NBER WORKING PAPER SERIES

THE EFFECT OF MULTITASKING ON EDUCATIONAL OUTCOMES AND ACADEMIC DISHONESTY

Victor Lavy

Working Paper 31699
http://www.nber.org/papers/w31699

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2023, revised January 2024

I thank the staff at the Ministry of Education for making an effort to recover the schedule of bagrut exams for 2000-2005 and for providing the data for this study. Special thanks to Assaf Yancu and Omry Yoresh for excellent research assistance and Hadar Avivi, who worked on this project in the early stages a few years ago when the treatment measures and methodology differed. I benefitted from comments from Hadar Avivi, Michael Amior, Netanel Ben Porath, Adi Finkelstein, Yoav Goldstein, Helena Skyt Nielsen, Nitzan Machlis, Genia Rachkovski, Assaf Yancu, Omri Yoresh, participants at the 2022 Oslo Education Conference, the CESifo 2023 Economics of Education conference, and seminars at Hebrew University, and the University of Warwick. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

The Effect of Multitasking on Educational Outcomes and Academic Dishonesty
Victor Lavy
NBER Working Paper No. 31699
September 2023, revised January 2024
JEL No. I20,J0

ABSTRACT

School authorities, universities, and employers often schedule multiple tests on the same day or week, causing overlapping exam preparation and a dense testing schedule. This multitask learning can be intense, under pressure, and challenge the student's mental and physical perseverance. As a result, it can compromise performance relative to a more 'relaxed' schedule. This paper examines the consequences of multitasking for test scores and cheating in exams and its implications for the ability and gender cognitive gap. The empirical context is high-stakes exit exams in Israel, done at the end of high school. I leverage the empirical setting on two natural experiments to estimate the causal effect of this multitasking learning. The first exploits random variation in the number of weekly tests—the second hinges on days with multiple exams versus days with a single exam. The results show several important regularities. First, the number of exams in a day or a week harms test performance. Second, these effects are evidenced for high- and low-ability students, both boys and girls. They are much more extensive for immigrants than natives. Third, the harm of such multitasking is larger in tests later in the schedule, daily or weekly. Fourth, these effects are larger in tests of STEM subjects. Fifth, dense exam schedules increase the likelihood of students behaving dishonestly in exams.

Victor Lavy
Department of Economics
University of Warwick
Coventry, CV4 7AL
United Kingdom
and Hebrew University of Jerusalem
and also NBER
v.lavy@warwick.ac.uk

1. Introduction

Across the globe, standardized testing is now a major component of the education system. The results of these tests are used to inform all manner of high-stakes decisions. In education, these tests capture short-run growth in student skills that will eventually translate into success and well-being later in life, particularly in the labor market. Yet, there is a growing body of work that suggests that seemingly inconsequential design choices about the structure of the test can influence student performance for reasons that are disconnected from traits the tests were initially intended to measure (e.g., Brown et al., 2023; Beltran et al., 2022; Reyes, 2023).

This paper focuses on an important test feature that has received comparatively little attention: the timing of when tests occur within an examination period, i.e., the density of the testing schedule. A dense exam schedule leads to multitasking learning and preparation, which may negatively affect performance. When performance in these exams is high stakes, organizing them within a short period can have long-term implications for human capital and labor market outcomes. An example is the high school exit exams in many countries and the *international baccalaureate program*, often used to rank university applicants.[1] A similar setup exists in higher education, where in some countries (e.g., the U.K.), universities schedule the exams for all courses taught in the first and second terms at the end of the academic year, causing a dense exam schedule in a short period.[2] Such a heavy testing load can lower test performance and average GPA, bearing consequences in the labor market, for example, internship selection and hiring decisions. Multi-tests in a dense schedule are also used in professional licensing, for example, in the bar exams for lawyers.[3] Firms often use multitasking testing as employers increasingly screen job applicants using recruitment tests as an objective method of assessing the suitability of potential candidates.[4]

Given the significant weight placed on test scores in these examples and other contexts of dense testing schedules, the first part of the paper examines whether systematic factors derived from the exams'

[1] An example of such a dense exam schedule can be viewed on the website of the *international baccalaureate program*: https://www.ibo.org/. The May 2023 examination schedule often includes 4-6 daily exams, over six hours.
[2] In contrast, Israeli universities recognized long ago the potentially harmful aspect of the resulting multitasking learning for these exams and offer, at a cost, two dates for each exam (organized in two dense exam periods), allowing students to choose one or the other to have fewer exams in each period.
[3] In Israel, the bar exams are scheduled in one day, from morning to evening. They include three tests, lasting 40 minutes, 160 minutes, and 100 minutes, respectively.
[4] An assessment center puts a group of candidates through individual and group tasks and simulations based on the available position requirements to assess their competencies and how well they fit the role and company culture. It uses personality tests, aptitude and skills tests, job-related simulations and role-play scenarios, timed tasks, and interviews which last a full day or multiple days. Dense exam schedules may lead to allocative inefficiency. In education, a less optimal match between students and study programs. In the labor market, inefficient allocation of workers across occupations and a less productive workforce overall.

structure affect test scores in high-stakes exams. It also examines whether these factors have distributional consequences by individuals' background and gender. Addressing these questions poses data, identification, and estimation challenges. This may explain why the limited research on multitasking has been conducted in a lab setting. As such, this study is one of the first based in a real-world environment with high-stakes implications.

The second part of the paper studies the consequences of multi-tests on students' ethical behavior. The importance of results in these tests may induce dishonesty, for example, cheating in exams. Students who cheat in an exam without being caught and punished are likely to benefit at the expense of their peers concerning valued outcomes such as grades and recommendations, eventually affecting admission to prestigious or highly demanded study programs. These potential negative impacts deserve attention because the problem of academic cheating has become endemic.[5] Multitasking load can affect dishonest behavior by causing anxiety, a sense of pressure, and fear, all shown to be related to cheating in exams (DeVoss and Rosati, 2002; Sterngold, 2004). If students become over-extended, they may use inappropriate resources and strategies to manage (e.g., Sterngold, 2004). Increased stress levels can lead students to be more prone to dishonest behaviors.

This paper examines the potential effect of multitasking tests on cognitive performance during high-stakes exit exams in Israeli high schools. The exams are known as the Bagrut and are a critical component of Israel's college admissions system, acting as a gatekeeper for the most selective schools and elite universities, similar to the role played by high-stakes exams in other countries, such as the SATs in the USA or A-levels in England. The Bagrut test scores are our measures of cognitive outcomes. Cheating in an exam, which leads to disqualification, is the measure of students' ethics. This is an almost ideal context for several reasons. First, we can access a complete record of all Bagrut exams taken between 2000 and 2005 and the date they were given, providing us with a large sample of high-stakes exams to observe test outcomes and the testing schedule. Second, since students generally take between 8–12 separate exams, there is significant variation in the number of tests in a week and a day across the same student's, enabling us to estimate models with student-fixed effects. Our design is also aided by the fact that retaking Bagrut exams is costly. Since most exams are given at the end of twelfth grade, and Israelis begin a period of compulsory military service (three years for boys and two years for girls) after high school graduation, retaking the exam is only possible for most students several years after the relevant coursework and would require many additional testing days. This makes it more difficult to retake the exam, and students rarely retake any section. As such, a

[5] A growing number of high school and college students worldwide opt for easy ways to get good grades, thus breaking the ethical rules of their educational institutions. Although the honor codes are getting stricter, the number of students putting their academic integrity at risk remains exceptionally high. According to McCabe (2010), 95% of the students polled (graduates and undergraduates) admitted to cheating in one form or another.

negative Bagrut outcome during a student's first attempt is likely to affect a student's post-secondary academic options significantly.

I leverage the empirical setting on two parts of a natural experiment to estimate the causal effect of this multitasking learning. The first is a conditional random variation in the number of weekly exams. Identification hinges on a model with a student and an exam fixed effects. A student-fixed effect model is feasible because a student experiences variation in the number of weekly exams: one, two, three, or more. Exploiting this student's level variation guarantees against omitted variables or selection biases related to a student's characteristics, such as ability, study program, and parental background. It also controls for differences between schools, classes, and localities. A remaining concern would be the correlation between student-level variation in treatment exposure and exam-specific student ability (or other unobserved exam-specific factors affecting student performance). This threat is less plausible in the Bagrut context for two reasons. First, the tests are entirely without any opportunity for rescheduling; the dates of national Bagrut exams are determined by the Ministry of Education years before the actual exam; and all students must take their exam at their local high school. Second, we have a pre-determined measure of a student's ability in each exam taken, and we use it as a control in the regression analysis.

The second part of the same natural experiment is the random number of daily exams. We use the same student and exam fixed effect model for identifying the causal effect of such multitasking. We can include an exam fixed effect in this quasi-experimental model because a particular exam is sometimes scheduled alone and sometimes with more exams on the same day. Secondly, as discussed above, there is no sorting or endogeneity in exam taking in the sample used.

Combining data for six cohorts, 2000-2005, generates considerable variation across and within students in the number of exams a day and a week. This data structure allows for a compelling empirical setup for estimating the effect of multitasking learning and testing. However, it involves endogenous sorting and selection since students with high (low) ability study more (less) subjects and, therefore, have a higher (lower) likelihood of having days or weeks with more than one exam. Nevertheless, using these two 'within' student variations allows for neutralizing the concern about the endogeneity of the number of exams of each student and the number of weekly and daily exams. Therefore, the natural experimental 'within' pupil variation in the two treatments of interest is not endogenous and can be viewed as 'quasi-random,' permitting causal identification.

The results presented in this paper show three important regularities regarding the effect of multitasking load on test scores. First, the number of exams in a day or a week harms test performance. The causal effect on test scores is opposite to the positive OLS association with scores because of the positive

selection of tests in a day or a week. Controlling for students and tests' characteristics in the same subject/exam reduces this bias but does not eliminate it. However, the positive effect is reversed once I control for pupil-fixed effects. Two tests in a week and each on a different day have a relatively small adverse impact relative to one test in a week. However, if the two tests are scheduled on the same day, each score declines by 0.02 SD. Three or four weekly tests reduce test scores by 0.06 SD relative to one in a week. However, if two or three tests are in the same subjects and scheduled on the same day, the score declines only by 0.03 and 0.04, respectively. This negative effect increases with the test order. For example, the test score of the third daily exam is lower by 0.122 SD. These effect sizes are economically meaningful. For example, the effect of three weekly exams is equivalent to the impact of lowering weekly instruction time in each of these subjects by almost 1.5 hours throughout the schooling year (Lavy 2015, 2021). It is nearly equal to the effect of incentivizing teachers in a generous pay-for-performance scheme (Lavy 2009). This large adverse impact affects more than 15 percent of the students who have at least one week with three plus tests. It is much larger than the effect of extremely low air quality on exam day (Ebenstein, Lavy, and Roth, 2016).

The second regularity is that the adverse effect of the number of weekly exams is very similar for students with high and lower abilities. For example, having three exams in a week lowers the test scores of low-ability students by 0.0576 SD and high-ability students by 0.0662 SD. Since the incidence of such treatment is similar in both groups, this treatment effect does not affect the average test score gap between high and low-ability students. However, the adverse impact of exams multitasking in one day, especially three in a day, is concentrated among high-ability (above median) students. The estimated effect of three daily exams is -0.0123 SD among low-ability students and -0.0496 among high-ability students. This result implies that scheduling multiple daily exams has distributional consequences, reducing the average test score gap between high and low-ability students. It also means costly inefficiency in matching students to university fields of study that require higher ability.

A third regularity is that the number of *weekly* exams on different days impacts boys and girls similarly. Still, boys are more negatively affected by three or more exams packed in *one day* or on *different days.* For example, the effect of three plus exams when three are on different days is -0.06 for boys and -0.04 for girls. The impact of three plus exams, three on the same day, is -0.04 for boys and -0.02 or girls. Though these gender differences are not statistically significant, they suggest a pattern consistent with the evidence of girls' advantages in some relevant non-cognitive skills and the 'folk' belief that women are more efficient in multitasking. The section that discusses this gender difference in detail provides references from the psychology literature about the mechanisms that can explain these results.

To assess the validity of the identification and estimation design, I also estimate placebo tests. For each student, I randomly assigned one of the treatments he experienced to each exam. The placebo estimates are all small and insignificant, practically zero estimates. This evidence supports the causal interpretation of the regularities presented above.

The evidence presented in this paper also shows that multitasking load affects students' dishonest behavior. In particular, it induces behavioral distortions such as cheating in an exam. Two interesting regularities are evident from the results. Multiple weekly and daily exams significantly increase the likelihood of being caught cheating, leading to an exam disqualification. The effect size is large, from 30 to 50 percent. Multiple daily exams increase the possibility of an exam disqualification more than multiple weekly exams. Significant differences in these treatment effects are evident by the student's ability and gender.

The rest of the paper is organized as follows. Section 2 presents a literature review focusing on studies in economics and psychology. The second part of this section presents Israel's high school exit exams system. Section 3 describes the data, and Section 4 presents the empirical strategy. Finally, sections 5-7 present the results on the effect of multitasking on test scores, and section 8 the effect on students' dishonest behavior. Section 9 concludes.

2. Background

A. Design and Structure of the Tests and Student Performance

Multitasking and Performance: Multitasking is defined as a condition in which cognitive processes of two or more tasks overlap in time (Koch et al., 2017). It is a broad construct that can be conceptualized and assessed in various ways (e.g., Künzell et al., 2018). Two types of multitasking are prevalent: one in which two or more tasks are carried out simultaneously, termed "concurrent multitasking," and a second in which two or more tasks are carried out sequentially, termed 'serial or switching multitasking' (Burgess 2015). More specifically, in serial/switching multitasking, participants alternate between tasks that vary in priority, difficulty, and duration. For example, testing in multiple exams during a day could be an example of concurrent multitasking. This includes the possibility that a student studies for each test sequentially in blocks or for half-days.

An episode of multiple exams during a week is perhaps closely related to sequential multitasking, where students can shift "back and forth" between preparation for different forthcoming exams. Both examples meet several of the criteria that Burgess (2000, 2015) suggests for such multitasking: they

comprise multiple, discrete tasks that are interleaved but carried out one at a time, and there is no direct signal indicating when it is time to return to an already running task, the tasks differ in terms of characteristics, priority, and length of time, and there is no immediate feedback.

Individual, ability, and gender-related differences in multitasking can emanate from at least two sources: domain-specific differences in performance in a given task in isolation and the ability to coordinate and monitor component tasks independent of task-specific skills and experiences. These latter higher-order multitasking functions are closely related to executive functioning (Watson & Strayer, 2010; Mäntylä, 2013): the ability to maintain and update multiple task goals and coordinate spatial relations.

Recent studies examine the cost of media multitasking when students use two or more media sources concurrently. These studies find that inherent mental habits of media multitasking—dividing attention, switching attention, and maintaining multiple trains of thought— have significant implications and consequences for students' academic performance. The research indicates that media multitasking interferes with attention and working memory, negatively affecting GPA, test performance, recall, reading comprehension, note-taking, self-regulation, and efficiency. See May and Elder (2018) for a review of these studies.

The evidence about gender differences in multitasking is inconsistent. Several studies do not find gender differences in multitasking abilities, while others report a male advantage (Mäntylä, 2013, Persson et al., 2013). The third line of findings supports the notion that females excel in multitasking. For example, Stoet et al. (2013) found a female advantage in an extensive study on gender differences in multitasking. Some of these inconsistent findings stem from differences in the methods and tasks used in these studies. Second, these studies are based on recruited student populations, making it difficult to generalize their findings (Stoet et al., 2013).

Cognitive and Non-Cognitive Correlates of Differences in Multitasking: Several lines of research in economics and psychology produced evidence relevant to understanding and interpreting the findings presented in this paper. I briefly summarize them by topic.

***Time to Learn*:** Do students benefit from more time for learning before exams? The answer to this question is of first-order importance for understanding our evidence about the effect of multiple exams in a day or a week. The theoretical hypothesis that learning takes time has been considered in at least three related fields — educational (Bloom, 1974), cognitive (Ericsson and Charness, 1994), and neural (McClelland, McNaughton, and O'Reilly, 1995). Overall, the conclusion is that learning occurs within an extended period during which information is encoded, rehearsed, elaborated, and consolidated. In economics and psychology, past research has supported this hypothesis and shown that school length, instruction time, and

preparation time affect students' test achievements. However, while most of the previous literature has primarily been devoted to identifying the effect of increased time spent in class or school[6], there is little evidence of the impact of the preparation time that high-school students have before an exam. Taraban et al. (2001) found a positive relationship between preparation time and exam scores. Goulas and Megalokonomou (2020) found a negative relationship between STEM scores and the number of days since the first exam.

***Time Pressure*:** Sustaining effort and performance during the five weeks of the exams in 12th grade in Israel, particularly in high density in a day or a week, can involve time pressure and mental and physiological stress. Pressure harms decision quality (Busemeyer and Diederich 2002) because it worsens reasoning processes and enhances individuals' tendency to ignore important information and rely on heuristics (Gigerenzer et al. 1999; Rieskamp and Hoffrage 2008). Despite the importance of time pressure in many economic decisions, economists only recently started paying attention to this issue (Kocher et al., 2003; Sutter et al., 2003; Kocher and Sutter, 2006; Bollard et al., 2007; Leder et al., 2013). The findings suggest that pressure changes individual attitudes toward risk and changes individual behavior by raising physiological stress, increasing risk-taking, and inhibiting strategic thinking.

***Exam Length*:** More insight into the potential effect of schedule density of exams is gained from recent studies that examine the effect of exam length on test scores. Particularly relevant for us is whether performance in a test deteriorates with the order of questions. Borghans and Schils (2012) document that performance substantially drops during the test, while the correlation between this drop and the test scores is small. Balart et al. (2018) report that females experience a lower decline in performance during the test regardless of the topic being assessed. Since effort declined during the test equally for men and women, they concluded that their findings are not driven by a difference in effort but rather by the efficacy of the mental processes that translate effort inputs into a correct answer. This explanation is consistent with gender differences in boredom: males have been found to experience higher levels of boredom when performing activities that have a long duration, and individuals who experience boredom have impaired performance on various tasks (for recent work on this topic (Eastwood et al., 2012).

B. *High-Stakes Examinations in Israel's High School*

The Israeli high school exit exam system represents a novel opportunity to examine these questions.[7] Most matriculation exams in Israel are held at the end of high school, with some done at the end of 9th and 10th

[6] See, For example, Wößmann, 2003; Pischke, 2007; Marcotte and Hemelt, 2008; Lavy, 2015 and 2020; and Rivkin and Schiman, 2015.

[7] Many countries administer a similar setup of exams at the end of high school to assess the three years of secondary school and as a doorway to higher education. The format and structure of high school exit exams vary significantly

grade.[8] In the Israeli context and some other countries with end-of-high school exit exams, students take multiple high-stakes exams in a relatively short period, often on the same day. To be awarded a matriculation certificate, students should pass exams in seven mandatory subjects and two or more elective subjects. When a student passes an exam, she is awarded credit units that increase with the exam's proficiency level. In particular, 21 credit units grant a student a certificate. Nevertheless, a student may choose advanced programs that grant him more credit units in most subjects. Hence, students in different programs face different exams in the same subject.

The exams occur about eight weeks each year, during May and June, and are administered towards the end of 12th grade. Most students have their exams scheduled for 5-6 weeks, held at the student's high school, without the opportunity for rescheduling or changing the testing site. Since students take multiple exams, there is significant variation in the number of exams in a day or a week, implying differences in the number of preparation days before exams, enabling us to estimate models with student-fixed effects. It is important to note that all exams scheduled on the same day are always in the same subject, each covering different parts of the material. For example, advanced math may involve separate algebra, geometry, and trigonometry exams. The English advanced study program includes an exam based on reading a text, a second exam based on listening to a reader, and a third oral exam. In the physics advanced study program, students have four exams: two written (mechanics and electricity) and two lab exams (regular and research labs). Each of the four exams is of two hours duration. These exams have a unique code and can be held on the same day or on different days. Table A1 in the appendix presents the exam schedule of four students in 2000. These examples demonstrate the variation in multi-week and multi-day exams, which we exploit for estimating the effect of multitasking. For example, student 1 has nine exams. In week 1 of the exam period, he has three exams. On Monday, two exams in chemistry, and on Thursday, an exam in Hebrew. Student 4 has 12 exams. In week 5, there are three exams in physics, all on Wednesday, and in week 7, two math

across countries. These exams have different names in different countries: *Baccalauréat* (France), *A-levels* (UK), *Abitur* (Germany), *Selectividad* (Spain), *bagrut* (Israel), and the *Matura* in many European countries. Other countries have similar high school exit exams, including Albania, Austria, Bosnia and Herzegovina, Bulgaria, Croatia, the Czech Republic, Hungary, Italy, Kosovo, Liechtenstein, Macedonia, Montenegro, Poland, Serbia, Slovakia, Slovenia, Switzerland, Ukraine. In the US, the *SAT* (scholastic assessment test) and its close competitor, the *ACT* (American College Testing), are standardized college admissions tests rather than all-around final-year exam qualifications – the US High School diploma tends to fulfill the latter function. These exams are taken mainly at the end of high school. The number of subjects covered and the number of exams varies considerably across nations; in some, they include three exams, and in others, the program extends to 8-10 exams. The exams usually occur during the end month of the last year of high school, in most countries, from mid-May to mid or end of June.

[8] Schools decide independently how many exams and subjects are scheduled for the end of 10th and 11th grade versus the end of 12th grade. Typically, one exam is taken at the end of the 10th and two at the end of the 11th grade. Students choose their study program at the beginning or end of 10th grade, at least two years before the schedule of *bagrut* exams at the end of 12th grade is announced.

exams, both on Wednesday.

Our design is also aided by the fact that retaking exams is costly. Since most exams are given at the end of the last year of high school, and Israelis begin a period of compulsory military service (a minimum of three years for boys and two years for girls) after high-school graduation, for most students, retaking the exam is only possible for several years after the relevant coursework and would require many additional days of studying and testing. A negative outcome during a student's first attempt is likely to affect a student's post-secondary academic options significantly.

The certificate is a pre-requisite for studying at universities and most academic and teachers' colleges.[9] Students are admitted to university programs based on their average scores and a separate psychometric examination (the Israeli equivalent of the SAT). Each university ranks applicants according to the same formula, thus producing an index based on a weighted average of the student's average scores on all their exams and the psychometric examination. This ranking determines students' eligibility for university admission and the major they can choose within the university. Therefore, multitasking situations resulting from exam schedules can affect students' university schooling by affecting their probability of passing the exams and their average scores. In summary, the mechanisms by which multitasking in exam testing can affect long-term economic outcomes are through their effect on (a) the probability of pursuing higher education, (b) the type of higher education pursued, and (c) the quality of higher education institutions ultimately attended.

3. Data

The data set I use in this study is generated by combining three primary data sources: Israeli test score data from 2000- 2005 (data for other years are not available for this study), information on the exact dates of all exams, and students' demographic and socio-economic characteristics. The Israeli Ministry of Education provides each test taker's exam and demographic data. These files also contain rich demographic information on the student and the student's family, such as parental education level, number of siblings, country of origin, and ethnicity.

[9] The post-secondary education system in Israel consists of nine research universities that grant PhDs (as well as other degrees), approximately 50 academic colleges that offer undergraduate degrees (of which a minimal subset offer masters degrees), and a set of non-university institutions of higher education that confer teaching and vocational certificates. In addition, practical engineering colleges run two-year programs awarding degrees (or certificates) in electronics, computers, and industrial production. An additional two years of study in an engineering school is required to complete a BSc in engineering.

The raw data comprises 6.8 million exams and 668,800 students from 1,028 schools in Israel between 2000 and 2005. However, I restrict the study sample to only students meeting four main requirements. First, to focus on a homogenous curriculum, we limit our attention to students from Jewish public schools.[10] Second, we focus on tests taken at the end of 12th grade. Third, we include students enrolled in a full *bagrut* study program to have the minimum set of compulsory and elective exams.[11] The sample is also restricted to students without missing values in the variables used in this study.[12] Fourth, in the primary sample, I exclude students who, at the end of high school, are retaking at least one exam that they have taken in earlier grades in high school or who are skipping an exam that they later take in a second chance test offered in the summer following 12th grade. These sample restrictions are intended to avoid measurement errors in weekly/daily tests. These steps reduced the estimation sample to about 1.2 million exam papers and 163,000 students from 529 schools. However, I present results in the paper when these sample restrictions are not imposed. The results from these samples are identical to those obtained from the primary sample. More discussion on this comparison appears in a later section.

A second data set that we use includes information for only 2002-2002 and has the same format and variables as the 2000-2005 data file. However, it also includes an exam-level indicator of disqualified test scores because of cheating. Such disqualification is based on inappropriate behavior during the exam, observed by the exam supervisors or by the assessor/marker of the exam. The Ministry of Education takes one or more disciplinary or criminal measures against an examinee found to have violated the exams' purity or one or more of the instructions of the exam procedure. First is the test's disqualification after the process of suspecting it. Secondly, according to the severity of the case, the matter will be transferred to the Disciplinary Committee or the police, as the Exam Purity Committee deems appropriate. The committee can take one or more of these steps regarding an examinee whose exam was disqualified: (i) Suspension of the examinee from the examinations in this subject for a fixed period of up to 3 years, (ii) Suspension of the examinee from the examinations in all subjects remaining for him for a period of up to 3 years (iii) The disqualification of the exams in which the examinee has been tested to date (iv) Filing a police complaint. The Ministry of Education can decide on additional disciplinary measures against the school if the exam purity procedures have been violated. For example, disqualifying all scores in a test because of irregularities or deficiencies caused by the teaching staff. Another common sanction is the temporary cancellation of the

[10] Public education in Israel includes the Jewish and Arab systems. These two have separate schools, curricula, and *bagrut* exams. Hebrew is the language used in Jewish schools, and Arabic is the languish of instruction in Arab schools. The study curriculum shares core subjects but has some unique subjects and bagurt exams.

[11] Since 21 *bagrut* credits are the minimum for obtaining a *bagrut* diploma, we chose the attainment of a 15 credits as the threshold for inclusion in the study sample. We note, however, that the results presented in the paper are not sensitive to increasing or lowering this threshold.

[12] For demographic variables, particularly parental years of schooling and number of siblings, missing values were replaced by the school by cohort average.

school's recognition for giving annual grades in the year in which there was a serious violation of the purity of the exams.

The 2000-2005 and 2000-2002 cannot be merged because each includes a different fictitious individual I.D. I use the first data set as the primary data set for estimating the effect of multitasking on test scores and the second for estimating the effect on cheating behavior. I obtained similar results for the effect of multitasking on test scores when using the 2000-2002 data set, which is presented in the appendix table.

The summary statistics for our sample are presented in Table 1 in six panels; Colum 1 shows means for the total sample and columns 2 and 3 for samples of students above and below median ability, which I measure based on average test scores in the school matriculation exams. Columns 4 and 5 present the information separately for girls and boys. The sample of girls is larger than that of boys due to the higher dropout rate of boys during high school.

Panel A reports the weekly sample means based on student-level data. The mean number of tests in a week is 1.73, and the number of weeks with more than one test is 1.8, relative to a mean of a five-week exam period. Three out of four students have at least one week with two tests, and over a third (36 percent) have a week with three tests. It is higher for high-ability students (39 percent) and marginally higher for boys.

Panel B reports the daily sample means. The mean number of tests per day is 1.3, and the number of days with more than one test is 1.06.[13] These statistics are slightly higher among high-ability students and boys. About 63 percent of the students have at least one day with two tests, and a tenth has a day with three. Again, the latter two means are higher among high-ability students (65 percent versus 61 percent among low-ability students) and boys (13 percent versus 8 percent among girls).

Panel C presents descriptive statistics for disqualified exams, which account for almost a third of a percent of all exams. This rate is higher in exams in weeks with more than one exam (0.4 percent) and higher in exams in days with more than one exam (0.55 percent). Over two percent of all students have at least one disqualified exam because of cheating. Cheating and disqualification of exams are much more prevalent among low-ability students. For example, disqualification in an exam with multiple daily exams is almost one percent among low-ability students and less than a fifth percent among high-ability students. Over three percent of low-ability students have at least one disqualified exam due to cheating, while the respective mean for high-ability students is less than a percent.

[13] The number of daily tests ranges from one to three. The data for estimation consists of 6 students with four tests daily. We omitted these students from the study sample.

The cheating in an exam is likely much higher than the rate of students caught in this act. Using the event of being disqualified due to cheating as an outcome provides an unbiased estimate of the effect of multitasking as long as the extent of cheating and disqualification are highly correlated. Another related issue is whether being caught is a good proxy for behavior. It will be more of a concern if high-ability students are caught cheating more easily, and it is correlated with multitasking. I will address this issue when discussing the results.

In panel D, I report the student-level means of national exam test scores and the school scores for each exam. The school scores in each subject are based on exams similar to the national ones and are administered in the weeks before the external exams start. Test scores are normalized to mean zero and unit standard deviation. This standardization is done for each exam and every cohort separately. The mean difference in the national exam test scores between the above and below median samples is 0.83 SD. The respective difference in the school exam scores is higher, 1.12 SD. Girls' mean national score is higher than that of boys by 0.13 SD. The gender gap in the school scores is much larger, 0.23 SD.

Panel E presents the means of students' characteristics. The mother and father's mean year of schooling is about 13.5, and the mean number of children is 2.6. These means are not significantly different by the students' gender but very different by students' mean ability. Parental schooling in the above median ability group is higher by about 1.2 years. However, family size is the same in the two ability sub-samples.

We also compute the number of days before an exam as follows: (a) if it is the second or later exam, we calculate the days between the current test and the previous exam; (b) in a day with multiple tests, we assign all observations the same number of days before; for the first exam in the testing period, we count the number of days from Friday of the previous week because Friday is a weekend day. We exclude from the study sample the last exam of every student because it is unique relative to all other exams as it does not have another exam that follows it. Results are not sensitive to this exclusion, nor how we compute the number of days before an exam. Panel E presents test characteristics accordingly. The mean number of days before a test is 8.2. The mean test duration is 135 minutes. These statistics are not different in the samples defined by ability and by gender. The sample consists of 1,095,853 exam papers and about 164,000 students from the same 529 schools.

4. Empirical Strategy – Exams' Multitasking and Test Scores

In the first section of our analysis, we examine the partial correlation between the two treatments of interest, the number of weekly and daily exams, and test scores in the sample of exam-level data. For identification,

we rely on the data's panel structure and the Bagrut exam's repeated nature. Since each student takes multiple exams, student-fixed effects can be added as controls in the regression. These fixed effects absorb unobserved variation in individuals. This is the key identification assumption because we have shown in the data section that the two treatment variables correlate with invariant student characteristics, such as ability. This point is discussed further later in this section.

Formally, a model with linear treatment effects is as follows:

$$(1) \quad Y_{ijt} = \alpha + \epsilon_w (N.\ Exams\ in\ a\ Week)_{ijt} + \epsilon_d (N.\ Exams\ in\ a\ Day)_{ijt} + \beta X_i + \Phi (School\_FE)_i + \Omega (Cohort\_FE)_j + \mu (Test\_Charac)_j + \Upsilon (School\_Score)_{ijt} + \phi (Student\ FE)_i + \pi (\text{Exam}\_FE)_j + \pi (\text{Exam\_Date}\_FE)_j + \varepsilon_{ijt}$$

where Y_{ist} is the test score (z-score, mean zero, and unit standard deviation) of student *i* in exam *j* at year *t*. $N.\ Exams\ in\ a\ Week_{ijt}$ *and* $N.\ Exams\ in\ a\ Day_{ijt}$ are the two treatment variables of interest, the number of exams in a week and the number of exams in a day. Since the equation (1) model imposes linear treatment effects, I would instead present estimates only from a model that relaxes this assumption by including dummy variables for each treatment level. The weekly exams have three treatment levels represented by dummy indicators: two, three, and four plus weekly exams. The daily exam treatment includes two treatment indicators, one for two exams and one for three. The omitted group is for one weekly exam. X_i is a vector of observable individual characteristics potentially related to testing outcomes, including parental education in years, a dummy for sex, and the number of siblings. These controls are dropped when a pupil fix effect is included in the regression. $Test\_Charac_j$ is a vector of test characteristics that includes a dummy variable for the weekly exam day[14], the order of exams within the general exam schedule, and a set of dummy indicators for the exam's code (exam fixed effect). It also includes the number of days since the previous exam, capturing the effect of preparation time before each exam. This effect can be viewed as a mechanism for the impact of weekly and daily exams, as discussed in the literature review section. Therefore, I will examine the estimated treatment effects without controlling for this exam characteristic to assess its importance. When the exam fixed effect is included as a control, the exam fixed characteristics are omitted. Note that the exam fixed effect also accounts for a subject fixed effect.[15] $School\_Score_{ijt}$ is the student school test score that is exam-specific. It is based on a school-level exam administered a few weeks

[14] The distribution of indicators of number of exams in a day or a week by days of the week does not show a particular pattern (Table A2 in online appendix). However. However, we prefer to add a dummy indicators for each day of the week to controls for any day of the week effect even if it is not clearly observed. Similary for the distribution of week of the number of exams in a week or a day by order of the week in the exam schedule (Tale A3 in the online appendix).

[15] Each exam awards students with *bagrut* credits. An exam at the basic level yields three credits, intermediate level four credits, and advanced level 5 credits. Twenty-two *bagrut* credits are needed for *bagrut* certification.

before the national exam, covering the same material and having the same format. *Student F.E.$_i$* is a fixed effect for the individual. Note that the student-level characteristics are dropped in specifications with individual fixed effects. Exam_Date*_FE* is a date-fixed effect, ensuring that the variation I leverage for identification is not pure time series variation within a student's exam schedule across weeks in a given year. ε_{ist} is an idiosyncratic error term.

Thus, ϵ_w and ϵ_d are the two parameters of interest. Standard errors are clustered at the school level to allow for correlation in the error term within and between students in the same school. Since the two treatments are likely correlated, and their effects vary with their interaction, I also estimate the following model, which includes five treatment indicators that saturate all the possible interaction terms between the three weekly exam indicators and the two daily exam indicators. These interaction terms are two weekly tests on different days, three or more weekly tests on different days, two weekly days on the same day, three or more weekly tests two on the same day, and three or more weekly tests three on the same day. I combine three and four weekly exams into one treatment because, as will be shown based on equation (1) estimates, the marginal effect of adding a fourth weekly exam is relatively small.

The following specification notes these treatment indicators as T_τ, τ runs from 1 to 5.

$$(2) \quad Y_{ijt} = \alpha + \sum \epsilon_\tau T_{ij\tau\tau} + \beta X_i + \Phi\,(School\_FE)_i + \Omega\,(Cohort\_FE)_j + \mu\,(Test\_Charac)_j + \Upsilon\,(School\_Score)_{ijt} + \phi\,(Student\ FE)_i + \beta\,(\text{Exam}\_FE)_j + \varepsilon_{ijt}$$

Two key assumptions are needed to interpret the effect of the two treatments of interest. The first is that the dates of exams, which determine the two treatment variables, are exogenous to the student's study program and choice of elective subjects.[16] This assumption holds since the study program of each student is determined in the beginning or during 10th grade, almost three years before the exam schedule is published. However, the student's study program determines her number of exams. Therefore, conditional on the exogenous exam dates, the study program determines the number of exams per day or week. A related assumption is that there is no set order in which exams are scheduled, meaning that the number of exams in a week/day is not always higher at the end or beginning of the exam period. Otherwise, there might be a bias, although the equation includes a control for the exam's order, which should address this concern.

[16] There have been no particular regularities in the Ministry of Education's exam scheduling over the years, with the following few exceptions. English is scheduled as the first exam in three of the six years, Math is scheduled second in two of the six years.

The second identifying assumption is that conditional on student fixed effect, unobserved student test performance determinants are uncorrelated with these two structural exam schedule parameters. The student-fixed effect is crucial here for several reasons. First, students choose their study program primarily by choosing their academic tracks based on unobservable characteristics. The study program determines the subsequent possible choice of subjects and the resulting number and level of exams. They are likely correlated with the student's characteristics, such as ability. Table 1 shows an example of such correlation: lower-ability students (below median) have fewer exams during the exam period. As a result, they have fewer weekly exams and fewer days with more than one exam. Therefore, the OLS estimate of ϵ_w and ϵ_d will likely have the opposite than expected sign, positive instead of negative. This adverse selection is likely eliminated by adding the student's background characteristics (X_{it}) and school score ($School\_Score)_{ijt}$, which is exam-specific, as controls. However, as shown below, adding pupil fixed effect changes further the estimate of ϵ_w and ϵ_d, indicating that these controls are not enough to control for selection choice of the study program and subjects of exams.[17] [18]

To assess the validity of the identification and estimation design, I also estimate placebo tests. For each student, I randomly assigned one of the treatments he experienced to each exam. For example, for a student who had ten exams and was exposed to four of the five treatments possible, I randomly assigned one of the four treatments to each of the ten exams. More details on this placebo treatment are in the following section.

5. Empirical Results

Table 2 presents estimates of the effect of the two treatments of interest based on the non-linear version of equation (1). As noted above, the treatments are measured by indicators that allow the estimation of non-linear effects. Four different regression specifications are used. In column 1, the estimated regression includes only cohort-fixed effects as controls; in the second, student characteristics are added; the third specification includes test characteristics and exam-specific school test scores. The fourth specification includes student-fixed effects while the student characteristics are dropped. In column 5, the exam date and week fixed effects are added.

[17] The exam fixed effect also neutralizes potential concerns that some exams are scheduled to allow more preparation time. However, there is no evidence that this is the case in the Israeli context.

[18] The importance of including a student fixed effect in a regression that estimates the effect of inputs in the education production function is also documented in Lavy (2020). This study estimated the effect of instructional time in school on test scores, and adding pupil fixed effect yielded estimates much lower than a regression that only controlled for observed student characteristics.

All three estimates of weekly exams treatment indicators are positive and statistically significant when no controls are included in the regressions (column 1). Adding students' characteristics as controls lowers all estimates, though they are all still positive and significant (column 2). Adding test characteristics and school scores reverses one estimate from positive to negative. However, the within-student estimates (a regression that includes a student-fixed effect) are all negative and statistically significant. Adding the exam date fixed effects (column) does not change the estimates.

The pattern of the estimates of the two dummy indicators for the number of daily tests, two or three, is different, being negative in column 1 and changing signs irregularly in columns 2 and 3. However, in columns 4 and 5, the two within-student estimates are negative and significant. Therefore, all five estimates in column 1 are biased and cannot be interpreted as causal effects. Similarly, the estimates presented in columns 2 and 3 are likely biased, and the focus should be on the estimates in column 5.

The harmful effect of multiple weekly exams on different days seems very non-linear. Adding a second weekly exam lowers test scores by 0.0062 relative to one weekly exam, but adding a third exam lowers test scores by 0.0228 SD relative to one weekly exam. This monotonic increase continues when adding a fourth exam but with a modest change relative to three weekly tests: a decline in test scores by 0.0349 SD. So, an increase from two to three weekly exams is much more harmful than a change from three to four.

The estimated effect of an additional daily exam, holding constant the number of weekly exams, is harmful and relatively large, with a 0.0329 SD decline in each daily exam. Adding a third exam on the same day seems not worse than having two daily exams in this model. I note that the estimated standard errors in the last column and the last row are bigger because there are few instances of this treatment, so the cell is small.

Next, I estimate a non-linear specification, including weekly and daily exams in the regressions and allowing for their interaction effects (estimation of equation 2). As noted in the previous section, I include 3 and 4 plus weekly exams in the same dummy indicator based on the estimates presented in Table 2 and for parsimonious specification. The omitted group includes one weekly exam. These results are presented in Table 3. All five dummy treatment indicator estimates in column 1 are positive and statistically significant, revealing how biased the simple OLS estimates are. Adding student and school characteristics still leaves some estimates positive. Adding the test characteristics and the school score reverses the sign of four of the five estimates. However, adding the pupil fixed effects (column 4) yields negative and significant estimates for all five treatment indicators, which are not changed when adding the date fixed effects.

An interesting pattern emerges. Two tests in a week, each scheduled on a different day, negatively affect test scores relative to one weekly exam, with a decline of 0.010 SD in both test scores. However, when both tests are on the same day, the decline relative to one test is doubled, as the score is lower by 0.021 relative to the score of one weekly exam. It was noted in section 3 that multiple daily exams are always in the same subject. Since exams-fixed effects are included, the above negative effect is net of a subject-fixed effect. However, the estimate of -0.021 is not significantly different from -0.097, so one should be careful in concluding that loading two tests in a day, even if they are on the same subject, is worse than two weekly tests on different days.

Another emerging pattern is the large effect of three or more weekly tests. Adding a third weekly exam, each on a different day, is three times worse than moving from one to two weekly exams on separate days. This non-linear jump lowers test scores of all three or four exams by 0.0647 SD. This effect is not changed when eliminating students with four weekly exams from the sample because these cases comprise only ten percent of the group of three or four weekly exams.

However, if two or three weekly exams are on the same day, it is much *less* detrimental than when the three tests are scheduled on different days. If two of the three are on the same day, it lowers the test score of each test by 0.034 SD. If three of the three plus weekly exams are on the same day, it lowers the test score of each test by 0.045 SD. However, these estimates are larger (and statistically different at 10 percent) than the negative effect of two weekly exams on two different days (-0.010). They are also larger than the effect of two exams on the same day (-0.0214) but not statistically different. The implication is that the marginal effect of adding a third weekly exam when the other two are on the same day is almost equal to the effect of the second weekly exam (the difference between -0.034 and -0.0214 is almost -0.010). These results enhance the confidence in the consistency of the other findings presented in Table 3.

The estimates presented in Table 3 show that the marginal effect of an additional weekly or daily exam is negative and economically meaningful. When moving from one to two exams, the negative effect is larger when the second exam is on the same day, not being mitigated by being on the same subject. But, the benefit of complementarity in the material covered (because it is related to the same subject) is evident in the smaller effect of three plus weekly exams when two or three are on the same day. The conclusion is that the non-linear monotonic increase in the detrimental impact of more than two weekly exams can be mitigated when related to the same subject.

Framing these results regarding the two types of multitasking discussed in the literature review section suggests that the cost of concurrent multitasking, when all daily exams are on the same subject, is smaller than the cost of serial multitasking (the number of weekly exams). However, we cannot extrapolate

this result when the daily exams are on different subjects because we do not have such heterogeneity in the Israeli matriculation exams because the estimates of the effect of daily exams are a lower bar under the condition that it is harder to study different subjects.

Robustness Concerning Sample Selection. The sample used until now includes all students who participated in the matriculation exams at the end of 12th grade. Some of these students repeated an exam they had taken earlier, in the winter exams of 12th grade or in 10th or 11th grade. Few students skipped an exam at the end of 12th grade and opted to take the second chance exam directly in the summer following 12th grade. These exams are offered only in English and math. Appendix Table A4 presents estimates derived from samples that sequentially omit the students who repeat or skip an exam in the testing round at the end of 12th grade. Column 2 presents estimates based on a sample excluding students who missed one or more exams at the end of 12th grade. In column 3, the sample excludes students who retook an exam taken earlier in 12th grade. In column 4, the sample excludes students who retook an exam that was taken also in 10-11th grade. In column 5, all these sample restrictions are imposed jointly. For comparison purposes, column 1 presents the estimates from the full sample (copied from column 4 in Table 3). The estimates in columns 2-5 are identical to those in column 1, implying that the three types of sample selection are not 'selective' and, therefore, do not affect the estimates presented in Table 3.

Placebo Regressions. To support the identification strategy and further validate the causal interpretation of the evidence presented in Table 3, I ran regressions with placebo treatments designed as follows. For each student, I randomized his treatments across his set of exams. The randomized exposure to treatments differs more from the actual exposure for students with more exams and treatments. For example, the placebo assignments for a student with ten exams and all five treatments will deviate with a higher likelihood from the actual treatment than that of a student with only five exams and two treatments. A situation with no difference between the placebo and the actual treatment exposure is the case of a student with only one treatment type. Thirteen-plus percent of the students are in this situation. Forty percent of the students have only two treatment types—however, such cases 'work' against finding a null placebo effect, so our placebo exercise is 'conservative'.[19]

Table 4 shows placebo regression estimates for each of the five specifications used in Table 3. Commenting first on the results in columns 4-5, based on our preferred specification (with a student fixed effect), all the estimates are small and not statistically different from zero. Four of the five estimates are even positive. This pattern is very different from the estimates shown in the respective columns of Table 3.

[19] Completing the information about the distribution of treatments in the sample, thirty five percent of the students have 3 different treatments, 5 percent has 4 treatments, and only a 0.2 percent has five treatments.

Second, it is interesting that the estimates in columns 1-2 of Table 4 are mostly positive and significant, resembling those in the same columns of Table 3 and reflecting the same selection pattern; high-ability students will have more exposure to multitasking. This is an expected result because the placebo randomization is performed within students, so the overall number of exams of each student is not changed; the placebo assignment changes only the exam date, resulting in a random multitasking assignment.

6. Heterogeneity in Treatment Effects

It is important to estimate the heterogeneous effect of test multitasking by the ability and gender of students since, as noted above, the Bagrut exams are an important part of Israel's university admission system. Furthermore, extracurricular activities and student essays do are not considered in the admission process. Consequently, bagrut scores can affect an individual's academic career and subsequent labor market outcomes. This factor is particularly important for high-ability students who apply to the programs that select high-performing students. It matters less for lower-performing students who often apply to study programs that do not require a minimum GPA or grade in a particular subject. A similar rationale is behind the interest in heterogeneity by gender. The gender earning gap is related to gender differences in the field of study in higher education. The higher enrollment of men in STEM fields may be related to gender heterogeneity in the effect of multitasking on bagrut test scores.

Heterogeneity by Student's Ability. Several alternative measures of ability can be used in this study, including the average score in all school exams, the overall number of credit units in the *bagrut* program, and the number of subjects studied at an advanced level chosen during 10^{th} grade. All three measures are highly correlated, but the preferred one is the first because it represents the average student's knowledge in the exams included in the study sample for each student. The sample was divided into two, below or equal and above the median of this average score. Naturally, the latter sample is larger.

Panels A and B in Table 5 present the results for these two ability-based samples based on the same specifications as Table 3. The general pattern from comparing the estimates in both panels is that multitasking learning due to dense exam scheduling mostly hurts students with high ability. However, it is remarkable that the sorting/selection pattern seen from column 1 is of the opposite sign for the low and high-ability groups. For the low-ability group, four of the five estimates in panel A, column 1, are positive; for the second group, these estimates except one (panel B) are negative, large, and statistically significant.

Adding controls up to a within-student estimation reverses the sign in panel A and lowers the negative estimated effects in panel B. Once a student-fixed effect is added to the regression, all treatment

estimates in both samples are negative. Overall, it seems that the effect of the number of weekly exams is not very different by ability, especially for three weekly exams held on different days. For example, the negative effect of three plus weekly tests, each on another day, is -0.058 in the below-median ability sample and -0.072 in the above-median ability sample. The two estimates are not statistically different. However, a striking difference is seen when comparing the estimated effect of two or three tests on the same day. The negative effect of two daily tests in the below-median ability sample is -0.0074 and not statistically different from zero. The above-median ability estimate is -0.036, significantly different from zero and the below-median ability estimate. The negative effect of three daily tests in the below-median ability sample is -0.019. In the above-median ability sample, it is -0.061. These two estimates are also significant and statistically different.

The increase in the estimated effect of the number of daily exams as we move up the ability ladder is interesting. We noted above that this effect is likely due to increased mental and physical fatigue and perhaps more complicated planning, time, and task management. On the other hand, it may increase with ability because of the sharp increase in the level and scope of the material covered in advanced courses, similar to advanced placement (A.P.) classes in the U.S. I add two additional, perhaps more important, explanations for this finding. The first relates to time-binding constraint: high-ability students will likely have higher test scores because they study more. Hence, another test constrains their time more than it constrains low-ability students' time. The second explanation relates to the relative importance of the test scores. High-ability students may view the exams as more high-stakes than low-ability students because of the competition for admission to high-demand fields of study such as medical school or prestigious computer science programs. Marginally higher test scores in the *major* subjects in high school may be crucial for securing admission to these programs and, therefore, may increase stress and intimidation, lowering test performance. We demonstrate the importance of this channel by showing below evidence on heterogeneity by STEM subjects, for example, completion of 5-credit exams in math and physics.

Heterogeneity by Gender. Table 6 presents the results for girls (panel A) and for boys (panel B). The 'naïve' estimated effect of the weekly and the daily number of tests is positive for all treatment indicators for both genders, implying a similar selection pattern. The within-student estimates in column 5 are negative, and most are statistically significant. There are, however, some noticeable gender differences. The effect of two weekly tests on different days is negative, small, and insignificant for both boys and girls. Two of the other three treatment estimates, three plus exams on different days or three of them on the same day, have larger effects on boys, though not statistically different from that of girls. The largest gender difference is in the effect of three plus weekly tests scheduled on the same day: -0.026 for girls versus -0.052 (se=0.0135) for boys, though the two estimates are not statistically different. When three exams are scheduled on different

days, the estimates are also quite different: -0.0334 for girls and -0.0645 for boys. To summarize, the heterogeneous results by gender suggest that for mild multitask pressure, the adverse effect for girls and boys is similar, but for intense multitask pressure, the negative impact for boys is higher. However, this pattern is not statistically significant, perhaps because of a lack of power.

The literature in psychology provides several explanations for this finding that very dense daily exam schedules (intense multitasking pressure) are less harmful to girls. First, it is documented that both in high school and college, women report greater use of strategies for coping and time management than men (e.g., Misra and McKean, 2000; Xu, 2006 and 2007). These strategies positively correlate with better academic performance and grades (Claessens, Van Eerde, Rutte, and Roe, 2007). There is also evidence that girls have more self-discipline (e.g., Duckworth and Seligman, 2006) than boys. Using cash incentives for high-school students in Israel to increase the matriculation certification rate, Angrist and Lavy (2009) found that such intervention led to a substantial increase in girls' certification rate and the likelihood of college attendance but had no effect on boys. They suggest that the female matriculation rates increased partly because treated girls devoted extra time to exam preparation.

A second explanation suggests that boys have a higher discount rate than girls (Warner and Pleeter 2001). People with high discount rates are likely to only work on a project close to the deadline since they discount future outcomes. Conversely, a higher return to preparation days might signal a low discount rate and better use of the days before the exam. Finally, the literature suggests that adolescent girls are more likely to delay gratification (e.g., Silverman, 2003; Duckworth and Seligman, 2006) than teenage boys.

A third additional explanation relies on gender differences in the reaction to pressure. The literature in psychology shows that increasing the pressure beyond a certain level can lead to a decline in performance, commonly referred to as "choking under pressure" (Baumeister 1984). This behavior depends on personal traits such as competitiveness and ego-relevant traits like the belief that a task is diagnostic of an inherent characteristic (such as intelligence) one cares about (Ariely et al. 2009). Females have more effective time management behaviors than males but experience higher academic stress and anxiety (Misra and McKean, 2000). Our results show that when students have severe time pressure of more than one daily test, the effect of these treatments on boys' test scores is higher than on girls. Azmat, Calsamiglia, and Iriberri (2016) also find heterogeneous results across genders for the effect of increased pressure. They exploit the variation in the test stakes, ranging from 5% to 27% of the final grade, and find that female students outperform male students in all tests. However, the female test score advantage declines when the stakes are high.

Finally, recent lab experiments show that women find it easier than men to multitask, switch tasks, set priorities, and adapt to changing conditions. For example, Kuptsova et al. 2015 found in a lab experiment

that men require more brainpower than women when multitasking, that women find it easier than men to switch attention, and that their brains do not need to mobilize extra resources in doing so, as opposed to male brains.

***Heterogeneity by Immigration Status*.** The effect of multitasking learning and testing might be more extensive for immigrants who may bear the additional burden of languish deficiencies that may impair their ability to grasp large quantities of reviewed material quickly and in pressured situations. Our sample includes one and a half percent of immigrants (students who immigrated to Israel in the last five years). I added an interaction term between each treatment effect and a dummy indicator of being an immigrant to the regression to estimate this heterogeneous effect. The results are presented in Table 7. All the interaction terms between the treatment variables and the immigrant's indicator are negative and large; all but one are statistically significant. The estimated effects of multi-week or multi-day exams are twice to three times larger than the effect on native students. Columns 2-3 present the estimates in samples by student's ability. The treatment effects are larger for both groups of immigrants but larger for low-ability immigrants. Columns 4-5 present the results by gender, showing no systematic treatment differences for immigrant boys and girls. However, we should note that the interaction effects in columns 2-5 are based on small samples, which may not allow precise estimation.

***Heterogeneity by STEM and Non-STEM*.** The effect of a dense exam period may vary by the test subject, particularly humanities/social science subjects, versus STEM subjects. The latter, which includes more analytical subjects, such as math, physics, biology, chemistry, and computer science, are considered more difficult for most students.[20] Thirty percent of all exams are in STEM subjects. I added an interaction term between each treatment effect and a dummy indicator of STEM subjects to the regression to estimate this heterogeneous effect.[21] The exam-fixed effect absorbs the main effect of this indicator. Table 8, column 1, reports estimates of the effect of non-STEM subjects and the interaction term of treatment and STEM subjects. Two estimated interaction terms are negative and significant: three plus exams when two or three are on the same day. For the first interaction term, the estimate of this treatment on the non-STEM subjects is negative, small (-0.0227), and significant (SE=0.008). If it is a STEM subject, the effect increases by -

[20] In a recent poll of hundreds of secondary school pupils, teachers, and parents that Accenture, Ireland, conducted in 2019, many students reported struggling with these subjects. About 69 percent of teachers and 56 percent of students agree that the STEM subject curriculum is too complicated and takes too much study time. Teachers believe students are most likely to drop out of higher-level maths, physics, and chemistry because the subjects are too complex for students (69 percent) and take up too much time (44 percent) [The Irish Times, Dec 4, 2019].

[21] An alternative to the interaction model is to split the sample into STEM and non-STEM exams. However, the latter has the disadvantage of estimating a different fixed effect for each student in the two regressions. I, therefore, preferred the model with all parameter estimates equal to STEM and non-STEM except the treatment effect.

0.0252, more than doubled. For the first interaction term of three daily exams, the estimate of this treatment on the non-STEM subjects is negative, small (-0.0158), but not statistically significant (SE=0.0126). If it is a STEM subject, the effect increases by -0.0373 and is precisely measured (SE= 0.0142). These results mean that in the multitasking combinations we explore in this study's context, the cost of exam density in STEM subjects is higher than in non-STEM subjects.

In columns 2-3 of Table 8, I present the estimates in sub-samples by ability. The large negative effect of three daily tests in STEM subjects is evident for high and low-ability students. The two estimated interaction terms are statistically significant and are similar, -0.0654 and 0.0754. For high-ability students, the same treatment has a negative effect, though more minor, also on non-STEM subjects. However, for the high-ability group, the estimated effect on STEM subjects is also larger for the three plus - two indicator on the same day.

Columns 4-5 in Table 8 present the evidence for boys and girls in STEM versus non-STEM subjects. Multiple daily STEM exams cause a decline in test scores of boys and girls. However, for boys, this negative effect is more evident when the treatment is three daily exams, and for girls, when the treatment is two daily exams.

These results imply that educational authorities, universities, and other organizations should be particularly concerned with dense exam schedules when the test material is complex. STEM subjects are such examples.

Heterogeneity by Test Order on the Day. The model we estimate includes a control for the test order in the whole testing program. However, it is possible to check whether the negative impact of multiple daily or weekly exams relates to the test order in the day or the week. This treatment will likely capture the effect of mental, cognitive, and physical fatigue. In column 1 of Table 9, we present estimates from a specification that includes two additional dummy indicators. The first indicates the second exam in a day with two or three exams. I am comparing the effect of a second exam to the first exam in a day with two or three exams. For the sake of a parsimonious specification, my implicit assumption is that the effect of a second exam in a day with two exams is the same as the effect of a second exam in a day with three exams. The second indicator is a dummy for an exam third in a day. The estimates of these indicators show a sharp change in the detrimental effect of adding daily exams beyond one a day. Adding a second exam lowers the test score in this test by 0.081 SD relative to the score in the first exam in a day with two or three exams. Adding a third exam lowers the test score of this exam by 0.144 SD relative to the score in the first exam in a day of three exams. A monotonic decline of such magnitude from the first to the third daily exam is consistent

with a fatigue mechanism, cognitive and physical. But it can also reflect less optimal preparation for the exams scheduled beyond the first day. Both explanations are part of multitasking complexity.

An interesting result in column 1 is that the estimated effects of the number of weekly exams, when scheduled on different days, are not changed when the indicators of the daily order of exams are added as treatments. The estimated effect of two weekly exams on different days is -0.0107 in column 1 of Table 8 and -0.0109 in column 4 of Table 3. The estimated effect of three weekly exams on different days is -0.0619 in column 1 of Table 8 and -0.0633 in column 5. This reflects the practically zero correlation between the number of weekly exams and the daily order of exams scheduled on the same day.

Columns 2-3 of Table 8 present the estimates with daily order of exams for low and high-ability students and columns 4-5 for boys and girls. The same pattern discussed above is replicated in each of these subsamples.

Does the Effect of Exam Density Depend on the Test Order of Days in the Week? Performance in exams scheduled on Sunday may differ from exams on later days in the week. Students may be more relaxed and 'fresh' earlier in the week because of the weekend rest (Friday and Saturday). But the weekend may have the opposite effect, taking students out of the track of studying, perhaps of leisure activity, less sleep, and more fatigue. As noted regarding equation (1), its specification includes a set of dummy variables for each exam's day of the week. Exams held on Sunday are the omitted group. The main effect of these indicators shows negative and statistically significant coefficients for exams held on Tuesdays (-0.0149, se=0.0069) and Thursdays (-0.0121, se=0.0064). The estimates of Mondays and Wednesdays are either zero or small and not statistically different from zero. Very few exams are held on Friday (less than 2 percent). This is a mixed pattern, but one can still conclude that relative to exams held on Sundays and Mondays, the test scores of exams held later in the week are significantly lower. Again, this result is consistent with mental, cognitive, and physical fatigue, as with several exams piled in a day. But it is also consistent with the difficulty of preparing for multiple exams in a week, and the 'weak link' is the chain's last exam.

Heterogeneity by Available Preparation Time? The literature reviewed in section 2 suggests several channels of the effect of multitasking learning and testing. Some relate to mental fatigue and pressure, and others to available preparation time before each exam. We can examine the latter mechanism's importance by the estimated effect of the number of days since the last exam before an exam, which is included as a control in the regression. This estimate is practically zero in all four regression specifications of Table 3.

Placebo Regressons With Heterogeneous Treatment Effects

In the online appendix, Table A5, I present placebo regressions in the sub-samples by ability and gender. The regressions are based on the full specification with student-fixed effect and control for the exam date. Column 1 replicates the complete sample estimates from column 5 in Table 4. All estimates in columns 2-5 are very small, close to zero, and not statistically significant. These precise zero null estimated effects reaffirm that random exposure to multitasking treatment load yields zero effect, in sharp contrast to the real treatment assignments.

7. The Effect of Multitasking on Students' Academic Dishonesty: Cheating in Exams

Academic dishonesty is related to three important conditions: (i) opportunity, (ii) incentive, pressure, or need, and (iii) rationalization or attitude (e.g., Becker et al., 2006; Ramos, 2003, Holden, Norris, and Kuhlmeier, 2021). These three conditions positively predict student cheating behavior (Becker et al., 2006). An opportunity occurs when students perceive the ability to cheat without being caught. Incentives, pressure, or needs can come from various sources, such as the self, parents, peers, teachers, and schools. The pressure felt by students to get good grades and the desire to be viewed as successful can create the incentive to cheat. Lastly, the rationalization of cheating behavior can occur when students view cheating as consistent with their ethics and believe that their behavior is within the bounds of acceptable conduct (Becker et al., 2006; Ramos, 2003). This study focuses on the second condition related to multitasking load: multiple weekly and daily exams may lead to anxiety, a sense of pressure, and fear (DeVoss and Rosati, 2002; Sterngold, 2004). If students become over-extended, they may use inappropriate resources and strategies to manage (e.g., Sterngold, 2004). Increased stress levels can lead students to be more prone to dishonest behaviors. Whitely (1998) reviews the results of 107 studies on cheating among American and Canadian college students published between 1970 and 1996. He finds that when individuals can perform a task well, they cheat less. In contrast, when the exam is more important, and the pressure to achieve high grades is higher, as in the case of matriculation exams, students are more likely to cheat. Carrell, Malmstrom, and West (2008) find that higher levels of peer cheating result in a substantially increased probability that an individual will cheat.

Cheating in exams can take several forms, including copying from others, having or using notes, formulas, or other information in a programmable calculator or other electronic devices without explicit permission, having or using a communication device such as a cell phone, pager, or electronic translator to send or obtain unauthorized information, taking an exam for another student, or permitting someone else to take a test for someone else and asking another to give you improper assistance, including offering money or other benefits, etc.

However, being disqualified because of an exam cheating is a relatively rare event. Therefore, I collapse the number of weekly and daily exams into two treatments: an indicator for two or more exams per week and an indicator for two or more exams per day. To estimate the effect of multitasking load on the likelihood of exam dishonesty, I ran equation (2) with the following modifications: (i) replacing the test score outcome in a particular exam as the dependent variable with an indicator getting a value of 1 if the exam was disqualified. (ii) including as treatments only the two indicators described above. (iii) Dropping the exam fixed effect to allow enough within-student variation.[22]

Table 9 presents the estimated effect of weekly and daily exams on the likelihood of cheating in an exam. Each of the estimates of the two dummy treatment indicators presented in column 1 is statistically significant and changes as controls are added to the regression. The within-pupil estimates in column 4 differ from those in column 1, revealing how biased the latter is. Both estimates in column 4 are positive and statistically significant. Having two or more weekly exams, each scheduled on a different day, positively affects the likelihood of cheating in an exam, with an increase of 0.0009, relative to a baseline of 0.00178 percent (the mean of the proportion of disqualified exams in a sample that includes only cases of one weekly exam), implying a 50 percent effect size increase. The treatment indicator of two or more daily exams also has a positive effect, and larger: a significant 0.0018 relative to a baseline of 0.0021 (the mean proportion of disqualified exams in a sample that includes only cases of one daily exam), implying a 90 percent effect size increase in the likelihood of cheating in an exam. Multiple daily exams deteriorate students' ethical norms more than multiple weekly exams.

A concern to note about an alternative and potentially equally valid interpretation of these results is that the amount of cheating is not changing, but instead that the detection probability goes up on days/weeks with multiple exams. For example, perhaps it is easier to detect cheating when examiners mark a student's scripts for three tests in a row (either on the same day or week). However, the institutional features of the marking system, detailed in section 2, make this unlikely. Exams are marked centrally. Each examiner marks exams in a school subject he teaches regularly. For example, math teachers mark math exams, and history teachers mark exams in history. Therefore, this rules out that an examiner marks multiple exams scheduled on different days of the week because these exams will be on different subjects. Each examiner also marks only one exam in each batch received. Each multiple daily exam will be placed in a

[22] I note again that this estimation is based on data for 2000-2002 because the information on disqualified exams is unavailable for 2003-2005. In appendix Table A5, I present estimates of the effect of multitasking on test scores based only on data for 2000-2002. The results in this table are very similar to those shown in Table 3, which are based on data for 2000-2005.

different pack, randomly allocated to markers. Therefore, it rules out the possibility that an examiner is marking two or three exams of the same student scheduled on the same day.

Heterogeneity by Student Ability and Gender: Table 10 presents respective estimates from stratified samples by high and low ability and gender. Overall, the patterns described above are maintained: multiple weekly or daily tests negatively affect exam disqualification. However, the estimates are heterogeneous. The positive effect of weekly tests on cheating is similar in the two ability groups[23] and is all driven by boys, with no effect on girls. On the other hand, the impact of daily tests on cheating is derived entirely from low-ability students, and it is larger for girls than boys.

However, it is worth noting that the heterogeneity of the treatment effect of multiple daily exams by ability is sharper with even a negative estimate, though not statistically significant, for the high-ability group. But this needs a qualification since cheating is measured here by being caught in dishonest behavior. The mean of ever being caught cheating in the sample of lower-ability students is almost six times higher than in the higher-ability students. Therefore, the heterogenous treatment estimates by ability may reflect better 'cheating techniques' among high-ability students, reducing the likelihood of being caught. But it can also reflect more need and, therefore, a higher incentive for dishonest behavior. We cannot distinguish between these potential explanations, so we cannot draw a conclusive interpretation of the heterogeneous evidence by the student's ability of background.

Another possible reason that can explain the lower cheating rate among high-ability students when exposed to multiple daily exams can be related to the importance of these exams for this group of students. The cost of being disqualified in one of the daily exams is much higher for the high ability students since, as noted earlier, all daily tests are on the same subject and always in one of the 'majors' in high school, which are studied at the advanced level. For high-ability students, the test score in this major (often multiple majors) is a requirement for admission to a highly demanded university field of studies. For example, admission to computer science university programs requires high school math and physics majors. Therefore, disqualifying in one of these majors will carry a heavy cost in terms of not enrolling in a desired study program at the university. This cost may deter high-ability students from cheating in exams that are part of these majors.

[23] The results are not changed when stratifying the sample by parental education instead of student ability, likely because these two variables are highly correlated.

8. Conclusions

Across the globe, standardized testing is now a major component of the education system. The results of these tests are used to inform all manner of high-stakes decisions. In theory, these tests are meant to capture short-run growth in student skills that will eventually translate into success and well-being later in life. Yet, there is a growing body of work that suggests that seemingly inconsequential design choices about the structure of the test can influence student performance for reasons that are disconnected from traits the tests were originally intended to measure (e.g. Brown et al., 2023; Beltran et al., 2022; Reyes, 2023). This paper adds to this literature by documenting an important test feature that has received comparatively little attention: the density of tests within an examination period. After conditioning on year of test fixed effects, exam fixed effects, student fixed effects, and additively separable controls for test order, number of days since the previous exam, day of the week, and exam-specific school average scores, the evidence presented in this paper show that student performance is lower on days/weeks with multiple tests. The findings have similar patterns along several margins of heterogeneity, including gender, baseline student ability, and immigrant status (where effects are slightly larger). Finally, the results suggest that conditional on a slightly less rich set of controls, having multiple exams per day/week also predicts that students are more likely to be found cheating.

Ebenstien, Lavy, and Roth (2016) have shown that even a slight decline in these test scores due to random shocks (in their paper, it is a shock to pollution and air quality) causes long-term negative consequences in post-secondary schooling, earnings, and allocative inefficiency in the labor market. The negative impact of multitasking load estimated in this paper is much more significant than the effect of pollution; therefore, one can assume its long-term impact will be more considerable accordingly. Furthermore, although the short-term impact on test scores is similar for low and high-ability students, its long-term impact will be more harmful to the latter group because high test scores open the gate to highly sought-after study programs and elite academic institutions. Therefore, this differential long-term impact by ability will cause allocative inefficiencies in the labor market (Ebenstien, Lavy, and Roth (2016).

The findings I present in this paper are relevant beyond the education sectors. Most firms today use a set of tests to screen and select job applicants. An industry provides this service to employers who aim to measure and attract talent and make better hiring decisions. The screening and selection process includes multiple daily tests over several days. The decisions are often among equally talented individuals with a small difference in measured criteria. These margins are likely shaped by the burden of multitasking tests and the differential effects among candidates in their impact. Yet, they can result in less optimal matching

between workers and firms. This conclusion, of course, should be qualified, perhaps irrelevant, for some firms who screen workers also based on good multitasking ability.

Another angle of the importance of the findings reported in this paper is related to the wide use of exam test scores in various contexts to evaluate interventions and policies. For example, much of the literature in economics of education on the effect of treatments, in experiments or with observational data, is based on test scores in national or school-level exams. Often, these situations involve multiple exams, like in the *bagrut* national exams. Relying on multitasking testing may make the test scores more noisier.[24] Perhaps more importantly, they may bias the results regarding the impact of these interventions because they might lower test scores enough to make a program seem ineffective.

An interesting and policy-relevant finding is that studying for multiple exams in different subjects is more harmful than studying for multiple exams in related subjects, even though the latter is condensed into one day. The 'advantage' of exams on the same subject likely results from a lower 'set up cost' when switching tasks. For example, the cost of switching from geography to math exam preparation is probably larger than that of switching from algebra to geometry. This example is consistent with the idea that switching costs increase with increased levels of multitasking (Crenshaw 2008, Coviello et al. 2014). Another insightful policy finding is that 'piling' exams in STEM subjects carries a more harmful impact. The short-term negative 'effect size' is larger for low-ability students because their mean outcomes are much lower. However, the long-run effect is larger for high-ability students because test scores count more for post-secondary schooling admission.

The findings that multitasking load increases dishonest behavior should be of concern since studies show that people who cheat on exams in high school are considerably more likely to be dishonest later in life (The Josephson Institute of Ethics Report, 2009). This report shows that habits formed in childhood persist: those who cheated in high school are more likely as adults to lie to customers, inflate an insurance or expense claim, cheat on taxes, and lie to their spouse. Students who engage in academic dishonesty in high school or university are likelier to engage in work-related dishonesty (Sims, 1993; Nonis and Swift, 2001; Carpenter et al., 2004). Other potential negative consequences are that cheating creates a culture of mistrust and devalues education (Sims 1993, Nonis and Swift 2001).

Finally, the evidence presented in this study points to several policy options to avoid some of the adverse effects of multitasking testing. The general policy implication, with external validity ranging from high school high-stakes exams to firms screening and selecting workers based on personality psychometric

[24] This statement should be qualified since sometimes exams are densely scheduled in purpose where the objective is to assess performance under pressure.

exams, is to avoid dense testing schedules in a week and a day. In high school, for example, by shifting some exams to the 12th-grade mid-year and 11th-grade mid- and end-year exam periods. A second policy option is to reduce the number of high school exit exams. In France, for example, the number of BAC exams is three (philosophy, Math, French), while in Israel, students have over ten *bagrut* exams.

9. References

Anaya, Lina, , Nagore Iriberri b, Pedro Rey-Biel c 1, Gema Zamarro. "Understanding performance in test taking: The role of question difficulty order." Economics of Education Review 90 (2022): 102293.

Angrist, J., and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. The American Economic Review, 99(4), 1384-1414.

Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large Stakes and Big Mistakes. The Review of Economic *Studies*, 76(2), 451-469.

Azmat, G., Calsamiglia, C., and Iriberri, N. (2016). Gender Differences in Response to Big Stakes. *Journal of the European Economic Association*, 14(6), 1372-1400.

Balart Pau , Matthijs Oosterveen, and Dinand Webbink (2018). "Test scores, non-cognitive skills and economic growth" *Economics of Education Review*, Volume 63, April, Pages 134-153

Baumeister, R. F. (1984). I am choking Under Pressure: Self-Consciousness and Paradoxical Effects of Incentives on Skillful Performance. *Journal of personality and social psychology*, 46(3), 610.

Becker, D., Connolly, J., Lentz, P., and Morrison, J. (2006). Using the Business Fraud Triangle to Predict Academic Dishonesty Among Business Students. Acad. Educ. Leadersh. J. 10 (1), 37–52.

Bloom, B. S. (1974). Time and learning. American Psychologist, 29(9), 682.

Bollard A., R Liu, A. Nursimulu, A. Rangel, P.. "Neurophysiological evidence on perception of reward and risk: Implications for trading under time pressure". No. REP_WORK. University of Zurich, 2007.

Brown Christina, Supreet Kaur, Geeta Kingdon & Heather Schofield. "Cognitive Endurance as Human Capital" Working Paper 30133, 2022.

Burgess, P. W. (2000). Real-world multitasking from a cognitive neuroscience perspective. In S. Monsell & J. Driver (Eds.), Control of cognitive processes: Attention and performance XVIII (pp. 465–472). Cambridge, MA: MIT Press.

Burgess, P. W. (2015). Serial multitasking: from lab to life. In J. Fawcett, E. F. Risko & A. Kingstone (Eds.), The Handbook of Attention, chapter 19 (pp. 443-462). MIT Press.

Busemeyer, J. R., and Diederich, A. (2002). "Survey of decision field theory." *Mathematical Social Sciences*, 43, 345–370.

Carpenter, Donald D., Harding, Trevor S., Finelli, Cynthia J., and Passow, Honor J. (2004). "Does academic dishonesty relate to unethical behavior in professional practice? An exploratory study." *Science and Engineering Ethics*, vol. 10, pp. 311–324.

Carrell, Scott E., Frederick V. Malmstrom, and James E. West. (2008). "Peer Effects in Academic Cheating." *The Journal of Human Resources*, vol. 43(1), pp. 173–207.

Claessens, B. J., Van Eerde, W., Rutte, C. G., & Roe, R. A. (2007). "A Review of the Time Management Literature." *Personnel Review*, 36(2), 255-276.

CNNMoney.com (2006). Women CEOs for FORTUNE 500 companies. http://money.cnn.com/magazines/fortune/fortune500/womenceos/

Coviello, D., A. Ichino, and N. Persico. 2014. "Time Allocation and Task Juggling." *American Economic Review*, 104 (2): 609-23.

Crenshaw, Dave (2008) The Myth of Multitasking: How "Doing It All" Gets Nothing Done. San Francisco, John Wiley, and Sons.

DeVoss, D., and Rosati, A. C. (2002). "It Wasn't Me, Was it?" Plagiarism and the Web. Comput. Compost. 19, 191–203. doi:10.1016/s8755-4615(02)00112-3.

Duckworth, A., Lee, and Martin P. Seligman. 2006. "Self-Discipline Gives Girls the Edge: Gender in Self-Discipline, Grades, and Achievement Test Scores." *Journal of Educational Psychology*, 98: 198–208.

Eastwood, J. D., A. Frischen, M. J. Fenske, and D. Smilek (2012). "The Unengaged mind: Defining Boredom in Terms of Attention." *Perspectives on Psychological Science* 7(5), 482–495.

Ebenstein A., V. Lavy, and S. Roth "The Long Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution" *American Economic Journal: Applied Economics*, 2016, 8(4): 36–65.

Ericsson, K. A., & Charness, N. (1994). Expert Performance: Its Structure and Acquisition. *American psychologist*, 49(8), 725.

Holden, Olivia L., Meghan E. Norris and Valerie A. Kuhlmeier Academic Integrity in Online Assessment: A Research Review Frontiers in Education July 2021 | Volume 6 |

Gigerenzer, G., Todd, P.M., & the ABC Research Group. (1999). Simple Heuristics That Make Us Smart. New York: Oxford University Press.

Goulas Sofoklis and Rigissa Megalokonomou (2020). "Marathon, Hurdling or Sprint? The Effects of Exam Scheduling on Academic Performance", *The B.E. Journal of Economic Analysis and Policy*.

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2017). Cognitive structure, flexibility, and plasticity in human multitasking—an integrative review of dual-task and task-switching research. *Psychological Bulletin.*

Kocher, M. G., & Sutter, M. 2006. Time is money–Time pressure, incentives, and the quality of decision-making. *Journal of Economic Behavior and Organization*, 6 (3), 375-39

Kocher, M. G., Pahlke, J. & Trautmann, S. T. 2003. Tempus Fugit: Time Pressure in Risky Decisions, *Management Science*, 59(10), 380- 390.

Künzell, S., Bröker, L., Dignath, D., Ewolds, H., Raab, M., & Thomaschke, R. (2018). What is a task? An ideomotor perspective. *Psychological Research*, 81(1), 4-11.

Lavy, V. 2009. "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics" *American Economic Review*, Vol. 99, No. 5, December: 1979-2011.

Lavy, V. 2015. "Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence From Developed and Developing Countries", *The Economic Journal*, 125(588), F397-F424.

Lavy, V. 2021. "Expanding School Resources and Increasing Time on Task: Effects of a Policy Experiment in Israel on Student Academic Achievement and Behavior", *Journal of the European Economic Association*. February 2020, 18(1):232–265.

Leder J., Häusser J. A., Mojzisch A. (2013). Stress and strategic decision-making in the beauty contest game. *Psychoneuroendocrinology* 38, 1503–1511.

Mäntylä Timo. (2013) "Gender differences in multitasking reflect spatial ability" *Psychol Sci*. Apr;24(4):514-20.

Marcotte, D. E., & Hemelt, S. W. (2008). Unscheduled School Closings and Student Performance", *Education Finance and Policy*, Vol 3, 316-338.

McCabe Donald. L. (2010). Academic Integrity Survey Report, Texas Tech University, June.

McClelland, J. L., McNaughton, B. L., & O'reilly, R. C. (1995). "Why There Are Complementary Learning Systems in The Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory." *Psychological review*, 102(3), 419.

May Kaitlyn E and Anastasia D. Elder. 2018. Efficient, helpful, or distracting? A literature review of media multitasking in relation to academic performance. *International Journal of Educational Technology in Higher Education*, Volume 15: 13.

Misra, R., & McKean, M. (2000). College students' academic stress and its relation to their anxiety, time management, and leisure satisfaction. *American Journal of Health Studies*, 16(1), 41.

Nonis, S., & Swift, C. O. 2001. An examination of the relationship between academic dishonesty and workplace dishonesty: A multicampus investigation. The Journal of Education for Business 76: 69-77.

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

Ramos, M. (2003). Auditors' Responsibility for Fraud Detection. *J. Accountancy* 195 (1), 28–35.

Rieskamp J1, Hoffrage U. (2008). Inferences under time pressure: how opportunity costs affect strategy selection. *Acta Psychol* (Amst). Feb;127(2):258-76.

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

Persson, J., Herlitz, A., Engman, J., Morell, A., Sjölie, D., Wikström, J., et al. (2013). Remembering our origin: Gender differences in spatial memory are reflected in gender differences in hippocampal lateralization. *Behavioural Brain Research*, 256(0), 219–228.

Silverman, Irwin W. 2003. "Gender Differences in the Delay of Gratification: A Meta-Analysis." *Sex Roles*, 49(9–10): 451–63.

Sims, R. L. 1993. The relationship between academic dishonesty and unethical business practices. *Journal of Education for Business* 68(4): 207-13.

Sterngold, A. (2004). Confronting Plagiarism: How Conventional Teaching Invites Cyber-Cheating. Change Mag. Higher Learn. 36, 16–21.

Stoet, Gijsbert, Daryl B O'Connor, Mark Conner & Keith R Laws. (2013). Are women better than men at multitasking? *BMC Psychology* volume 1, Article number: 18.

Sutter Matthias, Martin Kocher, Sabine Strauß. (2003). "Bargaining under time pressure in an experimental ultimatum game" *Economics Letters*, Volume 81, Issue 3, December, Pages 341-347.

Taraban, R., Rynearson, K., & Stalcup, K. A. (2001). Time as a variable in learning on the Worldwide Web. Behavior Research Methods, Instruments, & Computers, 33(2), 217-225.

Warner, John T., and Saul Pleeter. (2001). "The Personal Discount Rate: Evidence from Military Downsizing Programs." *American Economic Review*, 91(1): 33–53.

Whitley, Bernard E. (1998). 'Factors associated with cheating among college students: A review', *Research in Higher Education*, vol. 39, pp. 235–274.

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

Xu, J. (2006). Gender and homework management reported by high school students. *Educational Psychology*, 26(1), 73-91.

Table 1: Descriptive Tests Statistics, By Sample

	Full Sample	Ability Below Median	Ability Above Median	Girls	Boys
Number of Tests in the Exams Period	8.360	8.107	8.613	8.289	8.456
	(2.024)	(2.036)	(1.979)	(1.972)	(2.088)
A. Weekly Tests Statistics					
Number of Tests	1.731	1.692	1.770	1.698	1.776
	(0.878)	(0.849)	(0.905)	(0.853)	(0.909)
Weeks with More Than One Test	1.836	1.753	1.919	1.794	1.892
	(0.992)	(0.992)	(0.986)	(0.983)	(1.003)
% of Students with Two Tests in a Week	74.29	73.01	75.67	74.53	73.96
% of Students with Three Tests in a Week	35.86	33.06	38.87	34.79	37.32
B. Daily Statistics					
Number of Tests	1.293	1.285	1.300	1.272	1.320
	(0.536)	(0.525)	(0.547)	(0.513)	(0.565)
Days with More Than One Test	1.064	1.029	1.100	1.015	1.131
	(0.843)	(0.847)	(0.837)	(0.840)	(0.842)
% of Students with Two Tests in a Day	62.87	60.92	64.98	61.48	64.79
% of Students with Three Tests in a Day	10.04	8.65	11.53	8.00	12.83
C. Extent of Disqualified Exams					
% of Disqualified Exams	0.29	0.45	0.14	0.28	0.32
% of Disqualified Exams When More Than One Exam Per Week	0.41	0.67	0.17	0.40	0.44
% of Disqualified Exams When More Than One Exam Per Day	0.56	0.94	0.19	0.57	0.56
% of Students With a Disqualified Exam	2.19	3.23	1.10	2.06	2.38

Notes: The table presents means and standard deviations (in parentheses) of test characteristics in terms of weeks and days. Sections A-B and D-F are based on the 2000–2005 sample. Section C is based on different samples covering the years 2000–2002.

Table 1: Descriptive Statistics (*Continued*)

	Full Sample	Ability Below Median	Ability Above Median	Girls	Boys
D. High School Outcomes					
Average National Exams' Score	0.059	-0.352	0.470	0.116	-0.017
	(0.635)	(0.563)	(0.388)	(0.618)	(0.649)
Average School Exams' Score	0.047	-0.513	0.608	0.145	-0.085
	(0.699)	(0.482)	(0.341)	(0.670)	(0.715)
E. Student Characteristics					
Father's Years of Schooling	13.56	13.00	14.12	13.29	13.91
	(3.380)	(3.380)	(3.504)	(3.397)	(3.324)
Mother's Years of Schooling	13.53	13.05	14.01	13.30	13.84
	(3.029)	(2.866)	(3.110)	(3.074)	(2.937)
Number of Siblings	2.614	2.636	2.591	2.630	2.591
	(1.482)	(1.478)	(1.485)	(1.515)	(1.435)
Observations	1,280,329	640,156	640,173	736,422	543,907
Students	163,971	84,990	78,981	94,843	69,128
F. Test Characteristics					
Days Before Test	8.243	8.513	7.977	8.286	8.184
	(5.991)	(6.143)	(5.825)	(5.992)	(5.990)
Test Duration in Minutes	135.1	135.6	134.7	134.5	136.0
	(40.29)	(40.29)	(40.18)	(40.56)	(39.89)
Observations	1,095,853	543,515	552,338	631,393	464,460
Students	163,807	84,859	78,948	94,758	69,049

Notes: The table presents means and standard deviations (in parentheses) of high school outcomes, student characteristics, and test characteristics. Number of Days Before denotes the number of days before the current test. To deal with the complexity of analyzing the effect of the number of days before the first exam day, we assign it the number of days since the Friday before the exam season. In addition, to deal with the complexity of analyzing the effect of the number of days after the last exam day, the last exam day of each student was omitted. Hence, the statistics in Section E exclude the exams that are scheduled on the first and last days of the exam period.

Table 2: The Effect of the Number of Weekly and Daily Tests on a National Exams' Score

	(1)	(2)	(3)	(4)	(5)
Weekly Tests					
Two	0.0312***	0.0255***	-0.00835*	-0.00781*	-0.00621
	(0.00695)	(0.00652)	(0.00455)	(0.00427)	(0.00461)
Three	0.0452***	0.0325***	-0.00672	-0.0219***	-0.0228***
	(0.00949)	(0.00894)	(0.00632)	(0.00617)	(0.00656)
Four Plus	0.0890***	0.0623***	0.00443	-0.0259**	-0.0349***
	(0.0145)	(0.0137)	(0.00999)	(0.0103)	(0.0105)
Daily Tests					
Two	-0.0118	0.00403	0.00105	-0.0284***	-0.0329***
	(0.00914)	(0.00872)	(0.00931)	(0.00847)	(0.00884)
Three	-0.0445**	-0.0333*	0.0293*	-0.0262*	-0.0302**
	(0.0182)	(0.0180)	(0.0156)	(0.0140)	(0.0143)
N=1,095,853					
Year of Test Fixed Effect	Yes	Yes	Yes	Yes	Yes
Student Characteristics	No	Yes	Yes	Yes	Yes
Test Characteristics and School Score	No	No	Yes	Yes	Yes
Student Fixed Effect	No	No	No	Yes	Yes
Date and Week Fixed Effect	No	No	No	No	Yes

Notes: The table presents the estimated coefficients of the effect of two, three, four, or more tests in a week and of two or three tests in a day on national exams' z-score. In column 1, the model includes only the year-fixed effect. In column 2, the model also includes student characteristics: gender, father's and mother's education, and number of siblings. In column 3, the model also includes test characteristics: test order, number of days before the test, dummy indicator for exam's code (an exam fixed effect), dummy indicator for the days of the week, and the exam's specific school z-score. In column 4, the model includes the student fixed effect. In column 5, the date (day-month) and the order of the week in the year are entered. Standard errors clustered by school X exam's code (in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: The Effect of the Number of Weekly and Daily Tests on a National Exams' Score

	(1)	(2)	(3)	(4)	(5)
Weekly Tests					
Two - On Different Days	0.0288***	0.0192***	-0.00756	-0.0109**	-0.00977**
	(0.00707)	(0.00663)	(0.00477)	(0.00453)	(0.00497)
Two - On Same Day	0.0232**	0.0395***	-0.00930	-0.0194***	-0.0214***
	(0.00973)	(0.00923)	(0.00780)	(0.00729)	(0.00760)
Three Plus - On Different Days	0.0378*	0.0295	-0.0530***	-0.0633***	-0.0647***
	(0.0197)	(0.0188)	(0.0152)	(0.0155)	(0.0157)
Three Plus - Two On Same Day	0.0372***	0.0328***	-0.00446	-0.0300***	-0.0339***
	(0.00924)	(0.00873)	(0.00683)	(0.00664)	(0.00707)
Three Plus - Three On Same Day	0.0480***	0.0378***	0.0309***	-0.0364***	-0.0449***
	(0.0148)	(0.0144)	(0.0115)	(0.0109)	(0.0111)
N=1,095,853					
Year of Test Fixed Effect	Yes	Yes	Yes	Yes	Yes
Student Characteristics	No	Yes	Yes	Yes	Yes
Test Characteristics and School Score	No	No	Yes	Yes	Yes
Student Fixed Effect	No	No	No	Yes	Yes
Date and Week Fixed Effect	No	No	No	No	Yes

Notes: The table presents the estimated coefficients of the effect of two, three, four, or more tests in a week and of two or three tests in a day on national exams' z-score. In column 1, the model includes only the year-fixed effect. In column 2, the model also includes student characteristics: gender, father's and mother's education, and number of siblings. In column 3, the model also includes test characteristics: test order, number of days before the test, dummy indicator for exam's code (an exam fixed effect), dummy indicator for the days of the week, and the exam's specific school z-score. In column 4, the model includes the student fixed effect. In column 5, the date (day-month) and the order of the week in the year are entered. Standard errors clustered by school X exam's code (in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: The Effect of Placebo Treatment: Randomizing Within Students the Treatment Indicators

	(1)	(2)	(3)	(4)	(5)
Weekly Tests					
Two - On Different Days	0.0444***	0.0349***	0.00666**	0.00218	0.00162
	(0.00370)	(0.00350)	(0.00265)	(0.00211)	(0.00209)
Two - On Same Day	-0.0475***	-0.0313***	-0.0291***	0.00219	0.00224
	(0.00508)	(0.00483)	(0.00346)	(0.00251)	(0.00251)
Three Plus - On Different Days	0.0744***	0.0666***	-0.000587	0.00454	0.00395
	(0.0137)	(0.0130)	(0.0100)	(0.00888)	(0.00889)
Three Plus - Two On Same Day					
	0.0601***	0.0557***	0.00719**	-0.00143	-0.00164
	(0.00464)	(0.00441)	(0.00324)	(0.00238)	(0.00237)
Three Plus - Three On Same Day					
	0.152***	0.142***	0.0523***	0.00572	0.00600
	(0.00855)	(0.00830)	(0.00545)	(0.00414)	(0.00413)
N=1,095,853					
Year of Test Fixed Effect	Yes	Yes	Yes	Yes	Yes
Student Characteristics	No	Yes	Yes	Yes	Yes
Test Characteristics and School Score	No	No	Yes	Yes	Yes
Student Fixed Effect	No	No	No	Yes	Yes
Date and Week Fixed Effect	No	No	No	No	Yes

Notes: The table presents the estimated coefficients of the effect of two, three, four, or more tests in a week and of two or three tests in a day on national exams' z-score. In column 1, the model includes only the year-fixed effect. In column 2, the model also includes student characteristics: gender, father's and mother's education, and number of siblings. In column 3, the model also includes test characteristics: test order, number of days before the test, dummy indicator for exam's code (an exam fixed effect), dummy indicator for the days of the week, and the exam's specific school z-score. In column 4, the model includes the student fixed effect. In column 5, the date (day-month) and the order of the week in the year are entered. Standard errors clustered by school X exam's code (in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: The Effect of the Number of Weekly and Daily Exams on National Exams' Scores, Samples By Ability

	(1)	(2)	(3)	(4)	(5)
A. Below Median (N=543,515)					
Two - On Different Days	-0.00855	-0.0120	-0.0200***	-0.0203***	-0.0196***
	(0.00764)	(0.00739)	(0.00601)	(0.00573)	(0.00630)
Two - On Same Day	0.136***	0.146***	0.00330	-0.00528	-0.00736
	(0.0115)	(0.0114)	(0.00973)	(0.00922)	(0.00953)
Three Plus - On Different Days	0.0134	-0.000899	-0.0655***	-0.0624***	-0.0584***
	(0.0250)	(0.0246)	(0.0217)	(0.0222)	(0.0224)
Three Plus - Two On Same Day	0.0548***	0.0542***	0.00153	-0.0260***	-0.0254***
	(0.0116)	(0.0115)	(0.00881)	(0.00849)	(0.00888)
Three Plus - Three On Same Day	0.0197	0.0107	0.0565***	-0.0117	-0.0191
	(0.0206)	(0.0204)	(0.0160)	(0.0148)	(0.0151)
B. Above Median (N=552,338)					
Two - On Different Days	0.00393	0.00117	-0.00151	-0.00513	-0.00470
	(0.00612)	(0.00600)	(0.00490)	(0.00484)	(0.00536)
Two - On Same Day	-0.0809***	-0.0755***	-0.0322***	-0.0348***	-0.0362***
	(0.00883)	(0.00871)	(0.00796)	(0.00776)	(0.00815)
Three Plus - On Different Days	-0.0502***	-0.0619***	-0.0459***	-0.0645***	-0.0720***
	(0.0174)	(0.0173)	(0.0148)	(0.0164)	(0.0167)
Three Plus - Two On Same Day	-0.0662***	-0.0669***	-0.0227***	-0.0369***	-0.0456***
	(0.00789)	(0.00781)	(0.00674)	(0.00696)	(0.00761)
Three Plus - Three On Same Day	-0.0767***	-0.0847***	-0.00191	-0.0512***	-0.0611***
	(0.0121)	(0.0122)	(0.0113)	(0.0111)	(0.0115)
Year of Test Fixed Effect	Yes	Yes	Yes	Yes	Yes
Student Characteristics	No	Yes	Yes	Yes	Yes
Test Characteristics and School Score	No	No	Yes	Yes	Yes
Student Fixed Effect	No	No	No	Yes	Yes
Date and Week Fixed Effect	No	No	No	No	Yes

Notes: The table presents the estimated coefficients of the effect of two, three, four, or more tests in a week and of two or three tests in a day on national exams' z-score. In column 1, the model includes only the year-fixed effect. In column 2, the model also includes student characteristics: gender, father's and mother's education, and number of siblings. In column 3, the model also includes test characteristics: test order, number of days before the test, dummy indicator for exam's code (an exam fixed effect), dummy indicator for the days of the week, and the exam's specific school z-score. In column 4, the model includes the student fixed effect. In column 5, the date (day-month) and the order of the week in the year are entered. Standard errors clustered by school X exam's code (in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: The Effect of the Number of Weekly and Daily Exams on National Exams' Score, by Gender

	(1)	(2)	(3)	(4)	(5)
A. Girls (N=631,393)					
Two - On Different Days	0.0284***	0.0220***	-0.00906*	-0.0119**	-0.00571
	(0.00778)	(0.00739)	(0.00540)	(0.00525)	(0.00581)
Two - On Same Day	-0.00603	0.00929	-0.00999	-0.0245***	-0.0243***
	(0.0109)	(0.0104)	(0.00876)	(0.00844)	(0.00882)
Three Plus - On Different Days	0.0786***	0.0507**	-0.0298*	-0.0389**	-0.0334*
	(0.0240)	(0.0233)	(0.0172)	(0.0189)	(0.0192)
Three Plus - Two On Same Day	0.0395***	0.0412***	-0.00253	-0.0245***	-0.0245***
	(0.0102)	(0.00975)	(0.00765)	(0.00784)	(0.00827)
Three Plus - Three On Same Day	0.0131	-0.00260	0.0456***	-0.0197	-0.0262*
	(0.0160)	(0.0175)	(0.0128)	(0.0133)	(0.0136)
B. Boys (N=464,460)					
Two - On Different Days	0.0274***	0.0158*	-0.000454	-0.00346	-0.00394
	(0.00920)	(0.00851)	(0.00632)	(0.00592)	(0.00645)
Two - On Same Day	0.0690***	0.0763***	-0.0137	-0.0138	-0.0180*
	(0.0119)	(0.0112)	(0.0104)	(0.00969)	(0.0101)
Three Plus - On Different Days	0.0617**	0.0252	-0.0509**	-0.0604***	-0.0645***
	(0.0240)	(0.0230)	(0.0200)	(0.0198)	(0.0200)
Three Plus - Two On Same Day	0.0381***	0.0221**	0.000571	-0.0206**	-0.0243***
	(0.0114)	(0.0108)	(0.00916)	(0.00863)	(0.00927)
Three Plus - Three On Same Day	0.109***	0.0729***	0.0163	-0.0429***	-0.0521***
	(0.0172)	(0.0164)	(0.0147)	(0.0132)	(0.0135)
Year of Test Fixed Effect	Yes	Yes	Yes	Yes	Yes
Student Characteristics	No	Yes	Yes	Yes	Yes
Test Characteristics and School Score	No	No	Yes	Yes	Yes
Student Fixed Effect	No	No	No	Yes	Yes
Date and Week Fixed Effect	No	No	No	No	Yes

Notes: The table presents the estimated coefficients of the effect of an additional weekly or daily exam on national exams' z-score. In column 1, the model includes only the year-fixed effect. In column 2, the model also includes student characteristics: gender, father's and mother's education, and number of siblings. In column 3, the model also includes test characteristics: test order, number of days before the test, a dummy indicator for the exam's code (an exam fixed effect), a dummy indicator for the days of the week, and the exam's specific school z-score. In column 4, the model includes a student-fixed effect. Standard errors clustered by school X exam's code (in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: The Effect of the Number of Weekly and Daily Exams on National Exams' Score, by Immigrant Status, Full Sample and Samples by Ability, and Gender

	Full Sample	Ability Below Median	Ability Above Median	Girls	Boys
	(1)	(2)	(3)	(4)	(5)
Weekly Tests					
Two - On Different Days	-0.00974*	-0.0190***	-0.00524	-0.00595	-0.00345
	(0.00501)	(0.00634)	(0.00539)	(0.00585)	(0.00650)
Two - On Different Days X Immigrant	0.0135	-0.0110	0.0415	0.0258	-0.0148
	(0.0233)	(0.0321)	(0.0272)	(0.0279)	(0.0358)
Two - On Same Day	-0.0170**	-0.000740	-0.0339***	-0.0202**	-0.0129
	(0.00765)	(0.00958)	(0.00820)	(0.00889)	(0.0101)
Two - On Same Day X Immigrant	-0.232***	-0.302***	-0.138***	-0.197***	-0.301***
	(0.0282)	(0.0379)	(0.0315)	(0.0328)	(0.0423)
Three Plus - On Different Days	-0.0634***	-0.0569**	-0.0713***	-0.0328*	-0.0628***
	(0.0160)	(0.0228)	(0.0170)	(0.0196)	(0.0203)
Three Plus - On Different Days X Immigrant	-0.0575	-0.0558	-0.0370	-0.0239	-0.0949
	(0.0696)	(0.106)	(0.0811)	(0.0817)	(0.101)
Three Plus - Two On Same Day	-0.0324***	-0.0235***	-0.0446***	-0.0231***	-0.0227**
	(0.00711)	(0.00892)	(0.00764)	(0.00830)	(0.00931)
Three Plus - Two On Same Day X Immigrant	-0.0760***	-0.0910**	-0.0689**	-0.0705**	-0.0906*
	(0.0295)	(0.0409)	(0.0337)	(0.0332)	(0.0482)
Three Plus - Three On Same Day	-0.0406***	-0.0124	-0.0588***	-0.0205	-0.0491***
	(0.0112)	(0.0151)	(0.0116)	(0.0137)	(0.0135)
Three Plus - Three On Same Day X Immigrant	-0.293***	-0.380***	-0.194***	-0.336***	-0.232***
	(0.0487)	(0.0648)	(0.0633)	(0.0653)	(0.0669)

Notes: The table presents the estimated coefficients of the effect of an additional weekly or daily exam on an external exams' z-score, excluding the last exam day. All regressions include: year fixed effect, student characteristics (gender, father's and mother's education, and number of siblings), test characteristics (test order, number of days before the test, dummy indicator for the exam's code (an exam fixed effect), dummy indicator for the days of the week, and the exam's specific school z-score), student-fixed effect, and date and week-fixed effects. Standard errors clustered by school X exam's code (in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: The Effect of the Number of Weekly and Daily Exams on National Exams' Score, by STEM, in Full Sample and samples by Ability, and Gender

	Full Sample	Ability Below Median	Ability Above Median	Girls	Boys
	(1)	(2)	(3)	(4)	(5)
Weekly Tests					
Two - Different Days	-0.00874	-0.0200***	-0.00216	-0.00356	-0.00424
	(0.00542)	(0.00670)	(0.00593)	(0.00631)	(0.00700)
Two - Different Days X STEM	-0.00761	0.00145	-0.0150	-0.0144	-0.00131
	(0.0114)	(0.0161)	(0.0114)	(0.0132)	(0.0149)
Two - Same Day	-0.0117	-0.00148	-0.0243**	-0.0113	-0.00888
	(0.00943)	(0.0115)	(0.0109)	(0.0113)	(0.0134)
Two - Same Day X STEM	-0.0276*	-0.0129	-0.0370**	-0.0415**	-0.0204
	(0.0158)	(0.0204)	(0.0164)	(0.0182)	(0.0207)
Three Plus - Different Days	-0.0638***	-0.0468**	-0.0813***	-0.0405**	-0.0532**
	(0.0161)	(0.0230)	(0.0175)	(0.0196)	(0.0212)
Three Plus - Different Days X STEM	-0.00753	-0.0418	0.0191	0.0150	-0.0364
	(0.0337)	(0.0502)	(0.0336)	(0.0394)	(0.0422)
Three Plus - Two Same Day	-0.0227***	-0.0251**	-0.0256***	-0.00499	-0.0203*
	(0.00801)	(0.00981)	(0.00882)	(0.00944)	(0.0108)
Three Plus - Two Same Day X STEM	-0.0373***	-0.00585	-0.0597***	-0.0610***	-0.0192
	(0.0142)	(0.0192)	(0.0143)	(0.0164)	(0.0179)
Three Plus - Three Same Day	-0.0206	0.00372	-0.0351***	-0.0130	-0.0206
	(0.0127)	(0.0171)	(0.0134)	(0.0159)	(0.0157)
Three Plus - Three Same Day X STEM	-0.0693***	-0.0654**	-0.0754***	-0.0492*	-0.0691***
	(0.0220)	(0.0317)	(0.0216)	(0.0266)	(0.0267)

Notes: The table presents the estimated coefficients of the effect of an additional weekly or daily exam on an external exams' z-score, excluding the last exam day. All regressions include: year fixed effect, student characteristics (gender, father's and mother's education, and number of siblings), test characteristics (test order, number of days before the test, dummy indicator for the exam's code (an exam fixed effect), dummy indicator for the days of the week, and the exam's specific school z-score), student-fixed effect, and date and week-fixed effects. Standard errors clustered by school X exam's code (in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 9: The Effect of the Number of Weekly and Daily Tests on a National Exams' Score, Including Exam Order in a Day

	Full Sample	Ability Below Median	Ability Above Median	Girls	Boys
	(1)	(2)	(3)	(4)	(5)
Weekly Tests					
Two - On Different Days	-0.00894*	-0.0183***	-0.00418	-0.00488	-0.00297
	(0.00497)	(0.00629)	(0.00535)	(0.00580)	(0.00645)
Two - On Same Day	0.0167**	0.0305***	-0.00101	0.0198**	0.0176*
	(0.00787)	(0.0101)	(0.00852)	(0.00936)	(0.0105)
Three Plus - On Different Days	-0.0614***	-0.0552**	-0.0688***	-0.0303	-0.0613***
	(0.0157)	(0.0224)	(0.0166)	(0.0192)	(0.0200)
Three Plus - Two On Same Day	-0.00768	0.00135	-0.0220***	0.00692	-0.000928
	(0.00725)	(0.00918)	(0.00784)	(0.00854)	(0.00959)
Three Plus - Three On Same Day	0.0150	0.0502***	-0.0124	0.0398***	0.00595
	(0.0120)	(0.0164)	(0.0125)	(0.0151)	(0.0145)
Second Test in a Day	-0.0812***	-0.0822***	-0.0731***	-0.0893***	-0.0801***
	(0.00861)	(0.0108)	(0.00920)	(0.0103)	(0.0109)
Third Test in a Day	-0.144***	-0.174***	-0.112***	-0.158***	-0.140***
	(0.0174)	(0.0245)	(0.0178)	(0.0206)	(0.0217)
N	**1,095,853**	**543,515**	**552,338**	**631,393**	**464,460**

Notes: The table presents the estimated coefficients of the effect of an additional weekly or daily exam on an external exams' z-score, excluding the last exam day. All regressions include: year fixed effect, student characteristics (gender, father's and mother's education, and number of siblings), test characteristics (test order, number of days before the test, dummy indicator for the exam's code (an exam fixed effect), dummy indicator for the days of the week, and the exam's specific school z-score), student-fixed effect, date and week-fixed effects, and an indicator for the second exam in a day and an indicator for the third exam in a day. Standard errors clustered by school X exam's code (in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 10: The Effect of the Number of Weekly and Daily Tests on Cheating in the Exam, Full Sample

	(1)	(2)	(3)	(4)	(5)
More Than One Exam in a Week	0.00110***	0.00118***	0.000967***	0.000949***	0.000726*
	(0.000318)	(0.000317)	(0.000334)	(0.000357)	(0.000411)
More Than One Exam in a Day	0.00332***	0.00317***	0.00161***	0.00167***	0.000598
	(0.000517)	(0.000513)	(0.000587)	(0.000589)	(0.000787)
N=619,263					
Year of Test Fixed Effect	Yes	Yes	Yes	Yes	Yes
Student Characteristics	No	Yes	Yes	Yes	Yes
Test Characteristics and School Score	No	No	Yes	Yes	Yes
Student Fixed Effect	No	No	No	Yes	Yes
Date and Week Fixed Effect	No	No	No	No	Yes

Notes: The table presents the estimated coefficients of the effect of more than one exam in a week and more than one exam in a day on cheating in the exam. In column 1, the model includes only the year-fixed effect. In column 2, the model also includes student characteristics: gender, father's and mother's education, and number of siblings. In column 3, the model also includes test characteristics: test order, number of days before the test, dummy indicator for exam's code (an exam fixed effect), dummy indicator for the days of the week, and the exam's specific school z-score. In column 4, the model includes the student fixed effect. In column 5, the date (day-month) and the order of the week in the year are entered. Standard errors clustered by school X exam's code (in parentheses).* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 11: The Effect of the Number of Weekly and Daily Tests on Cheating in the Exam, Samples By Student's Ability and Gender

	Ability: Below Median	Ability: Above Median	Girls	Boys
	(1)	(2)	(3)	(4)
More Than One Exam in a Week	0.000533	0.000665*	0.000464	0.00100*
	(0.000629)	(0.000363)	(0.000461)	(0.000547)
More Than One Exam in a Day	0.00142	-0.000596	0.000519	0.000707
	(0.00124)	(0.000548)	(0.000872)	(0.00106)
N	308,120	311,143	361,936	257,327

Notes: The table presents the estimated coefficients of the effect of more than one exam in a week and more than one exam in a day on cheating in the exam. The model includes test characteristics: test order, number of days before the test, dummy indicator for the days of the week, the exam's specific school z-score, student-fixed effect, and date and week-fixed effects. Standard errors clustered by school X exam's code (in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

ONLINE APPENDIX

NOT FOR PUBLICATION

Table A1: Four Examples of a Student Exams Schedule, 2000 Exam Period

Student ID	Subject	Exam Code	Week	Date	Day of Week	Start Time	Duration minutes
1	Accounting	855301	1	5/8/2000	Monday	9:00 AM	180
1	Hebrew	905031	1	5/11/2000	Thursday	2:30 PM	180
1	Bible Studies	5102	3	5/21/2000	Sunday	9:00 AM	90
1	Civil Studies	913051	4	5/29/2000	Monday	2:30 PM	90
1	Bible	5103	5	6/4/2000	Sunday	9:00 AM	75
1	English	908643	6	6/13/2000	Tuesday	2:30 PM	180
1	History	913091	6	6/16/2000	Friday	9:00 AM	90
1	Literature	9112	7	6/18/2000	Sunday	4:30 PM	75
1	Mathematics	35204	7	6/21/2000	Wednesday	########	180
1	Bible	2213	8	6/25/2000	Sunday	9:00 AM	150
1	Hebrew	11101	8	6/30/2000	Friday	9:00 AM	150
2	Chemistry	918651	1	5/8/2000	Monday	9:00 AM	150
2	Chemistry	37201	1	5/8/2000	Monday	12:00 PM	150
2	Hebrew	905031	1	5/11/2000	Thursday	2:30 PM	180
2	Civil Studies	913051	4	5/29/2000	Monday	2:30 PM	90
2	Bible Studies	5105	5	6/4/2000	Sunday	########	75
2	English	908653	6	6/13/2000	Tuesday	2:30 PM	180
2	History	913091	6	6/16/2000	Friday	9:00 AM	90
2	Mathematics	35101	7	6/21/2000	Wednesday	12:30 PM	75
2	Bible	2212	8	6/25/2000	Sunday	9:00 AM	150
3	Hebrew	905031	1	5/11/2000	Thursday	2:30 PM	180
3	Arabic	910503	2	5/14/2000	Sunday	9:00 AM	165
3	History	913061	2	5/16/2000	Tuesday	2:30 PM	120
3	Sociology	931291	2	5/18/2000	Thursday	9:00 AM	120
3	Civil Studies	913051	4	5/29/2000	Monday	2:30 PM	90
3	English	908643	6	6/13/2000	Tuesday	2:30 PM	180
3	Mathematics	35302	7	6/21/2000	Wednesday	9:00 AM	180
3	Mathematics	35101	7	6/21/2000	Wednesday	12:30 PM	75
3	Bible Studies	900091	8	6/25/2000	Sunday	9:00 AM	135
3	Psychology	69201	9	7/3/2000	Monday	9:00 AM	120
4	Hebrew	905031	1	5/11/2000	Thursday	2:30 PM	180
4	Bible	6104	3	5/21/2000	Sunday	9:00 AM	90
4	Civil Studies	913051	4	5/29/2000	Monday	2:30 PM	90
4	Bible	6001	5	6/4/2000	Sunday	9:00 AM	180
4	Physics	917531	5	6/7/2000	Wednesday	9:00 AM	90
4	Physics	917521	5	6/7/2000	Wednesday	########	90
4	Physics	917551	5	6/7/2000	Wednesday	1:00 PM	90
4	English	908653	6	6/13/2000	Tuesday	2:30 PM	180
4	Mathematics	35302	7	6/21/2000	Wednesday	9:00 AM	180
4	Mathematics	35101	7	6/21/2000	Wednesday	12:30 PM	75
4	Bible Studies	2212	8	6/25/2000	Sunday	9:00 AM	150
4	Bible Studies	38202	10	7/9/2000	Sunday	9:00 AM	120

Notes: The table presents a schedule of exams for four students in 2000. The table links each exam to its subject and individual code, the week and date it was administered, and its start time and duration.

Table A2: The Distribution of the Treatments by the Weeks of Summer Exam Periods

	The Ordinal Number of the Week in the Exam Period								
Weekly Tests	1	2	3	4	5	6	7	8	9
One	80.07	67.45	57.68	44.45	27.77	52.51	31.62	46.94	45.46
Two - On Different Days	14.58	24.53	20.20	10.47	12.93	27.63	15.36	36.18	28.81
Two - On Same Day	0.99	5.54	12.61	21.02	27.08	5.32	9.68	3.33	0.96
Three Plus - On Different Days	0.75	0.87	1.41	0.13	0.57	0.47	1.92	0.88	0.37
Three Plus - Two On Same Day	3.61	1.57	7.52	10.17	22.03	8.76	38.16	9.11	22.20
Three Plus - Three On Same Day	0.00	0.04	0.58	13.75	9.62	5.30	3.26	3.56	2.20

The table presents the distribution of the six treatment variables by the ordinal number of the week in the 2000–2005 exam periods. The distribution is based on the observations included in the regression analysis.

Table A3: The Distribution of the Treatments by the Days of the Week

	Day of Week					
Weekly Tests	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday
One	46.50	44.88	41.56	42.63	59.23	64.53
Two - On Different Days	28.99	10.82	24.14	16.49	14.39	25.06
Two - On Same Day	4.87	26.69	13.19	8.23	8.33	2.36
Three Plus - On Different Days	1.14	0.36	1.56	0.45	0.68	2.45
Three Plus - Two On Same Day	16.26	9.63	12.83	21.84	15.42	5.57
Three Plus - Three On Same Day	2.24	7.60	6.72	10.36	1.96	0.03

The table presents the distribution of the six treatment variables by the days of the week. The distribution is based on the observations included in the regression analysis.

Table A4: The Effect of the Number of Weekly and Daily Tests on a National Exams' Score, By Sample Exclusion Restrictions

	Full Sample (copied from column 5 in Table 3)	Excluding Students Who Skipped One Exam	Excluding Students Who Retook an Exam Taken Earlier in 12th Grade	Excluding Students Who Retook an Exam Taken in 10-11th Grade	Excluding Students Who Were Excluded in Columns 2-4
	(1)	(2)	(3)	(4)	(5)
Weekly Tests					
Two - On Different Days	-0.00977**	-0.00938*	-0.00996**	-0.00869*	-0.00787
	(0.00497)	(0.00503)	(0.00503)	(0.00512)	(0.00522)
Two - On Same Day	-0.0214***	-0.0258***	-0.0197**	-0.0228***	-0.0264***
	(0.00760)	(0.00769)	(0.00775)	(0.00795)	(0.00812)
Three Plus - On Different Days	-0.0647***	-0.0649***	-0.0633***	-0.0631***	-0.0623***
	(0.0157)	(0.0158)	(0.0159)	(0.0162)	(0.0163)
Three Plus - Two On Same Day	-0.0339***	-0.0369***	-0.0316***	-0.0335***	-0.0346***
	(0.00707)	(0.00712)	(0.00716)	(0.00726)	(0.00736)
Three Plus - Three On Same Day	-0.0449***	-0.0485***	-0.0387***	-0.0373***	-0.0366***
	(0.0111)	(0.0112)	(0.0117)	(0.0117)	(0.0122)
N	**1,095,853**	**1,067,769**	**1,056,616**	**997,216**	**954,472**

Notes: The table presents the estimated coefficients of the effect of an additional weekly or daily exam on a national exam's z-score. All regressions include: year fixed effect, student characteristics (gender, father's and mother's education, and number of siblings), test characteristics (test order, number of days before the test, dummy indicator for the exam's code (an exam fixed effect), dummy indicator for the days of the week, and the exam's specific school z-score), student-fixed effect, and date and week-fixed effects. Standard errors clustered by school X exam's code (in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: The Effect of Placebo Treatment, in Full Sample and samples by Ability, and Gender

	Full Sample	Ability Below Median	Ability Above Median	Girls	Boys
	(1)	(2)	(3)	(4)	(5)
Weekly Tests					
Two - On Different Days	0.00162	0.00289	0.000782	0.00161	0.000463
	(0.00209)	(0.00334)	(0.00252)	(0.00259)	(0.00337)
Two - On Same Day	0.00224	0.00357	0.00170	0.00158	0.00435
	(0.00251)	(0.00384)	(0.00306)	(0.00323)	(0.00389)
Three Plus - On Different Days	0.00395	0.0229	-0.0109	-0.00735	0.00771
	(0.00889)	(0.0152)	(0.0106)	(0.0140)	(0.0117)
Three Plus - Two On Same Day	-0.00164	-0.00230	-0.00134	-0.00413	-0.000331
	(0.00237)	(0.00398)	(0.00281)	(0.00297)	(0.00389)
Three Plus - Three On Same Day	0.00600	0.00974	0.00220	0.00135	0.00907
	(0.00413)	(0.00728)	(0.00457)	(0.00602)	(0.00562)
N	**1,095,853**	**543,515**	**552,338**	**631,393**	**464,460**

Notes: The table presents the estimated coefficients of the effect of an additional weekly or daily exam on an external exams' z-score, excluding the last exam day. All regressions include: year fixed effect, student characteristics (gender, father's and mother's education, and number of siblings), test characteristics (test order, number of days before the test, dummy indicator for the exam's code (an exam fixed effect), dummy indicator for the days of the week, and the exam's specific school z-score), student-fixed effect, date and week-fixed effects, and an indicator for the second exam in a day and an indicator for the third exam in a day. Standard errors clustered by school X exam's code (in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: 2000–2002 Early Data, The Effect of the Number of Weekly and Daily Tests on a National Exams' Score, Full Sample

	(1)	(2)	(3)	(4)	(5)
Weekly Tests					
Two - On Different Days	0.0287***	0.0216**	0.00434	-0.0128**	-0.0111*
	(0.00925)	(0.00862)	(0.00650)	(0.00604)	(0.00631)
Two - On Same Day	0.0424***	0.0705***	0.000591	0.00170	0.00433
	(0.0125)	(0.0119)	(0.0108)	(0.0101)	(0.0104)
Three Plus - On Different Days	0.0274	0.0280	-0.0324*	-0.0495***	-0.0432**
	(0.0267)	(0.0248)	(0.0181)	(0.0178)	(0.0179)
Three Plus - Two On Same Day	0.0469***	0.0470***	0.00640	-0.0323***	-0.0299***
	(0.0117)	(0.0111)	(0.00927)	(0.00884)	(0.00906)
Three Plus - Three On Same Day	-0.0383*	-0.0436**	-0.00684	-0.0785***	-0.0787***
	(0.0198)	(0.0194)	(0.0187)	(0.0162)	(0.0165)
N=619,263					
Year of Test Fixed Effect	Yes	Yes	Yes	Yes	Yes
Student Characteristics	No	Yes	Yes	Yes	Yes
Test Characteristics and School Score	No	No	Yes	Yes	Yes
Student Fixed Effect	No	No	No	Yes	Yes
Date and Week Fixed Effect	No	No	No	No	Yes

Notes: The table presents the estimated coefficients of the effect of two, three, four, or more tests in a week and of two or three tests in a day on national exams' z-score. In column 1, the model includes only the year-fixed effect. In column 2, the model also includes student characteristics: gender, father's and mother's education, and number of siblings. In column 3, the model also includes test characteristics: test order, number of days before the test, dummy indicator for the days of the week, and the exam's specific school z-score. In column 4, the model includes the student fixed effect. In column 5, the date (day-month) and the order of the week in the year are entered. Standard errors clustered by school X exam's code (in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.