Time of Day and High Stakes Cognitive Assessments

Alessio Gaggero[1] and Denni Tommasi[2]

[1]University of Granada

[2]Monash University and IZA

July 2021

Abstract

A variety of external conditions may affect individual performances in high stakes cognitive assessments, with potentially lasting consequences on earnings and career. We provide the first causal evidence that the time of the day is an important condition affecting the performance at the moment of an evaluation. Exploiting a setting in which cognitive assessments are quasi-randomly assigned at a different time-of-day, we find that peak performance occurs in the early afternoon. The estimated time-of-day effects follow specific patterns consistent with the circadian rhythm, which suggests that biological factors are important determinants of performances even in economically meaningful settings.

JEL Codes: I20, I24, J22, J24.

Keywords: time of day, cognitive assessments, high stakes, performance, circadian rhythm.

Alessio Gaggero: Department of Quantitative Methods for Economics and Business, University of Granada. Campus Universitario de Cartuja, 18071 Granada, Spain. E-mail: alessiogaggero@go.ugr.es. Denni Tommasi: Department of Econometrics and Business Statistics, Monash University. Building H, Caulfield Campus 900 Dandenong Road Melbourne, VIC 3145, Australia. E-mail: denni.tommasi@monash.edu. This is a substantial revision of an earlier paper that circulated under the title "Time of Day, Cognitive Tasks, and Efficiency Gains" (IZA DP No. 13657). We have benefited from discussions with Josh Angrist, Sascha Becker, Sarah Bridges, Bart Cockx, Filippo De Marco, Christian Dustmann, Daniel Hamermesh, Eric Hanushek, Rey Hernandez-Julian, Jennifer Hunt, Andrea Ichino, Hannah Klauber, Claudio Labanca, Pedro Luis Silva, Lester Lusher, Eva Moreno, Lorenzo Neri, Trudy Owens, John Pencavel, Jan Sauermann, Kevin Williams, Hannah Klauber, Giulio Zanella, Yves Zenou, and seminar participants at the Melbourne Institute, University of Adelaide, University of Bologna, University of Hawaii, University of New South Wales, University of Nottingham, University of Trento, the EALE SOLE AASLE World Conference 2020, SOLE 2021, Essen Health Conference 2021 and the SMYE 2021. We thank all those involved in providing the data. We thank medical biologists Kristen Knutson and Joseph Takahashi for helpful comments and Giulia Zerbini for extensive and enlightening conversations. Kevin Wang provided excellent research assistance. This article has been screened to ensure no confidential information is revealed. All errors are our own.

1 Introduction

Assessments of cognitive abilities, such as school admission tests, job interviews, or job performance evaluations, are regularly conducted by organisations.[1] Since performance in such assessments can have lasting consequences on earnings and career,[2] it is important to understand under which circumstances cognitive ability can be measured best.[3] Recent works have shown that a variety of external conditions can affect performance in high stakes assessments, including air pollution (Ebenstein et al., 2016), pollen exposure (Bensnes, 2016), summertime heat exposure (Cho, 2017) and temperature (Park, 2020). These papers conclude that individuals evaluated in unfavourable conditions may be disadvantaged when competing for higher education or jobs.

Research on the circadian rhythm, which is a biological process governing our sleep-wake cycles (Cardinali, 2008),[4] suggests a novel potential condition: time of day. It is well-known that alertness and mental focus vary substantially at different times throughout the day, also affecting cognitive functioning (Higuchi et al., 2000). However, it is unclear to what extent these time-of-day effects generalise to actual performances in economically meaningful settings.[5] This is because, outside controlled laboratory experiments, individuals have strong incentives to minimise the effects of a decrease in cognitive functioning through behavioural responses (higher effort). Establishing the relevance of time-of-day effects in high stakes assessments may reveal insights that could change the way in which we assess individuals' cognitive ability in school, as well as in the workplace.

In this paper, we provide the first causal evidence of the relationship between the time

[1]This is due in large part to the increasing importance of cognitive skills in the workplace (e.g., Goldin and Katz, 2009; Acemoglu and Autor, 2011; Hanushek and Woessmann, 2012). Among the possible types of assessments, test scores are widely used as proxies for cognitive ability, human capital or individual productivity (Jacob and Rothstein, 2016).

[2]There is extensive literature showing that young labour market entrants are particularly vulnerable to initial conditions and quality of the first employment (von Wachter, 2020).

[3]This is particularly the case if retaking or rescheduling are costly due to frictions. For example, in the case of college admission tests (e.g., SAT or ACT), retaking may take up to one year, potentially creating high opportunity costs. Similarly, job interviews or job performance assessments require a significant amount of coordination, which make rescheduling costly.

[4]Circadian rhythms (or circadian clocks) are internal cycles that repeat circa every 24 hours and regulate many biological processes, including peak and through wakefulness cognitive functioning. "Cognitive functioning" refers to a person's ability to acquire knowledge, manipulate information, and process thoughts (Kiely, 2014). In any population, there are significant inter-individual differences in the timing of optimal cognitive functioning denoted chronotypes (Roenneberg et al., 2007a). For example, some individuals reach peak cognitive function in the morning (early-type), while others are more alert in the evening (late-type) (Schmidt et al., 2007).

[5]In controlled laboratory experiments, medical psychologists have shown that cognitive abilities are susceptible to time of day and chronotype effects (Goldstein et al., 2007). Some works have looked specifically at chronotype effects in a school setting and found a negative correlation between being a late-type and academic achievements (e.g., van der Vinne et al., 2015; Itzek-Greulich et al., 2016; Zerbini and Merrow, 2017).

of the day and performance in high stakes assessments by young adults. We exploit an ideal setting provided by the higher education system in the United Kingdom (UK) to identify the parameter of interest. In our framework, university students can sit no more than one examination per day at one of three timeslots: 'morning', 9 am; 'early afternoon', 1.30 pm; and 'late afternoon', 4.30 pm. Our identification strategy relies on a within-student quasi-random assignment of the times of day for the examinations. The Student Services Office is in charge of scheduling examinations. In doing so, it follows a set of clearly specified rules to simultaneously schedule examinations at different timeslots by means of a software program. The schedules are then published online for both students and professors to access a few weeks before the session commences. Hence, causal identification rests on the assumption that, controlling for the conditions and student-fixed effects, within-student variations in day-to-day times of exams are not correlated with unobserved exam characteristics and unobserved selection effects on educational performance.

We use administrative data on university students from one of the largest public universities in the UK from 2014–15 to 2018–19. The data contains the complete records of half a million student-examination level observations. Taking advantage of this detailed information, we are able to compare the performance of students sitting examinations at different times of the day in the final two weeks of the semester. As a student's future is largely determined by their academic performance, test scores are often used as proxies for cognitive skills that are valued in the labour market (Bradley and Green, 2020).[6] Additionally, from a biological standpoint, individuals enter adulthood at around 20 years of age (Roenneberg et al., 2004). As our sample comprises university students who largely satisfy this definition, they have, on average, similar biological characteristics to other adults performing cognitive assessments in the labour market. This makes our findings interesting to the broader population of skilled workers to whom we wish to speak.

We motivate our empirical analysis by writing a simple model of individual cognitive performance and optimal effort at different times of the day. The model captures the two po-

[6]This is because university examinations 'require a person to mentally process new information (i.e., acquire and organise knowledge/learn) and allow them to recall, retrieve that information from memory and to use that information at a later time in the same or similar situation (i.e., transfer)' (Kester and Kirschner, 2012). In addition to being a significant determinant of wage determination (Murnane et al., 1995) and labor force quality (Hanushek and Kimko, 2000), test scores are also a significant predictor of economic growth (Hanushek and Woessmann, 2008).

tential mechanisms driving a change in performance. On one side, alertness and mental focus vary during the day due to circadian processes and sleep deprivation; hence a change in cognitive functioning may trigger a direct change in performance. Conversely, students face a marginal disutility of effort (study strategy or food intake) as well as the returns to performance as a function of time. The key implication is that, as the stakes of a cognitive assessment increase, the association between time and performance is less likely to be driven solely by a change in behaviour, suggesting that the biological component may affect individual performances even in a high stakes environment.

The primary result is that, on average, the performance profile of students quasi-randomly assigned to high stakes examinations at different times of day has an inverse U-shape relationship, at which peak performance occurs in the early afternoon (1.30 pm). At this time of day, students' marks increase by 0.068 standard deviations (SD) with respect to the morning examination (9 am), which we use as a control group, and decrease again in the late afternoon (4.30 pm). The association between the time of the day and actual performance is consistent with the association between the time of the day and cognitive functioning. The implication is that, all else being equal, individuals sitting an important cognitive assessment at an unfavourable time are disadvantaged. The results are robust to a variety of sensitivity checks, including (i) different sample restrictions and specifications of the conditional independence assumption (CIA); (ii) specifications with exam fixed effects, which control for potential differences in exam attributes; and (iii) specifications with day fixed effects, which control for accumulated fatigue or warm-up effects during the exam session.

Next, we investigate the effect of heterogeneity to better understand when time-of-day effects on performance are strongest and who is more likely to be affected. Drawing from the literature on the circadian rhythm, we formulate three tests of hypothesis. First, time-of-day effects are stronger in seasons with limited sunlight exposure (Kantermann et al., 2007). In our sample, students sitting an examination in the early afternoon in January increase their performance by an SD of 0.094, whereas students sitting an examination in the early afternoon from May to June only increase their performance by half this amount. Second, time-of-day effects are stronger for cognitive assessments involving fluid intelligence (see e.g.

Ghisletta and Lecerf (2018) for a definition), such as problem-solving, logical thinking and abstract-reasoning tasks (Goldstein et al., 2007; Zerbini et al., 2017). Accordingly, we find that students taking a STEM examination in the early afternoon in autumn increase their performance by 0.139 SD;[7] however, no similar effects are found for students taking a non-STEM examination. Third, younger students aged 20 or below are more affected by time-of-day effects (Roenneberg et al., 2004). To test this hypothesis, we compare time-of-day effects between students aged 20 or below and those over 20. For a young student in our sample taking a STEM exam in autumn, there is a 0.21 SD increase in performance when moving an exam from morning to the early afternoon. In contrast, for students older than 20 years of age, the estimated time-of-day effects are significantly milder.

The overall effect we estimate is a combination of the two main mechanisms outlined previously. However, the fact that all our estimates are consistent with the predictions of the circadian rhythm provides suggestive evidence that the main mechanism is likely to operate through the circadian rhythm's effects on performance, with the behavioural effects likely mitigating the biological effects. This is an important result because it provides a rationale for the formulation of new policies that could increase the efficiency of universities. Indeed, our results suggest that schools have an additional tool that they can use to measure their student's performance while they are at their best: to optimise the daily examination schedules. Furthermore, since high stakes cognitive assessments are used in a variety of other settings as well, the implications of our results may extend beyond academia, including to students applying for college or graduate school, as well as to young graduates transitioning into the labour market. In particular, the scheduling of the assessments may produce welfare losses and unfairness, which, in turn, might have negative, lasting consequences on the labour market. In the final discussion section of this paper, we apply our insights to one important context that is likely to benefit from our findings (i.e., college admission tests) and suggest that a simple rescheduling can increase students' permanent income.

This paper is related to the economics literature studying the importance of time in economically relevant settings. Using the education setting, much of the research in this area has

[7] "STEM" education is a term used to refer collectively to the teaching of the disciplines within its umbrella: Science, Technology, Engineering and Mathematics.

looked at the effects of class scheduling and school start times on students' performance (e.g., Dills and Hernández-Julián, 2008; Carrell et al., 2011; Edwards, 2012; Pope, 2016; Heissel and Norris, 2017; Cotti et al., 2018; Shin, 2018; Williams and Shapiro, 2018; Lusher et al., 2019). Our study differs considerably from previous research because it examines the effects of the time of the day of the examinations (the assessment itself) rather than the scheduling of classes, school start times, learning, or knowledge. Hence, the parameter we are after is the change in performance (measured cognitive ability or productivity) of an individual taking a high stakes assessment at different times. Moreover, since we focus on young adults, our results can be used to generalise to cognitive assessments used in job interviews or job performance assessments, which are useful to understand the labour markets.[8]

The remainder of this paper is set out as follows. Section 2 describes the institutional setting, presents the data, and discusses the theoretical framework and identification strategy. Section 3 outlines the results, including several robustness and falsification tests, explains the results, and discusses the implications of our findings. Finally, Section 4 concludes the paper. The appendix contains additional analysis and robustness checks.

2 Background, Data, and Identification

This section is organized as follows: first, we provide information on the institutional setting, the allocation rules of exam scheduling, and evidence in favor of the conditional independence assumption (CIA) underlying our approach. Next, we describe the sample selection and the data used. Finally, we outline the theoretical framework and the empirical model.

2.1 Institutional Setting and Conditional Random Assignment

We use administrative data covering the population of students enrolled in one of the largest public universities in the U.K. The data span five academic years from 2014-15 to 2018-19. In line with the U.K. higher education system, the university offers undergraduate (UG), postgraduate taught (PGT), and postgraduate research (PGR) programs. Students enrolled in

[8] A smaller number of papers has focused on the effects of examination scheduling on student performance (e.g., Pope and Fillmore, 2015; Bensnes and Strom, 2019; Goulas and Megalokonomou, 2020). However, these studies investigate different measures of across-day scheduling, whereas the present study examines the within-day effects of time on performance.

taught programs (either UG or PGT) must choose a combination of previously approved compulsory and elective modules to qualify for an academic degree. While the overall assessment of certain modules solely depends on the final exam, assessments of other modules may also affect students' marks in assignments preceding the final exam (e.g., coursework, problem sets, presentations, etc).[9]

Exams are split into two main exam sessions –fall and spring– as well as an additional resit session in summer. The fall session is held in January and lasts approximately 12 days. The spring session begins toward the end of May and continues in June, lasting approximately 15 days. Students who fail one or more exams are able to resit at the end of August.[10] Exams take place between Monday and Saturday during each main session and are scheduled in one of three different time slots: a 9 am morning session, a 1.30 pm early afternoon session, and a 4.30 pm late afternoon session. The University Student Services office is in charge of scheduling the exams. Exams are scheduled in different time slots and available locations using Exam Scheduler software. The office takes students availability into account to prevent exam clashes and follows a set of clearly specified rules to ensure the integrity of the process. Three rules are crucial for our identification strategy: (i) student timetables depend on the number of exams that students are taking in a given exam period, whether a school has requested a particular date or time for an exam,[11] and the duration of the exam.[12] Otherwise the time slot is randomized. (ii) Chief examiners cannot modify the scheduling of exams unless they obtain formal approval following the submission of a motivated request to the office.[13] (iii) Exam schedules are published online for students and professors a few weeks prior to the exam session.[14]

[9]Note that, as it is common in the U.K. system, the final marks of first-year UG exams do not count toward the degree. In Section 3.2, we test the sensitivity of our findings by excluding first-year UG students.

[10]The resit session is also intended for students who, for a variety of reasons (e.g., personal, medical, religious), either could not take the exams on the scheduled date or believe their exam performance was negatively affected by "extenuating circumstances".

[11]The office has informed us that, from time to time, schools request specific times for their exams, but "it is not often that we can accommodate them, so this is not always a factor when scheduling examinations." For this reason, this is considered a minor condition.

[12]Exams longer than 2.5 hours are scheduled in the morning; otherwise, the time slot is randomized. In our sample, only 8% of exams were longer than 2.5 hours. For this reason, this condition is considered minor.

[13]Unfortunately, information about which exams have been re-scheduled following formal requests is unavailable. However, based on data from exam office administrators, this number is negligible.

[14]The other rules are: (iv) exams longer than 2.5 hours can be scheduled for the afternoon slot only on Saturdays. (v) The examination of modules taught both in the main and satellite university campuses must be scheduled in conjunction to ensure the integrity of the exam. (vi) Similarly, the schedules must be in coordinated with students who have alternative examination arrangements and are required to take their exam in a different location (e.g., smaller room, computer room). (vii) Students should have at least 24 hours between consecutive exams.

Therefore, our CIA entails that, conditional on the allocation rules used by the university's exam office, the times at which exams take place are randomly assigned within-student and, consequently, are orthogonal to both exams' and students' characteristics. To provide evidence in favor of this assumption, we regress each exam time on four within-student characteristics, along with student and year fixed effects (FE) and the set of conditions defining the allocation rule, namely the number of exams per session, school FE, and exam weight (in credits) FE. The latter variable is used because the data on exam duration are limited to the last two academic years of the sample. Hence, since exam weight is highly correlated with exam duration (over 90%), it enables the test to be run on the entire sample. The exam times are defined as a dummy variable for whether exams are taken in the morning (9 am), early afternoon (1.30 pm), or late afternoon (4.30 pm).

Given the available data, we can look at the following set of within-student characteristics: (1) class size of each exam, to check whether students enrolled in larger classes were systematically more likely to have an exam scheduled at a specific time of day;[15] (2) the number of days between each exam and the first exam scheduled in the session by the student; (3) the number of days between consecutive exams; the last two variables allow to examine whether the student examination's schedule is correlated with exam time; and (4) student's grade point average (GPA) from the previous semester, to assess whether high/low ability students select specific exam times.[16]

The results of the balancing tests are reported in Table 1. In Columns (1), (4), and (7), we report the estimates on the main sample used in estimation. In Columns (2), (5), and (8), we investigate balancing after including previous semester's GPA, which entails smaller samples than those used in estimation. Finally, in Columns (3), (6), and (9), we report the results of the balancing tests after including in the regression a dummy for exams longer than 2.5 hours.[17] Overall, they provide compelling evidence that our main identifying assumption

[15]This is important to check because one could argue that class size is correlated with unobserved learning and knowledge acquired during the semester and driven by differential peer effects. We also experimented balancing tests using class size FE. The results are equivalent and are available upon request.

[16]We also experimented balancing tests using the GPA from the previous year. The results are equivalent and are available upon request.

[17]Specifically, since we do not have information on duration for all five years, we use data on exam duration from the last two years to impute the duration data for the same exams that appeared in the first three years of the sample. Hence, the underlying (implicit) assumption of this balancing check is that, if an exam is longer than 2.5 hours in a given year, it is highly likely that it was longer than 2.5 hours a few years earlier. We then estimate the same regression model, including a dummy variable with a value of 1 if an exam is longer than 2.5 hours and a value of 0 if otherwise.

is likely to be satisfied.[18] The main limitation of our balancing tests is that we do not have more information about exam attributes or characteristics varying within students. We come back to this potential concern in the robustness checks and show that, by adding exam FE to the main specification, all the main results hold.

Table 1: Balancing Tests by Exam Time

	Morning Exam			**Early Afternoon Exam**			**Late Afternoon Exam**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Class Size	-0.000	0.000	-0.000	0.000*	0.000	0.000	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Days Since First	0.004	0.002	0.001	-0.000	0.002	0.002	-0.003*	-0.004**	-0.003
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Days Between	-0.000	0.001	0.002	0.002	0.000	-0.000	-0.002	-0.001	-0.001
	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.002)	(0.003)	(0.003)
Average GPA s-1		-0.002	-0.003		0.003	0.004		-0.002	-0.001
		(0.008)	(0.008)		(0.008)	(0.008)		(0.006)	(0.007)
Conditions:	✓	✓	✓	✓	✓	✓	✓	✓	✓
Student FE:	✓	✓	✓	✓	✓	✓	✓	✓	✓
Duration FE:			✓			✓			✓
Clusters:	7626	6183	5379	7626	6183	5379	7626	6183	5379
Observations:	500959	353090	311674	500959	353090	311674	500959	353090	311674
Adjusted R^2:	0.124	0.130	0.180	0.057	0.070	0.100	0.065	0.078	0.086
F-Statistics:	1.519	0.564	0.444	1.245	0.343	0.453	3.894	2.983	1.876

Notes: Dependent variable: dummy for whether the exam was taken at 9am (Morning Exam), dummy if it was taken at 1.30pm (Early Afternoon Exam), dummy if it was taken at 4.30pm (Late Afternoon Exam). Standard errors are clustered by exam-year. Observations are at the student-exam-year level. In each specification, from Column (1)-(9), we control for student fixed effects (FE), year FE, and the conditions defining the allocation rule, namely: number of exams per session fixed effects (FE), School FE, and exam weight (in credits) FE. In Columns (1), (4), and (7), we report estimates for the full sample. In Columns (2), (5), and (8), we investigate balancing when including previous semester's GPA, which entails smaller samples than those used in estimation. Finally, in Columns (3), (6), and (9) we report results of the balancing tests when including in the regression a dummy for exams longer than 2.5 hours. The last row in the Table reports the F-statistic resulting from a joint significance test. $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Finally, later in the robustness checks, we also consider the possibility that the allocation process is not completely (conditionally) random. Specifically, we assess the degree of omitted variable bias by implementing the coefficient stability test developed by Altonji et al. (2005) and generalized by Oster (2019). Also in this case, all the main results hold. Therefore, based on the tests and the way in which the timetables are scheduled, we believe that the allocation of students to exams held at particular times of the day is quasi-random, conditional on the characteristics used by the exam office.

[18] We assessed balancing extensively. Some further results are provided in Table A1 and Table A2 in Appendix. In the former, we show that balancing works even without student FE. In the latter, we show that even between-student quasi-randomness is at work.

2.2 Sample Selection and Data Description

The sample selection proceeds in three steps: first, we include only the exams of taught programs (UG and PGT), particularly exams that comprised a final assessment. Second, we focus on the two main sessions of fall and spring because the August resit session could be systematically different (in addition, it accounted for only 0.06% of the sample). Third, we consider exams scheduled within officials sessions and time slots (i.e., morning, early and late afternoon), as only 2.31% of the sample are scheduled off session or at different time slots.[19] The final sample covers 43,913 UG and 7,642 PGT full-time students over five academic years. There are 503,358 student exam observations and 7,669 exam-year clusters.

The data contain the full profile of students' academic performance, including exam marks separated from the specific marks obtained in various assignments (if any) that contributed to the final mark. In addition, the dataset contain specific information about scheduling, namely the date, time, and room in which each exam took place in the university. Lastly, the data include several student-level characteristics (e.g., gender, age, ethnicity) as control variables. By linking student-level exam marks with exam schedules, the dataset is uniquely suited to investigate the effect of the time of day of high stakes cognitive assessments on student performance.

In Table 2, we report the summary statistics of the main variables of interest. The outcome variable "Exam mark" denotes students' cognitive performance on the day of an exam.[20] In line with the U.K. higher education system, exam marks range between 0 and 100.[21] The average exam mark is 60 (median 62),[22] the average mark students achieved in assignments is 69, and the average final mark (a weighted average of exam mark and assignment mark) is 62 (median 63).[23] In total, 45% of exams are scheduled in the morning (9 am), 34% are scheduled in the early afternoon (1.30 pm), and 21% are scheduled in the late afternoon

[19]The results do not change when these observations are included and are available upon request.

[20]We use students' exam mark rather than their final mark as the outcome variable because it enables the performance of young adults engaged in a complex cognitive task to be observed with a precise and objective measure of performance. In addition, from a labor economics perspective, exams marks are of particular relevance as they involve a series of skills that are directly transferable to the workplace, including the ability to work under pressure, manage time effectively, and efficiently communicate ideas both in writing and orally. As discussed before, several papers use exam marks as a measure of cognitive task performance.

[21]Exam marks are double-marked internally, and a sample is sent to an external examiner for quality control purposes. The scores obtained are typically translated into distinct degree classifications according to whether the student is an UG or PGT.

[22]However, in the analysis, we follow the literature and use the standardized exam mark achieved by students, which is obtained by normalizing exam marks by their SD.

[23]The summary statistics of assignment marks refer to the sample of exams that included assignments as part of their coursework.

Table 2: Descriptive Statistics

	Mean	S.D.	Min	Max
Key Variables:				
Exam mark	60.32	16.66	0	100
Assignments mark	43.58	35.40	0	100
Final mark	62.02	14.10	0	100
Morning Exam [0,1]	0.45	0.50	0	1
Early Afternoon Exam [0,1]	0.34	0.47	0	1
Late Afternoon Exam [0,1]	0.21	0.41	0	1
Exam Characteristics:				
1-hour Exam [0,1]	0.14	0.35	0	1
2-hours Exam [0,1]	0.48	0.50	0	1
2.5-hours+ Exam [0,1]	0.08	0.26	0	1
Other Exam Duration [0,1]	0.30	0.46	0	1
10 Credits Exam [0,1]	0.47	0.50	0	1
20 Credits Exam [0,1]	0.41	0.49	0	1
30 Credits Exam [0,1]	0.03	0.16	0	1
30+ Credits Exam [0,1]	0.02	0.14	0	1
Number of Assignments per Exam	0.99	1.05	0	11
Class size	151.79	103.34	1	602
Student Characteristics:				
Age	20.78	2.35	17	64
Female [0,1]	0.47	0.50	0	1
Black [0,1]	0.07	0.25	0	1
Asian [0,1]	0.11	0.31	0	1
White [0,1]	0.63	0.48	0	1
Other Ethnicities [0,1]	0.19	0.39	0	1
Undergraduates [0,1]	0.93	0.25	0	1
First Year UG [0,1]	0.37	0.48	0	1
STEM [0,1]	0.50	0.50	0	1
Number of exams per year	6.32	2.12	1	12
Spring exam session [0,1]	0.55	0.50	0	1
Number of exams per session	3.70	1.40	1	9
Total credits per year	91.25	24.49	10	200
Total credits per session	56.39	28.02	10	400
Academic Year:				
2014/15	0.19	0.40	0	1
2015/16	0.19	0.40	0	1
2016/17	0.20	0.40	0	1
2017/18	0.19	0.39	0	1
2018/19	0.23	0.42	0	1
Observations	503358			

Notes: The Table reports the mean, standard deviation, minimum and maximum values of the main variables used in the paper. There is a total of 503,358 student-exam level observations, covering 51,555 students. Statistics on assignments mark are for the sample of exams which do include assignments as part of their coursework. Statistics for exam duration are limited to the last 2 academic years of the sample.

(4.30 pm).[24] The majority of exams (48%) last two hours. In terms of exam weight, 47% and 41% of exams are worth 10 and 20 credits, respectively. The average age of the students is 21. A total of 47% of students are female. In terms of ethnicity, 64%, 11%, and 7% of students are white, Asian, and black, respectively, and the remainder consists of other ethnicities. The vast majority of the sample are UG students (93%), and 50% of students are enrolled in a STEM school. Students sit an average of 6 (3) exams per year (session), amounting to 91 (56) credits in total.

2.3 Theoretical framework

To motivate the empirical analysis, we develop a simple model of individual's cognitive performance and optimal effort at different time of day. In practice, we adapt the model by Park (2020) to fit our context. Let a be true cognitive ability, knowledge or stock of human capital of an individual. Let $e \in [0,1]$ be the level of effort exerted to perform a cognitive task, such as an assessment or a skill-intensive assignment on the job. For example, an individual may choose a specific study strategy or food intake strategy. When $e = 1$, she is exerting maximum effort, and 0 otherwise. Let $t \in [0,1]$ be the time of day in which the task is performed. For simplicity, $t = 1$ when she is performing the task at her optimal time of day, and 0 at the worst time. What is "optimal" for an individual is not her choice; rather, it is a biological feature determined by nature. According to the circadian rhythm literature, t is itself a function of environmental conditions (see Section 3.3 for more details). For simplicity, in this section we abstract from it.

In our setting, an individual performs one high stake cognitive assessment per day, which is scheduled at a specific time of day. Her performance can be expressed as $y(e;t,a)$, where e is the endogenous input, t is the exogenous input, and together they jointly determine her performance for a given level of ability a. Since the latter is commonly unobserved, y can be used as a signal for a in the job market. The function $y(.)$ is such that $\frac{\partial y}{\partial e} > 0$ and $\lim_{e\to 1 t\to 1} y(e;t,a) = a$. The latter implies that the observed cognitive performance reflects individual true ability when she exerts maximal effort at the optimal time of day.

[24] Note that an exam is more likely to be allocated to a morning slot because it allows for the conjugate examination of modules taught in different satellite university campuses.

We assume that her utility function is $U(x,c)$, so that she derives a benefit from the consumption of a composite good x and experiences a cost c from undertaking the task ($\frac{\partial U}{\partial x} > 0$ and $\frac{\partial U}{\partial c} < 0$). This allows us to make a link between her task performance and the economic stakes. She needs to decide the optimal amount of effort:

$$\begin{aligned} \max_{e} \;& U(x,c) \\ \text{s.t. } & x = w(y(e;t,a)) \\ & c = c(e,t) \end{aligned} \tag{1}$$

where w is the monetary compensation, which depends positively on the realized performance ($\frac{\partial w}{\partial y} > 0$), whereas the cost is increasing in effort and decreasing in the difference from optimal time ($\frac{\partial c}{\partial e} > 0$ and $\frac{\partial c}{\partial t} < 0$). The compensation could be the return to education resulting from college admission, the wage offer after a successful job interview or the wage rise after a job performance evaluation.

Let e^* be the optimal effort that solves (1). Replacing the constraints into the utility function and setting $\frac{dU}{de} = 0$ yields the following first order condition:

$$\frac{\partial U}{\partial x}\frac{\partial w}{\partial y}\frac{\partial y}{\partial e^*} = -\frac{\partial U}{\partial c}\frac{\partial c}{\partial e^*} \tag{2}$$

which says that the chosen level of effort equalizes the trade-off between the marginal benefit on her performance and the marginal cost of taking a cognitive assessment. The higher the economic stakes, the larger is the marginal return on the performance and the larger is the level of effort chosen when taking the assessment at a non-optimal time of day. Re-arranging equation (2), we obtain the effect of effort on performance:

$$\frac{\partial y}{\partial e^*} = -\frac{\frac{\partial U}{\partial c}\frac{\partial c}{\partial e^*}}{\frac{\partial U}{\partial x}\frac{\partial w}{\partial y}} \tag{3}$$

which is positive, as $\frac{\partial U}{\partial c} < 0$.

Our research may be interpreted as an investigation of the overall effect of time of day on cognitive performance, which is the policy relevant variable. That is, we aim to estimate:

$$\frac{dy}{dt} = \frac{\partial y}{\partial t} + \frac{\partial y}{\partial e^*}\frac{\partial e^*}{\partial t} \tag{4}$$

which is a combination of two effects: the direct and indirect effects of time of day on performance. The direct effect can be thought as the effect of cognitive functioning, whereas the indirect effect is mediated by effort.

In practice, $\frac{\partial y}{\partial e^*}\frac{\partial e^*}{\partial t}$ is unobserved. Hence, even if time of day is randomized, the total effect that we estimate will be a combination of these two components. Substituting Equation (3) into (4), we obtain:

$$\frac{dy}{dt} = \frac{\partial y}{\partial t} - \frac{\frac{\partial U}{\partial c}\frac{\partial c}{\partial e^*}}{\frac{\partial U}{\partial x}\frac{\partial w}{\partial y}}\frac{\partial e^*}{\partial t} \tag{5}$$

Equation (5) allows to better visualize the underlying mechanisms behind our investigation. Even if the time of the assessment is not optimal, the individual will increase her effort as long as the cost is compensated by the high economic stakes. Our null hypothesis is that time of day has no effect on actual performance ($\frac{dy}{dt} = 0$). If we estimate a positive effect of time, it implies that the biological component affects performances even in an high stakes environment. This suggests that an unfavorable time of examination may disadvantage an individual in the long term when competing for higher education or jobs. Finally, under the mild assumption that effort and time are weak substitute ($\frac{\partial e^*}{\partial t} \leq 0$),[25] the indirect effect in Equation (5) has a negative sign, hence her performance will be a lower bound of what she could have achieved had she taken the assessment at the optimal time of day.

2.4 Empirical model

Given Equation (5), the reduced form specification that we estimate is simply:

$$Y_{i,e,t} = \alpha_i + \boldsymbol{T}_{e,t}^{'}\boldsymbol{\beta} + \boldsymbol{X}_{i,e,t}^{'}\boldsymbol{\gamma} + \epsilon_{i,e,t}, \tag{6}$$

[25]When t is close to 1, the brain is more efficient, then e is low because the worker does not have to exert so much effort. Whereas, if t is low, the worker needs to compensate the lowering of brain capacity, hence e has to be high in order to boost the performance. For example, an individual may choose a specific study strategy depending on the time of day that an assessment is scheduled. Similarly, an individual may employ a particular food intake strategy for a morning assessments, while she may choose a different strategy if an assessment is scheduled in the early or late afternoon.

where $Y_{i,e,t}$ is the standardized mark achieved by student i, in exam e, in academic year t. α_i is the individual FE, which captures students' time-invariant characteristics, both observable and unobservable, that might affect their performance. The term $\boldsymbol{T_{e,t}}$ is a vector of exam-specific time of day variables, namely: A dummy equal to 1 if an exam is in the early afternoon (1.30 pm), and 0 otherwise, and a dummy equal to 1 if an exam is in the late afternoon (4.30 pm), and 0 otherwise. Hence, the main parameters of interest, denoted by the vector $\boldsymbol{\beta}$, capture the average treatment effect (ATE) of time of day for students taking an early or late afternoon exam relatively to a morning exam, which is used as control group.[26] Our null hypothesis is that $\boldsymbol{\beta} = 0$ (time of day does not matter). The vector of covariates $\boldsymbol{X_{i,e,t}}$ contains the set of conditions for the validity of the conditional random assignment as well as a large set of student and exams' characteristics that might affect exam marks.[27] Finally, the error term $\epsilon_{i,e,t}$ captures common unobservable shocks. In our research design, the within student quasi-randomization of time of day ensures that $\boldsymbol{T}^{'}_{\boldsymbol{e,t}}$ is not correlated with $\epsilon_{i,e,t}$ in Equation (6). Further details about the identification are provided in Appendix A.3.

3 Empirical Analysis

This section is organised as follows: first, we estimate the effect of the time of the day on student performance using Equation (6) and present the results of a series of robustness checks to support the primary findings. Standard errors are always clustered by exam year.[28] Second, we draw from the literature on the circadian rhythm and examine effect heterogeneity to understand under which conditions time-of-day effects on performance are strongest. Lastly, we provide a simple back-of-the-envelope calculation of the implications of the present study for the rescheduling of college admission tests, which is an important context that could benefit from the results of our research.

[26] Given our research design, the ITT is equal to the ATE because we have full compliance.

[27] This includes: exam weight (in credits), class size, students' age, gender, ethnicity, nationality, a dummy for whether the student is undergraduate, a dummy for whether the exam belongs to STEM schools, academic year FE, year of enrollment FE, stage of study FE, school FE, the number of assignments and the average assignments mark, the number of exam credits, and the total number of credits the students achieved in a particular examination session and in the year.

[28] Results do not change when clustering at exam level and are available upon request.

3.1 Primary Results

Table 3 reports the main findings of the paper. In Column (1), we regress the standardised exam mark only on the main variables of interest, $\boldsymbol{T}_{e,t}$. The average effect of taking an exam in the early (late) afternoon is 0.046 (0.050) standard deviations (SD) higher than taking an exam in the morning. Column (2) includes the conditions (i.e., number of exams per session fixed effects (FE), class size, school FE and exam weight (in credits) FE) to enable the CIA to be satisfied. The estimated coefficients remain at 0.050 (0.042) SD for an early (late) afternoon exam. In Column (3), we include student FE, α_i, which absorbs students' time-invariant characteristics, including their individual study plan, combinations of exams taken, family background, motivation and ability. While the adjusted R-square increase considerably, the average effect remain positive and relatively stable, with 0.068 (0.043) SD for exams taken in the early (late) afternoon compared to the morning. Finally, Column (4) includes the remaining control variables, $\boldsymbol{X}_{i,e,t}$, that vary between students and may affect a student's performance. The average effect of an early afternoon exam remains 0.068, while the effect of a late afternoon exam is 0.036 SD.

We use the last specification, which is our preferred one, to perform a test of equality for early and late afternoon exams. The null is rejected at the 5% level. (i) The latter result and (ii) the fact that the parameter attached to the early afternoon is consistently larger in magnitude than the parameter attached to the late afternoon in all specifications estimated in this paper, suggests an inverse U-shaped relationship between the time of the day and cognitive performance. Overall, the results indicate a positive effect of the time of the day on student performance and suggest a significant improvement in scores for exams quasi-randomly assigned in the early afternoon time slot.

3.2 Robustness Checks

To explore the sensitivity of the results, we perform several robustness checks. The results are reported in Columns 5–10 of Table 3.

First, the regression model in Equation (6) satisfies the CIA if we can control for exam duration. As previously mentioned, data on exam duration are only available for the final

Table 3: Effects of Time of Day on Student Performance

	Primary Results				Robustness					Placebo
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Early Afternoon [0,1]	0.046***	0.050***	0.068***	0.068***	0.061***	0.032**	0.062***	0.070***	0.068***	0.009
	(0.017)	(0.016)	(0.015)	(0.015)	(0.017)	(0.012)	(0.014)	(0.014)	(0.015)	(0.015)
Late Afternoon [0,1]	0.050***	0.042**	0.043***	0.036**	0.026	0.019	0.022	0.034**	0.034**	-0.013
	(0.018)	(0.017)	(0.016)	(0.015)	(0.017)	(0.014)	(0.016)	(0.015)	(0.015)	(0.015)
Test of equality of Early and Late Afternoon exams in linear regressions above:										
$Pr > F$	0.815	0.623	0.129	0.036	0.032	0.313	0.014	0.021	0.033	0.171
Conditions:		✓	✓	✓	✓	✓	✓	✓	✓	✓
Student FE:			✓	✓	✓	✓	✓	✓	✓	✓
Covariates:				✓	✓	✓	✓	✓	✓	✓
Duration:					✓					
Exam FE:						✓				
Day FE:								✓		
Room FE:									✓	
Clusters:	7665	7665	7626	7626	6475	7578	7373	7626	7613	7626
Observations:	503359	503359	500959	500959	432185	500906	312103	500959	500920	500959
Adjusted R^2:	0.000	0.022	0.439	0.462	0.471	0.531	0.464	0.466	0.473	0.461

Notes: Dependent variable: standardized exam mark. Key variables: dummies for whether the exam was taken at 1.30pm (Early Afternoon Exam) or at 4.30pm (Late Afternoon Exam), where the 9am (Morning Exam) is the excluded category. Standard errors are clustered by exam-year. Observations are at the student-exam-year level. In Column (1), we report the unconditional estimates. In Column (2), we include only the conditions defining the allocation rule. In Column (3), we include student fixed effects (FE), and in Column (4) we include all the covariates that change within students. This represents our preferred specification. In Column (5), we report the results when including a dummy for exams longer than 2.5 hours. In Columns (6), we account for exam FE. In Column (7), we report the results excluding first year UG students. In Columns (8) and (9), we include, respectively, day and room FE. In Column (10), we report the results of a placebo test. Below the test of equality of Lunch and Afternoon exams is reported the p-value of the test statistic. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

two years of the sample; therefore, in the main analysis, we estimated the model using a proxy - namely, the exam weight FE (which is highly correlated with exam duration). This may represent a potential limitation. As a robustness check, we use data on exam duration from the last two years to impute the duration data for the same exams that appeared in the first three years of the sample.[29] We then estimate the same regression model, including a dummy variable with a value of 1 if an exam is longer than 2.5 hours and a value of 0 if otherwise. As shown in Column 5, the estimated peak performance in the early afternoon time slot is robust to this inclusion.

Second, one concern may be that exams held in the morning are systematically different from those held later in the day, which could explain why students perform differently. This concern may be motivated by the fact that we do not have information about the specific content of the exams. Hence, even though we know the allocation rules adopted by the

[29]The estimation sample in this case is smaller than the original sample because we cannot impute this information for all exams. We wish to highlight that only 8% of the exams in the sample are longer than 2.5 hours. The majority of these exams were sat by first-year UG students. This means that it is likely that these exams are compulsory exams that are offered each year. Hence, the underlying (implicit) assumption of this robustness check is that if an exam is longer than 2.5 hours in a given year, it is highly likely that it was longer than 2.5 hours a few years earlier.

exam office, in finite samples, the conditional random assignment of exam time may not guarantee the balancing of their characteristics at a different time of day. Unfortunately, the only information we have about the exams' characteristics are class size, weight and duration. To show that potential differences in exams' characteristics do not overthrow the results, we proceed by re-estimating Model (6), also including exam FE, which controls for time-invariant observable and unobservable exam attributes. The identifying assumption is that the content or characteristics of the exams do not vary over the years. As shown in Column 6, the time-of-day effect estimated in the early afternoon is still significant, though smaller in magnitude. When we study effect heterogeneity in Section 3.3, we show that this reduction is driven by the performances in the spring session.

There are two possible explanations for this result. On one side, there might be some imbalance, particularly in the spring session, in the content or characteristics of exams that are not picked up by the balancing tests. While this is possible in principle, unfortunately, we cannot provide evidence to discard the concern. Conversely, exam FE may also control exam difficulty, which could be as well an outcome of exam time. In the latter case, exam FE is potentially a bad control (e.g., Ding and Miratrix, 2015; Cinelli et al., 2020), causing a bias (downward) in the result. This is because it accounts for much more than what is needed to identify the parameter we are after, including part of the total effect we wish to measure. Even though (i) including a potential bad control invalidates the CIA and (ii) *all* the results of the paper do not change in a fundamental way when we include exam FE, we believe it is still important to report this robustness check both for the primary results and for the heterogeneous analysis.

Third, the differential effects of the time of the day might be contaminated by first-year UG students. This is because, in line with the UK higher educational system, exam marks for these students do not count towards the final degree. Hence, this sub-group of individuals, which accounts for 37% of the observations, may be intrinsically different from the rest of the sample because the stakes are not as high for their exams. Moreover, first-year UG students may lack experience in university exam procedures or be unaccustomed to the rhythm of study and preparation, so they might respond to different incentives. To account for this concern, we re-estimate Model (6) on a restricted sample excluding this sub-group and find that, as shown in Column 7, the estimated peak performance in the early afternoon time slot is still robust.

Fourth, another potential concern could be that comparing students taking exams across days throughout an exam session may pick up other confounding effects besides the time of day, including accumulated cognitive fatigue or warm-up effects. Accordingly, we include in Model (6) a set of days of exam session FE to take these potential confounders into account. The results in Column 8 show that this is not a concern. Along the same line, fatigue, unrest, stress, warm-up, and the length of a working day can also affect the productivity patterns *within* a day (e.g., Hamermesh, 1990, 1996). In our context, one could argue that students perform worse in the morning because they have less time to recover from a previous task, or they perform better in the early or late afternoon because they have more time to warm-up for the task. However, given our research design, individuals perform one cognitive task per day only. Since this is a high stakes assessment, it is likely that they organise their day to maximise the performance of that task.[30] Moreover, any individual factor potentially correlated with these mechanisms is already absorbed by the individual FE (e.g., having a part-time job). Hence, although we cannot exclude completely that these alternative factors co-vary with the time of exams, we argue that they are likely to play only a minor role in explaining the difference in performance between students taking exams scheduled at different times of the day.

Fifth, it could be argued that the type of room or the environment in which students take their exams may also affect the results. In our context, students taking an exam in the morning may perform worse simply because, for example, some rooms or buildings are colder at this time of day, particularly in January. Hence, we include exam room FE to control for potential unobserved room-specific effects that may affect student performance during an exam, such as cold/heat, light/dark, size, air quality, seat disposition or distance from other students, and

[30] For a similar reason, students who sit an exam in the morning could be affected by sleep inertia compared with those who sit exams later in the day. Sleep inertia refers to the transitional state between sleep and wake, which is marked by impaired performance if we wake up too early and we are still in our biological night. Hence, it could be argued that the observed effects are driven by the length of time that students have been awake. Although information on sleep patterns is unavailable, this explanation could be relevant if students sitting exams at 9 am were tested within a very small window of time after waking up (the first half an hour), which is unlikely in our case.

the average number of invigilators (which is correlated with class size and room size). The results in Column 9 are again robust.

Lastly, we conduct a placebo test to show that the estimated effects do not exist when, in fact, they should not. Specifically, we randomise exam time across the three conditions. Column 10 reports the results, which confirm that, indeed, there is no effect. Alternatively, we also consider the possibility that the allocation process is not completely (conditionally) random. This could occur if students could choose their exams according to their time-of-day preference, which would lead to a correlation between students' unobserved characteristics and the number of exams taken at a specific time of day. Therefore, we assess the degree of omitted variable bias in our estimates by implementing the coefficient stability test developed by Altonji et al. (2005) and generalized by Oster (2019). The rationale behind this method is the idea that both the movements of the main coefficients after the inclusion of the observed controls and the movements in the adjusted R-square should be tested to evaluate the degree of omitted variable bias. In Table A7 in Appendix, we present the validation results of the analysis. As can be observed, under the most restrictive test scenario, all the bounding sets are close to the estimated coefficients. This confirms that the omitted variable bias from the possible correlation between students' unobserved preferences and the time of day is not a concern for us.

3.3 Heterogeneous Time of Day Effects

Having established a causal link between time of day and high stakes cognitive performances, it is important to understand under which conditions time of day effects are strongest. To answer this question, we exploit the literature on the circadian rhythm. As previously mentioned, circadian rhythms are ingrained biological processes that regulate many aspects of human life, including cognition (Schmidt et al., 2007). In a given population, peak sleep and wake times show a near-Gaussian distribution, with extreme early types waking up when extreme late types fall asleep (Cardinali, 2008). Drawing from the literature, the salience of this primordial clock varies depending on (mainly) three key factors: (1) sunlight exposure, (2) type of (cognitive) task, and (3) age. In what follows, we exploit the richness of our data and present the results of a heterogeneity analysis that allows us to gauge at the extent to which these factors matter in explaining the observed time of day effects on performance.

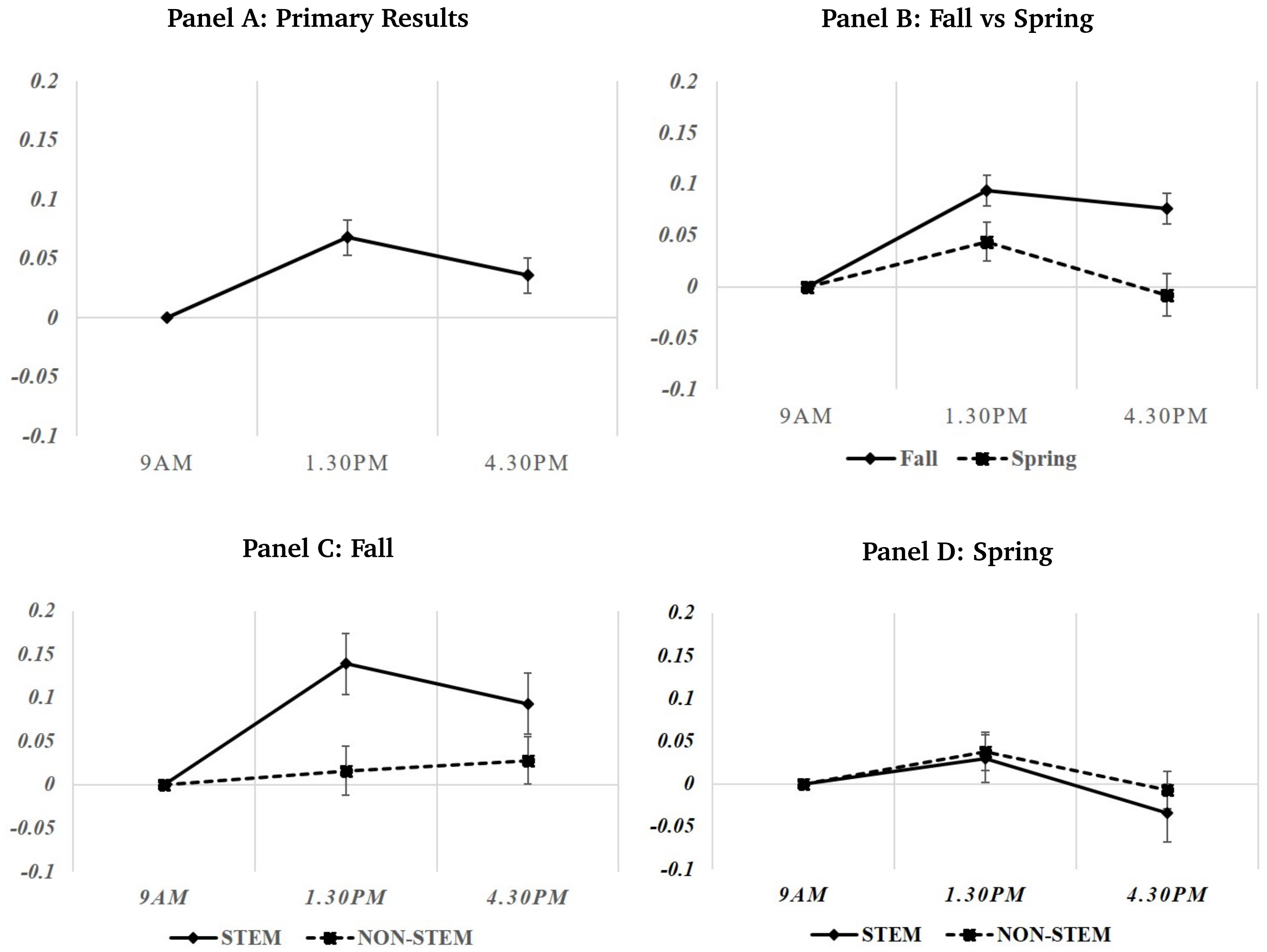
Sunlight exposure. First, a host of external signals, often called Zeitgebers, act as cues capable of synchronizing this inner clock. Sunlight exposure is among the most important factors (Duffy and Wright, 2005; Roenneberg et al., 2007b; Wright et al., 2013). Indeed, compelling evidence suggests that chronotypes are significantly affected by seasonal changes (Shawa et al., 2018). Hubert et al. (1998) first noted that earlier chronotypes were more common in summer compared with winter. This result was later confirmed in an influential research by Kantermann et al. (2007), who showed how the timing of sleep followed the seasonal progression of dawn. Therefore, from a biological point of view, different effects of the time of day between the fall and spring examination sessions are expected: in the fall session, the average student may be severely disrupted by early morning exams due to a lack of light exposure, while this effect should be mitigated by the earlier sunrise in the spring exam session.

To test this hypothesis, we split the sample into fall and spring exam sessions. Balancing checks (similar to those in section 2.1) for these samples are provided in Panel A of Table A3 in Appendix. In Panel B of Figure 1, we provide a graphical representation of the results. The full set of estimates are provided in Panel A of Table A8 in Appendix. As one can see, the figure confirms the expectations and shows that the effect of the time of day is significantly stronger in the fall exam session when morning sunlight is limited.[31] For example, sitting an exam in the early (late) afternoon increase performance by 0.094 (0.076) SD compared with sitting an exam in the morning. Whereas, for the spring exam session, the results confirm that students perform better early in the afternoon, although the point estimate is less than half the effect observed in fall. In contrast, no significant effect is observed for late afternoon exams.[32] These results are robust to the same battery of sensitivity checks outlined for the

[31]The interaction effect is significant and it is omitted for simplicity, but it is available upon request.

[32]A potential concern is that the different results between fall and spring could have been driven by other confounders occurring in May–June but not in January. For example, research by Bensnes (2016) suggest a connection between pollen exposure and decreases in exam performance. If students who sat exams early or late in the afternoon in spring are more affected by daytime pollen exposure, then pollen levels should be controlled for in the model. For this reason, in the present study, we conduct a robustness analysis that includes daily city-level pollen data obtained from the British Meteorological Office. The results are identical to the benchmark and are omitted.

Figure 1: Time of Day and Performances: by Season and School



Notes: The Figures provide a graphical representation of the main findings of the paper. In Panel A, we report the estimated coefficients, with associated standard error, of the effect of time of day on performance. In Panel B, we report the estimated coefficients, with associated standard error, of the effect of time of day on performance comparing sample Fall and Spring exam session. In Panel C, we report the estimated coefficients, with associated standard error, for the sample of students enrolled in the Fall exam session (January), by school. Similarly, in Panel D, we report the estimated coefficients, and associated standard errors, for the sample of students in the Spring session (May-June), by school. Exams taken at 9 am represents the control group. One can grasp this inverse-U shape relationship between time-of-day and performances.

primary results and reported in Table A9 in Appendix.

Type of task. Second, Goldstein et al. (2007) and Zerbini et al. (2017) show that time of day and chronotype effects vary systematically depending on the type of task being performed. Specifically, both papers show significant time of day effects for cognitive tasks involving fluid intelligence (e.g., working memory, logical thinking, problem-solving, and abstract reasoning), but not for those involving crystallized intelligence (e.g., knowledge and vocabulary).

To test this hypothesis, we split the sample between STEM exams, which tend to rely more on fluid intelligence, and non-STEM exams, which tend to rely more on crystallized intelligence. Balancing checks for these samples are provided in Panel B and C of Table A3 in Appendix. In Panel C and D of Figure 1, we provide a graphical representation of the results. The full set of estimates are provided in Panel A of Table A8 in Appendix. Once again, the results confirm our expectations, as they indicate that time of day matters almost exclusively for STEM exams.[33] Specifically, we find that that students taking a STEM examination in the early (late) afternoon in the fall increase their performance by 0.139 (0.093) SD; however, no similar effects are found for students taking a non-STEM examination. Overall, these results suggest that tasks involving fluid intelligence benefit from later exam timing. Also in this case, they are robust to the same battery of sensitivity checks outlined for the primary results and reported in Table A9 in Appendix.

Age. Finally, age is one of the main determinants of a person's chronotype (Roenneberg et al., 2004).[34] Children are generally earlier chronotypes, progressively delaying their body clock during development and reaching a maximum "lateness" at around age 20, before subsequently advancing their clock with increasing age. Therefore, from a biological point of view, it is expected that the performance of the average (biologically speaking) student in exams scheduled early in the morning should be worse compared with their performance later in the day.

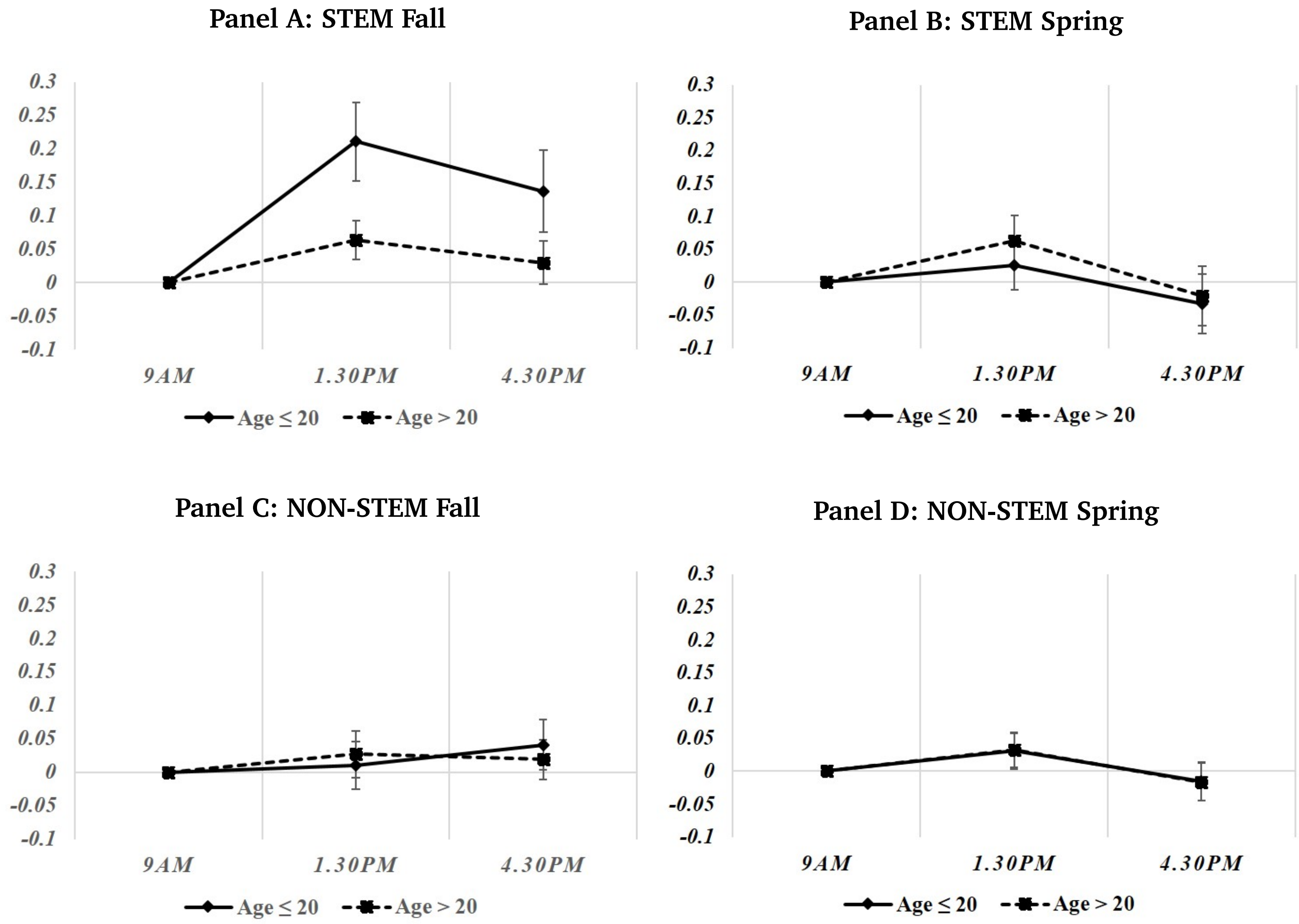
To test this hypothesis, we compare time of day effects between students aged ≤ 20 and those aged > 20. Balancing checks for these samples are provided in Table A3 in Appendix. In Figure 2, we provide a graphical representation of the results. The full set of estimates are provided in Panel B and C of Table A8 in Appendix. Importantly, in the regression analysis we control also for year of study, in order to isolate the effect of age from other confounding factors, such as experience in university exam procedures.[35] The figure confirms the findings from the circadian rhythm literature and shows that the estimated time of day effects are significantly larger for students 20 years old or below. For example, for a student in this age

[33]As before, the interaction effect is significant and it is omitted for simplicity, but it is available upon request.

[34]Toh et al. (2001), Vink et al. (2001) and Archer et al. (2003) show that chronotypes depend on genetic factors and Roenneberg et al. (2003) on environmental factors.

[35]As before, the interaction effect is significant and it is omitted for simplicity, but it is available upon request.

Figure 2: Time of Day and Performances: by Age



Notes: The Figures provide a graphical representation of the main findings of the paper. In Panel A (B), we report the estimated coefficients, with associated standard error, for the sample of students enrolled in the Fall (Spring) exam session for STEM schools. Exams taken at 9 am represents the control group. Similarly, in Panel C (D), we report the estimated coefficients, and associated standard errors, for the sample of students in the Fall (Spring) session for non-STEM schools. One can grasp this inverse-U shape relationship between time-of-day and performances.

bracket, taking a STEM exam in the Fall, there is a 0.21 SD increase in performance when moving an exam from morning to early afternoon. In contrast, for students older than 20 years of age, the estimated time of day effects are significantly milder. Again, the results are robust to the same battery of robustness checks outlined for the primary results and reported in Table A10 in Appendix.

3.4 Implications

The main implication of our study for economics is that, organizations around the world, seeking to measure individual cognitive ability at best, should re-weight the exam marks or re-schedule their assessments according to the pattern of the circadian rhythm. For simplicity, in this section we focus on one implication only: re-scheduling. In general, for young adults, cognitive assessments involving problem-solving skills should be moved to early afternoon at times of year when sunlight exposure is limited. Conversely, other types of cognitive assessments could be moved to the beginning or the end of the workday as they are less impacted by time of day effects.

In relation to our specific setting, the reschedule of assessments can be primarily implemented by schools to improve the conditions under which students are tested. Several papers in the education literature have investigated the effectiveness of various interventions aimed at improving student outcomes. For example, studies have examined the effects of reducing class sizes (Bandiera et al., 2010), delaying school start time (Carrell et al., 2011), and evaluating teachers' performance (Taylor and Tyler, 2012). Many of these policies, while effective, can also be quite expensive to implement. Our results suggest that schools have an additional and simple tool that they can use to achieve this goal, that is: to optimize the daily examination schedules.[36] This new policy may be implemented in a variety of ways, depending on the context. For example, since the scheduling of examinations is centrally managed, moving STEM examinations to the early afternoon and other examinations to the morning or late afternoon could increase test scores. However, there are constraints to the extent to which universities can implement such a policy, as universities will be constrained by the supply of examination rooms at each given hour. However, this issue could be solved by a more careful optimization of the capacity constraints.

In relation to other settings, the reschedule of cognitive assessments could be implemented to improve the quality of information (i) contained in school admission tests or (ii) obtained through job interviews and job performance evaluations. Particularly for organizations wanting to hire new promising candidates or to promote productive employees, there is little doubt that innovations that improve efficiency gains and that have little or no increase in costs would be a goldmine. Indeed, the adoption of superior ideas may lead to improvements in performance or higher mark-ups for firms. Studies on personnel, labor economics and management

[36]It should be noted that the results are interpreted as indicating higher performances and we argue that students could do better with different examination schedules. We do not interpret the results as higher learning or knowledge. Indeed, in our context, students have, on average, the same knowledge at the moment of the examination, as there is (conditional) random assignment.

science have long examined possible solutions for achieving efficiency gains.[37] Our results suggest that, for workers performing cognitive tasks, the scheduling of their job interviews or job performance evaluations should be re-arranged as well.

3.5 Rescheduling of college admission tests: A numerical example

To understand the relevance of our findings, we apply our insights to an important context that could greatly benefit from our research: college admission tests. The Scholastic Assessment Test (SAT) is a standardized test widely used for college admissions in the U.S. The SAT is typically taken towards the end of high school, testing begins at 9 am, takes three hours to finish, and it is intended to measure cognitive abilities that are needed in college. This is a particularly suited example because students can take the test at different times of the year (sessions), in different locations (latitudes) and at different ages (chronotypes). All this gives rise to potential discrimination in the assessment of student ability because the SAT does not standardize the test results according to some important features determining student performance.

We draw on the 2020 SAT suite of assessments annual report, which presents data on over 2 million students in the U.S. who took the SAT test in that year. Accordingly, we implemented a back of the envelope calculation using our estimate of 21 percent SD increase in performance associated with a cognitive assessment performed at 1.30 pm, relative to 9 am, for young students (age $\leq$ 20). The formula for the effect size states that $ES = (SAT_l - SAT_m)/SD$, where SAT_l (SAT_m) is the average total SAT score for a test scheduled in the early afternoon (morning) and SD is the standard deviation of the overall test score. Given our raw data, scores range from 400 to 1600 and $SD = 211$. We maintain the simplifying assumption that the effect size found in our paper applies to this context as well ($ES \approx 0.21$). This assumption is arbitrary and can be relaxed depending on the data at hand.

Extrapolating from this relationship implies that, if an average student could have moved her time test from morning to the early afternoon, she would have obtained, on average, 44

[37] For example, via better compensation schemes (Lazear, 2000), better workplace practices, information technology, human capital investments, innovations (Black and Lynch, 2001, 2004), team incentives (Hamilton et al., 2003), higher wages (Charness and Kuhn, 2007), limiting task juggling (Coviello et al., 2014), or promoting on-the-job trainings (Konings and Vanormelingen, 2015).

additional points on her SAT. Based on the 2018 National University Rankings of U.S. News, this magnitude allows you to move from the 75th percentile SAT Score of a Tufts University student to the 75th percentile SAT Score of a Harvard student. Using data from the Equality of Opportunity Project (Chetty et al., 2017), this implies an average earnings loss of $8,000 by the age of 34. Thus, our findings suggest that a simple innovation in the scheduling of college admission tests could lead to potentially large efficiency gains. Of course, issues of external validity and maintained assumption suggest that the exact result from this back of the envelope calculation should be viewed with caution.

4 Conclusion

The time of day at which a cognitive assessment is conducted may affect the performances; however, to date, economists have paid little attention to this relationship. In our paper, we exploit a unique setting to provide the first causal estimate of this link. Specifically, in the context of university education, students sit different examinations at the end of each semester, the scores of which generally make up a large portion of their final grades. The identification strategy relies on a clear assignment rule and on quasi-random assignment of the time of day of the examinations. We find that the performance profile of students during the day has an inverse U-shape relationship, whereby peak performance occurs in the early afternoon (1.30 pm) and is poorer in the morning (9 am) as well as in the late afternoon (4.30 pm). Furthermore, these time of day effects are much larger (i) in January, when sunlight exposure is limited, than May–June, (ii) for students sitting STEM examinations, which are tasks requiring more fluid intelligence than crystallized intelligence, than those sitting non-STEM examinations, and (iii) for students aged below 20. We support this causal relationship with a battery of robustness and falsification tests.

These results suggest that the time of day is a crucial determinant of students' performance. These gains in performance may be caused by different mechanisms. We do not have information that allows us to identify, separately, the magnitude of each specific mechanism driving this relationship; however, our results can be rationalized by recent findings in the circadian rhythm literature. Since the results of high-stakes cognitive assessments can have lasting consequences on earnings and career, it is important to measure individual ability under the best circumstances. We conclude that efficiency gains can be obtained in a simple way and, potentially, for many sectors of the economy. The main policy implication is that cognitive assessments involving problem-solving skills are more affected by time of day and should be moved to the early afternoon at times of year when sunlight exposure is more limited. The rescheduling of the assessments could be primarily implemented by schools to increase student outcomes. Similar implications apply to other organizations seeking to measure individual cognitive ability at best.

References

ACEMOGLU, D. AND D. AUTOR (2011): "Skills, Tasks and Technologies: Implications for Employment and Earnings," in *Handbook of Labor Economics, Volume 4. Amsterdam: Elsevier-North*, 1043–1171. [2]

ALTONJI, J. G., T. E. ELDER, AND C. R. TABER (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184. [9], [20]

ARCHER, S. N., D. L. ROBILLIARD, D. J. SKENE, M. SMITS, A. WILLIAMS, J. ARENDT, AND M. VON SCHANTZ (2003): "A length polymorphism in the circadian clock gene Per3 is linked to delayed sleep phase syndrome and extreme diurnal preference," *Sleep*, 26, 413–415. [23]

BANDIERA, O., V. LARCINESE, AND I. RASUL (2010): "Heterogenous Class size effects: New evidence from a panel of university students," *The Economic Journal*, 120, 1365–1398. [25]

BENSNES, S. S. (2016): "You sneeze, you lose:: The impact of pollen exposure on cognitive performance during high-stakes high school exams," *Journal of Health Economics*, 49, 1 – 13. [2], [21]

BENSNES, S. S. AND B. STROM (2019): "Earning or Learning? How Extending Closing Time in the Retail Sector Affects Youth Employment and Education," *Oxford Bulletin of Economics and Statistics*, 81, 299–327. [6]

BLACK, S. E. AND L. M. LYNCH (2001): "How to Compete: The Impact of Workplace Practices and Information Technology on Productivity," *The Review of Economics and Statistics*, 83, 434–445. [26]

——— (2004): "What's Driving the New Economy?: The Benefits of Workplace Innovation," *The Economic Journal*, 114, F97–F116. [26]

BRADLEY, S. AND C. GREEN (2020): *The Economics of Education (Second Edition)*, Academic Press, second edition ed. [3]

CARDINALI, D. P. (2008): *Chronoeducation: How the biological clock influences the learning process*, Cambridge University Press, 110–126. [2], [20]

CARRELL, S. E., T. MAGHAKIAN, AND J. E. WEST (2011): "A's from Zzzz's? The Causal Effect of School Start Time on the Academic Achievement of Adolescents," *American Economic Journal: Economic Policy*, 3, 62–81. [6], [25]

CHARNESS, G. AND P. KUHN (2007): "Does Pay Inequality Affect Worker Effort? Experimental Evidence," *Journal of Labor Economics*, 25, 693–723. [26]

CHETTY, R., J. N. FRIEDMAN, E. SAEZ, N. TURNER, AND D. YAGAN (2017): "Mobility Report Cards: The Role of Colleges in Intergenerational Mobility," Working Paper 23618, National Bureau of Economic Research. [27]

CHO, H. (2017): "The effects of summer heat on academic achievement: A cohort analysis," *Journal of Environmental Economics and Management*, 83, 185 – 196. [2]

CINELLI, C., A. FORNEY, AND J. PEARL (2020): "A Crash Course in Good and Bad Controls," . [18]

COTTI, C., J. GORDANIER, AND O. OZTURK (2018): "Class meeting frequency, start times, and academic performance," *Economics of Education Review*, 62, 12 – 15. [6]

COVIELLO, D., A. ICHINO, AND N. PERSICO (2014): "Time Allocation and Task Juggling," *American Economic Review*, 104, 609–23. [26]

DILLS, A. K. AND R. HERNÁNDEZ-JULIÁN (2008): "Course scheduling and academic performance," *Economics of Education Review*, 27, 646 – 654. [6]

DING, P. AND L. W. MIRATRIX (2015): "To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias," *Journal of Causal Inference*, 3, 41–57. [18]

DUFFY, J. AND K. WRIGHT (2005): "Entrainment of the Human Circadian System by Light," *Journal of biological rhythms*, 20, 326–38. [21]

EBENSTEIN, A., V. LAVY, AND S. ROTH (2016): "The Long-Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution," *American Economic Journal: Applied Economics*, 8, 36–65. [2]

EDWARDS, F. (2012): "Early to rise? The effect of daily start times on academic performance," *Economics of Education Review*, 31, 970 – 983. [6]

GHISLETTA, P. AND T. LECERF (2018): *Crystallized and Fluid Intelligence*, Oxford Bibliographies. [5]

GOLDIN, C. AND L. KATZ (2009): *The race between education and technology*, Harvard University Press. [2]

GOLDSTEIN, D., C. S. HAHN, L. HASHER, U. WIPRZYCKA, AND P. D. ZELAZO (2007): "Time of day, intellectual performance, and behavioural problems in morning versus evening type adolescents: is there a synchrony effect?" *Personality and Individual Differences*, 42, 431–440. [2], [5], [22]

GOULAS, S. AND R. MEGALOKONOMOU (2020): "Marathon, Hurdling or Sprint? The Effects of Exam Scheduling on Academic Performance," *The B.E. Journal of Economic Analysis & Policy*. [6]

HAMERMESH, D. (1996): *Workdays, Workhours and Work Schedules: Evidence for the United States and Germany*, Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. [19]

HAMERMESH, D. S. (1990): "Shirking or Productive Schmoozing: Wages and the Allocation of Time at Work," *ILR Review*, 43, 121–S–133–S. [19]

HAMILTON, B. H., J. A. NICKERSON, AND H. OWAN (2003): "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation," *Journal of Political Economy*, 111, 465–497. [26]

HANUSHEK, E. A. AND D. D. KIMKO (2000): "Schooling, Labor-Force Quality, and the Growth of Nations," *The American Economic Review*, 90, 1184–1208. [3]

Hanushek, E. A. and L. Woessmann (2008): "The Role of Cognitive Skills in Economic Development," *Journal of Economic Growth*, 46, 607–668. [3]

——— (2012): "Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation," *Journal of Economic Growth*, 17, 267–321. [2]

Heissel, J. and S. Norris (2017): "Rise and Shine: The Effect of School Start Times on Academic Performance from Childhood through Puberty," *Journal of Human Resources*. [6]

Higuchi, S., Y. Liu, T. Yuasa, A. Maeda, and Y. Motohashi (2000): "Diurnal Variation in the P300 Component of Human Cognitive Event Related Potential," *Chronobiology International*, 17, 669–678. [2]

Hubert, M., M. Dumont, and J. Paquet (1998): "Seasonal and Diurnal Patterns of Human Illumination Under Natural Conditions," *Chronobiology International*, 15, 59–70. [21]

Itzek-Greulich, H., C. Randler, and C. Vollmer (2016): "The interaction of chronotype and time of day in a science course: Adolescent evening types learn more and are more motivated in the afternoon," *Learning and Individual Differences*, 51, 189 – 198. [2]

Jacob, B. and J. Rothstein (2016): "The Measurement of Student Ability in Modern Assessment Systems," *Journal of Economic Perspectives*, 30, 85–108. [2]

Kantermann, T., M. Juda, M. Merrow, and T. Roenneberg (2007): "The Human Circadian Clock's Seasonal Adjustment Is Disrupted by Daylight Saving Time," *Current Biology*, 17, 1996 – 2000. [4], [21]

Kester, L. and P. A. Kirschner (2012): *Cognitive Tasks and Learning*, Boston, MA: Springer US, 619–622. [3]

Kiely, K. M. (2014): *Cognitive Function*, Dordrecht: Springer Netherlands, 974–978. [2]

Konings, J. and S. Vanormelingen (2015): "The Impact of Training on Productivity and Wages: Firm-Level Evidence," *The Review of Economics and Statistics*, 97, 485–497. [26]

Lazear, E. P. (2000): "Performance Pay and Productivity," *American Economic Review*, 90, 1346–1361. [26]

Lusher, L., V. Yasenov, and P. Luong (2019): "Does schedule irregularity affect productivity? Evidence from random assignment into college classes," *Labour Economics*, 60, 115 – 128. [6]

Murnane, R. J., J. B. Willett, and F. Levy (1995): "The Growing Importance of Cognitive Skills in Wage Determination," *The Review of Economics and Statistics*, 77, 251–266. [3]

Oster, E. (2019): "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 37, 187–204. [9], [20]

Park, J. (2020): "Hot Temperature , High Stakes Exams , and Avoidance Behavior : Evidence from New York City Public Schools," *Journal of Human Resources*. [2], [12]

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. [6]

Pope, N. G. (2016): "How the Time of Day Effects Productivity: Evidence from School Schedules," *Review of Economics and Statistics*. [6]

Roenneberg, T., T. Kuehnle, M. Juda, T. Kantermann, K. Allebrandt, M. Gordijn, and M. Merrow (2007a): "Epidemiology of the human circadian clock," *Sleep Medicine Reviews*, 11, 429–438. [2]

Roenneberg, T., T. Kuehnle, P. P. Pramstaller, J. Ricken, M. Havel, A. Guth, and M. Merrow (2004): "A marker for the end of adolescence," *Current Biology*, 14, 1038–1039. [3], [5], [23]

Roenneberg, T., C. Kumar, and M. Merrow (2007b): "The Human Circadian Clock Entrains to Sun Time," *Current biology : CB*, 17, R44–5. [21]

Roenneberg, T., A. Wirz-Justice, and M. Merrow (2003): "Life between clocks: daily temporal patterns of human chronotypes." *Journal of biological rhythms*, 18 1, 80–90. [23]

Schmidt, C., F. Colette, C. Cajochen, and P. Peigneux (2007): "A time to think: circadian rhythms in human cognition," *Cognitive Neuropsychology*, 24, 755–789. [2], [20]

Shawa, N., D. E. Rae, and L. C. Roden (2018): "Impact of seasons on an individual's chronotype: Current perspectives," *Nature and Science of Sleep*, 10, 345–354. [21]

Shin, J. (2018): "Sleep More, Study Less? The Impact of Delayed School Start Time on Sleep and Academic Performance," working paper. [6]

Taylor, E. S. and J. H. Tyler (2012): "The Effect of Evaluation on Teacher Performance," *The American Economic Review*, 102, 3628–3651. [25]

Toh, K. L., C. R. Jones, Y. He, E. J. Eide, W. A. Hinz, D. M. Virshup, L. J. Ptáček, and Y. H. Fu (2001): "An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome," *Science*, 291, 1040–1043. [23]

van der Vinne, V., G. Zerbini, A. Siersema, A. Pieper, M. Merrow, R. A. Hut, T. Roenneberg, and T. Kantermann (2015): "Timing of Examinations Affects School Performance Differently in Early and Late Chronotypes," *Journal of Biological Rhythms*, 30, 53–60. [2]

Vink, J. M., J. M. Vink, A. S. Groot, G. A. Kerkhof, and D. I. Boomsma (2001): "Genetic analysis of morningness and eveningness," *hronobiology International*, 18, 809–822. [23]

von Wachter, T. (2020): "The Persistent Effects of Initial Labor Market Conditions for Young Adults and Their Sources," *Journal of Economic Perspectives*, 34, 168–94. [2]

Williams, K. M. and T. M. Shapiro (2018): "Academic achievement across the day: Evidence from randomized class schedules," *Economics of Education Review*, 67, 158 – 170. [6]

Wright, K., A. McHill, B. Birks, B. Griffin, T. Rusterholz, and E. Chinoy (2013): "Entrainment of the Human Circadian Clock to the Natural Light-Dark Cycle," *Current biology : CB*, 23. [21]

Zerbini, G. and M. Merrow (2017): "Time to learn: How chronotype impacts education," *PsyCh Journal*. [2]

Zerbini, G., V. van der Vinne, L. K. M. Otto, T. Kantermann, P. W. Krijnen, T. Roenneberg, and M. Merrow (2017): "Lower school performance in late chronotypes: underlying factors and mechanisms," *Scientific Reports*. [5], [22]

A Appendix

A.1 Additional Balancing for Primary Results

Table A1: Balancing Tests, No Student FE

	Morning Exam			Early Afternoon Exam			Late Afternoon Exam		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Class Size	-0.000**	-0.000	-0.000	0.000***	0.000	0.000	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Days Since First Exam	0.004	0.003	0.002	-0.000	0.001	0.001	-0.003*	-0.004**	-0.003
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Days Between Exams	-0.000	0.000	0.000	0.002	0.000	0.000	-0.002	-0.001	-0.000
	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
Average GPA s-1		0.000	-0.002		-0.001	0.002		0.000	0.000
		(0.003)	(0.003)		(0.003)	(0.003)		(0.002)	(0.002)
Clusters:	7665	6233	5424	7665	6233	5424	7665	6233	5424
Observations:	503359	355636	315284	503359	355636	315284	503359	355636	315284
Adjusted R^2:	0.090	0.113	0.170	0.054	0.067	0.098	0.065	0.078	0.092
F-Statistics:	2.672	0.553	0.481	2.863	0.527	0.433	4.191	2.876	1.864
p-value:	0.046	0.697	0.750	0.035	0.716	0.785	0.006	0.022	0.114

Notes: Dependent variables: dummy for whether the exam was taken at 9am (Morning Exam), dummy if it was taken at 1.30pm (Early Afternoon Exam), dummy if it was taken at 4.30pm (Late Afternoon Exam). Standard errors are clustered by exam-year. Observations are at the student-exam-year level. In each specification, from Column (1)-(9), we control only for the conditions and year FE: number of exams per session fixed effects (FE), School FE and exam weight (in credits) FE. We report only the coefficients on the main variables of interest for the balancing tests. In Columns (1), (4), and (7), we report estimates for the full sample. In Columns (2), (5), and (8), we investigate balancing when including previous semester's GPA, which entails smaller samples than those used in estimation. Finally, in Columns (3), (6), and (9) we report results of the balancing tests when including in the regression a dummy for exams longer than 2.5 hours.*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

Table A2: Balancing Tests, Between students, No Student FE

	Morning Exam			Early Afternoon Exam			Late Afternoon Exam		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	-0.002	-0.003	-0.001	0.004	0.006*	0.004	-0.002	-0.003	-0.004
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)
Black	-0.000	0.001	0.001	-0.001	-0.001	-0.002	0.002	0.000	0.001
	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)
Asian	-0.003	-0.013	-0.013	-0.001	0.007	0.007	0.004	0.007	0.006
	(0.010)	(0.012)	(0.012)	(0.010)	(0.012)	(0.012)	(0.007)	(0.008)	(0.008)
White	-0.000	0.000	0.003	0.000	-0.002	-0.005	0.000	0.002	0.001
	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)
Average GPA s-1		0.001	-0.002		-0.001	0.002		0.001	0.000
		(0.003)	(0.003)		(0.003)	(0.003)		(0.002)	(0.002)
Clusters:	7665	6233	5424	7665	6233	5424	7665	6233	5424
Observations:	503359	355636	315284	503359	355636	315284	503359	355636	315284
Adjusted R^2:	0.087	0.112	0.170	0.050	0.066	0.098	0.063	0.075	0.090
F-Statistics:	0.138	0.340	0.531	0.810	0.980	0.936	0.543	0.737	0.784
p-value:	0.968	0.889	0.753	0.519	0.429	0.456	0.704	0.596	0.561

Notes: Dependent variable: dummy for whether the exam was taken at 9am (Morning Exam), dummy if it was taken at 1.30pm (Early Afternoon Exam), dummy if it was taken at 4.30pm (Late Afternoon Exam). Standard errors are clustered by exam-year. Observations are at the student-exam-year level. In each specification, from Column (1)-(9), we control for the conditions identifying the allocation rule, namely: number of exams per session fixed effects (FE), School FE, exam weight (in credits) FE, Student FE, and Year FE. In Columns (1), (4), and (7), we report estimates for the full sample. In Columns (2), (5), and (8), we investigate balancing when including previous semester's GPA, which entails smaller samples than those used in estimation. Finally, in Columns (3), (6), and (9) we report results of the balancing tests when including in the regression a dummy for exams longer than 2.5 hours. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

A.2 Additional Balancing for Heterogeneity Analysis

Table A3: Balancing Tests by Exam Time

	Morning Exams			Early Afternoon Exams			Late Afternoon Exams		
	(1) **All**	(2) **Fall**	(3) **Spring**	(4) **All**	(5) **Fall**	(6) **Spring**	(7) **All**	(8) **Fall**	(9) **Spring**
Panel A: Full-Sample									
Class Size	0.000	-0.000	0.000	0.000	0.000	0.000	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average GPA s-1	-0.002	-0.007	-0.016*	0.003	-0.011	0.020**	-0.002	0.017	-0.004
	(0.008)	(0.017)	(0.010)	(0.008)	(0.016)	(0.010)	(0.006)	(0.014)	(0.008)
Days Since First Exam	0.002	0.007	0.002	0.002	-0.003	0.002	-0.004**	-0.004	-0.004*
	(0.003)	(0.007)	(0.003)	(0.003)	(0.006)	(0.003)	(0.002)	(0.006)	(0.002)
Days Between Exams	0.001	-0.002	0.002	0.000	0.007	-0.002	-0.001	-0.005	-0.000
	(0.004)	(0.008)	(0.004)	(0.003)	(0.008)	(0.004)	(0.003)	(0.007)	(0.003)
Clusters:	6183	2123	4127	6183	2123	4127	6183	2123	4127
Observations:	353090	89828	258947	353090	89828	258947	353090	89828	258947
Adjusted R^2:	0.130	0.130	0.167	0.070	0.050	0.085	0.078	0.082	0.096
F-Statistics:	0.564	0.830	1.380	0.343	0.436	1.227	2.983	1.374	2.655
p-value:	0.689	0.506	0.238	0.849	0.783	0.297	0.018	0.240	0.031
Panel B: STEM									
Class Size	0.000	-0.000	0.000	-0.000	0.000	-0.000	-0.000	0.000	-0.000
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average GPA s-1	0.014	0.012	0.005	-0.003	-0.011	0.007	-0.011	-0.001	-0.012
	(0.010)	(0.021)	(0.012)	(0.009)	(0.019)	(0.012)	(0.008)	(0.016)	(0.010)
Days Since First Exam	-0.001	0.002	-0.001	0.005	-0.003	0.005	-0.004	0.001	-0.004
	(0.004)	(0.008)	(0.004)	(0.004)	(0.008)	(0.004)	(0.003)	(0.007)	(0.003)
Days Between Exams	0.002	-0.010	0.005	-0.003	0.018*	-0.007	0.001	-0.008	0.002
	(0.005)	(0.009)	(0.005)	(0.005)	(0.009)	(0.005)	(0.004)	(0.008)	(0.004)
Clusters:	2849	1038	1841	2849	1038	1841	2849	1038	1841
Observations:	180117	47174	131729	180117	47174	131729	180117	47174	131729
Adjusted R^2:	0.132	0.111	0.179	0.062	0.057	0.079	0.081	0.049	0.109
F-Statistics:	0.860	0.523	0.965	0.545	1.447	0.727	1.383	0.462	1.451
p-value:	0.487	0.719	0.426	0.702	0.216	0.573	0.237	0.763	0.215
Panel C: NON-STEM									
Class Size	-0.000	-0.000	0.000	0.000	0.000	0.000	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average GPA s-1	-0.016	-0.023	-0.027*	0.004	-0.022	0.016	0.011	0.045***	0.010
	(0.012)	(0.018)	(0.016)	(0.011)	(0.017)	(0.015)	(0.010)	(0.016)	(0.013)
Days Since First Exam	0.007*	0.014	0.006	-0.002	-0.003	-0.002	-0.005	-0.011	-0.004
	(0.004)	(0.011)	(0.004)	(0.004)	(0.011)	(0.004)	(0.003)	(0.010)	(0.003)
Days Between Exams	-0.001	0.007	-0.002	0.005	-0.005	0.005	-0.004	-0.002	-0.003
	(0.005)	(0.012)	(0.005)	(0.005)	(0.013)	(0.005)	(0.004)	(0.012)	(0.004)
Clusters:	3314	1062	2272	3314	1062	2272	3314	1062	2272
Observations:	170841	41501	125322	170841	41501	125322	170841	41501	125322
Adjusted R^2:	0.151	0.200	0.183	0.094	0.069	0.114	0.077	0.122	0.089
F-Statistics:	1.804	2.463	1.405	0.632	0.744	0.673	3.114	3.009	2.147
p-value:	0.125	0.044	0.230	0.640	0.562	0.611	0.014	0.018	0.073

Notes: Dependent variables: dummy for whether the exam was taken at 9am (Morning Exam), dummy if it was taken at 1.30pm (Early Afternoon Exam), dummy if it was taken at 4.30pm (Late Afternoon Exam). Standard errors are clustered by exam-year. Observations are at the student-exam-year level. In each specification, from Column (1)-(9), we control only for the conditions and year FE: number of exams per session fixed effects (FE), School FE and exam weight (in credits) FE. We report only the coefficients on the main variables of interest for the balancing tests. In Panel A, we report estimates of the full sample. In Panel B, we report estimates for STEM schools only. In Panel C, we report estimates for Non-STEM schools only. Furthermore, in Column (1), (4) and (7), we report the results on the pooled sessions. In Column (2), (5) and (8), we report the results for the Fall exam session only. In Column (3), (6) and (9), we report the results for the Spring exam session only. $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

Table A4: Balancing Tests by Exam Time, with Exam Duration

	Morning Exams			Early Afternoon Exams			Late Afternoon Exams		
	(1) **All**	(2) **Fall**	(3) **Spring**	(4) **All**	(5) **Fall**	(6) **Spring**	(7) **All**	(8) **Fall**	(9) **Spring**
Panel A: Full-Sample									
Class Size	-0.000	-0.000	0.000	0.000	0.000	0.000	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average GPA s-1	0.000	0.002	-0.001	-0.001	-0.006	0.001	0.000	0.003	0.000
	(0.003)	(0.005)	(0.003)	(0.003)	(0.004)	(0.003)	(0.002)	(0.004)	(0.002)
Days Since First Exam	0.003	0.007	0.002	0.001	-0.002	0.001	-0.004**	-0.005	-0.003
	(0.003)	(0.007)	(0.003)	(0.003)	(0.006)	(0.003)	(0.002)	(0.006)	(0.002)
Days Between Exams	0.000	-0.004	0.002	0.000	0.006	-0.001	-0.001	-0.002	-0.000
	(0.004)	(0.008)	(0.004)	(0.003)	(0.007)	(0.004)	(0.003)	(0.007)	(0.003)
Clusters:	6233	2153	4174	6233	2153	4174	6233	2153	4174
Observations:	355636	93915	261721	355636	93915	261721	355636	93915	261721
Adjusted R^2:	0.113	0.093	0.152	0.067	0.056	0.087	0.078	0.088	0.093
F-Statistics:	0.553	0.697	0.587	0.527	0.865	0.267	2.876	1.043	2.069
p-value:	0.697	0.594	0.672	0.716	0.484	0.899	0.022	0.384	0.082
Panel B: STEM									
Class Size	0.000	-0.000	0.000	0.000	0.000	0.000	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average GPA s-1	0.005*	0.004	0.003	-0.004	-0.002	-0.003	-0.000	-0.002	-0.000
	(0.003)	(0.006)	(0.003)	(0.003)	(0.005)	(0.003)	(0.002)	(0.004)	(0.003)
Days Since First Exam	-0.001	-0.002	-0.001	0.005	0.000	0.005	-0.003	0.002	-0.004
	(0.004)	(0.008)	(0.004)	(0.004)	(0.008)	(0.004)	(0.003)	(0.007)	(0.003)
Days Between Exams	0.002	-0.006	0.005	-0.003	0.012	-0.007	0.001	-0.007	0.002
	(0.005)	(0.009)	(0.006)	(0.005)	(0.009)	(0.005)	(0.004)	(0.008)	(0.004)
Clusters:	2865	1050	1855	2865	1050	1855	2865	1050	1855
Observations:	181456	48422	133034	181456	48422	133034	181456	48422	133034
Adjusted R^2:	0.114	0.091	0.150	0.060	0.053	0.079	0.084	0.086	0.097
F-Statistics:	0.759	0.676	0.537	1.228	0.989	0.802	0.781	0.344	0.880
p-value:	0.552	0.609	0.709	0.297	0.413	0.524	0.537	0.849	0.475
Panel C: NON-STEM									
Class Size	-0.000	-0.000	0.000	0.000	0.000	0.000	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average GPA s-1	-0.005	0.001	-0.007	0.005	-0.011**	0.007	-0.001	0.010	0.010
	(0.005)	(0.007)	(0.005)	(0.005)	(0.005)	(0.006)	(0.004)	(0.007)	(0.013)
Days Since First Exam	0.008**	0.020*	0.007	-0.004	-0.005	-0.004	-0.005	-0.015	-0.004
	(0.004)	(0.011)	(0.004)	(0.004)	(0.010)	(0.004)	(0.003)	(0.009)	(0.003)
Days Between Exams	-0.003	-0.004	-0.003	0.005	-0.000	0.005	-0.002	0.004	-0.003
	(0.005)	(0.013)	(0.005)	(0.005)	(0.011)	(0.005)	(0.004)	(0.011)	(0.004)
Clusters:	3368	1103	2319	3368	1103	2319	3368	1103	2272
Observations:	174180	45493	128687	174180	45493	128687	174180	45493	125322
Adjusted R^2:	0.122	0.115	0.165	0.082	0.069	0.104	0.075	0.093	0.089
F-Statistics:	1.678	1.687	1.212	0.901	1.109	0.861	2.556	1.778	2.147
p-value:	0.152	0.151	0.303	0.462	0.351	0.486	0.037	0.131	0.073

Notes: Dependent variables: dummy for whether the exam was taken at 9am (Morning Exam), dummy if it was taken at 1.30pm (Early Afternoon Exam), dummy if it was taken at 4.30pm (Late Afternoon Exam). Standard errors are clustered by exam-year. Observations are at the student-exam-year level. In each specification, from Column (1)-(9), we control only for the conditions and year FE: number of exams per session fixed effects (FE), School FE and exam weight (in credits) FE. We report only the coefficients on the main variables of interest for the balancing tests. In Panel A, we report estimates of the full sample. In Panel B, we report estimates for STEM schools only. In Panel C, we report estimates for Non-STEM schools only. Furthermore, in Column (1), (4) and (7), we report the results on the pooled sessions. In Column (2), (5) and (8), we report the results for the Fall exam session only. In Column (3), (6) and (9), we report the results for the Spring exam session only. $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

Table A5: Balancing Tests by Exam Time, No Student FE

	Morning Exams			Early Afternoon Exams			Late Afternoon Exams		
	(1) **All**	(2) **Fall**	(3) **Spring**	(4) **All**	(5) **Fall**	(6) **Spring**	(7) **All**	(8) **Fall**	(9) **Spring**
Panel A: Full-Sample									
Class Size	-0.000	-0.000	-0.000	0.000	0.000	0.000	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average GPA s-1	-0.003	-0.010	-0.014	0.004	-0.008	0.016	-0.001	0.018	-0.002
	(0.008)	(0.018)	(0.010)	(0.008)	(0.017)	(0.010)	(0.007)	(0.015)	(0.009)
Days Since First Exam	0.001	0.008	0.001	0.002	-0.004	0.003	-0.003	-0.004	-0.003
	(0.003)	(0.007)	(0.003)	(0.003)	(0.007)	(0.003)	(0.002)	(0.006)	(0.002)
Days Between Exams	0.002	-0.001	0.003	-0.000	0.006	-0.002	-0.001	-0.005	-0.001
	(0.004)	(0.008)	(0.004)	(0.004)	(0.008)	(0.004)	(0.003)	(0.007)	(0.003)
Clusters:	5379	1913	3532	5379	1913	3532	5379	1913	3532
Observations:	311674	81529	225471	311674	81529	225471	311674	81529	225471
Adjusted R^2:	0.180	0.151	0.242	0.100	0.050	0.132	0.086	0.092	0.112
F-Statistics:	0.444	0.828	0.866	0.453	0.211	0.971	1.876	1.249	1.428
p-value:	0.777	0.507	0.483	0.770	0.932	0.422	0.112	0.288	0.222
Panel B: STEM									
Class Size	0.000	-0.000	0.000	-0.000	0.000	-0.000	-0.000	0.000	-0.000
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average GPA s-1	0.011	0.005	0.004	0.002	-0.004	0.008	-0.013	-0.001	-0.012
	(0.010)	(0.022)	(0.012)	(0.009)	(0.022)	(0.012)	(0.008)	(0.015)	(0.011)
Days Since First Exam	-0.002	0.004	-0.002	0.004	-0.005	0.005	-0.002	0.001	-0.003
	(0.004)	(0.009)	(0.004)	(0.004)	(0.009)	(0.004)	(0.003)	(0.008)	(0.003)
Days Between Exams	0.003	-0.010	0.005	-0.004	0.019*	-0.008	0.001	-0.008	0.002
	(0.005)	(0.010)	(0.006)	(0.005)	(0.010)	(0.006)	(0.004)	(0.009)	(0.005)
Clusters:	2451	937	1544	2451	937	1544	2451	937	1544
Observations:	152762	41556	109647	152762	41556	109647	152762	41556	109647
Adjusted R^2:	0.158	0.137	0.217	0.086	0.049	0.117	0.092	0.069	0.131
F-Statistics:	0.785	0.314	0.815	0.471	1.107	0.875	0.955	0.448	0.705
p-value:	0.535	0.869	0.515	0.757	0.352	0.478	0.431	0.774	0.589
Panel C: NON-STEM									
Class Size	-0.000**	-0.000	-0.000	0.000**	0.000	0.000**	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average GPA s-1	-0.016	-0.030	-0.017	0.000	-0.023	0.005	0.015	0.053***	0.012
	(0.012)	(0.019)	(0.018)	(0.012)	(0.017)	(0.017)	(0.011)	(0.016)	(0.015)
Days Since First Exam	0.005	0.013	0.004	-0.001	-0.002	-0.001	-0.004	-0.011	-0.003
	(0.004)	(0.011)	(0.004)	(0.004)	(0.011)	(0.004)	(0.003)	(0.010)	(0.004)
Days Between Exams	-0.000	0.010	-0.001	0.004	-0.008	0.004	-0.003	-0.002	-0.004
	(0.005)	(0.013)	(0.005)	(0.005)	(0.014)	(0.005)	(0.004)	(0.012)	(0.004)
Clusters:	2914	953	1980	2914	953	1980	2914	953	1980
Observations:	156837	38820	113959	156837	38820	113959	156837	38820	113959
Adjusted R^2:	0.220	0.216	0.291	0.126	0.074	0.167	0.088	0.134	0.104
F-Statistics:	2.856	2.664	1.586	1.448	0.757	1.237	2.379	3.528	1.470
p-value:	0.022	0.031	0.175	0.216	0.553	0.293	0.050	0.007	0.209

Notes: Dependent variables: dummy for whether the exam was taken at 9am (Morning Exam), dummy if it was taken at 1.30pm (Early Afternoon Exam), dummy if it was taken at 4.30pm (Late Afternoon Exam). Standard errors are clustered by exam-year. Observations are at the student-exam-year level. In each specification, from Column (1)-(9), we control only for the conditions and year FE: number of exams per session fixed effects (FE), School FE and exam weight (in credits) FE. We report only the coefficients on the main variables of interest for the balancing tests. In Panel A, we report estimates of the full sample. In Panel B, we report estimates for STEM schools only. In Panel C, we report estimates for Non-STEM schools only. Furthermore, in Column (1), (4) and (7), we report the results on the pooled sessions. In Column (2), (5) and (8), we report the results for the Fall exam session only. In Column (3), (6) and (9), we report the results for the Spring exam session only. $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A6: Balancing Tests by Exam Time, Between Comparison, No Student FE

	Morning Exams			Early Afternoon Exams			Late Afternoon Exams		
	(1) All	(2) Fall	(3) Spring	(4) All	(5) Fall	(6) Spring	(7) All	(8) Fall	(9) Spring
Panel A: Full-Sample									
Average GPA s-1	0.001	0.003	-0.001	-0.001	-0.006	0.000	0.001	0.003	0.000
	(0.003)	(0.005)	(0.003)	(0.003)	(0.004)	(0.003)	(0.002)	(0.004)	(0.002)
Female	-0.003	-0.005	-0.002	0.006*	0.004	0.006	-0.003	0.000	-0.003
	(0.003)	(0.005)	(0.004)	(0.003)	(0.005)	(0.004)	(0.002)	(0.004)	(0.003)
Black	0.001	0.008	-0.002	-0.001	0.001	-0.001	0.000	-0.009	0.003
	(0.004)	(0.008)	(0.005)	(0.004)	(0.007)	(0.005)	(0.004)	(0.007)	(0.004)
Asian	-0.013	-0.040**	0.005	0.007	0.020	-0.008	0.007	0.020	0.003
	(0.012)	(0.016)	(0.014)	(0.012)	(0.014)	(0.014)	(0.008)	(0.014)	(0.009)
White	0.000	0.007	-0.004	-0.002	0.000	-0.001	0.002	-0.007	0.005*
	(0.004)	(0.007)	(0.004)	(0.003)	(0.006)	(0.004)	(0.003)	(0.006)	(0.003)
Clusters:	6233	2153	4174	6233	2153	4174	6233	2153	4174
Observations:	355636	93915	261721	355636	93915	261721	355636	93915	261721
Adjusted R^2:	0.112	0.090	0.152	0.066	0.055	0.086	0.075	0.085	0.090
F-Statistics:	0.340	2.300	0.414	0.980	1.019	0.923	0.737	0.944	1.129
p-value:	0.889	0.043	0.840	0.429	0.405	0.465	0.596	0.451	0.343
Panel B: STEM									
Average GPA s-1	0.005*	0.006	0.003	-0.005	-0.003	-0.003	-0.000	-0.003	-0.000
	(0.003)	(0.006)	(0.003)	(0.003)	(0.006)	(0.003)	(0.002)	(0.004)	(0.003)
Female	-0.007*	-0.013	-0.004	0.007*	0.008	0.006	0.000	0.004	-0.002
	(0.004)	(0.008)	(0.005)	(0.004)	(0.008)	(0.005)	(0.003)	(0.006)	(0.004)
Black	0.006	0.017	0.000	-0.005	-0.017	0.002	-0.001	-0.000	-0.002
	(0.006)	(0.011)	(0.007)	(0.006)	(0.011)	(0.007)	(0.005)	(0.009)	(0.005)
Asian	0.007	-0.027	0.026**	-0.011	0.018	-0.026**	0.004	0.009	0.000
	(0.012)	(0.023)	(0.013)	(0.011)	(0.021)	(0.012)	(0.009)	(0.019)	(0.010)
White	-0.002	0.005	-0.007	-0.004	-0.005	-0.001	0.006	0.000	0.008*
	(0.006)	(0.011)	(0.007)	(0.005)	(0.010)	(0.006)	(0.004)	(0.009)	(0.005)
Clusters:	2865	1050	1855	2865	1050	1855	2865	1050	1855
Observations:	181456	48422	133034	181456	48422	133034	181456	48422	133034
Adjusted R^2:	0.114	0.088	0.149	0.059	0.050	0.077	0.082	0.083	0.095
F-Statistics:	1.578	1.595	1.722	1.618	0.990	1.567	0.588	0.230	0.980
p-value:	0.163	0.159	0.126	0.152	0.422	0.166	0.710	0.950	0.428
Panel C: NON-STEM									
Average GPA s-1	-0.005	0.001	-0.007	0.006	-0.011**	0.008	-0.000	0.010	-0.000
	(0.005)	(0.007)	(0.006)	(0.005)	(0.006)	(0.006)	(0.004)	(0.007)	(0.004)
Female	0.002	0.003	0.000	0.004	0.000	0.004	-0.006	-0.003	-0.004
	(0.005)	(0.007)	(0.006)	(0.005)	(0.006)	(0.006)	(0.004)	(0.006)	(0.004)
Black	-0.007	-0.003	-0.008	0.005	0.014	0.002	0.002	-0.011	0.007
	(0.006)	(0.010)	(0.006)	(0.005)	(0.009)	(0.006)	(0.005)	(0.009)	(0.005)
Asian	-0.036*	-0.050**	-0.019	0.025	0.012	0.013	0.011	0.038**	0.006
	(0.021)	(0.019)	(0.027)	(0.022)	(0.018)	(0.027)	(0.013)	(0.019)	(0.015)
White	0.003	0.008	-0.001	-0.001	0.005	-0.002	-0.002	-0.013*	0.002
	(0.004)	(0.008)	(0.005)	(0.004)	(0.007)	(0.005)	(0.004)	(0.008)	(0.004)
Clusters:	3368	1103	2319	3368	1103	2319	3368	1103	2319
Observations:	174180	45493	128687	174180	45493	128687	174180	45493	128687
Adjusted R^2:	0.118	0.104	0.162	0.080	0.068	0.102	0.071	0.086	0.089
F-Statistics:	1.586	1.774	0.770	0.820	1.519	0.564	0.952	1.674	0.673
p-value:	0.161	0.115	0.571	0.535	0.181	0.728	0.446	0.138	0.644

Notes: Dependent variables: dummy for whether the exam was taken at 9am (Morning Exam), dummy if it was taken at 1.30pm (Early Afternoon Exam), dummy if it was taken at 4.30pm (Late Afternoon Exam). Standard errors are clustered by exam-year. Observations are at the student-exam-year level. In each specification, from Column (1)-(9), we control only for the conditions and year FE: number of exams per session fixed effects (FE), School FE and exam weight (in credits) FE. We report only the coefficients on the main variables of interest for the balancing tests. In Panel A, we report estimates of the full sample. In Panel B, we report estimates for STEM schools only. In Panel C, we report estimates for Non-STEM schools only. Furthermore, in Column (1), (4) and (7), we report the results on the pooled sessions. In Column (2), (5) and (8), we report the results for the Fall exam session only. In Column (3), (6) and (9), we report the results for the Spring exam session only. $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

A.3 Identification: Details

For simplicity, and slight abuse of notation, we re-write Equation (6) as follows:

$$Y(h,\boldsymbol{\theta}) = T(h,\boldsymbol{\theta}) + \epsilon(h,\boldsymbol{\theta}) \tag{A1}$$

where $Y(h,\boldsymbol{\theta})$ is student performance, which is a function of hour of the day h and a vector of assessment and individual's characteristics $\boldsymbol{\theta}$. Furthermore, $T(h,\boldsymbol{\theta})$ is the individual time-of-day effect and $\epsilon(h,\boldsymbol{\theta})$ is a standard random unobserved error with mean zero.

In our dataset, we observe each individual performing multiple and different assessments. For the moment, suppose we observe two performances. Ideally, we wish to identify:

$$E[\Delta_Y(h,\boldsymbol{\theta})] = E[Y(h,\boldsymbol{\theta}) - Y(0,\boldsymbol{\theta})] \tag{A2}$$

which is the average treatment effect (ATE) of performing an assessment $\boldsymbol{\theta}$ at hour h on performance with respect to a baseline hour (9am in our case). Since each individual is observed performing assessments with different characteristics $\boldsymbol{\theta}$, we need to make some identifying assumptions.

Assumption 1. *Random assignment:*

i. $E[Y(h,\boldsymbol{\theta})] = E[Y(h,\boldsymbol{\theta}')] = E[Y(h)]$,

ii. $E[Y(0,\boldsymbol{\theta})] = E[Y(0,\boldsymbol{\theta}')] = E[Y(0)]$.

Assumption 1 says that the potential outcomes of the individuals are independent of the vector of characteristics $\boldsymbol{\theta}$. This means that, on average, individuals need to perform similar assessments at different time-of-day. Under assumption 1, the population model simplifies to:

$$Y(h) = T(h) + \epsilon(h), \tag{A3}$$

where $T(h) \perp \epsilon(h)$ by design.

Assumption 2. *Time-of-day:*

$$T(h) = B(h) + E(h).$$

Assumption 2 says that time-of-day is made of two components: a "biological" and an "economic" component. $B(h)$ captures the direct effect of the circadian rhythm on individual's performance. That is, a brain is more productive at performing a cognitive assessment at specific hours of the day. $E(h)$ captures the individual's behavior to increase her performance (e.g. effort, work strategy, food intake). Under assumption 2, the population model becomes:

$$Y(h) = W(h) + B(h) + \epsilon(h). \tag{A4}$$

We take expectations over the difference in performance between two cognitive assessments using the following mild assumption:

Assumption 3. *Random shocks:*

$$E[\epsilon(h)] = E[\epsilon(0)] = 0.$$

Hence we have:

$$\begin{aligned} E[\Delta_Y(h)] &= E[Y(h) - Y(0)] \\ &= E[B(h) - B(0)] + E[E(h) - E(0)] \\ &= E[\Delta_B(h)] + E[\Delta_E(h)] \end{aligned} \tag{A5}$$

The third line of Equation (A5) shows that the effect that we wish to identify is affected by two components that are functions of the same factor: we call $E[\Delta_B(h)]$ the expected brain reaction effect to the hour of the day and $E[\Delta_E(h)]$ the expected individual reaction effect to the hour of the day.

A.4 Additional Robustness Check

Table A7: Coefficient Stability Tests

	(1) **Conditions**	(2) **Student FE, Covariates**	(3) **Bounding Set**
Early Afternoon [0,1]	0.050***	0.068***	[0.036, 0.068]
	(0.016)	(0.015)	
Late Afternoon [0,1]	0.042**	0.036**	[0.018, 0.036]
	(0.017)	(0.015)	
Clusters:	7665	7626	7626
Observations:	503359	500959	500959
Adjusted R^2:	0.022	0.462	0.462

Notes: The table presents the validation results for the analysis of the effect of time-of-day on performance. Dependent variable: standardized final exam mark. Key variables: dummies for whether the exam was taken at 1.30pm (Early Afternoon Exam) or at 4.30pm (Late Afternoon Exam), where the 9am (Morning Exam) is the excluded category. Column (1) presents estimates when controlling for the conditional independence assumption variables, while in Panel Column (2) we include in addition other covariates as well as student FE. Finally, Column (3) reports estimates of the bounding set assuming a Rmax equals to one. Standard errors are clustered by exam-year.*p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01.

A.5 Additional Heterogeneous Effects

Table A8: Effect of Time of Day on Student Performance: Heterogeneity Analysis

	Full-Sample			STEM			NON-STEM		
	(1) **Pooled Sessions**	(2) **Fall Session**	(3) **Spring Session**	(4) **Pooled Sessions**	(5) **Fall Session**	(6) **Spring Session**	(7) **Pooled Sessions**	(8) **Fall Session**	(9) **Spring Session**
Panel A: Benchmark									
Early Afternoon [0,1]	0.068***	0.094***	0.044**	0.081***	0.139***	0.031	0.037**	0.016	0.037*
	(0.015)	(0.024)	(0.019)	(0.022)	(0.035)	(0.028)	(0.017)	(0.028)	(0.022)
Late Afternoon [0,1]	0.036**	0.076***	-0.009	0.045*	0.090***	-0.034	0.011	0.029	-0.007
	(0.015)	(0.022)	(0.021)	(0.025)	(0.035)	(0.034)	(0.017)	(0.027)	(0.022)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.036	0.440	0.009	0.148	0.191	0.047	0.153	0.671	0.053
Clusters:	7626	3434	4287	3465	1639	1885	4145	1780	2388
Observations:	500959	219208	274704	250363	109496	138440	248208	107718	134324
Adjusted R^2:	0.462	0.457	0.499	0.482	0.483	0.516	0.450	0.444	0.495
Panel B: Age $\leq$ 20									
Early Afternoon [0,1]	0.070***	0.121***	0.035	0.093***	0.211***	0.027	0.033	0.012	0.030
	(0.020)	(0.036)	(0.025)	(0.032)	(0.059)	(0.037)	(0.022)	(0.035)	(0.028)
Late Afternoon [0,1]	0.039*	0.101***	-0.015	0.063*	0.137**	-0.032	0.012	0.044	-0.015
	(0.021)	(0.035)	(0.026)	(0.036)	(0.060)	(0.045)	(0.022)	(0.038)	(0.028)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.124	0.532	0.048	0.383	0.202	0.166	0.377	0.366	0.116
Clusters:	5286	2378	3016	2277	1021	1317	2999	1346	1686
Observations:	272438	111136	157577	127836	49540	76961	143468	60582	79551
Adjusted R^2:	0.476	0.470	0.525	0.499	0.501	0.545	0.465	0.458	0.516
Panel C: Age > 20									
Early Afternoon [0,1]	0.061***	0.063***	0.043**	0.066***	0.064**	0.035	0.029	0.025	0.032
	(0.015)	(0.024)	(0.020)	(0.019)	(0.029)	(0.027)	(0.022)	(0.035)	(0.026)
Late Afternoon [0,1]	0.017	0.033	-0.016	0.013	0.029	-0.038	-0.003	0.017	-0.016
	(0.017)	(0.023)	(0.024)	(0.025)	(0.032)	(0.036)	(0.020)	(0.030)	(0.025)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.012	0.259	0.008	0.038	0.315	0.033	0.136	0.809	0.044
Clusters:	7448	3337	4177	3424	1618	1860	3996	1689	2292
Observations:	227383	103769	114270	122299	58656	60677	103550	43815	52375
Adjusted R^2:	0.484	0.489	0.517	0.506	0.511	0.536	0.471	0.479	0.509

Notes: Dependent variables: dummy for whether the exam was taken at 9am (Morning Exam), dummy if it was taken at 1.30pm (Early Afternoon Exam), dummy if it was taken at 4.30pm (Late Afternoon Exam). Standard errors are clustered by exam-year. Observations are at the student-exam-year level. In each specification, from Column (1)-(9), we control only for the conditions and year FE: number of exams per session fixed effects (FE), School FE and exam weight (in credits) FE. We report only the coefficients on the main variables of interest for the balancing tests. In Panel A, we report estimates of the full sample. In Panel B, we report estimates students with age less or equal than 20 only. In Panel C, we report estimates for students with aged greater than 20. Furthermore, in Column (1), (4) and (7), we report the results on the pooled sessions. In Column (2), (5) and (8), we report the results for the Fall exam session only. In Column (3), (6) and (9), we report the results for the Spring exam session only. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

Table A9: Effect of Time of Day on Student Performance: By Season and School

	Full-Sample			STEM			NON-STEM		
	(1) **Pooled Sessions**	(2) **Fall Session**	(3) **Spring Session**	(4) **Pooled Sessions**	(5) **Fall Session**	(6) **Spring Session**	(7) **Pooled Sessions**	(8) **Fall Session**	(9) **Spring Session**
Panel A: Benchmark									
Early Afternoon [0,1]	0.068***	0.094***	0.044**	0.081***	0.139***	0.031	0.037**	0.016	0.037*
	(0.015)	(0.024)	(0.019)	(0.022)	(0.035)	(0.028)	(0.017)	(0.028)	(0.022)
Late Afternoon [0,1]	0.036**	0.076***	-0.009	0.045*	0.090***	-0.034	0.011	0.029	-0.007
	(0.015)	(0.022)	(0.021)	(0.025)	(0.035)	(0.034)	(0.017)	(0.027)	(0.022)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.036	0.440	0.009	0.148	0.191	0.047	0.153	0.671	0.053
Clusters:	7626	3434	4289	3465	1639	1887	4145	1780	2388
Observations:	500959	219208	274715	250363	109496	138451	248208	107718	134324
Adjusted R^2:	0.462	0.457	0.499	0.482	0.483	0.516	0.450	0.444	0.495
Panel B: Duration									
Early Afternoon [0,1]	0.061***	0.086***	0.049**	0.068***	0.126***	0.048	0.039**	0.027	0.031
	(0.017)	(0.025)	(0.022)	(0.025)	(0.036)	(0.033)	(0.019)	(0.029)	(0.024)
Late Afternoon [0,1]	0.026	0.058**	-0.013	0.019	0.041	-0.034	0.015	0.046	-0.020
	(0.017)	(0.023)	(0.023)	(0.028)	(0.035)	(0.039)	(0.019)	(0.029)	(0.025)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.032	0.247	0.006	0.077	0.020	0.032	0.208	0.515	0.034
Clusters:	6475	2891	3672	2900	1372	1580	3556	1502	2084
Observations:	432185	184391	239673	207810	88606	115769	222120	93883	121997
Adjusted R^2:	0.471	0.473	0.506	0.493	0.504	0.524	0.457	0.454	0.499
Panel C: No First Year UG									
Early Afternoon [0,1]	0.062***	0.068***	0.047**	0.063***	0.080***	0.038	0.042**	0.022	0.047**
	(0.014)	(0.021)	(0.020)	(0.020)	(0.028)	(0.029)	(0.018)	(0.030)	(0.023)
Late Afternoon [0,1]	0.022	0.043**	-0.012	0.005	0.033	-0.024	0.025	0.043*	-0.005
	(0.016)	(0.020)	(0.024)	(0.025)	(0.031)	(0.036)	(0.017)	(0.024)	(0.024)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.014	0.268	0.010	0.020	0.159	0.090	0.353	0.485	0.027
Clusters:	7373	3326	4126	3390	1614	1829	3960	1687	2275
Observations:	312103	140127	163590	157129	71419	83012	153090	67126	79100
Adjusted R^2:	0.464	0.468	0.491	0.485	0.493	0.512	0.452	0.454	0.484
Panel D: Exam FE									
Early Afternoon [0,1]	0.032**	0.066***	-0.009	0.055***	0.111***	-0.003	-0.011	0.001	-0.023
	(0.012)	(0.020)	(0.015)	(0.019)	(0.029)	(0.024)	(0.013)	(0.025)	(0.017)
Late Afternoon [0,1]	0.019	0.049**	-0.019	0.023	0.053*	-0.027	0.012	0.032	-0.012
	(0.014)	(0.019)	(0.019)	(0.023)	(0.029)	(0.031)	(0.014)	(0.023)	(0.019)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.313	0.346	0.567	0.111	0.034	0.376	0.102	0.202	0.534
Clusters:	7578	3407	4257	3458	1635	1879	4102	1758	2365
Observations:	500906	219176	274678	250356	109492	138442	248160	107693	134296
Adjusted R^2:	0.531	0.534	0.559	0.551	0.562	0.574	0.512	0.505	0.552
Panel E: Day FE									
Early Afternoon [0,1]	0.070***	0.092***	0.047**	0.087***	0.140***	0.048*	0.040**	0.015	0.044**
	(0.014)	(0.023)	(0.018)	(0.022)	(0.033)	(0.028)	(0.016)	(0.027)	(0.019)
Late Afternoon [0,1]	0.034**	0.080***	-0.019	0.047*	0.097***	-0.036	0.009	0.028	-0.014
	(0.015)	(0.022)	(0.020)	(0.025)	(0.035)	(0.033)	(0.017)	(0.026)	(0.021)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.021	0.593	0.001	0.108	0.248	0.011	0.071	0.625	0.008
Clusters:	7626	3434	4289	3465	1639	1887	4145	1780	2388
Observations:	500959	219208	274715	250363	109496	138451	248208	107718	134324
Adjusted R^2:	0.466	0.461	0.504	0.490	0.488	0.526	0.456	0.450	0.502
Panel F: Room FE									
Early Afternoon [0,1]	0.068***	0.084***	0.041**	0.099***	0.126***	0.045	0.025	0.026	0.024
	(0.015)	(0.023)	(0.019)	(0.022)	(0.033)	(0.028)	(0.018)	(0.028)	(0.023)
Late Afternoon [0,1]	0.034**	0.073***	-0.003	0.051**	0.075**	-0.024	0.008	0.029	-0.004
	(0.015)	(0.021)	(0.021)	(0.025)	(0.033)	(0.034)	(0.017)	(0.025)	(0.023)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.033	0.649	0.034	0.061	0.173	0.035	0.357	0.903	0.242
Clusters:	7613	3432	4279	3462	1639	1885	4132	1778	2375
Observations:	500920	219186	274686	250345	109482	138435	248174	107702	134291
Adjusted R^2:	0.473	0.469	0.513	0.500	0.502	0.536	0.460	0.456	0.513

Notes: Dependent variable: standardized final exam mark. Key variables: dummies for whether the exam was taken at 1.30pm (Early Afternoon Exam) or at 4.30pm (Late Afternoon Exam), where the 9am (Morning Exam) is the excluded category. Standard errors are clustered by exam-year. Observations are at the student-exam year level. In each Column (1)-(9) we use the preferred specification which controls for student FE and all covariates. In Columns (1)-(3) we report estimates of the full sample, while in Columns (4)-(6) and Columns (7)-(9) we report results for STEM and NON-STEM schools, separately. Specifically, in Columns (1), (4), and (7) are reported the results for the pooled sample. In Columns (2), (5), and (8) we report the results for the Fall exam session only. In Columns (3), (6), and (9) we report the results for the Spring exam session only. Panel (A) presents the benchmark results. Panel (B) report the results when including in the regression a dummy for exams longer than 2.5 hours. Finally, In Panel (C), we exclude the sample of first year UG students. In Panels, (D), (E) and (F) we include, respectively, exam, day, and room FE. Below the test of equality of Early and Late afternoon exams is reported the p-value of the test statistic. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A10: Effect of Time of Day on Student Performance: By Season and Age

	Full-Sample			STEM			NON-STEM		
	(1) **Pooled Sessions**	(2) **Fall Session**	(3) **Spring Session**	(4) **Pooled Sessions**	(5) **Fall Session**	(6) **Spring Session**	(7) **Pooled Sessions**	(8) **Fall Session**	(9) **Spring Session**
Panel A: Benchmark									
Early Afternoon [0,1]	0.068***	0.094***	0.044**	0.070***	0.121***	0.035	0.061***	0.063***	0.043**
	(0.015)	(0.024)	(0.019)	(0.020)	(0.036)	(0.025)	(0.015)	(0.024)	(0.020)
Late Afternoon [0,1]	0.036**	0.076***	-0.009	0.039*	0.101***	-0.015	0.017	0.033	-0.016
	(0.015)	(0.022)	(0.021)	(0.021)	(0.035)	(0.026)	(0.017)	(0.023)	(0.024)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.036	0.440	0.009	0.124	0.532	0.048	0.012	0.259	0.008
Clusters:	7626	3434	4289	5286	2378	3016	7448	3337	4179
Observations:	500959	219208	274715	272438	111136	157577	227383	103769	114281
Adjusted R^2:	0.462	0.457	0.499	0.476	0.470	0.525	0.484	0.489	0.517
Panel B: Duration									
Early Afternoon [0,1]	0.061***	0.086***	0.049**	0.061***	0.118***	0.048*	0.056***	0.062**	0.038*
	(0.017)	(0.025)	(0.022)	(0.023)	(0.037)	(0.028)	(0.017)	(0.026)	(0.022)
Late Afternoon [0,1]	0.026	0.058**	-0.013	0.031	0.089**	-0.004	0.006	0.029	-0.033
	(0.017)	(0.023)	(0.023)	(0.023)	(0.035)	(0.029)	(0.019)	(0.025)	(0.027)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.032	0.247	0.006	0.145	0.361	0.055	0.009	0.248	0.005
Clusters:	6475	2891	3672	4461	1978	2564	6325	2801	3585
Observations:	432185	184391	239673	231804	89635	137255	198801	90012	98851
Adjusted R^2:	0.471	0.473	0.506	0.488	0.496	0.535	0.493	0.501	0.523
Panel C: No First Year UG									
Early Afternoon [0,1]	0.062***	0.068***	0.047**	0.038*	0.058*	0.019	0.064***	0.060**	0.047**
	(0.014)	(0.021)	(0.020)	(0.022)	(0.032)	(0.035)	(0.015)	(0.024)	(0.021)
Late Afternoon [0,1]	0.022	0.043**	-0.012	0.008	0.068**	-0.036	0.015	0.022	-0.016
	(0.016)	(0.020)	(0.024)	(0.024)	(0.031)	(0.034)	(0.017)	(0.023)	(0.025)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.014	0.268	0.010	0.200	0.770	0.106	0.006	0.140	0.008
Clusters:	7373	3326	4126	4844	2160	2727	7083	3169	3949
Observations:	312103	140127	163590	108594	44329	59689	202348	91510	101194
Adjusted R^2:	0.464	0.468	0.491	0.496	0.504	0.534	0.471	0.483	0.499
Panel D: Exam FE									
Early Afternoon [0,1]	0.032**	0.066***	-0.009	0.042**	0.104***	-0.015	0.023*	0.041**	-0.005
	(0.012)	(0.020)	(0.015)	(0.017)	(0.029)	(0.020)	(0.013)	(0.018)	(0.019)
Late Afternoon [0,1]	0.019	0.049**	-0.019	0.040**	0.100***	-0.004	-0.005	0.012	-0.045**
	(0.014)	(0.019)	(0.019)	(0.018)	(0.027)	(0.023)	(0.015)	(0.020)	(0.023)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.313	0.346	0.567	0.901	0.861	0.598	0.049	0.160	0.052
Clusters:	7578	3407	4257	5141	2295	2927	7377	3302	4133
Observations:	500906	219176	274678	272287	111046	157474	227306	103729	114229
Adjusted R^2:	0.531	0.534	0.559	0.545	0.554	0.582	0.549	0.556	0.576
Panel E: Day FE									
Early Afternoon [0,1]	0.070***	0.092***	0.047**	0.071***	0.118***	0.031	0.064***	0.064***	0.047**
	(0.014)	(0.023)	(0.018)	(0.019)	(0.035)	(0.023)	(0.015)	(0.022)	(0.020)
Late Afternoon [0,1]	0.034**	0.080***	-0.019	0.033*	0.109***	-0.034	0.015	0.033	-0.021
	(0.015)	(0.022)	(0.020)	(0.020)	(0.034)	(0.025)	(0.017)	(0.022)	(0.024)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.021	0.593	0.001	0.053	0.780	0.007	0.005	0.217	0.002
Clusters:	7626	3434	4289	5286	2378	3016	7448	3337	4179
Observations:	500959	219208	274715	272438	111136	157577	227383	103769	114281
Adjusted R^2:	0.466	0.461	0.504	0.484	0.477	0.533	0.488	0.493	0.521
Panel F: Room FE									
Early Afternoon [0,1]	0.068***	0.084***	0.041**	0.070***	0.095***	0.032	0.062***	0.063***	0.038**
	(0.015)	(0.023)	(0.019)	(0.021)	(0.033)	(0.025)	(0.015)	(0.024)	(0.019)
Late Afternoon [0,1]	0.034**	0.073***	-0.003	0.040*	0.095***	-0.003	0.016	0.033	-0.024
	(0.015)	(0.021)	(0.021)	(0.021)	(0.033)	(0.027)	(0.017)	(0.022)	(0.023)
Test of equality of Early and Late Afternoon exams in linear regressions above:									
$Pr > F$:	0.033	0.649	0.034	0.142	0.984	0.185	0.009	0.241	0.006
Clusters:	7613	3432	4279	5261	2376	2990	7430	3335	4165
Observations:	500920	219186	274686	272388	111117	157532	227340	103750	114241
Adjusted R^2:	0.473	0.469	0.513	0.491	0.489	0.540	0.496	0.499	0.534

Notes: Dependent variable: standardized final exam mark. Key variables: dummies for whether the exam was taken at 1.30pm (Early Afternoon Exam) or at 4.30pm (Late Afternoon Exam), where the 9am (Morning Exam) is the excluded category. Standard errors are clustered by exam-year. Observations are at the student-exam year level. In each Column (1)-(9) we use the preferred specification which controls for student FE and all covariates. In Columns (1)-(3) we report estimates of the full sample, while in Columns (4)-(6) and Columns (7)-(9) we report results for students age less or equal than 20 and age greater than, separately. Specifically, in Columns (1), (4), and (7) are reported the results for the pooled sample. In Columns (2), (5), and (8) we report the results for the Fall exam session only. In Columns (3), (6), and (9) we report the results for the Spring exam session only. Panel (A) presents the benchmark results. Panel (B) report the results when including in the regression a dummy for exams longer than 2.5 hours. In Panel (C), we exclude the sample of first year UG students. Finally, in Panels (D) (E) and (F) we include, respectively, exam, day, and room FE. Below the test of equality of Early and Late afternoon exams is reported the p-value of the test statistic. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.