 standard deviations. In other words, 95 % of the time, the potential benefit in terms of overall average grade from optimizing the exam schedule can be as big as 0.02 standard deviations.

To highlight the potential impact of optimizing the exam schedule, we use the marginal effects of other standard educational interventions, presented in Section 2 and summarized in Table 6 as benchmark. We find that optimizing the exam schedule has a benefit on average performance equivalent to decreasing class size by 0.07 of standard deviation (less than a student), raising teacher quality by 0.4 standard deviations, increasing weekly instructional time by more than an hour, or attending school for additional 14 days. Reducing class size, increasing teacher quality, increasing instructional time, or increasing school attendance may be potentially more costly than optimizing the exam schedule. It is important to note that while various educational inputs may influence test performance by changing student learning, exam scheduling is more likely to influence test performance only rather than learning during the year

## 8 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, previous studies have not simultaneously measured and compared those effects. In addition, 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, unraveling the causal effects of exam scheduling on student achievement has been difficult due to issues related to self-selection.

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 through a lottery. Randomized exam dates, mandatory attendance, stable curriculum and assessment protocols, along with extensive background data on students, allow 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, cohorts, and subjects, we disentangle three channels of scheduling effects on academic performance across STEM and non-STEM subjects, overall, by gender, and at different parts of the ability distribution. We find that the time between exams (Scheduling Effect I) has, on average, a negative and significant effect on students' productivity in non-STEM subjects, while the effect on STEM subjects is not statistically different from zero. Scheduling Effect I differs across the ability distribution. In particular, students with high prior performance exhibit positive and significant type I Scheduling Effects on their performance in non-STEM subjects. At the same time, Scheduling Effect I is negative and significant for students of lower prior performance in non-STEM subjects.

We also find that the length of time since the first exam (Scheduling Effect II) has a negative and significant effect on performance in STEM subjects, while the effect for non-STEM subjects is not significant. Scheduling Effect II is found to be positive and significant for 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. In contrast, non-STEM subjects exhibit a Scheduling Effect III that is not statistically different from zero, on average. Similar to the other scheduling effects, there are heterogeneous effects across prior performance. Low-achieving students experience positive and significant effects of Scheduling Effect III, while the corresponding effect is negative and significant for high achievers. High achievers, whose meta-cognition may already be high, may experience lower cognitive returns to practice from exams compared to low achievers, while additional practice may induce fatigue; this pattern is confirmed in our empirical findings. Moreover, we find that exam productivity in STEM subjects increases faster for boys than it does for girls as they take additional exams (the warm-up effect).

To understand the implications of our findings for designing the optimal exam schedule, we conduct a series of simulation experiments. Our simulations show that the higher the number of exams taken the higher the potential benefit from optimizing exam scheduling. Our simulations show that optimizing the exam schedule can improve overall performance by as much as 0.02 standard deviations.

Our findings have important implications for both education policy and task management alike. 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 a benefit equivalent to 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.

## Notes

1 Boksem, Meijman, and Lorist (2005); van der Linden, Frese, and Meijman (2003); Lorist et al. (2000); Hockey and Earle (2006); Rohrer and Taylor (2006, 2007); Rohrer (2009); Taylor and Rohrer (2010); Bjork, Dunlosky, and Kornell (2013).
2 STEM is an acronym for fields of study related to Science, Technology, Engineering, or Mathematics.
3 The process of knowing something so deeply that recollecting it becomes effortless is called "overlearning." For instance, someone who has overlearned, e. g. operations in algebra, does not have to use resources for algebra while learning a concept in physics (that makes use of algebraic operations). This means that there are more mental resources for new concepts in physics. High performing students may tend to overlearn material, while lower-performing students may learn the material just well enough to replicate it. Therefore, high-performing students may tend to have a substantial reduction in cognitive load from previous knowledge. Using our previous example of algebra and physics, a high-performing student is more likely to have learned algebra so well that studying physics is much easier. On the other hand, a lower-performing student is likely to have learned algebra just well enough to answer algebra problems, and therefore must exert greater effort when studying physics. We direct the interested reader to the work of Stephen Chew (Chew 2007), (www.samford.edu/how-to-study), who has also produced study videos for students.
4 Every student can take only the courses available for his or her grade level.
5 Students are alphabetically assigned to classrooms. In a given year, there are 4–5 classrooms per grade. Usually, different teachers teach the same subject in different classrooms. Thus, there are multiple teachers teaching in the same subject by grade configuration. For the sampled school, in a given year, 2.62 teachers (SD = 1.13) teach in the same subject by grade configuration, on average. Teachers only teach subjects relevant to their specialty.
6 See Laws of the Hellenic Republic 2525/1997 (A 188), and 2909/2001 (A 90) as amended by Presidential Decree 60/2006 published in the Government Gazette Issue 65 volume A.
7 Students in the Classics track take ancient Greek, philosophy, and Latin. Students in the Science track take advanced mathematics, advanced physics, and advanced chemistry. Students in the Information Technology track take advanced mathematics, advanced physics, and communications technology.
8 Our analysis excludes track electives, additional electives, as well as a compulsory course on religious education.
9 Nontraditional public schools make up around 10 % of public schools in Greece and consist of charter, evening and religious schools. There is only one non-traditional (evening) school out of a total of 16 in the sampled school's district. Private schools represent roughly 8 % of schools in Greece. There are also two private schools out of 16 in the sampled school's district.
10 The number of days between exams ranges between 1 and 3 for more than 90 % of the exams in our empirical data
11 The exam dates vary from 1 year to the next.
12 We also perform an analysis of overall scheduling effects across subjects, regardless of STEM or non-STEM designation. Table 19 in Appendix E reports our results on the overall analysis. As we demonstrate later, because STEM and non-STEM subjects exhibit scheduling effects of opposite signs, the scheduling effects appear to be rather muted in the overall analysis.
13 We also consider an alternative specification where the *exam order* variable is divided by the *days since first exam* variable. This specification allows for the estimation of the marginal effect of an additional exam *per* day since the first exam. The results are presented in Table 25 in Appendix J.
14 Students remain in the same classroom for all compulsory courses for the duration of each school day. Assignment of students to classrooms is done alphabetically based on surnames.
15 We report the estimated effects associated with each day of the week in Table 26 in Appendix K.
16 In Table 23 in Appendix H, we present an analysis of differential scheduling effects by exam history. Exam history is measured by the number of STEM and the number of non-STEM exams taken prior to attempting a subsequent exam. Overall, we find that interacting each scheduling effect in STEM or non-STEM subjects with the number of STEM or non-STEM exams taken prior does not provide evidence of significant differential scheduling effects by exam history on exam performance, reinforcing the validity of our analysis. The only exception is the interaction between the number of non-STEM exams taken and Scheduling Effect I for non-STEM subjects. In particular, having taken more non-STEM exams is associated with higher impact of the number of days between exams on non-STEM performance.
17 Table 24 in Appendix I shows estimates from restricted models across all subjects.

18 We assume that an increase or a decrease of an input by the same amount changes student performance by the same amount but in opposite direction.
19 We also explore nonlinear exam scheduling effects across the entire range of values of scheduling variables using specification 4 in Appendix F. Specification 4 does not rely on natural breaks in the distribution of exam scheduling variables but uses the quadratic form of the exam scheduling variables. The results are reported in Table 20 in Appendix F.
20 We also investigate heterogeneous scheduling effects with respect to prior performance by interacting each scheduling variance with midterm performance, our proxy for prior performance. The results are presented in Table 21 in Appendix G.
21 Sample size in the model of column 4 decreased by 166 observations compared to the model of column 1 of Table 9 because of excluded same-day exams.
22 The correlation between the number of days since the last exam and the number of days until the next exam is small and not statistically significant ($\rho = -0.081$ with p-value 0.313, suggesting that bunching of exams following a longer break between exams may not be a threat to identification.)
23 As the number of days between exams ranges between 1 and 3 for more than 90 % of the exams in our empirical data, we can be more confident that the estimated marginal effect of days between exams on performance is valid within the [1,3] range.

## References


Aaronson, D., L. Barrow, and W. Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of labor Economics* 25 (1): 95–135.

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.

Ambrose, S. A., M. W. Bridges, M. DiPietro, M. C. Lovett, and M. K. Norman. 2010. *How Learning Works: Seven Research-Based Principles for Smart Teaching*. Bringham: John Wiley & Sons. https://search.proquest.com/openview/86e61dd13a927ca84432570e1d8ec460/1?pq-origsite=gscholar&cbl=28962.

Askell-Williams, H., M. J. Lawson, and G. Skrzypiec. 2012. "Scaffolding Cognitive and Metacognitive Strategy Instruction in Regular Class Lessons." *Instructional Science* 40 (2): 413–43.

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: 417–44.

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): 107–16.

Buser, T., and N. Peter. 2012. "Multitasking." *Experimental Economics* 15 (4): 641–55.

Carrell, S. E., and J. E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy* 118 (3): 409–32.

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): 645–51.

Chew, S. L. 2007. "Study More! Study Harder! Students' and Teachers' Faulty Beliefs about How People Learn." *Essays from Excellence in Teaching Volume VII* 30, 22.

Coviello, D., A. Ichino, and N. Persico. 2014. "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): 528–54.

Eilam, B., and I. Aharon. 2003. "Students' Planning in the Process of Self-regulated Learning." *Contemporary Educational Psychology* 28 (3): 304–34.

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.

Goulas, S., and R. Megalokonomou. 2015. "Knowing Who You Are: The Effect of Feedback Information on Exam Placement." *University of Warwick, mimeo*.

Goulas, S., R. Megalokonomou, and Y. Zhang. 2018. "Does the Girl Next Door Affect Your Academic Outcomes and Career Choices?" *IZA Discussion Paper Number 11910*.

Halpern, D. F. 2004. "A Cognitive-Process Taxonomy for Sex Differences in Cognitive Abilities." *Current Directions in Psychological Science* 13 (4): 135–39.

Halpern, D. F. 2013. *Sex Differences in Cognitive Abilities*. New York: Psychology Press. https://doi.org/10.4324/9781410605290.

Hanushek, E. A., P. E. Peterson, and L. Woessmann. 2012. "Achievement Growth: International and Us State Trends in Student Performance. Pepg report no.: 12–03." *Program on Education Policy and Governance, Harvard University*.

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): 494–95.
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. "Investigating the Effects of Exam Length on Performance and Cognitive Fatigue." *PLOS ONE* 8 (8): 1–9.
Kane, T. J., J. E. Rockoff, and D. O. Staiger. 2008. "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City." *Economics of Education Review* 27 (6): 615–31.
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. Weaver III. 2007. "Repeated Testing Sessions and Scholastic Aptitude in College Students' Metacognitive Accuracy." *European Journal of Cognitive Psychology* 19 (4–5): 689–717.
Krueger, A. B. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics* 114 (2): 497–532.
Lavy, V. 2015. "Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries." *The Economic Journal* 125 (588): F397–F424.
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): 614–25.
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. "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: 134–38. 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): 10593–97.
Özsoy, G., A. Memiş, and T. Temur. 2017. "Metacognition, Study Habits and Attitudes." *International Electronic Journal of Elementary Education* 2 (1): 154–66.
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.
Rivkin, S. G., E. A. Hanushek, and J. F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–58.
Rockoff, J. E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94 (2): 247–52.
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): 1209–24.
Rohrer, D., and K. Taylor. 2007. "The Shuffling of Mathematics Problems Improves Learning." *Instructional Science* 35 (6): 481–98.
Rothstein, J. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *The Quarterly Journal of Economics* 125 (1): 175–214.
Taylor, K., and D. Rohrer. 2010. "The Effects of Interleaved Practice." *Applied Cognitive Psychology* 24 (6): 837–48.
Tversky, A., and D. Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124–31.
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.
Vygotskiĭ, L. S. 2012. *Thought and Language*. Cambridge, MA: MIT Press.
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): 181–95.
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.

# Appendices

## A 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

**Figure 7:** Example of exam schedule.
Note: The picture above shows the exam schedule for students in the 11th grade in May–June 2005. The first and second columns 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. On some days students in different concentrations take exams in different subjects (e. g. May 23, May 30, June 10). 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 2, three concentration subjects (one for each concentration) were testes for students of the same grade: Latin, chemistry, and communications technology.

## B Supplementary Descriptive Statistics

In this section, we provide supplementary descriptive tables of student-level data. Table 11, Table 12, and Table 13 provide student-level summary statistics for each cohort for students in the 10th grade, 11th grade, and overall, respectively. We find that students characteristics are substantially similar across cohorts and grades.

**Table 11:** Summary statistics for 10th graders.

| 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 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2007 | N | 108 | 108 | 107 | 108 | 108 | 108 |
| | 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 |
| 2008 | N | 116 | 116 | 111 | 116 | 116 | 116 |
| | 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 |
| 2009 | N | 98 | 98 | 98 | 98 | 98 | 98 |
| | 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 |
| 2010 | N | 92 | 92 | 92 | 92 | 92 | 92 |
| | 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 |
| Total | N | 113 | 113 | 112 | 113 | 113 | 113 |
| | 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 |

Notes: This table presents the mean, standard deviation, and number of observations for a set of variables by year (2002–2010) for students in 10th grade. Variables include gender, age, GPA (between 0 and 20), midterm score (between 0 and 20), final exam score (between 0 and 20), and a dummy that indicates whether a student is retained.

**Table 12:** Summary statistics for 11th graders.

| 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 |

Notes: This table presents the mean, standard deviation, and number of observations for a set of variables by year (2002–2010) for students in 11th grade. Variables include gender, age, GPA (between 0 and 20), midterm score (between 0 and 20), final exam score (between 0 and 20), and a dummy that indicates whether a student is retained.

**Table 13:** 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 |

Notes: This table presents the mean, standard deviation, and number of observations for a set of variables by year (2002–2010) for students in 10th and 11th grade. Variables include gender, age, GPA (between 0 and 20), midterm score (between 0 and 20), final exam score (between 0 and 20), and a dummy that indicates whether a student is retained.

Table 14 shows exam-level descriptive statistics for males and females for STEM subjects, non-STEM subjects, and overall. We find that males and females seem to have similar midterm and final exam scores in STEM subject, while females have slightly higher average midterm and final exam scores in non-STEM subjects than males.

**Table 14:** Summary statistics for midterm and final exam scores by gender.

| | | STEM exams | | Non-STEM exams | |
|---|---|---|---|---|---|
| Gender | | Midterm | Final exam | Midterm | Final exam |
| Male | Mean | 16.96 | 12.26 | 16.93 | 13.32 |
| | SD | 2.54 | 6.00 | 2.52 | 4.93 |
| | N | 5680 | 5036 | 5214 | 5233 |
| Female | Mean | 17.18 | 12.39 | 17.78 | 14.59 |
| | SD | 2.46 | 6.04 | 2.18 | 4.76 |
| | N | 6651 | 5995 | 7138 | 7150 |
| All | Mean | 17.08 | 12.33 | 17.42 | 14.05 |
| | SD | 2.50 | 6.02 | 2.36 | 4.87 |
| | N | 12331 | 11031 | 12352 | 12383 |

Note: This table presents the mean, standard deviation, and number of observations for STEM and non-STEM exams, for males and females (but also combined), separately, for the following variables: midterm score (between 0 and 20) and final exam score (between 0 and 20).

## C Balancing Tests

One potential concern is whether exam scheduling is consistent with a random process and orthogonal to student characteristics. We deploy balancing tests to examine whether there is any systematic association between the exam scheduling variables and student characteristics, such as age, gender, or prior performance in each subject. Exam scheduling varies across year and grade configurations. To answer whether student characteristics vary along the same dimensions as exam scheduling, we regress each student characteristic on the full set of year by grade dummies. Table 15 presents the results across all students. Columns 1, 2, and 3 show the variation of midterm score, gender, and age, respectively, across all year by grade configurations. Columns 3 controls for grade fixed effects, as age is anticipated to change with grade. The coefficients of the year by grade dummies across all columns of Table 15 are in general small and not statistically significant. The *F* test for the collective equivalence of the coefficients of the year by grade dummies to zero accepts the null hypothesis of equivalence in each column. We conclude that there is no significant association between exam scheduling and students' prior characteristics, reinforcing our confidence in the randomization process of exam scheduling.

**Table 15:** Balancing tests.

| **Variables** | **(1)**<br>**Midterm score** | **(2)**<br>**Female** | **(3)**<br>**Age** |
|---|---|---|---|
| 2002 × Grade 10 | 0.000 | 0.000 | 0.000 |
| | (0.000) | (0.000) | (0.000) |
| 2002 × Grade 11 | –0.058 | 0.003 | |
| | (0.141) | (0.108) | |
| 2003 × Grade 10 | –0.037 | –0.104 | 0.000 |
| | (0.092) | (0.089) | (0.000) |
| 2003 × Grade 11 | –0.038 | 0.027 | –0.028 |
| | (0.088) | (0.087) | (0.018) |
| 2004 × Grade 10 | –0.057 | 0.059 | 0.004 |
| | (0.123) | (0.073) | (0.004) |
| 2004 × Grade 11 | –0.015 | –0.085 | –0.015* |
| | (0.113) | (0.130) | (0.009) |
| 2005 × Grade 10 | –0.085 | –0.120* | 0.000 |
| | (0.105) | (0.067) | (0.000) |
| 2005 × Grade 11 | 0.003 | 0.052 | 0.041 |
| | (0.130) | (0.084) | (0.036) |
| 2006 × Grade 10 | –0.008 | –0.077 | –0.000 |
| | (0.131) | (0.060) | (0.000) |
| 2006 × Grade 11 | –0.065 | –0.104* | 0.014 |
| | (0.085) | (0.059) | (0.013) |
| 2007 × Grade 10 | –0.078 | –0.035 | –0.002 |
| | (0.132) | (0.056) | (0.004) |
| 2007 × Grade 11 | –0.006 | –0.041 | 0.009 |
| | (0.120) | (0.060) | (0.034) |
| 2008 × Grade 10 | –0.126 | –0.053 | –0.000 |
| | (0.128) | (0.064) | (0.000) |
| 2008 × Grade 11 | 0.035 | –0.022 | –0.007 |
| | (0.073) | (0.062) | (0.006) |
| 2009 × Grade 10 | –0.067 | –0.137* | –0.000 |
| | (0.108) | (0.073) | (0.000) |
| 2009 × Grade 11 | –0.012 | –0.016 | –0.021* |
| | (0.090) | (0.076) | (0.012) |
| 2010 × Grade 10 | –0.061 | –0.038 | –0.000 |
| | (0.086) | (0.058) | (0.000) |
| 2010 × Grade 11 | –0.020 | –0.138* | –0.010 |
| | (0.079) | (0.074) | (0.015) |
| Observations | 1,736 | 1,736 | 1,736 |
| *R*-squared | 0.002 | 0.014 | 0.939 |
| *F*-Stat *P*-value | 0.982 | 0.296 | 0.640 |
| Grade FE | No | No | Yes |

Notes: The dependent variable in column (1) is the standardized midterm score, in column (2) the gender of each student, and in column (3) the age of each student. Results in each column come from a separate OLS regression across all students. Cluster-robust standard errors at the classroom by year level are reported in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

## D Variation Sufficiency

Exam scheduling variables vary across years, grades, and subjects. There are two grades (10th and 11th) and nine cohorts. Each subject is tested once for each grade in a given year. Therefore, each scheduling variable takes 18 (two grades × nine cohorts) values for each subject in our data. We illustrate the variation in exam scheduling in Table 16, Table 17, and Table 18. Table 16 shows how *Days between Exams* varies across subjects. Each entry in Table 16 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 5 days, as shown in the first column of Table 16. As an illustration, algebra was tested on the same day as the previous exam zero times, 2 days after the previous exam seven times, and so on. History was tested on the same day as the previous exam zero times, 1 day after the previous exam twice, and so on. At the bottom of Table 16 we report the mean and standard deviation of the number of days elapsed since the previous exam for each subject. We observe considerable within-subject variation in the time since the previous exam. On average, English and modern Greek have the shortest average time since the previous exam compared to 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 16:** How Does *Days Between Exams* Vary across Subjects?

| Days between exams | Ancient Greek | Literature | Modern Greek | History | Algebra | Geometry | Physics | Chemistry | English |
|---|---|---|---|---|---|---|---|---|---|
| – | 4 | 0 | 2 | 4 | 2 | 0 | 2 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 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 | 2.57 | 1.67 | 1.63 | 2.50 | 2.94 | 2.83 | 2.89 | 1.83 | 1.33 |
| SD | 1.09 | 0.77 | 0.81 | 0.94 | 1.06 | 0.92 | 1.09 | 0.79 | 0.59 |

Notes: The table presents the variation of the scheduling variable *Days Between Exams* across subjects. Each entry shows how frequently the subject in that column was tested in the number of days since the previous exam shown in that row. The variable *Days Between Exams* takes values from 0 to 5, indicating that from 0 up to 5 days might intervene between two consecutive exams in the sample. For example, modern Greek was tested 1 day after the previous exam nine times and 3 days after the previous exam three times. The first line corresponds to the times each subject was tested first, and thus the *Days Between Exams* variable is set to missing.

**Table 17:** How Does *Days Since First Exam* Vary across Subjects?

| Days since first start | Ancient Greek | Literature | Modern Greek | History | Algebra | Geometry | Physics | Chemistry | English |
|---|---|---|---|---|---|---|---|---|---|
| – | 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 | 13.86 | 15.39 | 17.69 | 17.36 | 12.69 | 15.17 | 10.89 | 16.50 | 18.06 |
| SD | 7.64 | 8.92 | 6.07 | 8.16 | 7.24 | 7.26 | 6.83 | 9.15 | 7.00 |

Notes: The table presents the variation of the scheduling variable *Days Since the First Exam* across subjects. Each entry 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. The variable *Days Since the First Exam* takes values from 2 to 29, indicating that students take compulsory exams for a maximum duration of 29 days after the first exam. For example, modern Greek was tested 4 days after the exam season started once times and 25 days after the exam season started once. The first line corresponds to the times each subject was tested first, and thus the *Days Since the First Exam* variable is set to missing.

**Table 18:** 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 |

Notes: The table presents the variation of the scheduling variable *Exam Order* across subjects. Each entry shows how frequently the subject in that column was tested in the order shown in that row. The variable *Exam Order* takes values from 1 to 16 indicating the order of the tested subject. For example, modern Greek was tested first twice, while it was tested seventh three times. The first line corresponds to the times each subject was tested first, and thus the *Exam Order* variable is set to missing.

Table 17 shows how *Days lapsed since the Exam Season started* varies across subjects. Each entry in Table 17 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 29 days, after the first exam, as shown in the first column of Table 17. For illustrative purposes, algebra was tested on the first day twice, 2 days after the first exam three times, and so on. History was tested on the first day four times, 2 days after the first exam once, and so on. Physics was never tested more than 24 days after the first exam, and algebra was never tested more than 25 days after the first exam. At the bottom of Table 17 we report the mean and standard deviation of the number of days elapsed since the first exam for each subject was administered. We observe considerable within-subject variation in the number of days since the first exam for each subject. On average, physics and ancient Greek were tested closer to the first exam than the other subjects. We do not see any systematic differences in the testing patterns for STEM and non-STEM subjects in terms of the number of days elapsed since the first exam.

Table 18 shows how the scheduling variable *Exam Order* varies across subjects. Each entry in Table 18 shows how frequently the subject in that column was tested in the order shown in that row. Students take a maximum of 16 exams (including electives), as shown in the first column of Table 18. For illustrative purposes, 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. Physics and ancient Greek were the only subjects that were never tested later than the 11th place in the order of exams. At the bottom of Table 18 we report the mean and standard deviation of the place in the exam order at which each subject was tested. We observe considerable within-subject variation in the order in which each subject was tested across years and grades. On average, physics and ancient Greek were tested at a later place of order than the other subjects. We do not see any systematic differences in the testing patterns for STEM and non-STEM subjects.

## E Overall Scheduling Effects

In this section we provide estimates of overall scheduling effects, without differentiating between STEM and non-STEM subjects. Table 19 shows estimates from the following model:

$$S_{i,s,g,t} = \alpha_0 + \alpha_1 Days\ Between_{s,g,t} + \alpha_2 Days\ Since\ First_{s,g,t} + \alpha_3 Exam\ Order_{s,g,t} + \alpha_4 M_{i,s,g,t} + \alpha_5 X_{i,s,g,t} + \kappa_{sg} + \lambda_{gt} + \zeta_t + \eta_{i,s,g,t}, \quad (3)$$

where the coefficients $\alpha_1, \alpha_2, \alpha_3$ reflect the average impact of Scheduling Effects I, II, and III, respectively, across STEM and non-STEM subjects. As shown earlier, STEM and non-STEM subjects exhibit substantially different scheduling effects. These largely cancel out in an overall analysis of scheduling effects, with the exception of Scheduling Effect I. Table 19 shows that as the number of days between exams increases, the final exam score decreases, on average across subjects.

**Table 19:** Overall effect of exam timing on performance.

| Variables | (1) | (2) | (3) |
|---|---|---|---|
| Scheduling Effect III | 0.031 | 0.003 | –0.010 |
| | (0.023) | (0.003) | (0.011) |
| Scheduling Effect II | –0.013 | –0.001 | 0.001 |
| | (0.011) | (0.002) | (0.005) |
| Scheduling Effect I | –0.045** | –0.007** | –0.024*** |
| | (0.017) | (0.003) | (0.008) |
| Scheduling Effect III$^2$ | | | 0.000 |
| | | | (0.000) |
| Scheduling Effect II$^2$ | | | 0.000 |
| | | | (0.000) |
| Scheduling Effect I$^2$ | | | 0.003** |
| | | | (0.001) |
| Observations | 14,258 | 14,258 | 14,258 |
| *R*-squared | 0.026 | 0.985 | 0.985 |
| Student controls | No | Yes | Yes |

Notes: The dependent variable in each specification is the standardized 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 by year fixed effects, subject by grade fixed effects, day of the week fixed effects, and a full set of birth year by cohort fixed effects. Columns 2 and 3 include individual controls. Individual controls include indicators for students' gender, and a dummy that indicates whether a student is retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

## F Non-linear Scheduling Effects

In this section, we explore non-linear scheduling effects on exam performance using the quadratic form of each scheduling variable. Specification 4 is an augmented version of specification 1, and includes the square of each scheduling variable.

$$S_{i,s,g,t} = \alpha_0 + \alpha_{1c} Days\ Between_{s,g,t} + \alpha_{2c} Days\ Since\ First_{s,g,t} + \alpha_{3c} Exam\ Order_{s,g,t} + \alpha_{4c} Days\ Between^2_{s,g,t} + \alpha_{5c} Days\ Since\ First^2_{s,g,t} + \alpha_{6c} Exam\ Order^2_{s,g,t} + \alpha_7 M_{i,s,g,t} + \alpha_8 X_{i,s,g,t} + \kappa_{sg} + \lambda_{gt} + \zeta_t + \eta_{i,s,g,t}. \quad (4)$$

Column 1 of Table 20 shows the effects for STEM and non-STEM subjects separately (using specification 1), while column 2 focuses on the difference in each scheduling effect for STEM relative to non-STEM subjects. Intuitively, the main effects shown in column 2 correspond to the scheduling effects of non-STEM subjects in column 1, while the coefficients of the interaction terms reflect the additional (marginal) scheduling effects of STEM subjects compared to non-STEM subjects.

Column 3 of Table 20 shows the estimated nonlinear effects of three distinct channels of exam scheduling on performance separately for STEM and non-STEM subjects (using specification 4), while column 4 of Table 20 focuses on the differences of nonlinear scheduling effects between STEM and non-STEM subjects. Scheduling Effect I is found to have nonlinear effects only on exam performance in 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 further in days from the previous exam are associated with decreasingly lower performance, while controlling for other influences. Scheduling Effect III is found to have nonlinear effects only on exam performance in 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 place in the exam order are associated with increasingly higher performance, while controlling for other influences. In contrast to the other scheduling effects, Scheduling Effect II is found not to have nonlinear effects in either STEM or non-STEM subjects.

**Table 20:** The effect of exam scheduling 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) | |
| Scheduling Effect III | | –0.002 | | –0.004 |
| | | (0.002) | | (0.004) |
| Scheduling Effect II | | 0.002 | | 0.003 |
| | | (0.002) | | (0.002) |
| Scheduling Effect I | | –0.012*** | | –0.012*** |
| | | (0.004) | | (0.004) |
| STEM × Scheduling Effect III | | 0.019*** | | –0.015 |
| | | (0.005) | | (0.017) |
| STEM × Scheduling Effect II | | –0.008** | | –0.001 |
| | | (0.002) | | (0.007) |
| STEM × Scheduling Effect I | | 0.013*** | | 0.024 |
| | | (0.003) | | (0.022) |

| | | | | |
|---|---|---|---|---|
| STEM × Scheduling Effect III$^2$ | | | | 0.002* |
| | | | | (0.001) |
| STEM × Scheduling Effect II$^2$ | | | | –0.000 |
| | | | | (0.000) |
| STEM × Scheduling Effect I$^2$ | | | | –0.002 |
| | | | | (0.003) |
| Observations | 14,258 | 14,258 | 14,258 | 14,258 |
| *R*-squared | 0.985 | 0.985 | 0.985 | 0.985 |

Notes: The dependent variable in each specification is the standardized 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 midterm score, 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 a dummy that indicates whether a student is retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

## G Differential Scheduling Effects by Student Prior Performance

We interact each scheduling effect for STEM and non-STEM subjects in specification 1 with a continuous variable that captures the standardized prior performance of each student in each subject. Results are relegated to Table 21. For Scheduling Effect I, the interaction of interest is positive and statistically significant for non-STEM subjects, whereas it is negative and significant for STEM subjects. This indicates that the gap in the effect of an additional day between exams between STEM and non-STEM subjects decreases with prior performance.

The estimated coefficient of the interaction of interest for Scheduling Effect II in non-STEM subjects is zero, while the coefficient of the interaction for STEM subjects with prior performance is positive and significantly different from zero. For STEM subjects, higher-achieving students benefit more from an additional day since their first exam for STEM subjects; this is not the case for exams in non-STEM subjects.

The estimated coefficient of the interaction of Scheduling Effect III in non-STEM subjects with prior performance is zero, while the coefficient of the interaction for STEM subjects is negative and statistically 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.

**Table 21:** Differential 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 |

Notes: The dependent variable is the standardized 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 by year fixed effects, subject by grade fixed effects, day of the week fixed effects, a linear time trend, a full set of birth year by cohort fixed effects, and individual controls. Individual controls include indicators for students' gender, and a dummy that indicates whether a student is retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 22 shows the estimated Scheduling Effects by STEM and non-STEM subjects for four quantiles of prior performance, proxied by standardized midterm score. The estimates are presented graphically in Figure 5.

**Table 22:** Differential 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) |

Notes: The dependent variable is the standardized final exam score at the subject and grade level. The sample includes 14,258 observations. Cluster-robust standard errors at the classroom by year level are reported in parentheses. Specification includes 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' gender, and a dummy that indicates whether a student is retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

## H Differential Scheduling Effects By Exam History

In this section, we provide estimates of differential scheduling effects by exam history. Exam history is measured by the number of STEM and non-STEM exams taken prior to attempting a subsequent exam. We expand model 1 by adding interactions of each scheduling variable for STEM and non-STEM subjects with the number of STEM and non-STEM exams taken prior. Table 23 shows estimates for the interactions of interest.

**Table 23:** Differential scheduling effects by exam history.

| **Variables** | **(1)** |
|---|---|
| Scheduling Effect III for non-STEM × No. of STEM exams taken | 0.001 |
| | (0.001) |
| Scheduling Effect III for STEM × No. of STEM exams taken | –0.001 |
| | (0.003) |
| Scheduling Effect III for non-STEM × No. of non-STEM exams taken | 0.002 |
| | (0.001) |
| Scheduling Effect III for STEM × No. of non-STEM exams taken | 0.004* |
| | (0.002) |
| Scheduling Effect II for non-STEM × No. of non-STEM exams taken | 0.001 |
| | (0.001) |
| Scheduling Effect II for STEM × No. of STEM exams taken | 0.001 |

| | |
|---|---|
| | (0.001) |
| Scheduling Effect II for non-STEM × No. of non-STEM exams taken | –0.001 |
| | (0.001) |
| Scheduling Effect II for STEM × No. of non-STEM exams taken | –0.001 |
| | (0.001) |
| Scheduling Effect I for non-STEM × No. of non-STEM exams taken | –0.000 |
| | (0.001) |
| Scheduling Effect I for STEM × No. of STEM exams taken | –0.001 |
| | (0.002) |
| Scheduling Effect I for non-STEM × No. of non-STEM exams taken | 0.003*** |
| | (0.001) |
| Scheduling Effect I for STEM × No. of non-STEM exams taken | 0.001 |
| | (0.002) |
| Observations | 14,258 |
| *R*-squared | 0.986 |

Notes: This table presents differential scheduling effects by exam history. We interact each scheduling variable for STEM and non-STEM subjects with the number of STEM and non-STEM exams taken prior. The model includes the three scheduling variables by type of subject, controls for midterm score, 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 a female gender dummy, and a dummy that indicates whether a student is retained. This table reports only the coefficients of interactions of interest in the interest of space.

## I Restricted Models of Scheduling Effects

To increase the empirical intuition behind the size and direction of omitted variable bias when one or more exam scheduling variables are excluded from the regression model, we estimate models using subsets of the exam scheduling variables. Table 24 show the estimate from restricted models across all subjects.

**Table 24:** Subsets of the effects of exam timing on performance.

| Variables | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Scheduling Effect III | 0.002*** | | | 0.005* | | 0.002*** |
| | (0.001) | | | (0.003) | | (0.001) |
| Scheduling Effect II | | 0.001*** | | –0.002 | 0.001** | |
| | | (0.000) | | (0.001) | (0.000) | |
| Scheduling Effect I | | | –0.009*** | | –0.008*** | –0.008** |
| | | | (0.003) | | (0.003) | (0.003) |
| Observations | 14,258 | 14,258 | 14,258 | 14,258 | 14,258 | 14,258 |
| *R*-squared | 0.985 | 0.985 | 0.985 | 0.985 | 0.985 | 0.985 |

Notes: The dependent variable is the standardized 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 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' gender, and a dummy that indicates whether a student is retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

## J Alternative Modeling Approach to Scheduling Effect III

In this section, we explore an alternative modelling approach for Scheduling Effect III. In particular, we estimate a model using specification Table 20, where the *exam order* variable is replaced by the ratio of *exam order* over *days since first exam*. The estimated coefficient for the ratio of the variables associated with Scheduling Effect III and II is reported in Table 25. The results show that an additional exam *per* day since the first exam increases test performance by 0.047 standard deviations across all subjects, and by 0.075 standard deviations across STEM subjects. Performance in non-STEM subjects is not found to be significantly impacted by the number of exams *per* days since the first exam. The ratio of *exam order* (Scheduling Effect III) over the *number of days since the first exam* (Scheduling Effect II) has a positive and significant coefficient for STEM subjects, as does the *number of days since first exam* variable for STEM subjects. This suggests that providing students with more time to study may be optimal for STEM subjects.

**Table 25:** Scheduling Effect III relative to Scheduling Effect II.

| Variables | (1) | (2) |
|---|---|---|
| Scheduling Effect III/II for non-STEM | | 0.031 |
| | | (0.030) |
| Scheduling Effect III/II for STEM | | 0.075*** |
| | | (0.026) |
| Scheduling Effect II for non-STEM | | 0.001* |
| | | (0.000) |
| Scheduling Effect II for STEM | | 0.002*** |
| | | (0.000) |
| Scheduling Effect I for non-STEM | | –0.012*** |
| | | (0.004) |
| Scheduling Effect I for STEM | | –0.001 |
| | | (0.004) |
| Scheduling Effect III/II | 0.047** | |
| | (0.021) | |
| Scheduling Effect II | 0.001*** | |
| | (0.000) | |
| Scheduling Effect I | –0.006** | |
| | (0.003) | |
| Observations | 14,258 | 14,258 |
| *R*-squared | 0.985 | 0.985 |

Notes: The dependent variable is the standardized 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 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' gender, and a dummy that indicates whether a student is retained. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

## K Day of the Week Effects

One potentially interesting aspect of exam scheduling may be the differential day-of-the-week effects associated with test performance. Table 26 shows our estimated day-of-the-week effects from different specifications ranging from no controls to controlling for subject by grade unobservable effects and scheduling variables. The estimated effects correspond to the marginal effects on test scores associated with each day of the week with respect to Monday (the omitted category). The results show that, without controls for exam scheduling, Tuesday, Thursday, and Saturday may be associated with increases test scores on subject tested on those days, relative to Monday.

**Table 26:** Day of the week effects.

| Variables | (1) | (2) | (3) |
|---|---|---|---|
| Tuesday | 0.088*** | 0.090** | 0.050 |
| | (0.033) | (0.040) | (0.040) |
| Wednesday | 0.049 | 0.053 | –0.026 |
| | (0.036) | (0.036) | (0.036) |
| Thursday | 0.092** | 0.076* | –0.030 |
| | (0.039) | (0.039) | (0.046) |
| Friday | 0.042 | 0.038 | –0.079** |
| | (0.035) | (0.036) | (0.038) |
| Saturday | 0.289*** | 0.326*** | 0.223*** |
| | (0.047) | (0.069) | (0.073) |
| Scheduling variables | No | No | Yes |
| Subject by Grade FE | No | Yes | Yes |

Notes: Sample: 14,258 observations. The dependent variable is the standardized final exam score at the subject and grade level. Cluster-robust standard errors at the classroom by year level are reported in parentheses. Scheduling variables include *exam order, days between exams, days since first exam*. All specifications include grade by year fixed effects. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.