The effects of supplementary videos on learning in intermediate microeconomics: estimates from a field experiment.

Melissa Famulari and Zachary A. Goodman*

University of California, San Diego

This version: December 2022

Abstract

We estimate the effectiveness of videos to improve learning outcomes in intermediate microeconomics. In a field experiment involving about 400 students, who revealed

*mfamulari@ucsd.edu and zgoodman@ucsd.edu. The authors thank the students who took intermediate microeconomics in the fall of 2018 and 2019 for participating in this study. We thank UC San Diego's Teaching + Learning Commons for providing the anonymized data used for analysis. We gratefully recognize Jessica Morales, Yingjia Zhang, and Shuli Zhu for outstanding research assistance. We thank the applied microeconomics group at UC San Diego for their help with the experimental design. We thank Julian Betts, Julie Cullen, Gordon Dahl, Simone Galperti, Craig McIntosh, Katherine Meckel and Jennifer Murdock for their suggestions which greatly improved the paper. This research was approved under UC San Diego's Human Research Protections Program (IRB approval 170886 in fall 2018 and 2019). The paper investigates the use of Intermediate Microeconomics Video Handbook (IMVH) video lectures by UC San Diego students, some of which were developed by one of the authors, in collaboration with UC San Diego and the UC Office of the President. UC San Diego currently owns the rights to distribute the IMVH. The videos were provided to the students at no charge and neither author has a direct financial interest in the distribution of the IMVH within the University of California. As of fall 2020, one of the authors has a financial interest in the distribution of the IMVH outside of the University of California.

themselves as poor performers on an early assessment, we randomly assigned a grade-based incentive that induced treatment students to watch over 60% more videos than did control students. We observe significant reduced form effects: being assigned treatment caused students to score 0.18 standard deviations higher on midterm and final exams. Using an instrumental variables approach, we estimate that the marginal hour of video content increased exam scores between 0.05 and 0.16 standard deviations. We rule out large negative spillover effects to other courses taken concurrently, and we observe persistent take-up of video watching in the subsequent quarter when exogenous incentives to watch videos were removed.

1 Introduction

"You expect me to read the textbook? Ha!"

– Anonymous student

Every year, university students spend tens of thousands of dollars on tuition and course materials and hundreds of hours studying, in large part, to learn. Instructors can help their students learn more efficiently by providing and recommending pedagogical tools that have high returns per unit time and financial cost. Despite the value to students and instructors, little empirical work exists that estimates the effectiveness of different learning technologies (Allgood, Walstad, & Siegfried, 2015).

We measure the impacts of one such technology, instructional videos, on outcomes in intermediate microeconomics. This research is timely as, after two years of remote instruction due to the Covid-19 pandemic, most educators have numerous videos they could make available to future students as a supplemental resource and the question is whether they should. On the one hand, modern students, who have had unprecedented exposure to electronic media, may find videos more engaging and, perhaps, more effective at building human capital than conventional texts. Further, videos have near-zero marginal cost and are accessible anywhere via the internet, helping reduce financial and geographic barriers to high-quality educational materials. Finally, the perceived low cost of watching a video may be easier to overcome than perceived higher costs of other learning methods, potentially leading to

more frequent studying, which decades of psychological research has demonstrated leads to more long-term learning than does cramming (Cepeda et al., 2006; Kornell, 2009). On the other hand, students may substitute video watching for more effective learning methods and videos enable students to both procrastinate and cram.

Ultimately, the beneficial features of video-based learning tools are of value only if they can improve student learning outcomes, an empirical question we seek to answer in this paper. The learning videos we study are short, were professionally produced and were organized into a handbook, called the Intermediate Microeconomics Video Handbook (IMVH). To estimate the effects the IMVH, we administered a field experiment involving nearly 400 undergraduates enrolled in the same one-quarter-long microeconomics course over two years. Of note, these students all scored below the median on the first midterm exam, which suggests their study habits could be improved, and perhaps standing to gain the most from an intervention that targets studying. We randomly assigned a grade-based incentive to half of these lower-performing students to encourage take-up of the videos, which were made available to all students in the class, allowing identification of intent to treat (ITT) effects and local average treatment effects (LATEs) while maintaining equitable access to learning resources. We tracked video watching at the student level using the software platform that hosts the videos. We observe grades, GPA, and video watching in both the term of the experiment and the subsequent term when students typically take the second intermediate microeconomics class in the sequence.

The first-stage impact of the exogenous encouragement on video watching is significant and substantial. Students who receive the grade-based incentive watch over 28% more unique videos by the second midterm and 63% more unique videos by the final exam, or about 1.1 and 3.4 hours of content, respectively, than did their control peers. We find large reduced-form effects of treatment on exam scores: for students in the bottom half of the class as of the first midterm, being assigned treatment (ITT) increases midterm and final exam scores by about 0.18 standard deviations. Our estimates imply that the marginal hour of videos watched increases exam scores (LATE) by between 0.05 and 0.16 standard deviations. Reinterpreting the LATE, we find that it takes between one and three hours of video watching to raise exam grades by one step, e.g., a C minus to a C.

To better understand the spillover effects of treatment, we examine other forms of studying including class attendance, visits to a tutoring center (specific to this course), and interacting with the class discussion board. We do not find any statistically significant changes in any observed studying method, and we can rule out large changes. In nearly all cases, treatment students used other studying methods at directionally *greater* rates than did their control peers. We also investigate spillovers to other courses taken during the term of the experiment and similarly find that treatment students perform directionally *better* than their control peers. Though not statistically significant, we can rule out large negative effects, suggesting that treatment did not cause students to substitute away from studying for other courses.

Although we observe that treated students performed better on course assessments, it is unclear whether treatment increases student utility. An important piece of the welfare puzzle is whether treated students continue to use instructional videos at higher rates after exogenous incentives are removed. Persistent take-up in the absence of external prodding provides some confidence that students, now with updated priors, value the technology. Fortunately, we can observe video watching in the subsequent microeconomics course in the term following the experiment. Despite there being no direct incentives to watch videos in the subsequent course, treatment students persistently watched more videos than did control students, about 8 - 10 more unique videos, or 1.2 - 1.5 more hours of unique content. Our sample in the subsequent term is nearly half the original size, so we lack power to precisely estimate effects on exam scores; however, our confidence intervals include effect sizes consistent with those observed in the experiment term.

Collectively, we interpret our findings as evidence that providing a small grade incentive to watch videos is a net positive on underperforming students' academic achievement. Though formal welfare analysis is beyond the scope of this paper, we present suggestive evidence that incentivizing video-watching increases under-performing student utility, as our results are consistent with a poor-information model of student learning. Our findings justify paternalistic incentive structures in settings where a large portion of the class is at risk of failing and the instructor has more information about the usefulness of a novel teaching technology than do her students.

The rest of the paper is organized as follows. Section 2 provides background on existing related literature. Section 3 describes the study design and the videos used in the experiment. Section 4 presents the results of the experiment, and Section 5 presents competing models of studying behavior that may explain the observed phenomena. Section 6 discusses the contributions and limitations of our study, and Section 7 concludes.

2 Related Literature

Students have many time-consuming activities to help them learn including attending class, watching recorded lectures, reading the textbook, doing homework, completing practice exams, and attending tutoring labs. Several empirical challenges make it difficult to estimate the causal effects of such activities. First, researchers must address nonrandom selection into using study methods as the decision is likely influenced by unobservable student characteristics, such as motivation or ability, that also affect student learning outcomes. Second, empirical evidence suggests there is "dynamic selection" where students change their study strategies in response to a negative exam shock or being close to a letter grade threshold (Oettinger (2002), Krohn and O'Connor (2005), Stinebrickner and Stinebrickner (2008), Bonesrønning and Opstad (2012) and Bonesrønning and Opstad (2015).) Dynamic selection means that including student fixed effects in class performance regressions will not uncover causal effects. Given these empirical challenges, we focus our review on research that uses experiments or quasi-experiments to identify the causal effects of learning methods[1] that take students' time.[2] We organize these studies into two broad groups: *guided study*, where a content expert can immediately answer student questions and modify the presentation of course material, and *self study*, where the student cannot get immediate feedback from

[1]Even well-conducted experiments may not identify the causal effects of a learning method. First, study methods may be substitutes or complements in student learning and experimental inducements to use one study strategy may affect takeup of another. Second, experimental inducements to use a study method may change the total time students devote to a course. Experiments will then identify the causal effects of a study policy and all of the attendant changes in student behavior caused by that policy. While the causal effects of an educational policy are useful for educators considering how to design their classes, they are less useful for students wanting to know the most productive use of their study time.

[2]We do not include research that examines how to make studying methods more productive for a fixed quantity of time, though the intensive margin of pedagogical tools is certainly an important area of study with potentially clearer welfare implications.

an expert.[3] Guided study includes attending lecture, office hours, tutoring labs, discussion sections and supplemental instruction while self-study includes reading the textbook, doing homework/problem sets, taking online quizzes, doing practice exams and watching recorded lectures.

There is experimental evidence on the effectiveness of several guided-study learning methods including attending lectures (Kirby and McElroy (2003), Dobkin, Gil, and Marion (2010), Joyce et al. (2015), Tang et al. (2020)), discussion sections (Arulampalam, Naylor, and Smith (2012), Bratti and Staffolani (2013), Kapoor, Oosterveen, and Webbink (2020)) and peer tutoring labs (Munley, Garvey, and McConnell (2010)). Collectively, these studies consistently find that student performance is improved across all guided-study methods, but only if students do not substitute guided-study time for self-study time.

For the effectiveness of self-study, there is experimental evidence on overall self-study time (Stinebrickner and Stinebrickner (2008), Oreopoulos et al. (2019), Clark et al. (2020)), doing homework (Trost and Salehi-Isfahani (2012), Grodner and Rupp (2013)) and watching supplemental videos (Noetel et al. (2021)). The research on the causal effects of self-study overall and doing homework is mixed and ranges from no significant effect on exam scores to large positive effects. In contrast, Noetel et al. (2021)'s meta analysis reports that student performance is significantly improved when given access to videos with no other change to instruction.[4]

[3]Discussion boards, where students can ask questions and get help from fellow classmates and the instructional staff, blurr the lines between guided and self study.

[4]The average effect size across 34 RCTs where videos were supplementary, primarily conducted in medical education, was .88 standard deviations (Hedges g) with 88 percent of the studies finding a significant positive effect, 12 percent finding no statistically significant effects and no study finding a significantly negative effect. In contrast, across the 83 RCTs cited in Noetel et al. (2021) where videos replaced some aspect of teaching, the effect size on average was .28 standard deviations (Hedges g) with 50 percent finding a significant positive effect, 19 percent finding a significant negative effect, and 31 percent finding no significant effects. In economics, Expósito et al. (2020) find that University of Seville students taking a macroeconomics course in their second year did significantly better on a supplementary exam, which did not affect the student's GPA, when videos were substituted for lecture.

3 Study Design

3.1 Description of the sample and institution

We conducted the field experiment in an undergraduate intermediate microeconomics course taught during fall 2018 and fall 2019 by one of the authors. The university is a large, diverse and selective public research university in the United States.[5] At this institution, intermediate microeconomics is a three-quarter sequence required for students majoring in Economics. The experiment was conducted in the first course of the sequence, *Micro A*. We also observe grades and video watching in the second course of the sequence, *Micro B*, during the winter 2019 and winter 2020 quarters. The same instructor taught Micro B in both years (and was a different instructor than the one who taught Micro A). Both Micro A and B instructors created half of the videos relevant to their course in the IMVH.

The class structure is similar across the Micro sequence. Instructors teach two 80 minute lectures back-to-back on Tuesday and Thursday (e.g., one from 11–12:20pm and the other from 12:30–1:50pm). Students have the option to attend either lecture and lectures are not recorded. Two midterms and the final exam are held at a common time outside of lecture for the two classes. In addition to lecture, students across both classes have access to a common class web site, weekly one-hour discussion sections run by graduate teaching assistants (TAs) who are all Economics PhD candidates, including, at the time, one of the authors. In lieu of office hours, the graduate TAs and Undergraduate Instructional Assistants staff a tutoring lab open between three and four hours per day, six days per week. Students may also attend weekly Supplemental Instruction (SI) sessions offered by undergraduates majoring in Economics and trained by the university in SI. Besides the IMVH, students have access to a variety of online learning resources including a discussion board moderated by the instructional team, four years of previous exam questions, weekly ungraded problem

[5]The Carnegie Classification of Institutions of Higher Education classifies the university as an R1 (very high research activity) university. For the 2017-2018 academic year, the undergraduate student body had the following demographics: 49.1% female, 50.6% male; 75.0% in-state, 5.5% out-of-state, and 19.5% international; 59% students of color; 28.6% majoring in the social sciences, 26% of which major in Economics. Among newly admitted students, about one-third were transfer students. Average SAT scores were 652 and 605 for math and critical reading, respectively. About 34% of students are the first in their family to attend a four-year university.

sets, and semi-weekly graded online quizzes.

Students were told about the experiment during the first lecture, given a printed copy of the consent form in the second lecture, and provided a virtual copy of the consent form in the syllabus on the course webpage. At any time during the quarter, students could opt out of having their data included in the analysis.[6] Students below the age of 18 at the start of the course as well as students enrolled via the university's extension program were removed from the analysis dataset.[7] Ultimately, four students under 18, five extension students, and seven students who opted-out were removed from the analysis dataset, leaving a sample of 850 students.

There are two unique demographic features of the class worth noting. First, many non-economics majors take the class to either satisfy general education requirements or to explore majoring in economics. As there are many students in the experiment on the margin of majoring in economics, an important outcome is the likelihood the student takes Micro B. Second, about 37% of the class is transfer students, most of whom are transferring from a community college. For most transfer students, Micro A is their first experience with upper division coursework, their first class at a four-year research university, and their first time taking classes under the faster-paced quarter system.[8] We examine treatment effect heterogeneity to understand whether transfer students might differentially benefit from the IMVH.

3.2 Description of the Intermediate Microeconomics Video Handbook (IMVH)

The IMVH is a collection of 220 short videos that cover the material in a year-long intermediate microeconomics course sequence.[9] The videos were designed as a complement to

[6]Students could opt out via an online form visible to a third party university organization so that neither the instructor nor research team could observe which students elected to opt out.

[7]Students under the age of 18 were excluded per IRB protocol. We exclude extension students because of their potentially very different preparation for the course and our inability to observe pretreatment covariates and outcomes outside of Micro A.

[8]Community colleges, the most common previous institution for transfer students, are on the semester system in the state of the university.

[9]A preview of the IMVH can be found at https://iti.ucsd.edu/IMVH_Misc/Promo/IMVHPromo.html.

Table 1: Comparison of information transmission formats

Feature	Lecture	eText-book	Lecture Capture	IMVH
Instructor's time used	✓			
Instructor-learner interaction	✓			
Learner-learner interaction	✓			
Readable		✓	?	✓
Scalable	?	✓	✓	✓
Searchable		✓		✓
Skimmable		✓		✓
Stoppable	?	✓	✓	✓
Watchable	✓		✓	✓
Consumed on Demand		✓	After Lecture	✓

both lectures and the course textbook. Each topic typically has two videos, one with the graphical and verbal intuition and the other with the formal algebraic definitions and proofs, identified by having (Calculus) at the end of the video title.

The videos were created by six UC San Diego faculty members with professional videographer and production support. Many videos utilize the "learning glass," an innovative presentation technology where instructors write with neon markers on a large sheet of glass that has lights embedded along the glass edge to make the colors pop. The remaining videos feature faculty superimposed in front of slides that are written on during the presentation. Videos are closed captioned and were checked by graduate students for accuracy. The IMVH was designed to help students find material quickly and so (a) videos can be accessed either via the table of contents or the index (b) for each video, the web interface includes time stamps for each concept covered, and (c) while watching the video, captions are searchable which allows students to jump to the searched-for word.

While we do not know of another textbook completely comprised of videos, the IMVH is similar to the Khan Academy website, lecture capture, and textbook websites that incorporate instructional videos. Table 1 presents a classification of some options to present course material to students.[10] Besides the engaging viewable nature, the IMVH differs from a traditional textbook in that the instructors explain, graph, and derive mathematical results

[10]This table is a slight modification of the classification table Martin Osborne proposed to one of the authors in an email correspondence.

in much the same way one would in a conventional lecture. Compared to lecture, students control the pace of the IMVH (they can rewatch, speed up, or slow down the videos), students can both listen to and read the IMVH due to closed captioning, and the IMVH has better visibility than lectures in large rooms. Students have the option to watch the IMVH videos *before* lecture to prepare, and no recurring demand is placed on the instructor's time. The primary benefits of lecture over the IMVH is that students can receive immediate help when they have questions, student questions may have import externalities for the learning of other students and there is a social aspect as students can interact with each other before, during, and after lecture. Finally, compared to recorded lectures, the IMVH videos are much shorter, averaging under ten minutes, are focused on a single topic and do not include lecture components that often do not work well when recorded, such as group work, student questions and class discussion.

3.3 Experiment Design

The experiment began four weeks into the ten week term following grading the first midterm exam. All students who scored above the median on the first midterm, the *Above median* arm, had a conventional grading scheme that placed weight only on exams and quizzes. For students who scored below the median on the first midterm, we randomly assigned half to the *Control* arm who had a conventional grading scheme that placed weight only on exams and quizzes (the same grading scheme as the students scoring above the median) and half to the *Incentive* arm, whose grading scheme allots four percentage points conditional on watching at least 40 of 48 eligible videos in the IMVH.[11] These 48 videos cover new class content since the first midterm that would be assessed in the second midterm and final exam. All students could still view the 26 videos relevant to the first midterm and, as they could help students on the cumulative final exam, we include them in our measures of video watching despite not counting towards the grade incentive.

The two different grading schemes are outlined in Table 2. Notably, the four percentage

[11] Watched in standard speed, 40 videos would require students to spend between 5.5 and 7.1 hours, depending on the length of videos chosen (on average 9.7 minutes in length each). Watching all 48 incentivized videos in standard speed would require just shy of eight hours.

Table 2: Grade scheme by treatment arm for students who scored below the median on the first midterm. Differences between the two grade schemes in bold.

Assessment	Incentive	Control
>40 videos	**4%**	**0%**
Midterm 1	**18%**	**22%**
Midterm 2	22%	22%
Final Exam	50%	50%
Math Quiz	1%	1%
Best 5 of 6 Quizzes	5%	5%
Total	100%	100%

points for video watching come at the expense of reduced weight placed on the first midterm score, which had already occurred at the time of treatment assignment. Hence, at the time of treatment assignment, the video incentive is the sole forward-looking difference between treatment arms.

To improve balance between *Incentive* and *Control* arms and increase statistical power, we assigned students to treatment arms using paired randomization (Athey & Imbens, 2017), matching students by their first midterm scores before randomly assigning one member of each pair to *Incentive* and the other to *Control* (further details on treatment assignment can be found in Appendix A.1). We emailed each student letting them know their assignment and grading scheme. Students could also find their assignment listed in the online gradebook. To confirm that students knew their assignment, we surveyed students using an in-class attendance quiz, and 94% of students correctly identified their grading scheme. We emailed the students who responded incorrectly to clarify their assignments.[12]

We informed *Incentive* students that they must watch the entire video and only one video at a time to get credit towards their 40 required videos. It is not possible to observe "watching" as students could, for example, minimize their browser, walk away from their computer, or otherwise play a video without actively watching it. As a proxy for watching, we use data recorded by the IMVH software that captures the video ID, student ID, and the date and time when a student opens a video link. We define the following measures:

1. *Videos*: Number of links opened, including duplicates

[12] 11 of 164 *Incentive*, 23 of 167 *Control*, and 10 of 373 *Above median* students did not identify their grading schemes correctly. 146 students did not answer the quiz, several of whom had dropped the course following the first midterm.

2. *Unique videos*: Number of unique video links opened

3. *Hours of videos*: Total runtime of video links opened

4. *Hours of unique videos*: Total runtime of video links opened with duplicates removed

For expositional ease, we use "watching" to refer to the link-opening behavior as defined above. Although video watching in our data is a binary measure, watching behavior can vary greatly in intensity. Some students take notes, pausing and rewatching portions of the video as needed. Other students, we suspect, play videos in the background without absorbing much material. Exploring the intensive margin of video watching remains an area for future research that will benefit from new technologies that can quantify video engagement including interactive content embedded in videos, eye-tracking devices, and more.

We helped students keep track of their progress towards 40 videos by periodically updating the online gradebook with counts determined from the IMVH data. Although nearly all students followed our instructions to watch videos completely and sequentially,[13] in a few cases, students opened 40 or more video links within a few minutes. We manually adjusted their video counts in the gradebook and emailed them a reminder of the requirements for videos to count towards the grade incentive.[14] Though it is doubtful these strategic students gained much from opening so many videos so quickly, to maintain interpretability of our results, we do *not* remove these clicks from our video count measures.[15] Our *unique* video measures, however, are less sensitive to this behavior.

To ensure fairness, we informed students that final letter grades would *not* be affected by being in the experiment. We accomplished parity between *Control* and *Incentive* arms through curving final letter grades. First, we applied a curve to the *Control* and *Above median* arms as one group to achieve a final grade distribution in line with that of previous cohorts. Second, we curved the *Incentive* students' letter grades to match the grade distribution among the *Control* students.[16]

[13]The time between timestamps for video links opened in succession was almost always longer than the runtime of the video.

[14]Although additional email communication as a result of treatment could violate the exclusion restriction, the small number of affected students is unlikely to have much (if any) influence on our results.

[15]Our causal effects are per "links opened" rather than per "links opened subject to certain qualifying conditions". The cost of this interpretability is likely downward bias on our causal effect estimates.

[16]For example, assume the top 10 percent of the control group received an A final grade and the next 12

3.4 Empirical strategies

We estimate the effect of being assigned to the *Incentive* arm on our outcome variables of interest, Intent To Treat (ITT) effects, as well as the effect of watching videos for those induced by the incentive to watch more videos, a Local Average Treatment Effect (LATE) (Imbens & Rubin, 2015). We explore whether there is treatment effect heterogeneity across key demographic variables. Below we outline the empirical strategies for estimating each of these effects.

3.4.1 Intention To Treat (ITT)

Our experiment has two-sided non-compliance: some students in the treatment arm do not watch videos and some students in the control arm watch videos. Therefore, the causal effect of being assigned to the *Incentive* arm on outcomes of interest, such as exam scores, are ITT estimates and not average treatment effect estimates. Since the incentive itself in our setting is representative of how future instructors may induce their students to watch videos, the ITT estimates are policy-relevant for instructors considering a similar grade incentive to encourage video-delivered learning methods in their courses.

Our baseline ITT specification is the partially linear model:

$$Y_i = \beta Z_i + f(X_i) + \epsilon_i \tag{1}$$

where Y_i is an outcome of interest (e.g. videos watched or test scores) for student i, $Z_i \in \{0, 1\}$ is a treatment indicator with those in the *Control* arm having $Z_i = 0$ and those in the *Incentive* arm having $Z_i = 1$, $f()$ is a generic function through which X_i, a vector of controls, affects Y_i, and ϵ_i is an unobserved residual. β, our parameter of interest, is the causal effect of being assigned to the *Incentive* arm on the outcome of interest Y, assumed to be constant across the population.[17] Under unconfoundedness, $\widehat{\beta}$ is an unbiased estimate

percent of the control group received an A minus final letter grade. For the incentive group, the number of grade points they earned in the course was irrelevant outside of ranking the students: the top 10 percent of the incentive group received an A final grade and the next 12 percent of the incentive group received an A minus final grade. We did this for each letter grade.

[17] Our experiment takes place over two years, and we pool the sample across both years. Out of the 850 student-years, one student repeated the course and hence there are 849 unique students. For simplicity, we drop the subscript t from our specifications, treating the one repeating student as independent across years.

of the ITT effect (Imbens & Rubin, 2015).[18]

In our baseline estimation of Equation 1, we include in X_i a year indicator and first midterm score, following the advice of Bruhn and McKenzie (2009) to control for all covariates used in seeking balance. In a second model, we include additional controls chosen using the Post-Double-Selection (PDS) procedure of Belloni, Chernozhukov, and Hansen (2014b), explained in detail in Appendix A.3.[19]

3.4.2 Local Average Treatment Effect (LATE)

We estimate the LATE using two-stage least squares (2SLS) with assignment to the *Incentive* grade scheme as the instrument, Z_i.

$$v_i = \alpha Z_i + f(X_i) + e_i \tag{2}$$

$$Y_i = \gamma \widehat{v}_i + g(X_i) + u_i \tag{3}$$

where Y_i is an outcome of interest (e.g., exam scores) for a student i, X_i is a vector of pretreatment covariates, $f()$ and $g()$ are generic functions through which X_i affects v_i and Y_i, respectively, e_i and u_i are unobserved model residuals, and $\widehat{v}_i$ is instrumented videos estimated by Equation 2. We assume the influence of Z_i on v_i is monotonically increasing, that is, $v_I = \mathrm{E}(v_i|Z_i = 1) \geq \mathrm{E}(v_i|Z_i = 0) = v_C$. Hence, γ is the per-video average treatment effect, local to students induced by the incentive to watch on average $v_I - v_C$ additional videos.

Under the assumptions of unconfoundedness, excludability, monotonicity, and non-interference, $\widehat{\gamma}$ is an unbiased estimate of the LATE (J. D. Angrist & Imbens, 1995). Unconfoundedness requires that Z_i be independent of potential outcomes, a reasonable assumption given random assignment of students to the *Incentive* arm. Excludability assumes that outcomes

Dropping this student from the sample leaves the results virtually unchanged.

[18]Though we cannot test whether Z_i is confounded by unobservable covariates, we have confidence unconfoundedness holds given the random assignment of Z_i and the balance across observable covariates as demonstrated in Table **??** and **??**.

[19]We also estimated two additional models on the sample of students whose matched pair did not attrite: we fit Equation 1 including pair fixed effects and we estimate β as the mean difference in outcomes across pairs. The drop in observations increases the width of our confidence intervals, albeit modestly since including only matched pairs reduces unexplained variance in the outcome variables of interest. Results using those whose match pair did not attrite are available on GitHub.

(grades) are only affected by the instrument (incentive) through watching videos. This assumption could be violated if, for example, telling a student she is treated were to give her more confidence on subsequent exams during the quarter. Monotonicity, sometimes referred to as the "no defiers" assumption, is necessary because of two-sided noncompliance and requires that students assigned treatment watch weakly more videos than they would if they were assigned control. A violation of this assumption could occur if students get utility from rebelling against their assigned grade scheme. Non-interference, also known as the Stable Unit Treatment Value Assumption (SUTVA), assumes that each student's outcome depends only on their own treatment status and not the treatment status of their peers. Violations of SUTVA may include control students benefiting from having treatment students in the same class and, perhaps, studying together.

Although we believe unconfoundedness,[20] excludability,[21] and monotonicity[22] are reasonable assumptions, we have more concern about non-interference because of the potential for spillovers between students in the same class. If we had unlimited resources, a robust experimental design would assign treatment at the class (or coarser) level, reducing the chance for interactions between treated and control students. However, given our resource constraints, assigning treatment at coarser levels would have resulted in insufficient statistical power to detect reasonable effect sizes. Hence, we proceed acknowledging the potential for spillovers between students. We hypothesize that spillovers likely bias our estimates of the treatment effect *downwards* as we believe control students are more likely to benefit from having well-studied peers than they are to lose from, for example, having peers too busy watching videos to join a study group.[23]

[20] Although we randomly assigned treatment, one concern is nonrandom attrition. In the results section, we show that the *Incentive* and *Control* arms remain balanced across observables by the end of the experiment.

[21] While this assumption is not testable, we took care in the experimental design to make the treatment and control arms as similar as possible except for the grading schemes. Of course, watching videos inherently requires time that takes away from some other activity. Hence, the results should be interpreted as the causal effects of more videos and less of whatever else students would have been doing.

[22] Though not testable directly, one testable implication of monotonicity is that the cumulative distribution function of videos watched for each treatment arm should not cross. Indeed, Figure 2 shows that the two CDFs do not cross.

[23] Although spillovers are possible, we believe the magnitude of the spillovers are likely small given that students have for the most part not yet formed strong social networks. 47% of students in the *Incentive* or *Control* arms are transfer students in their first term at the university. The remaining students are predominantly sophomores taking their first upper division course. Social dynamics at the university facilitate networks within "colleges" more than majors for the very reason of encouraging academic diversity among

Similar to our estimates of Equation 1, we estimate Equation 3 with two sets of controls: only year and first midterm score and controls chosen using PDS.

3.4.3 Treatment Effect Heterogeneity

We also investigate the extent to which treatment effects vary along key demographic variables. In particular, we add an interaction term to Equation 1:

$$Y_i = \beta_1 Z_i + \beta_2 Z_i \mathbb{1}_{x_i=d} + f(X_i) + \epsilon_i \tag{4}$$

where $Z_i \mathbb{1}_{x_i=x}$ takes a value of 1 when treatment students have the value d for demographic variable $x \in X$. β_2 represents the difference in treatment effects for those with demographic d relative to those without. In practice, we estimate Equation 4 including in X_i dummies for year and the demographic characteristic of interest as well as first midterm score. While we observe many demographic variables, we are underpowered to detect reasonable effect sizes after adjusting for multiplicity, given the small sample sizes of our subgroups. As such, we focus on heterogeneity along our blocking variables and covariates we hypothesized *ex ante* may have treatment heterogeneity: levels of videos watched pretreatment and whether the student transferred from a community college, is an under-represented minority, or is of Asian ethnicity.[24]

peer groups. One example of a possible positive spillover is the online discussion board where students could ask questions about content covered in the IMVH.

[24]Levels of pretreatment video watching may matter either because those with greater experience watching videos may have greater treatment effects or, alternatively, they may have watched many videos during the experiment even if not treated, and hence the incentive may have little effect. As mentioned in Section 3, nearly all transfer students in the experiment are taking their first term at a four-year university and may not yet have optimized their studying practices. The university achievement gap between underrepresented and majority groups is well documented (Bowen, Chingos, & McPherson, 2009) and it is important to know how treatment may affect this gap. Finally, a large fraction of the international students in this university are Asian and the IMVH may be particularly helpful to those for whom English is not a native language, such as captions and the ability to reply videos. Unfortunately, we are unable to observe native language directly or better proxies such as country of home address or visa status.

4 Results

In this section, we first examine attrition and establish that the *Incentive* and *Control* arms in our analysis sample are balanced on observable characteristics. Second, we show that the grade encouragement worked: students in the *Incentive* arm watched significantly more videos than did their *Control* peers. Third, we estimate the effects of being assigned to the *Incentive* arm (ITT) and the effects of videos (LATE) on grade outcomes. Fourth, to better understand mechanisms, we examine spillovers to other studying methods for Micro A as well as grades in other courses taken during the experiment term. We then see whether behavior change persists after exogenous incentives are removed by estimating treatment effects on video watching and grades in the subsequent microeconomics course, Micro B. Finally, we discuss several thresholds in our experiment to clarify the LATE we estimate and to consider the relationship to the *population* ATE.

4.1 Attrition and balance

At the university where the experiment took place, Micro A is the first upper-division economics course and has higher withdrawal rates than most other economics courses, a product of both challenging course material and updated priors on interest in the field. In Micro A one year before the experiment, for students who scored below the median on the first midterm, 14.7% did not take the second midterm and 24.2% did not take the final exam.[25] In the present study, high attrition is not problematic, other than reducing statistical power, as long as attrition is independent of potential outcomes.

Since all students below the median on the first midterm had equal probabilities of being assigned to the *Control* and *Incentive* arms, treatment arms are balanced on covariates in expectation. In practice, due to chance and nonrandom attrition, treatment arms can be unbalanced on covariates, which can bias estimates if not addressed in the analysis, particularly in small samples (Athey & Imbens, 2017). Appendix Table A.2 details all sources of attrition and we check for balance on observable characteristics after attrition for both the

[25]The 2017 statistics are calculated from a sample that differs somewhat in inclusion criteria relative to the 2018 and 2019 samples. We provide these statistics to highlight the historically high rates of attrition and not to make comparisons with the experiment sample.

second midterm and final exam samples. As can be seen in Table 4, we find no statistically significant difference between the *Control* and *Incentive* arms in observable covariates including first midterm score, year, previous term's cumulative GPA, videos watched before the first midterm, ethnicity, gender, and transfer status. However, as discussed in Section 3.4, to correct for potential imbalance and to improve precision, we estimate models that include controls chosen via the post-double-selection method of Belloni, Chernozhukov, and Hansen (2014b).

4.2 Relevancy of the encouragement instrument

We use a Two-Stage Least Squares approach to estimate the LATE of watching videos on exam performance, as detailed in the Section 3.4. We must check that our instrument is both valid and relevant to ensure this method will produce an unbiased estimate of the LATE (Imbens & Rubin, 2015). The validity condition is met by assigning treatment at random conditional on midterm exam score and year of instruction. Balance across pretreatment observables, as demonstrated in Table 4, give us further confidence that treatment status is uncorrelated with demographics.

Next we check instrument relevancy, that is, whether treatment status generates significantly more video watching. In Table 5 we present estimates from Equation 1. We find that by the second midterm exam, being assigned to the *Incentive* arm induces students to watch 9.1 - 10.5 videos and 6.0 - 6.8 unique videos more than being assigned to the *Control* arm. The gap between treatment and control grows by the final exam to 38.4 - 39.2 videos and 20.5 - 21.6 unique videos. The larger gap by the final is unsurprising given that the deadline to earn the grade incentive was the day before the final exam. Following the recommendations of Andrews, Stock, and Sun (2019), we assess the strength of our instrument using the effective F-statistic of Montiel Olea and Pflueger (2013) which, in our just-identified setting, coincides with the Wald statistic of Kleibergen and Paap (2006). The effective F-statistic for the second midterm and final exam first-stage specifications are 18.6 and 194.6, respectively, both of which are greater than the Stock and Yogo (2005) critical value of 16.4 and the rule-of-thumb cutoff of 10.

Graphically, we depict the distribution of videos watched as a function of treatment in

Figure 2. Notably, the gap between treatment and control distributions remains significantly positive at every level of video watching by the final exam. The difference is most pronounced near the required number of videos to earn the grade incentive, after which the difference diminishes towards zero. For the second midterm sample, the difference is smaller but significantly different from zero between zero and 62 videos watched. We also show time series plots of video watching in Figure 3, which show no differences in video watching before the first midterm and marked increases before the second midterm and final exam. Collectively, given the highly significant first-stage regression results, large first-stage F-statistics, and monotonic increase in video watching across the sample, we have high confidence that our instrument meets the relevancy criterion.

4.3 Effects on exam scores

In this section, we estimate the causal effects of being assigned to the *Incentive* arm on exam scores (ITT). This estimate is relevant for educators interested in predicting how requiring videos will change exam scores in their classes using the same grade-based incentive implemented in our experiment. Additionally, we estimate the causal effect of watching videos on exam scores (LATE), which is of interest to educators deciding which teaching technologies to provide for their classes as well as to students choosing among different studying tools.

For both the ITTs and LATEs, we examine effects on the second midterm and final exams using both parametric methods (i.e. Equations 1 and 3) and nonparametric methods a la the repeated sampling framework of Neyman (1923). As results are similar across methods, we report parametric estimates here and the nonparametric results on GitHub. We check that our parametric results are robust to model specification by estimating Equations 1 and 3 with and without $f(X_i)$ as a vector of linear control variables chosen via PDS (Belloni, Chernozhukov, & Hansen, 2014b).

Table 6 presents estimates of the effects of treatment on second-midterm and final exam scores. Across our four specifications, we estimate reduced-form (RF) impacts of being assigned to the *Incentive* arm of 0.17 - 0.18 standard deviations on the second midterm. These estimates, along with our first-stage estimates (Table 5), imply LATEs of 0.26 - 0.30

standard deviations per 10 unique videos, or 0.16 - 0.18 standard deviations per hour of unique content. For the final exam, we estimate similar ITT effects: being assigned to the *Incentive* arm raises scores by 0.14 - 0.18 standard deviations. However, given the larger first stage effects for the final exam, we estimate smaller LATEs: 0.08 - 0.09 standard deviations per 10 unique videos, or 0.04 - 0.05 standard deviations per hour of unique content.[26]

4.4 Spillovers to concurrent courses

We next estimate spillover effects to other courses taken concurrently during the term of the experiment. Although we find positive effects on exam scores in Micro A, it is important to examine spillover effects to other classes. Although we do not observe student time use, an important proxy is whether grade outcomes declined in other courses, which would suggest that students reduced study time in other courses to watch videos in Micro A. On the other hand, if grades in other courses remain constant, it suggests that students substituted video watching for alternative studying methods within Micro A (a hypothesis we explore in the next section) or students increased total studying time in the quarter of the experiment.

Table 7 presents our estimates of Equation 1 where Y_i is GPA, number of classes passed, or portion of classes taken for a letter grade. Since video watching in Micro A may improve performance in other economics classes, we estimate the effects of treatment on term GPA calculated separately for all classes, all classes excluding Micro A, all classes outside of economics, and all economics classes excluding Micro A. In general, we find marginally significant or insignificant but directionally positive spillover effects on GPA.[27] We can rule out large negative spillover effects: in our worst-case specification for term GPA, our 95% confidence interval rules out negative spillover effects larger than -0.02 (on a 4.0 scale), or less than 1% of the mean term GPA among control students. There is no statistically significant difference between our estimates of spillover effects on GPA when restricting to

[26]Given the large F-statistics when estimating first-stage effects of our incentive instrument on videos watched, we are not particularly concerned about bias from weak instruments. However, following the advice of Andrews, Stock, and Sun (2019), we report Anderson-Rubin confidence sets, which are efficient regardless of the strength of our instrument. As can be seen in Appendix Table A3, we find that the weak-instrument-robust confidence intervals are very similar to those presented in Table 6.

[27]At this university, term GPA is affected only by classes taken for a letter grade. Hence, students may not have attrited from the sample but may have taken all courses Pass/No Pass and thus have no term GPA. As such, we report sample sizes for each GPA specification.

only economics or non-economics courses.

We additionally estimate the effects of treatment on the number of classes passed and find small, insignificant, but directionally positive effects. We find that treatment caused students to pass 0.02 - 0.09 more classes, or about 1% - 3% more classes than the control mean. We find no effect of treatment on number of classes not passed or withdrawn. Interestingly, treatment students were somewhat less likely to take Micro A for a letter grade than were control students, but this difference is only marginally significant for one of the four specifications and insignificant for the rest. We find no relationship between treatment and fraction of classes taken for a letter grade versus Pass/No Pass. Across all grade spillovers examined, we find mostly small, directionally positive effects, which gives us confidence that treatment is likely not harmful to academic success outside of Micro A.

4.5 Spillovers to other study methods in Micro A

We also examine spillovers to other studying methods within Micro A. Doing so helps us better understand mechanisms: do students substitute away from other studying when encouraged to watch more videos, or are they more likely to complement their video-watching with other unincentivized studying? In Table 8, we display the results of estimating equation 1 where Y_i is an alternative form of studying. We find that *Incentive* students are directionally less likely to attend class, though point estimates are near zero and not statistically significant. On the other hand, treatment students interacted with the online discussion board more than did control students, but again these estimates are not statistically distinguishable from zero. We do not find any significant relationship between treatment and tutoring attendance. Unfortunately, we do not observe the complete picture of student time use, but our evidence is consistent with students increasing self-study time while holding guided-study time nearly constant.

4.6 Spillovers to subsequent term

While we offered treatment students a grade incentive to watch videos during Micro A, students were not offered a grade incentive during the subsequent course in the intermediate

microeconomics sequence, Micro B. However, all students in Micro B maintained access to the IMVH, were given direction on which videos to watch each week in the syllabus, and were verbally encouraged to watch videos as a study method. Fortunately, we are able to observe video watching and grade outcomes in Micro B.

We present our estimates of spillover effects in the subsequent term in Table 9. In Panel A, we estimate the effect of treatment on videos watched during the subsequent term, for those who took Micro B. We find large and statistically significant effects: treatment caused students to watch 8.1 - 9.9 more unique videos and 1.2 - 1.5 hours of unique content compared to control students. In Panel B, we estimate equation 1 where Y_i is the first midterm, second midterm, or final exam score. Unfortunately given the small subsample of students who took Micro B, we are underpowered to detect effect sizes consistent with those observed during Micro A. Finally, in Panel C, we estimate the effect of treatment on taking Micro B and the number of classes passed and withdrawn. We find no effects statistically distinguishable from zero, though, as mentioned in section 4.1, treatment students were directionally less likely to take Micro B than were control students.

4.7 Treatment Effect Heterogeneity

Here we estimate Equation 4 to examine whether treatment effects vary along key demographic variables. We present our results in Table 10, which displays the coefficient estimates of β_2 from Equation 4. We find no evidence of treatment effect heterogeneity across our blocking variables, year and first midterm score. We are hesitant to make strong conclusions given the width of our confidence intervals, but this finding is consistent with stable treatment effects across the distribution of student abilities and years of the experiment. In Appendix Figures A4 and A5, we fit local linear regressions of videos watched, midterm 2 scores, and final exam scores as functions of midterm 1 score. We do not find any statistically significant differences along the midterm 1 score distribution; however, for the final exam, the point estimates are largest for the bottom quarter of midterm 1 scorers. Next, we examine heterogeneity by levels of videos watched pretreatment and find no significant differences in three of four specifications. We find marginally significant positive effects on the second midterm for those with higher levels of pretreatment video watching.

Moving to student demographics, we do *not* find statistically significant heterogeneous treatment effects for transfer students, and the point estimates for the two exams are similar in magnitude and opposite in direction. Interestingly, we *do* find marginally significant negative heterogeneous effects for female students on the final exam, but no significant effect on the second midterm, though the point estimate remains negative. This observation that male students may have benefited more from the intervention is consistent with the findings of Clark et al. (2020) as well as the literature on self-control more broadly,[28] which suggests male students, who tend to have less self-control than female students, may benefit from interventions that address self-control problems. We find significant negative effects for Asian students and positive effects for White students on the second midterm, but no effect on the final exam. However, after adjusting for multiplicity, either using a Bonferroni correction or the less conservative methods proposed by List, Shaikh, and Xu (2019), none of our heterogeneity results remain significant. Our results motivate future work investigating differences along gender and racial dimensions, which has implications for instructor recommendations and personalized education more generally.

4.8 Experiment design choices and *local* treatment effects

Our experiment contains several cutoffs: a midterm score cutoff below which students are eligible for the experiment, a video cutoff above which treatment students earn the grade incentive, and a date after which videos no longer count towards the incentive. All of these cutoffs influence the treatment effect estimates, which are *local* to compliers. Specifically, we estimate the effect of videos for those induced by the incentive to watch, so changing who is induced and the videos they are induced to watch will affect estimates of the LATE. Though the purpose of this study was not to identify optimal thresholds, we discuss some observations that may motivate future work and explain our intuition for the size of the *population* average treatment effect (ATE).

We required treatment students to watch 40 of 48 eligible videos to receive the incentive, providing students some agency in the videos they chose to watch. To better understand how the composition of watched videos differs between treatment arms, we plot video watch

[28] See, for example, Duckworth and Seligman (2006), Duckworth et al. (2015), and the works cited therein.

rates by video in Figure 4. This plot reveals that the videos students chose to watch are not random. Watch rates in the control group vary across videos from nearly 0% up to over 70%, demonstrating that the perceived intrinsic value of videos varies considerably. For treatment students, watch rates vary less across videos but are positively correlated with control watch rates. Notably, treatment watch rates are 70% or greater for videos whose control watch rates are between 5% and 20%. This gap has implications for the estimated local average treatment effect: if treatment encourages students to watch low value videos, then the LATE estimate will be diluted relative to the population ATE.

We investigate this gap first by considering whether treatment encouraged students to watch shorter videos. One may hypothesize that treatment students would choose to skip the longest videos since earning the incentive is not dependent on length of videos. If students picked the shortest 40 videos, they would watch 1.6 fewer hours of content versus watching the longest 40 videos. We fit a linear model that predicts a video's treatment watch rate using the duration of the video and its control watch rate.[29] Using this model, we find that each additional minute in duration is associated with a 0.3 percentage point decrease in watch rate by the treatment group. For the longest incentivized video, this corresponds with a predicted 3.7 percentage point lower watch rate compared to the average length video. We interpret this finding, given the strong correlation between control and treatment watch rates, as evidence that *content* is a more important driver than is minimizing time cost when choosing which videos to watch. However, it is plausible that some students prioritized shorter-length videos.

Next, we examine how watch rates varied over time. In Figure 5 we plot watch rates by video organized by week each video's content was covered in class. Interestingly, control group watch rates taper off towards the end of the term while treatment watch rates remain much higher. This observation is consistent with the video incentive serving as a commitment mechanism, perhaps encouraging students to spend more time studying for the final exam. Alternatively, perhaps control students watch fewer videos in weeks nine and ten as they shift their study time to other methods while preparing for the final. It is theoretically

[29] It is important to control for the control group's watch rate since the educational value of a video may be endogenously related to its length (e.g., harder concepts take longer to explain).

ambiguous whether the gap in watch rates towards the end of the term suggests a larger or smaller LATE relative to the population ATE.

Finally, we consider the influence of the due date on our treatment effect estimates. To earn the grade incentive, students had to watch 40 videos by the last day of instruction, which is the day before the final exam. This is an intuitive deadline, but it allows procrastination-prone students to delay watching many hours of videos until the final week of the term. It is plausible that these students could be harmed by treatment, or at the very least have very small treatment effects, which could explain part of the reason why our LATE estimates are smaller for the final exam than the second midterm. To get a sense of whether treatment students were more likely to "binge watch" videos, we plot the max number of videos watched in one week per student in Figure A6. Though some control students watch 40 or more videos in one week, this behavior seems more prevalent in the treatment group. While this may very well be a useful, rational studying strategy for some, research suggests that spreading learning over time may be more effective (Kornell, 2009). Ariely and Wertenbroch (2002) find that offering deadlines, though costly from a rational agent perspective, results in better grades. As an alternative to offering one deadline, offering multiple deadlines, perhaps weekly or shortly before both exams, may improve consistency between treatment effect estimates for the two exams.

Collectively, it is not obvious how the population ATE compares with the LATEs we estimate. Neoclassical economic theory suggests that those who select into watching videos regardless of exogenous incentives likely do so because they benefit more than those who only select into watching videos only in the presence of exogenous incentives. This view would suggest that the population ATE is greater than the LATE. However, this view is, perhaps, inconsistent with the alternative that students may not know how much studying is optimal, or what they should be studying. While it is not possible to estimate the population ATE given our experimental design, our intuition is that the ATE is likely higher than the LATE estimated for the final exam and closer to that estimated for the second midterm, which is less likely to be diluted from deadline-induced binge watching and selecting the shortest videos.

5 Models of Studying Behavior

To understand our empirical results in the context of economic theory, we discuss three models of student studying behavior: a neoclassical model, an imperfect information model, and a behavioral/procrastination model. For each model we consider the testable implications of a grade inducement to encourage adoption of a study method.[30] Neoclassical models of studying behavior assume that rational agents know their returns to studying using the methods available to them and allocate the optimal study time to each method given their utility function, which is increasing in leisure and grades and decreasing in time spent studying. College instructors have little knowledge about student utility functions and do not know student preferences over performance in other classes or other uses of the student's time which may also have large payoffs in the labor or marriage markets. In this model, there is no room for an instructor to increase student well-being by intervening in their study decisions. Both Oettinger (2002) and Kapoor, Oosterveen, and Webbink (2020) find empirical support for the neoclassical model. Oettinger (2002) finds that student effort responds rationally to nonlinear grade incentives: across 1200 students in a principles of economics class with absolute grading standards, he finds evidence of bunching just above letter grade cutoffs and student performance on the final exam is higher if the student is just below a grade threshold. Kapoor, Oosterveen, and Webbink (2020) explore the effects of a university policy that required students who performed poorly in their first year to attend a large fraction of the tutorials for each class in their second year. At the poor performance threshold, both tutorial and lecture attendance increased by over 50 percent with a concomitant decline in self-study hours. Grades for students at the policy threshold were *lower* by 0.16-0.26 standard deviations. At least for students at the margin of poor performance, the requirement to attend tutorials appears to have hurt students by not allowing them to pick their optimal study strategy across self study, tutorials, and lectures.

A key assumption of the neoclassical model is that students possess complete information about the returns across studying methods. However, there is evidence from psychology that

[30] For all three models, we do not address the issue that the IMVH is a relatively unique study tool in that, to our knowledge, it is the first instructional book to be created entirely of videos. However, given the availability of close substitutes to the IMVH (lecture capture, for example) we do not explore the added issues of inducing students to use a study tool whose usefulness is not known to the instructor.

college students do not know the return to various study options[31] and many universities fund "Teaching and Learning Centers" or "Academic Skills Centers," part of whose mission is to help undergraduates learn to study more productively.[32] Further, the "raison d'etre" of higher education is not only to teach students specific skills but to teach students how to learn. As an alternative to the neoclassical model, we hypothesize that students supply a quantity of study time that is optimal given their information constraints. In this 'imperfect information' model, students choose study methods and quantities that are suboptimal relative to those they would have picked in a full information setting. Hence, an intervention by an entity that has more information about returns to studying across various methods (i.e. an instructor) can enhance student utility.

A third model is a behavioral one in which students plan to study more than they end up studying when the time comes. This phenomenon is consistent with two-self models in which a person's "planner" self, the one who desires high grades at the expense of leisure, is at odds with her "doer" self who must choose between immediately gratifying leisure and delayed gratification from higher grades. Indeed, survey and experimental data suggest that many students study less than they report they "should" and finish the term with grades lower than what they had anticipated they would earn at the start of the term.[33] Clark et al. (2020) provide empirical support for this model by finding that setting tasked-based goals helps improve college student performance. As descriptive evidence in support of this model, Beattie et al. (2019) find that students that do much worse than expected in college are those who say they have poor time management or procrastination issues, including a tendency to cram and spending very little time studying.

We consider the testable implications of the three models applied to a setting where students are incentivized to use a time-consuming educational input, say, a set of instructional videos (or attending class, reading the textbook, working on homework, etc.). The incentive is structured such that students who consume the educational input receive a higher grade in the course by consuming a set level of the input. In this simple setting, students gain utility

[31] See, for example, Dunlosky et al. (2013), McCabe (2011), and Pashler et al. (2007).

[32] All nine University of California campuses have such a center. Examples outside the UC include Dartmouth's Academic Skills Center, Michigan's Center for Research on Teaching and Learning, UNC's Learning Center, and Yale's Teaching and Learning Center.

[33] See, for example, Chen et al. (2017), Ferrari (1992), and Lavecchia, Liu, and Oreopoulos (2016).

only from leisure and grades. We assume grades, a function of time spent studying, and utility are both continuous, smooth, and increasing and concave in their inputs. Students can choose to study using the incentivized educational input, v, or some outside option that is not directly incentivized, o, or a combination thereof.

Across all three models, before the first educational input is incentivized, students allocate time to the two studying methods until the marginal benefit of each (through higher grades) is equal to the marginal cost of forgone leisure. Consider the population of students initially consuming below the requisite level to earn the grade incentive. These students must decide if earning the grade incentive is worth forgone leisure and less time allocated to their outside studying option. Next we explore the differences in predictions across the three models for *compliers*, those for whom the incentive induces greater take-up of the incentivized input.

In the neoclassical model, as long as o and v are not perfect substitutes in the grades production function, the marginal return to grades of the incentivized input is less than that of the outside option for compliers. This model predicts bunching at the incentivized level cutoff since compliers would prefer to spend their marginal hours on leisure or studying with their other method. This model predicts a weak increase in video watching and weak decrease in other studying and leisure consumption. It is ambiguous whether cumulative study time increases or decreases as this depends on relative utility benefits of leisure and grades and the returns to studying by each method. However, if cumulative study time remains constant or decreases, then exam performance should strictly decrease since students are now suboptimally allocating study time versus their first-best allocation when considering only marginal returns to studying. On the other hand, if cumulative study time increases, students may earn greater exam grades but achieve lower utility compared to baseline. Importantly, this model predicts that in subsequent quarters students return to their pre-incentive levels of studying.

In the imperfect information model, students' *ex ante* allocations to each studying method are not necessarily first-best. The effect of the incentive on video-watching depends on whether students update their priors about the returns to watching videos as they work towards hitting the minimum required level for the grade incentive. At this cutoff, they make a decision whether to continue watching videos depending on their updated perceptions of

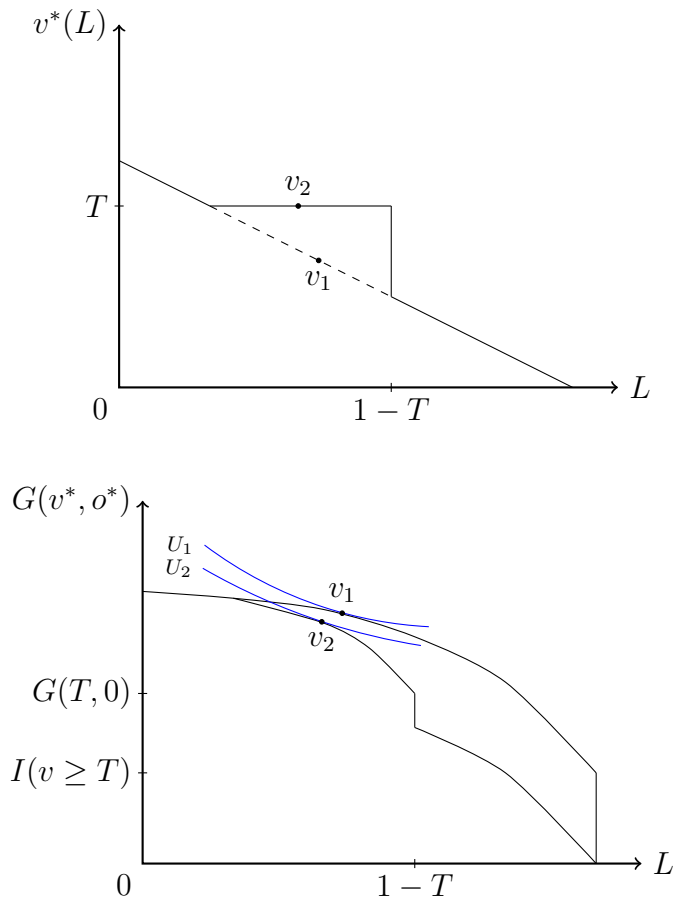


Figure 1: A student maximizes her utility U, a function of leisure hours, L, and grades, G. Grades are a function of video watching hours, v, and hours spent on the next best studying option, o. *Above*: Let $v^*(L)$ be the demand curve for video watching as a function of leisure $L \in [0, 1]$. Assume the student maximizes her utility by watching v_1 videos and spending o_1 hours on her other study method. Suppose an instructor offers a grade incentive worth I units, which a student can earn by watching $v \geq T$ hours of videos. Note that at $L = 1 - T$ hours of leisure, the student maximizes grades by spending all study time watching videos, that is, $v^* = T$. *Below*: Student's utility maximization problem for the neoclassical model. The grade incentive, I, is given to the student conditional on watching T hours of videos (inner budget constraint) or, in the unincentivized case, given regardless of video watching (outer budget constraint). Time bundles along the inner budget constraint are weakly less preferred since the incentive draws the student away from otherwise more efficient (v, o) combinations.

the marginal benefit. We do not expect bunching in this model unless students believe the cutoff for the grade incentive is the optimal level or the updated marginal benefit at the cutoff is lower than the marginal benefit of the next best studying option or the marginal utility of leisure. A sharp prediction is that video watching will continue at the incentivized level in the absence of the grade incentive as students have learned an effective study tool. We also expect the treatment effect to be greater for students with more information problems, perhaps students in their first semester/quarter at university.

Finally, in the behavioral/procrastination model, the instructor's inducement helps students stick to study plans up to the minimum required level for the grade incentive. As long as total study time does not fall, the inducement will increase exam performance (see Ariely and Wertenbroch (2002) where students with externally set deadlines had higher grades relative to students who choose their own deadlines). This model also predicts bunching at the incentivized level cutoff as long as using the incentivized input does not change the student's "planner" and "doer" selves. In the absence of the inducement, i.e., in future classes, a sharp prediction is that video watching will revert to pre-inducement levels.

We summarize the predictions across models in Table 3. One key empirical difference is whether or not students return to their pre-incentive level of video watching in the absence of the incentive. We find that treatment students continued watching videos in the absence of any grade incentive in Micro B, which is inconsistent with the predictions from both the neoclassical and behavioral models of learning in our setting. Another key empirical difference across the models is whether students watch exactly 40 videos, the number of videos required to earn the grade incentive. In Figure A2 we plot the distribution of incentive-eligible videos watched, and one can see that treatment students typically watched more 40 videos. Since students did not receive immediate feedback about the number of videos they had watched, the lack of bunching may be influenced by student uncertainty about whether they had met the 40-video threshold.[34] Nevertheless, we view our results as most consistent with the imperfect information model of learning.

[34]Anecdotally, many students reported keeping track of which videos they had watched by checking-off the list of incentive-eligible videos and/or taking notes on each video watched.

Table 3: Predictions Across Models for Compliers, those induced by incentive to consume more videos.

Outcome	Neoclassical	Imperfect Information	Behavioral/ Procrastination
Number of Videos, v	up	up	up
Other study method, o	?	?	?
Total studying, $v + o$	?	?	?
Bunching at 40 videos, T	Yes	No	Yes
Exam Performance	Down*	Up	Up
Video use, future classes**	return to baseline	remain at incentivized level	return to baseline

Notes: *As long as total studying stays the same or falls. **We assume future classes do not incentivize video watching.

6 Discussion

6.1 Contributions

Our results add to the literature on what motivates students to increase their performance in college. Previous research finds financial incentives have little effect (see papers cited in Gneezy, Meier, and Rey-Biel (2011)) but tend to work better if educational inputs as opposed to outputs are incentivized (Fryer Jr (2011), Gneezy, Meier, and Rey-Biel (2011)). There is mixed evidence on the effects of having students set goals on the use of educational inputs with no penalty for missing the goal (see Clark et al. (2020) who find positive grade effects of setting goals on number of practice quizzes to complete while Oreopoulos et al. (2019) find that setting a weekly schedule ahead of time and weekly reminders via a text message had only a small effect on study time and no effect on output as measured by grades, retention or credit accumulation.). Interestingly, Clark et al. (2020) find no effect of having students set goals on class outcomes, such as course grade or exam scores. We find that a small grade incentive is effective in motivating poorly performing college students to significantly take-up video watching, an educational input.[35] We also reduced the weight on an early assessment and allowed students to earn back the lost points fully by meeting the video watching requirement, which may also be an important motivator. Grade incentives

[35]It is possible that the grade incentive is not an important motivator as Dobkin, Gil, and Marion (2010)) required students to attend class if they performed poorly on an early assessment and, though TAs carefully recorded attendance, there was no penalty for not attending class. Nevertheless, poorly performing students significantly increased class attendance.

have the unappealing feature that grades are directly a function of input use. The grade incentive used in this study was small, at most four percent of the student's grade, which may help mitigate this concern.

Second, we find that inducing students who performed poorly on an early assessment to increase the amount of time they spend watching instructional videos increased their exam performance. Since there was no drop in either grades for other courses taken in the same quarter or a drop in the use of many other educational for the class, this suggests that total study time for the course increased for treatment students. Unfortunately, we were not able to determine where the added study time came from, which is important for student welfare calculations. Our results are consistent with other experimental and quasi-experimental studies that find positive effects of educational inputs on college student performance.

Universities often have policies that emphasize improving the performance poorly performing students, such as academic probation policies and the policy studied by Kapoor, Oosterveen, and Webbink (2020). We focus on students who perform poorly on the first midterm, a population similar to that of Dobkin, Gil, and Marion (2010). Some may wonder why we did not include the entire class in the experiment. While including the entire class would have increased statistical power, we were concerned that the additional precision could come at the expense of welfare losses by high performing students. The first midterm provides a signal of which students likely know for themselves how to study, both methods and duration. Coercing these high-type students to spend time with a potentially different studying method runs a greater risk of harming utility. Students who, through a low midterm score, both indicate a need for alternative studying practices and stand to benefit the most from instructor-provided guidance.

Finally, we find support for an imperfect information model of learning because treatment students frequently watched more than the 40 videos they were required to watch to earn the grade incentive and because treatment students continued watching videos at a significantly higher rate in the following class when watching videos was not exogenously incentivized. J. Angrist, Lang, and Oreopoulos (2009) also find continued higher use of academic support services after the incentivized year for women. An imperfect information model of learning can also account for why incentivizing educational inputs has been found to be more effective

than incentivizing grades or exam performance directly (see studies cited in Gneezy, Meier, and Rey-Biel (2011))

6.2 Limitations

The present study has several limitations that should be considered before, for example, creating one's own videos and incentivizing students to use them. First, the population studied is students who score below the median on the first midterm of an intermediate microeconomics course at a large, highly-selective public research university. The extent to which treatment effects vary by course, instructor, university, or along the top half of the midterm score distribution is important but beyond the scope of this paper.[36] Additionally, the causal effects of watching videos that we estimate are *local* to compliers, i.e. students induced by the grade incentive to watch additional videos. As discussed in Section 4.8, we cannot recover the *population* average treatment effect.

The positive effects we estimate are attributable to watching IMVH videos and spending more time studying for the course. Our experimental design does not allow us to separately identify the effects of these two mechanisms for improving exam performance. Would a similar incentive structure that induces greater takeup of an alternative study method have similar effects? We view this as an important question for future research.

Another question is whether *any* instructional videos would have the same impact on exam scores that we estimate for the IMVH or are their aspects of the IMVH that specifically contribute to student learning? Research in psychology identifies several aspects of videos that appear to improve learning transmission (see Mayer (2021) and papers cited therein) and so it is possible that inducing students to watch other videos may have larger or smaller effects than what we estimate for the IMVH.

The next consideration is the time frame during which the experiment took place, 2018 to 2019. About three months after the conclusion of our experiment, most students in the United States and all students at the studied university began remote learning as the

[36] As an example, Carrell and Kurlaender (2020) found that an intervention for poorly performing students (personalized e-mails from the professor with useful information about where to get help) that was effective in raising exams scores in introductory microeconomics classes at a large public research university did not raise exam scores across several different types of classes in a broader-access institution.

COVID-19 pandemic prompted stay-at-home orders. With increased experience learning via electronic media, it is possible that treatment effects will be higher in the future than we estimate in our paper. On the other hand, if students find online learning materials increasingly *less* engaging, we may find the opposite.

In addition to estimating the effects of video handbooks in other educational settings, future research should examine treatment effects in the presence of weekly deadlines instead of one final deadline at the end of the term. Given our observation of greater "binge watching" by treatment students and the smaller effect sizes before the final exam compared to the second midterm exam, we suspect weekly deadlines may reduce the deleterious effects of procrastination. Despite the rich literature on the advantages of spread-out studying (Cepeda et al., 2006; Kornell, 2009), we note that "binge watching" was not unique among treatment students. Indeed, most students within each treatment arm watched more videos the last week of the term than any other week.

7 Conclusion

We examine the effectiveness of an innovative educational technology, a video handbook composed of 220 brief instructional videos on intermediate microeconomic theory. We used random assignment of a grade-based incentive to experimentally vary takeup of the video handbook, and we found that greater takeup caused students to score significantly higher on exams. Specifically, we estimate that treatment causes students to score about 0.18 standard deviations higher on midterm and final exams. For students on the margin of watching videos, watching an additional hour of unique content causes students to score between 0.05 to 0.15 standard deviations higher on exams.

Instructors may have concerns about making a resource such as the IMVH available if they believe students may substitute away from lectures or other more productive studying methods Kay (2012). Another concern is that forcing students to spend more time studying in one's class may worsen performance in other classes. Our analysis provides some confidence that neither of these fears are first-order concerns. We do not find evidence that students decrease their consumption of other forms of studying, nor do we find that students perform

worse in other courses during the same quarter. Our point estimates of the effect of treatment on takeup of other studying methods, though not statistically significantly different from zero, are *positive* for most alternatives, suggesting that if any, students consider the videos complements to other forms of studying. A potential mechanism might be that the videos help students realize what they *don't* know, increasing the marginal benefit of subsequent studying.

A final concern is one of welfare. In a neoclassical model, instructors cannot make their students better off by exogenously incentivizing quantities of studying they would not otherwise have chosen for themselves. In an imperfect information model, which we think is more appropriate in our university classroom setting, instructors *can* improve student welfare through intervention when information barriers lead to suboptimal time allocation decisions. We observe two phenomena that support this model. First, treatment students do not bunch at the cutoff for the grade incentive. Second, video consumption remains much higher among treatment students in the term following conclusion of the experiment.

While there are many educational interventions that instructors could offer their students, the research on causal effects of educational interventions remains limited. Our study serves as an example of a feasible research design that runs a lower risk of generating welfare losses for high performing students than does a class-wide experiment. It is our hope, as educators ourselves, that more research will be conducted on the effectiveness of pedagogical technologies.

References

Allgood, S., Walstad, W. B., & Siegfried, J. J. (2015). Research on teaching economics to undergraduates. *Journal of Economic Literature*, *53*(2), 285–325.

Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, *11*, 727–753.

Angrist, J., Lang, D., & Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, *1*(1), 136–63.

Angrist, J. D., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, *90*(430), 431–442.

Ariely, D., & Wertenbroch, K. (2002). Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological Science*, *13*(3), 219–224.

Arulampalam, W., Naylor, R. A., & Smith, J. (2012). Am i missing something? the effects of absence from class on student performance. *Economics of Education Review*, *31*(4), 363–375.

Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments* (pp. 73–140, Vol. 1). Elsevier.

Beattie, G., Laliberté, J.-W. P., Michaud-Leclerc, C., & Oreopoulos, P. (2019). What sets college thrivers and divers apart? a contrast in study habits, attitudes, and mental health. *Economics letters*, *178*, 50–53.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29–50.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

Bonesrønning, H., & Opstad, L. (2012). How much is students' college performance affected by quantity of study? *International Review of Economics Education*, *11*(2), 46–63.

Bonesrønning, H., & Opstad, L. (2015). Can student effort be manipulated? does it matter? *Applied Economics*, *47*(15), 1511–1524.

Bowen, W. G., Chingos, M. M., & McPherson, M. S. (2009). *Crossing the finish line: Completing college at america's public universities*. Princeton University Press.

Bratti, M., & Staffolani, S. (2013). Student time allocation and educational production functions. *Annals of Economics and Statistics/ANNALES D'ÉCONOMIE ET DE STATISTIQUE*, 103–140.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics*, *1*(4), 200–232.

Carrell, S. E., & Kurlaender, M. (2020). *My professor cares: Experimental evidence on the role of faculty engagement* (Working Paper No. 27312). National Bureau of Economic Research.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, *132*(3), 354.

Chen, P., Chavez, O., Ong, D. C., & Gunderson, B. (2017). Strategic resource use for learning: A self-administered intervention that guides self-reflection on effective resource use enhances academic performance. *Psychological Science*, *28*(6), 774–785.

Clark, D., Gill, D., Prowse, V., & Rush, M. (2020). Using goals to motivate college students: Theory and evidence from field experiments. *The Review of Economics and Statistics*, *102*(4), 648–663.

Dobkin, C., Gil, R., & Marion, J. (2010). Skipping class in college and exam performance: Evidence from a regression discontinuity classroom experiment. *Economics of Education Review*, *29*(4), 566–575.

Duckworth, A. L., Shulman, E. P., Mastronarde, A. J., Patrick, S. D., Zhang, J., & Druckman, J. (2015). Will not want: Self-control rather than motivation explains the female advantage in report card grades. *Learning and individual differences*, *39*, 13–23.

Duckworth, A. L., & Seligman, M. E. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of educational psychology*, *98*(1), 198.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58.

Expósito, A., Sánchez-Rivas, J., Gómez-Calero, M. P., & Pablo-Romero, M. P. (2020). Examining the use of instructional video clips for teaching macroeconomics. *Computers & Education*, *144*, 103709.

Ferrari, J. R. (1992). Psychometric validation of two procrastination inventories for adults: Arousal and avoidance measures. *Journal of Psychopathology and Behavioral Assessment*, *14*(2), 97–110.

Fryer Jr, R. G. (2011). Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics*, *126*(4), 1755–1798.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of economic perspectives*, *25*(4), 191–210.

Grodner, A., & Rupp, N. (2013). The role of homework in student learning outcomes: Evidence from a field experiment. *The Journal of Economic Education*, *44*(2), 93–109.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Joyce, T., Crockett, S., Jaeger, D. A., Altindag, O., & O'Connell, S. D. (2015). Does classroom time matter? *Economics of Education Review*, *46*(100), 64–77.

Kapoor, S., Oosterveen, M., & Webbink, D. (2020). The price of forced attendance. *Journal of Applied Econometrics*.

Kay, R. H. (2012). Review: Exploring the use of video podcasts in education: A comprehensive review of the literature. *28*(3).

Kirby, A., & McElroy, B. (2003). The effect of attendance on grade for first year economics students in university college cork. *The Economic and Social Review*, *34*, 311–326.

Kleibergen, F., & Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of econometrics*, *133*(1), 97–126.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *23*(9), 1297–1317.

Krohn, G. A., & O'Connor, C. M. (2005). Student effort and performance over the semester. *The Journal of Economic Education*, *36*(1), 3–28.

Lavecchia, A. M., Liu, H., & Oreopoulos, P. (2016). Behavioral economics of education: Progress and possibilities. In *Handbook of the economics of education* (pp. 1–74, Vol. 5). Elsevier.

List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, *22*(4), 773–793.

Mayer, R. E. (2021). Evidence-based principles for how to design effective instructional videos. *Journal of Applied Research in Memory and Cognition*, *10*(2), 229–240.

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *3*(39), 462–476.

Montiel Olea, J. L., & Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, *31*(3).

Munley, V. G., Garvey, E., & McConnell, M. J. (2010). The effectiveness of peer tutoring on student achievement at the university level. *American Economic Review*, *100*(2), 277–82.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Statistical Science*, *5*(4), 465–472.

Noetel, M., Griffith, S., Delaney, O., Sanders, T., Parker, P., del Pozo Cruz, B., & Lonsdale, C. (2021). Video improves learning in higher education: A systematic review. *Review of Educational Research*, *91*(2), 204–236.

Oettinger, G. S. (2002). The effect of nonlinear incentives on performance: Evidence from "econ 101". *The Review of Economics and Statistics*, *84*(3), 509–517.

Oreopoulos, P., Patterson, R. W., Petronijevic, U., & Pope, N. G. (2019). Low-touch attempts to improve time management among traditional and online college students. *Journal of Human Resources*, 0919–10426R1.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin and Review*, *14*(2), 187–193.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366.

Stinebrickner, R., & Stinebrickner, T. R. (2008). The causal effect of studying on academic performance. *The B.E. Journal of Economic Analysis & Policy*, *8*(1), 1–55.

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear iv regression. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, *80*(4.2), 1.

Tang, L., Li, S., Auden, E., & Dhuey, E. (2020). Who benefits from regular class participation? *The Journal of Economic Education*, *51*(3-4), 243–256.

Trost, S., & Salehi-Isfahani, D. (2012). The effect of homework on exam performance: Experimental results from principles of economics. *Southern Economic Journal*, *79*(1), 224–242.

Table 4: Balance test, exam taker sample

	Midterm 2 Sample			P-values	Final Exam Sample			P-values
Variable	Above Median	Control	Incentive	(3) - (2)	Above Median	Control	Incentive	(6) - (5)
Midterm 1 score	0.838	-1.094	-1.172	0.398	0.840	-1.056	-1.153	0.291
	(0.025)	(0.063)	(0.068)		(0.025)	(0.061)	(0.069)	
Year = 2019	0.492	0.513	0.500	0.797	0.489	0.516	0.500	0.753
	(0.025)	(0.036)	(0.035)		(0.025)	(0.037)	(0.036)	
Cumulative GPA	3.445	2.944	2.948	0.965	3.445	2.946	2.959	0.864
	(0.029)	(0.043)	(0.058)		(0.029)	(0.044)	(0.060)	
No cum. GPA	0.230	0.368	0.332	0.452	0.231	0.359	0.332	0.583
	(0.021)	(0.035)	(0.033)		(0.021)	(0.035)	(0.034)	
Math quiz score	0.592	0.037	0.106	0.471	0.599	0.071	0.152	0.396
	(0.044)	(0.070)	(0.065)		(0.043)	(0.068)	(0.066)	
Tutoring visits	0.269	0.259	0.223	0.655	0.270	0.272	0.237	0.684
	(0.042)	(0.059)	(0.056)		(0.043)	(0.061)	(0.060)	
Videos watched	13.228	13.368	13.777	0.750	13.292	13.418	13.658	0.856
	(0.681)	(0.886)	(0.931)		(0.682)	(0.909)	(0.953)	
Videos, unique	9.746	9.689	10.188	0.554	9.793	9.783	10.111	0.704
	(0.431)	(0.580)	(0.611)		(0.432)	(0.598)	(0.622)	
Hours videos	2.476	2.527	2.588	0.805	2.488	2.535	2.572	0.880
	(0.127)	(0.170)	(0.176)		(0.127)	(0.173)	(0.180)	
Hours videos, unique	1.729	1.752	1.838	0.565	1.738	1.770	1.830	0.696
	(0.075)	(0.104)	(0.108)		(0.075)	(0.107)	(0.110)	
Asian	0.700	0.694	0.668	0.581	0.701	0.696	0.653	0.376
	(0.022)	(0.033)	(0.033)		(0.022)	(0.034)	(0.035)	
Latinx	0.060	0.135	0.158	0.506	0.060	0.141	0.158	0.654
	(0.012)	(0.025)	(0.026)		(0.012)	(0.026)	(0.027)	
White	0.151	0.114	0.124	0.765	0.149	0.109	0.132	0.497
	(0.018)	(0.023)	(0.023)		(0.018)	(0.023)	(0.025)	
Other ethnicity	0.089	0.057	0.050	0.741	0.089	0.054	0.058	0.882
	(0.014)	(0.017)	(0.015)		(0.014)	(0.017)	(0.017)	
Female	0.393	0.342	0.391	0.312	0.393	0.348	0.405	0.253
	(0.024)	(0.034)	(0.034)		(0.024)	(0.035)	(0.036)	
Male	0.592	0.653	0.604	0.316	0.593	0.647	0.584	0.215
	(0.024)	(0.034)	(0.034)		(0.024)	(0.035)	(0.036)	
Transfer	0.271	0.477	0.455	0.673	0.272	0.462	0.447	0.778
	(0.022)	(0.036)	(0.035)		(0.022)	(0.037)	(0.036)	
Observations	417	193	202		415	184	190	

Note: This table includes all students who completed the second midterm and final exams. Descriptions of each variable can be found in Table A4. *Female* and *Male* are coded zero for nine students who do not report a gender. *P-values* are reported for the Welch's t-test of equal means between the *Control* and *Incentive* arms. Standard errors in parentheses are robust to heteroskedasticity. ***, **, and * indicate significance at the 1, 5, and 10 percent critical levels, respectively.

Table 5: Effects of Grade Incentive on Video Watching

	Control Mean	(1)	(2)	(3)	(4)
Panel A: By Midterm 2					
Videos	33.91	10.19***	10.54***	9.08***	9.58***
		(2.85)	(3.12)	(2.03)	(2.19)
Unique videos	23.13	6.63***	6.79***	5.97***	6.11***
		(1.54)	(1.70)	(0.98)	(1.11)
Hours of videos	5.88	1.68***	1.72***	1.48***	1.55***
		(0.50)	(0.55)	(0.35)	(0.38)
Hours of unique videos	3.85	1.10***	1.13***	0.99***	1.02***
		(0.25)	(0.28)	(0.16)	(0.18)
Observations		395	362	395	362
Panel B: By Final Exam					
Videos	53.09	39.25***	39.07***	38.57***	37.99***
		(4.06)	(4.37)	(3.40)	(3.69)
Unique videos	33.95	21.55***	21.08***	21.28***	20.49***
		(1.55)	(1.66)	(1.22)	(1.27)
Hours of videos	8.93	6.30***	6.26***	6.18***	6.05***
		(0.69)	(0.75)	(0.57)	(0.62)
Hours of unique videos	5.54	3.43***	3.36***	3.38***	3.26***
		(0.25)	(0.27)	(0.20)	(0.21)
Observations		374	332	374	332
Treatment assignment controls		Yes	No	Yes	Yes
Demographic controls		No	No	Yes	Yes
Pair Fixed Effects		No	No	No	Yes

Note: Model (1) contains linear controls for midterm 1 score and year; (2) is the difference in means and standard errors calculated using the repeated sampling framework of Neyman (1923); (3) and (4) use the post-double-selection (PDS) procedure of Belloni, Chernozhukov, and Hansen (2014b) to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A5. Models (2) and (4) include only students whose matched-pair did not attrite from the experiment. *Control Mean* is the mean for the Control students included in models (1) and (3), which is nearly identical to the mean for the Control students included in models (2) and (4). Standard errors in parentheses are robust to heteroskedasticity. ***, **, and * indicate significance at the 1, 5, and 10 percent critical levels, respectively.

Table 6: Effects of Videos on Grades

	(1)	(2)	(3)	(4)
Panel A: Midterm 2 score				
RF: Incentive	0.176*	0.183*	0.176*	0.174*
	(0.090)	(0.094)	(0.090)	(0.096)
2SLS: 10 videos	0.266*	0.270*	0.300**	0.293**
	(0.146)	(0.150)	(0.151)	(0.145)
2SLS: 1 hour of videos	0.160*	0.163*	0.181**	0.170*
	(0.087)	(0.090)	(0.090)	(0.095)
Observations	395	362	395	362
Panel B: Final exam score				
RF: Incentive	0.175**	0.174*	0.175**	0.138
	(0.089)	(0.103)	(0.088)	(0.103)
2SLS: 10 videos	0.081**	0.082*	0.082**	0.087*
	(0.041)	(0.049)	(0.041)	(0.046)
2SLS: 1 hour of videos	0.051**	0.052*	0.052**	0.043
	(0.026)	(0.031)	(0.026)	(0.031)
Observations	374	332	374	332
Treatment assignment controls	Yes	No	Yes	Yes
Demographic controls	No	No	Yes	Yes
Pair Fixed Effects	No	No	No	Yes

Note: This table reports coefficients on $Incentive_i$ from Equation 1 (Reduced Form, *RF*) and $\hat{Video}_i$ from Equation 3 (Two-Stage Least Squares, *2SLS*). Test scores are measured in standard deviation units. Model (1) contains linear controls for midterm 1 score and year; (2) is the difference in means and standard errors calculated using the repeated sampling framework of Neyman (1923); (3) and (4) use the post-double-selection (PDS) procedure of Belloni, Chernozhukov, and Hansen (2014b) to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A5. Models (2) and (4) include only students whose matched-pair did not attrite from the experiment. Standard errors in parentheses are robust to heteroskedasticity. ***, **, and * indicate significance at the 1, 5, and 10 percent critical levels, respectively.

Table 7: Spillover Effects of Incentive on Other Course Grades

	Control Mean	(1)	(2)	(3)	(4)
Panel A: Effects on Term GPA					
All classes	2.59	0.13**	0.13*	0.11*	0.10
		(0.06)	(0.07)	(0.06)	(0.06)
		373	332	373	332
Excluding Micro A	2.75	0.10	0.11	0.09	0.10
		(0.07)	(0.08)	(0.07)	(0.08)
		370	329	370	329
Excluding econ classes	2.99	0.06	0.09	0.06	0.08
		(0.10)	(0.09)	(0.09)	(0.12)
		315	278	315	278
Econ classes ex. Micro A	2.44	0.07	0.02	0.07	-0.03
		(0.09)	(0.08)	(0.09)	(0.12)
		258	228	258	228
Panel B: Effects on classes passed					
Num. classes passed	3.28	0.08	0.09	0.05	0.02
		(0.09)	(0.10)	(0.09)	(0.09)
Num. classes not passed	0.31	0.01	-0.01	0.01	-0.01
		(0.06)	(0.06)	(0.06)	(0.06)
Num. classes withdrawn	0.05	0.01	0.01	0.01	0.01
		(0.03)	(0.02)	(0.03)	(0.02)
Panel C: Effects on class grade type					
Letter grade in Micro A	0.95	-0.04	-0.05*	-0.03	-0.04
		(0.03)	(0.03)	(0.02)	(0.03)
% classes taken for letter	0.93	-0.01	-0.01	-0.01	-0.01
		(0.01)	(0.02)	(0.01)	(0.02)
% classes taken P/NP	0.07	0.01	0.01	0.01	0.01
		(0.01)	(0.02)	(0.01)	(0.02)
Observations		374	332	374	332
Treatment assignment controls		Yes	No	Yes	Yes
Demographic controls		No	No	Yes	Yes
Pair Fixed Effects		No	No	No	Yes

Note: This table reports coefficients on $Incentive_i$ from Equations 1. GPA is measured on a 4.0 scale and is only affected by courses taken for a letter grade. Courses taken for Pass/No Pass (P/NP) have no bearing on GPA, nor do withdrawn courses. Model (1) contains linear controls for midterm 1 score and year; (2) is the difference in means and standard errors calculated using the repeated sampling framework of Neyman (1923); (3) and (4) use the post-double-selection (PDS) procedure of Belloni, Chernozhukov, and Hansen (2014b) to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A5. Models (2) and (4) include only students whose matched-pair did not attrite from the experiment. *Control Mean* is the mean for the Control students included in models (1) and (3), which is nearly identical to the mean for the Control students included in models (2) and (4). Standard errors in parentheses are robust to heteroskedasticity. ***, **, and * indicate significance at the 1, 5, and 10 percent critical levels, respectively.

Table 8: Spillover Effects of Incentive on Other Studying

	Control Mean	(1)	(2)	(3)	(4)
Attendance checks	5.91	-0.08	-0.09	-0.16	-0.10
		(0.18)	(0.17)	(0.17)	(0.18)
Discussion board views	49.81	10.64	8.51	10.64	3.69
		(7.64)	(8.25)	(7.60)	(8.05)
Discussion board days online	10.40	1.43	1.89	1.43	1.67
		(1.55)	(1.59)	(1.54)	(1.65)
Discussion board questions asked	0.53	0.32	0.30	0.32	0.30
		(0.25)	(0.30)	(0.25)	(0.31)
Discussion board answers	0.47	0.08	0.01	0.08	-0.02
		(0.26)	(0.28)	(0.26)	(0.28)
Tutoring visits	0.41	0.05	-0.01	0.07	0.00
		(0.13)	(0.14)	(0.12)	(0.12)
Observations		374	332	374	332
Treatment assignment controls		Yes	No	Yes	Yes
Demographic controls		No	No	Yes	Yes
Pair Fixed Effects		No	No	No	Yes

Note: This table reports coefficients on $Incentive_i$ from Equations 1. There were seven *Attendance checks* during the quarter. *Tutoring visits* includes those after the first midterm. Model (1) contains linear controls for midterm 1 score and year; (2) is the difference in means and standard errors calculated using the repeated sampling framework of Neyman (1923); (3) and (4) use the post-double-selection (PDS) procedure of Belloni, Chernozhukov, and Hansen (2014b) to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A5. Models (2) and (4) include only students whose matched-pair did not attrite from the experiment. *Control Mean* is the mean for the Control students included in models (1) and (3), which is nearly identical to the mean for the Control students included in models (2) and (4). Standard errors in parentheses are robust to heteroskedasticity. ***, **, and * indicate significance at the 1, 5, and 10 percent critical levels, respectively.

Table 9: Spillover Effects during Subsequent Quarter

	Control Mean	(1)	(2)	(3)	(4)
Panel A: Videos during subsequent quarter					
Num. of videos	25.46	14.00***	12.78*	11.70***	11.35
		(4.45)	(6.74)	(4.24)	(7.08)
Num. unique videos	19.77	9.87***	8.85**	8.25***	8.07**
		(3.03)	(4.04)	(2.92)	(4.12)
Hours of videos	3.82	2.14***	1.88*	1.79***	1.70
		(0.68)	(1.03)	(0.64)	(1.08)
Hours unique videos	2.90	1.51***	1.33**	1.27***	1.22**
		(0.45)	(0.60)	(0.44)	(0.61)
Observations		211	108	211	108
Panel B: Effects on classes passed					
Midterm 1 score		-0.04	-0.24	-0.04	-0.30
		(0.13)	(0.18)	(0.13)	(0.19)
		213	112	213	112
Midterm 2 score		0.00	-0.04	0.00	0.03
		(0.13)	(0.20)	(0.13)	(0.21)
		214	112	214	112
Final exam score		0.12	0.00	0.12	0.23
		(0.14)	(0.18)	(0.14)	(0.23)
		211	108	211	108
Panel C: Effects on class grade type					
Took Micro B	0.61	-0.07	-0.07	-0.07	-0.08
		(0.05)	(0.05)	(0.05)	(0.06)
Num. classes passed	3.46	-0.07	-0.05	-0.07	-0.04
		(0.11)	(0.12)	(0.11)	(0.12)
Num. classes not passed	0.23	0.07	0.08	0.07	0.07
		(0.06)	(0.06)	(0.06)	(0.06)
Num. classes withdrawn	0.06	0.04	0.04	0.04	0.03
		(0.03)	(0.03)	(0.03)	(0.03)
Observations		374	332	374	332
Treatment assignment controls		Yes	No	Yes	Yes
Demographic controls		No	No	Yes	Yes
Pair Fixed Effects		No	No	No	Yes

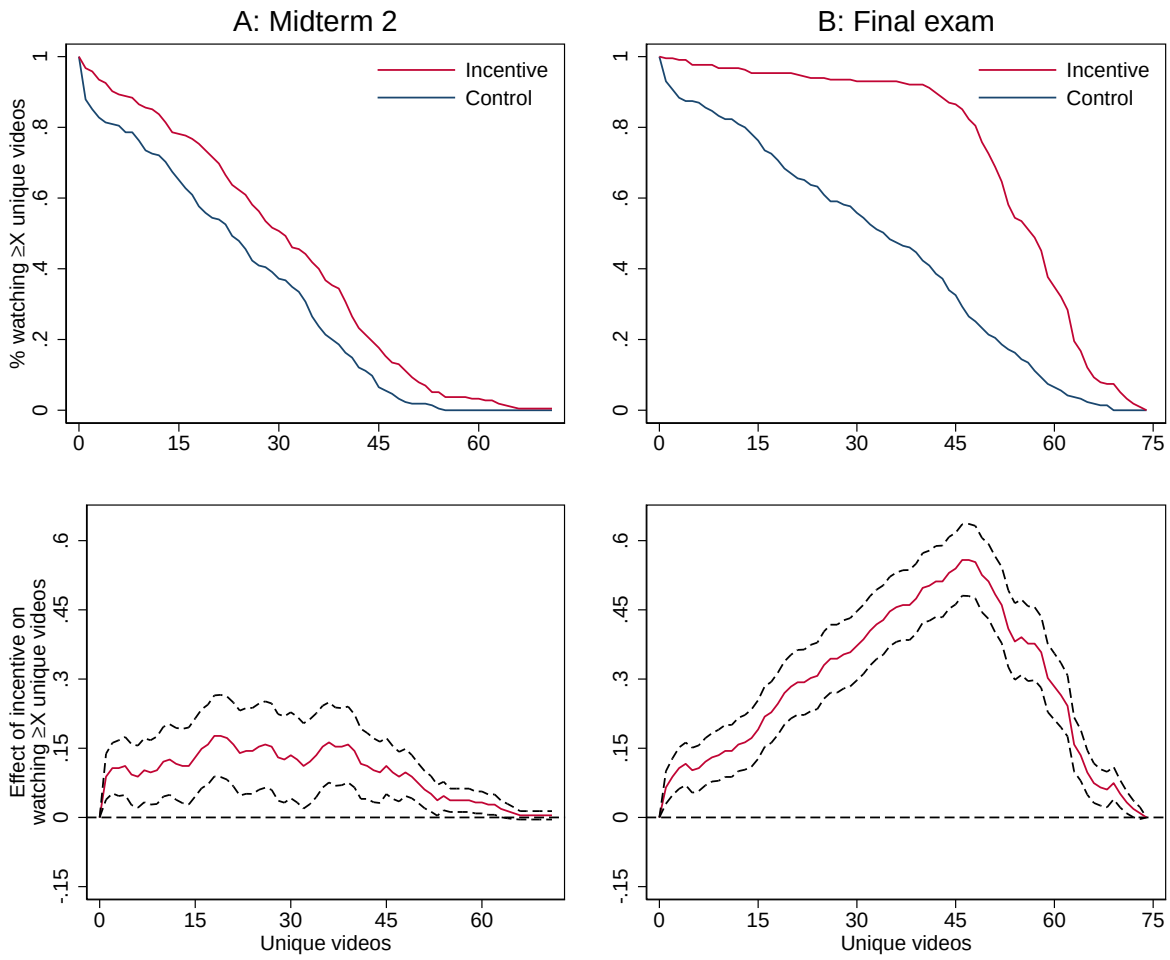
Note: This table reports coefficients on $Incentive_i$ from Equations 1. Panel A restricts the sample to those who completed both the first and second microeconomics courses (Micro A and B). Panel C includes those who completed the first microeconomics course (Micro A). Test scores are measured in standard deviation units. Model (1) contains linear controls for midterm 1 score and year; (2) is the difference in means and standard errors calculated using the repeated sampling framework of Neyman (1923); (3) and (4) use the post-double-selection (PDS) procedure of Belloni, Chernozhukov, and Hansen (2014b) to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A5. Models (2) and (4) include only students whose matched-pair did not attrite from the experiment. *Control Mean* is the mean for the Control students included in models (1) and (3), which is nearly identical to the mean for the Control students included in models (2) and (4). Standard errors in parentheses are robust to heteroskedasticity. ***, **, and * indicate significance at the 1, 5, and 10 percent critical levels, respectively.

Table 10: Heterogeneous Effects of Treatment

Interaction Variable	Midterm 2	Final Exam
Midterm 1 score	0.009	0.002
	(0.086)	(0.095)
Year = 2019	0.068	0.017
	(0.181)	(0.179)
Pretreatment videos	0.011	-0.004
	(0.007)	(0.007)
Pretreatment videos, unique	0.020*	0.001
	(0.011)	(0.011)
Transfer	-0.139	0.130
	(0.183)	(0.178)
Female	-0.208	-0.322*
	(0.185)	(0.188)
Asian	-0.402**	-0.215
	(0.189)	(0.180)
Latinx	0.237	0.022
	(0.254)	(0.227)
White	0.638**	0.271
	(0.260)	(0.250)
Other ethnicity	-0.113	0.276
	(0.307)	(0.347)
Observations	395	374

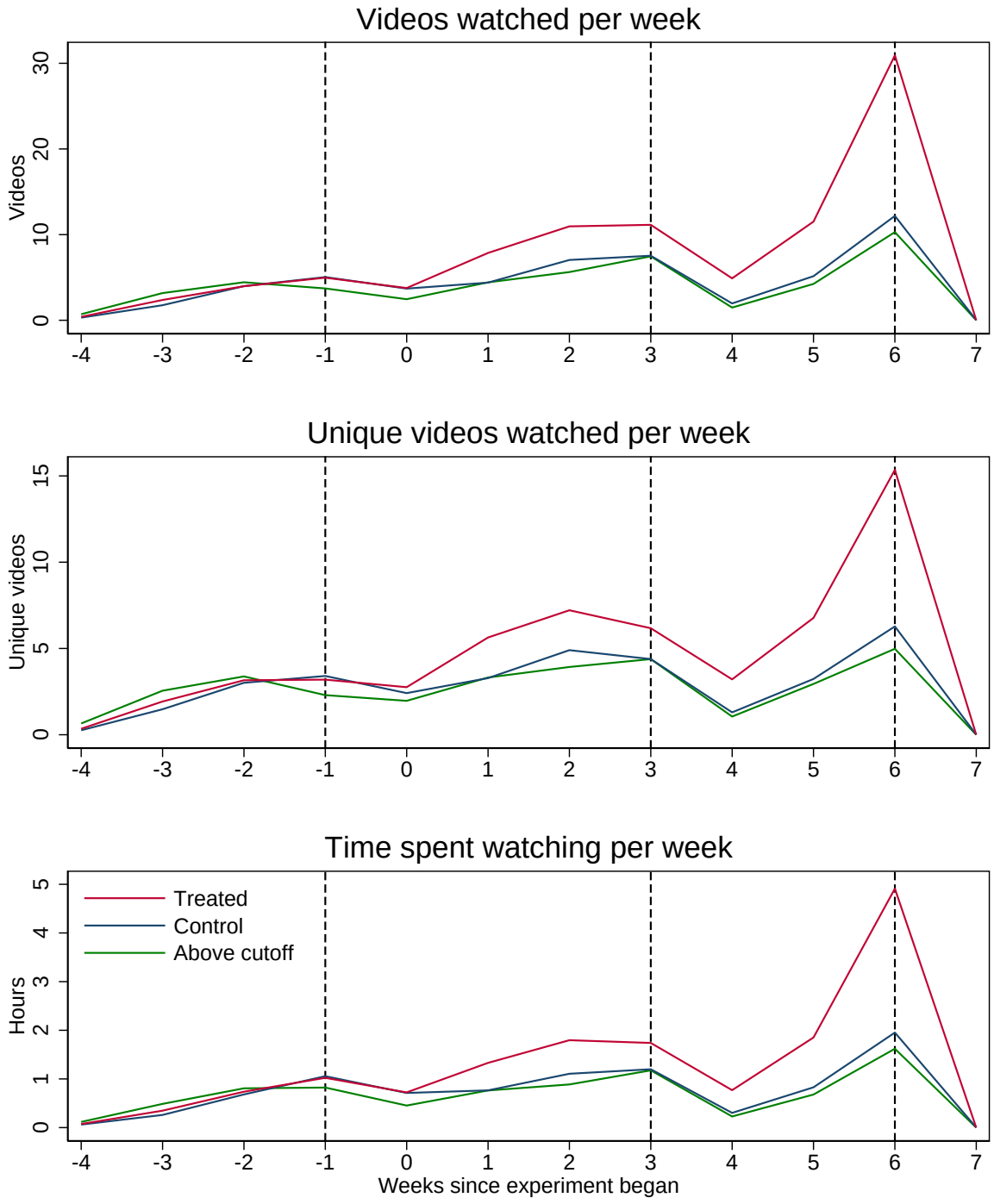
Note: This table reports estimates for β_2 from Equation 4. Test scores are measured in standard deviation units. Standard errors in parentheses are robust to heteroskedasticity. ***, **, and * indicate significance at the 1, 5, and 10 percent critical levels, respectively.

Figure 2: Effect of grade incentive on videos watched



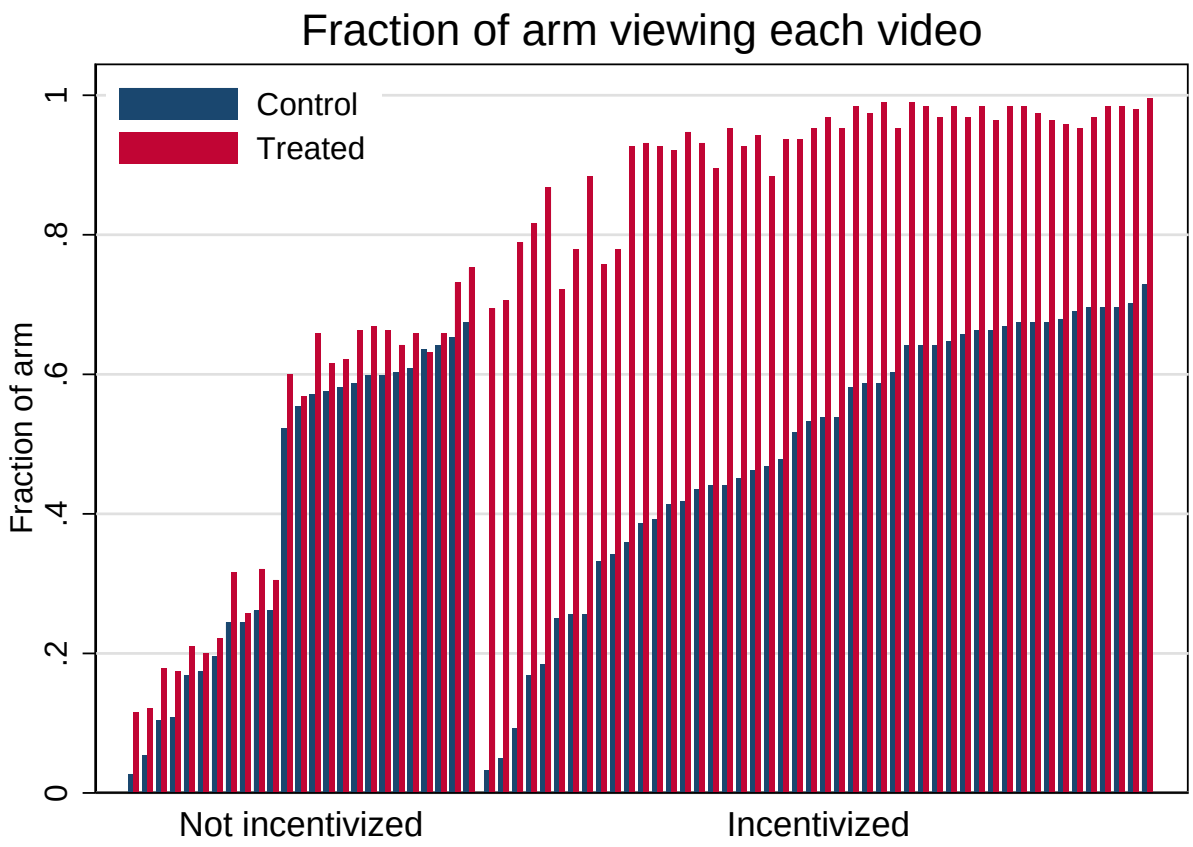
Top panels display the percent students in the *Control* and *Incentive* arms that watched at least X unique videos (left) or hours of unique videos (right). Bottom panels display the differences between the two arms in the top panels with 95% confidence intervals estimated by regressing an indicator for whether on the student watched at least $X \in \{0, \ldots, X_{\max}\}$ unique videos (or hours of unique videos) on the student's treatment status.

Figure 3: Weekly video watching by treatment arm



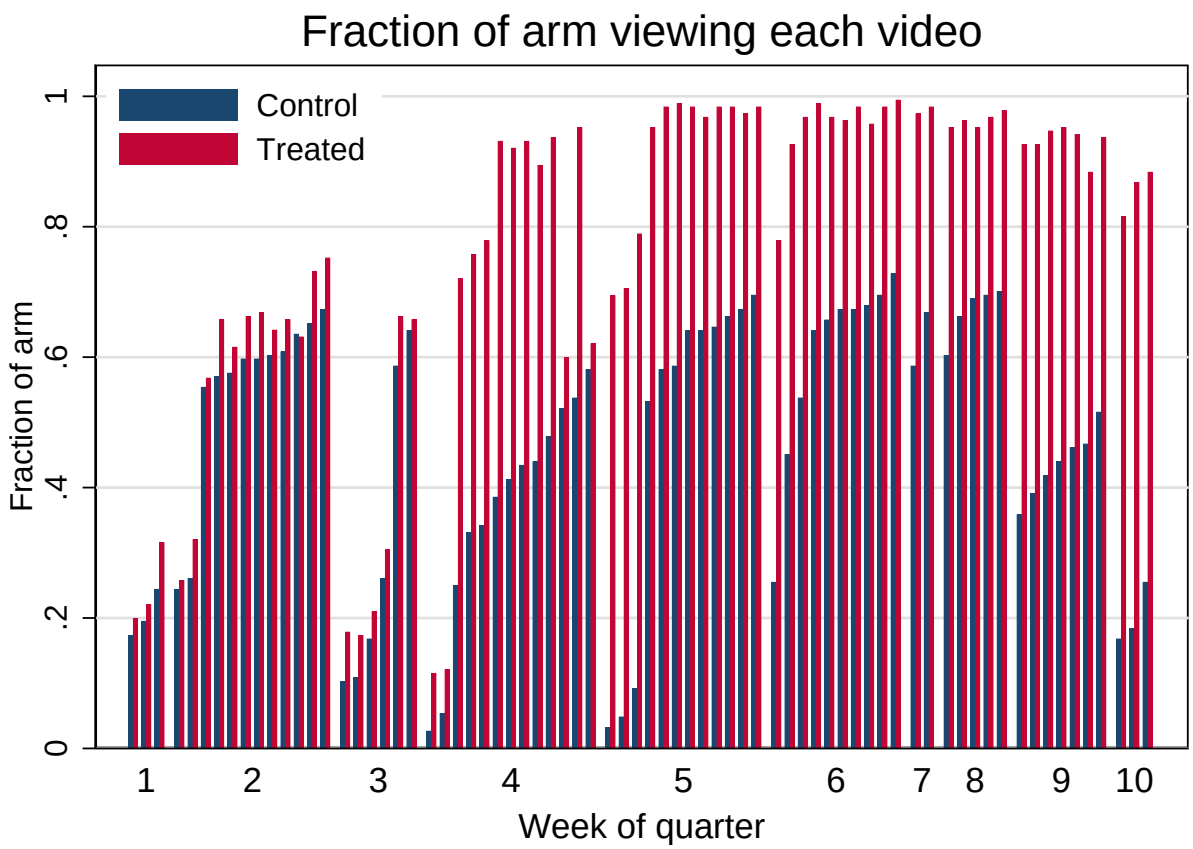
Dashed lines represent Midterm 1, Midterm 2, and Final exams

Figure 4: Video watch rates by video and treatment arm, grouped by incentive



Each bar represents the fraction of the treatment arm that watched a particular video. Bars are in order of control group watch rates separately for incentivized and non-incentivized videos.

Figure 5: Video watch rates by video and treatment arm, grouped by week



Each bar represents the fraction of the treatment arm that watched a particular video. Bars are in order of control group watch rates separately for each week the corresponding content was covered in lecture, as listed in the syllabus.

Appendix

A Additional experiment details

In this section we outline additional experiment details that could prove useful for replication or understanding our analysis choices.

A.1 Randomization

Students were assigned to treatment arms using a matched pairs design, a special case of blocked randomization in which each block contains exactly two units, one treated and one control. Several authors detail how matched pair designs can improve the *ex ante* precision of treatment effect estimates (versus complete randomization) by matching treatment units whose potential outcomes are similar (e.g. Athey and Imbens, 2017; Imbens and Rubin, 2015).

We were unable to observe most pretreatment covariates until after the experiment had concluded because of student privacy considerations, thereby making it impossible to block on these variables. We learned from the previous cohorts' data that between the first midterm score and math quiz score, both observable at the time of randomization, the midterm score predicted significantly more variation in the final exam score. Hence, we stratified on midterm score when assigning treatment. While we could have used an alternative method (e.g. matching methods) that take into consideration multiple covariates when assigning treatment, we opted for a simpler design given the high correlation between midterm and math quiz score and the comparatively high number of missing observations for the latter assessment (the math quiz was given on the second class day and so before some students enrolled in the class).

We assigned treatment shortly after issuing the first midterm exam grades, which occurred during the fourth week of the quarter. To assign treatment, we ordered the students by exam score, then paired students along this ordering for students below the median. Within pairs, we randomly assigned one student to *Incentive*, the other to *Control*. By construction, these two arms were *ex ante* balanced on midterm exam score, and we verified at time of treatment that the arms were also balanced on math quiz score. Since this random-

ization was performed independently across year cohorts, by construction, the samples were also balanced on year.

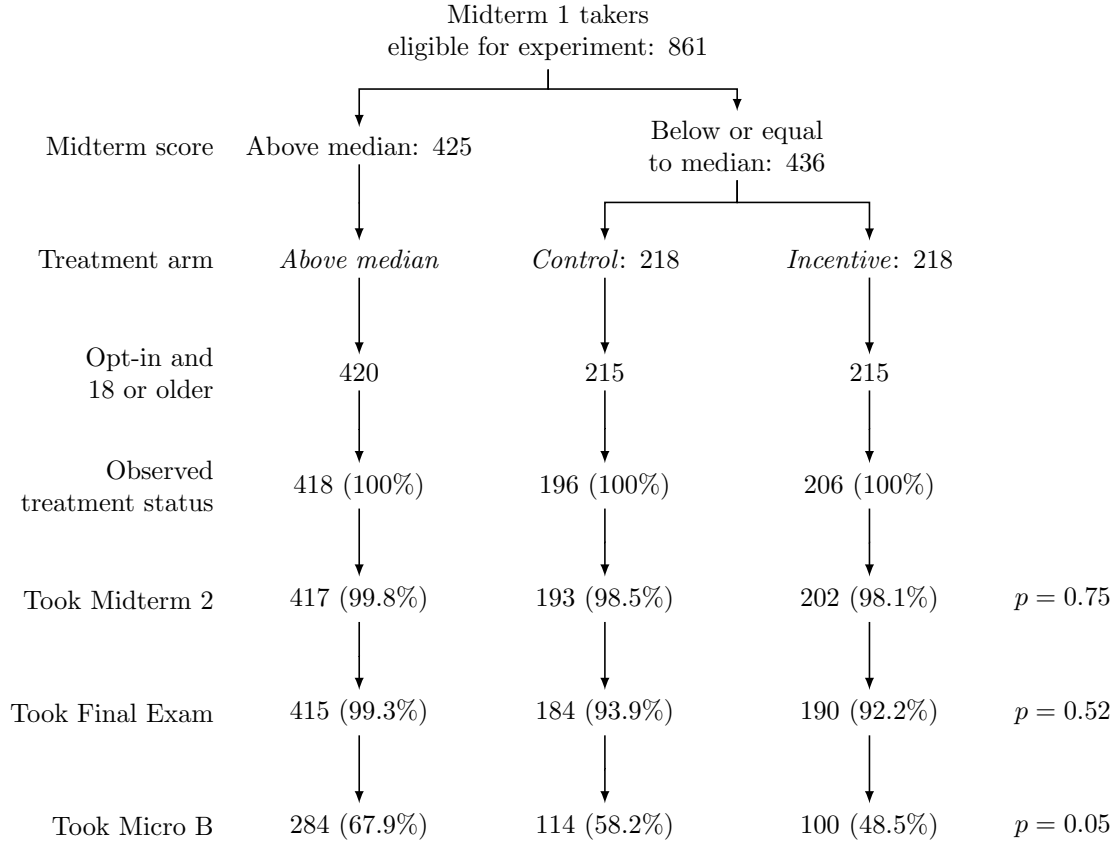
Although our treatment assignment method provides a better chance of balance than does simple random sampling, by random chance and through non-random attrition, it is possible that the two treatment arms vary on *ex post* observable and unobservable covariates that are correlated with the outcomes of interest, thereby confounding our treatment effect estimates. The primary cause of attrition was withdrawing from the course, which reduced our experiment sample by 35 students before the second midterm and an additional 21 students before the final exam. A 13% withdrawal rate is in line with those observed in previous quarters. Another cause of attrition, albeit not from the course, is age: four students under the age of 18 during the experiment were removed from the analysis dataset. Additionally, seven students opted out of having their data included in the experiment analysis.

Since neither the students' intent to withdraw, age, nor opt-out preferences were observable at the time of treatment assignment, we could not *ex ante* balance this attrition across treatment arms. If students attrited non-randomly, that is, decided to attrite depending on their treatment status, then our treatment effect estimates would be biased. Fortunately, despite 8% attrition before the second midterm and 13% before the final exam, the two treatment arms below the median are balanced on all observable pretreatment covariates, as shown in Table 4, which gives us confidence that the *Control* arm is a good counterfactual for the *Incentive* arm.

A.2 Attrition Details

Students could have attrited in three ways. The first, and largest source of attrition was withdrawing from the course. Most course withdrawals occurred *before* students learned about their treatment status (19 Controls and 9 Treated) as the university's penalty-free drop deadline was the Friday before the Monday when treatment status was revealed. Second, students under the age of 18 at the start of the experiment were excluded due to the IRB protocol. Finally, students who opted out of having their data included in the experiment were excluded. The analysis sample was prepared and anonymized by a campus-based inde-

Figure A1: Attrition by treatment arm



Percentages in parentheses are the portion of students who observe their treatment status and took the second midterm, final exam, and enrolled in Micro B, calculated separately for each treatment arm. The p-values are from a two-sample t-test of the equality of attrition rates between the *Control* and *Incentive* arms at each stage.

pendent education research organization, per IRB requirement, which removed four minors and seven opt-outs from the sample and merged demographic variables before returning the anonymized data to the research team.[37]

A.3 Selection of control variables

In this section we discuss how we select control variables included in our linear models.

Equation 1 includes a vector of control variables related linearly to the outcomes of interest. Although d_i, the treatment indicator is randomly assigned and in expectation d_i is orthogonal to all observed and unobserved pretreatment covariates, in small samples

[37] In total, five *Above median*, three *Control*, and three *Incentive* students were removed for age or opting-out.

stochastic imbalances can occur, which if controlled for can reduce bias of the treatment effect estimator (Athey & Imbens, 2017). Even if perfect balance is achieved, controlling for orthogonal covariates can improve precision of the treatment effect estimator if the covariates can predict unexplained variance in the outcome.

By definition it is not possible to guarantee balance on unobserved covariates. As discussed in Appendix A.1, we mechanically balanced the treatment arms on first midterm score, one of the few observables at the time of treatment assignment, with our knowledge from previous cohorts' data that the first midterm score explains a significant amount of variance in final exam score. Hence, in our estimation strategies including controls, we always include the first midterm score and year, following the recommendations of Bruhn and McKenzie (2009) to control for all covariates used to seek balance when assigning treatment.

For variables unobservable at time of randomization but observable at time of analysis, we lack the luxury of guaranteed balance by construction, nor is it clear *ex ante*, beyond our intuition, which will predict variation in the outcome variables of interest. On one hand, failing to control for valid predictors reduces statistical power. On the other hand, hand-picking control variables increases researcher degrees of freedom, risking increasing the prevalence of Type I errors (Simmons, Nelson, & Simonsohn, 2011). As such, in addition to a model without controls beyond the ones used for treatment assignment (year and midterm score), we fit a second model that includes a vector of linear controls chosen using the post-double-selection (PDS) procedure introduced by Belloni, Chernozhukov, and Hansen (2014b).

PDS is a two step process in which first, model covariates are selected in an automated, principled fashion, and second, the model coefficients of interest are estimated while controlling for those selected covariates. The first step involves predicting, separately, both the outcome of interest (e.g., videos watched) and treatment status using lasso regression, which shrinks coefficient estimates towards zero. Note that since treatment is randomly assigned, the lasso should shrink most, if not all, of the coefficients towards zero when predicting treatment status. Next, the researcher takes the union of all covariates with non-zero coefficients and includes these covariates as controls in her model. With her control variables selected, she can now estimate treatment effects with reduced bias relative to including controls with

less empirical rationale.

In Table A4, we describe all covariates observable in our study. In Table A5, we describe the covariates selected as controls for estimating the effect of treatment on each outcome variable of interest. All models include year and first midterm score as controls. To ensure these controls are "selected" by the PDS procedure, we partialed out these controls from the first step prediction models by residualizing both sides of the equation as described in Belloni, Chernozhukov, and Hansen (2014a).

B LATE estimators using Neyman's repeated sampling approach

In this section we derive LATE estimators using the repeated sampling approach of Neyman (1923), which considers each pair as an independent, completely randomized experiment.

Similar to a Wald estimator, the point estimate of the LATE is the mean within-pair difference in outcome divided by the mean within-pair difference in videos:

$$\widehat{\gamma} = \frac{\overline{\Delta y}}{\overline{\Delta v}} = \frac{\frac{1}{J}\sum_{j=1}^{J} \Delta y_j}{\frac{1}{J}\sum_{j=1}^{J} \Delta v_j} = \frac{\overline{y_I} - \overline{y_C}}{\overline{v_I} - \overline{v_C}} \tag{5}$$

where y is the outcome of interest (grades) and v is the number of videos, both indexed by pair $j \in J$ and treatment status C or I for *Control* or *Incentive*, respectively.

We use the delta method to calculate the approximate standard error of $\widehat{\gamma}$. First, we define the following normally-distributed random variables:

$$\begin{aligned} Y &= \overline{y_I} - \overline{y_C} \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \\ V &= \overline{v_I} - \overline{v_C} \sim \mathcal{N}(\mu_V, \sigma_V^2) \end{aligned} \tag{6}$$

Using a first-order Taylor expansion and letting $g() = \frac{Y}{V}$, we have:

$$\begin{aligned} \mathrm{Var}(g) &= \mathrm{E}[(g - \mathrm{E}(g))^2] \\ &\approx \mathrm{E}[(g(\theta) + (Y - \theta_Y)g'_Y(\theta) + (V - \theta_V)g'_V(\theta) - g(\theta))^2] \\ &= \mathrm{E}[(Y - \theta_Y)^2(g'_Y(\theta))^2 + (V - \theta_V)^2(g'_V(\theta))^2 + 2(Y - \theta_Y)(V - \theta_V)g'_Y(\theta)g'_V(\theta)] \\ &= \mathrm{Var}(Y)(g'_Y(\theta))^2 + \mathrm{Var}(V)(g'_V(\theta))^2 + 2\mathrm{Cov}(Y, V)g'_Y(\theta)g'_V(\theta) \end{aligned} \tag{7}$$

Expanding about $\theta = (\theta_Y, \theta_V) = (\mu_Y, \mu_V)$ and letting $g'_Y(\theta) = \mu_V^{-1}$ and $g'_V(\theta) = \frac{-\mu_Y}{\mu_V^{-2}}$:

$$\begin{aligned}\mathrm{Var}(g) &\approx \frac{1}{\mu_V^2}\mathrm{Var}(Y) + \frac{\mu_Y^2}{\mu_V^4}\mathrm{Var}(V) + 2\frac{-\mu_Y}{\mu_V^{-2}}\mathrm{Cov}(Y,V) \\ &= \frac{\mu_Y^2}{\mu_V^2}(\frac{\sigma_Y^2}{\mu_Y^2} + \frac{\sigma_V^2}{\mu_V^2} - 2\frac{\mathrm{Cov}(Y,V)}{\mu_Y\mu_V})\end{aligned} \tag{8}$$

$$\begin{aligned}\widehat{\mathrm{Var}}(Y) = \widehat{\sigma_Y^2} &= \frac{1}{J(J-1)}\sum_{j=1}^{J}(\Delta y_j - \overline{\Delta y})^2 \\ \widehat{\mathrm{Var}}(V) = \widehat{\sigma_V^2} &= \frac{1}{J(J-1)}\sum_{j=1}^{J}(\Delta v_j - \overline{\Delta v})^2 \\ \widehat{\mathrm{Cov}}(Y,V) = \widehat{\sigma_{YV}} &= \frac{1}{J(J-1)}\sum_{j=1}^{J}(\Delta y_j - \overline{\Delta y})(\Delta v_j - \overline{\Delta v})\end{aligned} \tag{9}$$

and the following estimators for the population means of Y and V:

$$\begin{aligned}\widehat{\mu_Y} &= \mathrm{E}(\mu_Y) = \overline{\Delta y} \\ \widehat{\mu_V} &= \mathrm{E}(\mu_V) = \overline{\Delta v}\end{aligned} \tag{10}$$

Substituting these variance and means estimators into the final step of 8, we arrive at the standard error estimator for $\widehat{\gamma}$:

$$\widehat{\sigma_\gamma} = \frac{\overline{\Delta y}}{\overline{\Delta v}}\sqrt{\frac{\widehat{\sigma_Y^2}}{\overline{\Delta y}^2} + \frac{\widehat{\sigma_V^2}}{\overline{\Delta v}^2} - 2\frac{\widehat{\sigma_{YV}}}{\overline{\Delta y}\overline{\Delta v}}} \tag{11}$$

Table A1: Balance test, full sample

	All Students			P-values
Variable	Above Median	Control	Incentive	(3) - (2)
Midterm 1 score	0.837	-1.127	-1.167	0.671
	(0.025)	(0.066)	(0.067)	
Year = 2019	0.490	0.505	0.490	0.767
	(0.024)	(0.036)	(0.035)	
Cumulative GPA	3.445	2.931	2.941	0.890
	(0.029)	(0.045)	(0.057)	
No cum. GPA	0.232	0.372	0.330	0.375
	(0.021)	(0.035)	(0.033)	
Math quiz score	0.589	0.016	0.093	0.424
	(0.044)	(0.071)	(0.065)	
Tutoring visits	0.268	0.255	0.223	0.691
	(0.042)	(0.058)	(0.055)	
Videos watched	13.196	13.321	13.704	0.763
	(0.680)	(0.875)	(0.915)	
Videos, unique	9.722	9.668	10.155	0.558
	(0.431)	(0.572)	(0.602)	
Hours videos	2.470	2.522	2.576	0.824
	(0.127)	(0.168)	(0.173)	
Hours videos, unique	1.725	1.752	1.834	0.578
	(0.075)	(0.103)	(0.106)	
Asian	0.701	0.694	0.670	0.607
	(0.022)	(0.033)	(0.033)	
Latinx	0.060	0.133	0.155	0.518
	(0.012)	(0.024)	(0.025)	
White	0.151	0.117	0.121	0.902
	(0.018)	(0.023)	(0.023)	
Other ethnicity	0.089	0.056	0.053	0.905
	(0.014)	(0.016)	(0.016)	
Female	0.392	0.337	0.393	0.241
	(0.024)	(0.034)	(0.034)	
Male	0.593	0.658	0.597	0.206
	(0.024)	(0.034)	(0.034)	
Transfer	0.273	0.485	0.447	0.445
	(0.022)	(0.036)	(0.035)	
Observations	418	196	206	

Note: This table includes all students who saw their treatment assignment. Descriptions of each variable can be found in Table A4. *Female* and *Male* are coded zero for nine students who do not report a gender. *P-values* are reported for the Welch's t-test of equal means between the *Control* and *Incentive* arms. Standard errors in parentheses are robust to heteroskedasticity. ***, **, and * indicate significance at the 1, 5, and 10 percent critical levels, respectively.

Table A2: Baseline balance test, matched pairs

	Midterm 2		P-values	Final		P-values
Variable	Control	Incentive	(2) - (1)	Control	Incentive	(4) - (3)
Midterm 1 score	-1.070	-1.078	0.933	-1.032	-1.039	0.938
	(0.065)	(0.066)		(0.064)	(0.065)	
Year = 2019	0.514	0.514	1.000	0.518	0.518	1.000
	(0.037)	(0.037)		(0.039)	(0.039)	
Cumulative GPA	2.942	2.992	0.487	2.929	3.001	0.346
	(0.045)	(0.056)		(0.047)	(0.059)	
No cum. GPA	0.365	0.320	0.377	0.367	0.313	0.299
	(0.036)	(0.035)		(0.038)	(0.036)	
Math quiz score	0.054	0.137	0.396	0.061	0.157	0.338
	(0.071)	(0.068)		(0.071)	(0.071)	
Tutoring visits	0.276	0.232	0.612	0.283	0.253	0.746
	(0.062)	(0.061)		(0.066)	(0.066)	
Videos watched	13.663	13.729	0.961	13.729	13.789	0.966
	(0.929)	(0.986)		(0.978)	(1.023)	
Videos, unique	9.845	10.116	0.760	9.795	10.181	0.674
	(0.606)	(0.644)		(0.630)	(0.665)	
Hours videos	2.601	2.584	0.946	2.605	2.600	0.985
	(0.178)	(0.185)		(0.187)	(0.192)	
Hours videos, unique	1.790	1.833	0.781	1.780	1.853	0.652
	(0.108)	(0.114)		(0.113)	(0.117)	
Asian	0.713	0.652	0.215	0.711	0.633	0.129
	(0.034)	(0.036)		(0.035)	(0.038)	
Latinx	0.133	0.166	0.377	0.139	0.169	0.448
	(0.025)	(0.028)		(0.027)	(0.029)	
White	0.105	0.138	0.336	0.102	0.145	0.244
	(0.023)	(0.026)		(0.024)	(0.027)	
Other ethnicity	0.050	0.044	0.804	0.048	0.054	0.804
	(0.016)	(0.015)		(0.017)	(0.018)	
Female	0.343	0.392	0.328	0.337	0.404	0.212
	(0.035)	(0.036)		(0.037)	(0.038)	
Male	0.652	0.602	0.329	0.657	0.584	0.176
	(0.036)	(0.036)		(0.037)	(0.038)	
Transfer	0.470	0.436	0.528	0.470	0.416	0.321
	(0.037)	(0.037)		(0.039)	(0.038)	
Observations	181	181		166	166	

Note: This table includes all matched pairs who both complete the second midterm or final exams. Descriptions of each variable can be found in Table A4. *Female* and *Male* are coded zero for nine students who do not report a gender. *P-values* are reported for the Welch's t-test of equal means between the *Control* and *Incentive* arms. Standard errors in parentheses are robust to heteroskedasticity. ***, **, and * indicate significance at the 1, 5, and 10 percent critical levels, respectively.

Table A3: Anderson-Rubin confidence sets

Outcome variable	Endogenous variable	Anderson-Rubin CI
Midterm 2 score	10 unique videos	[-0.001, 0.653]
Midterm 2 score	1 hour videos	[-0.001, 0.392]
Final exam score	10 unique videos	[0.000, 0.163]
Final exam score	1 hour videos	[0.000, 0.102]

Note: This table displays Anderson-Rubin confidence sets at the 95% confidence level for the 2SLS estimator $\hat{\gamma}$ from Equation 3 including year dummies and first midterm score as controls. Outcomes are measured in standard deviations. Instrumented endogenous variables are measured in 10s of unique videos or hours of unique content.

Table A4: Candidate control variables for post-double-selection

Variable	Description
Midterm 1 score	Score on the first midterm
Year = 2019	1 if course taken in 2019, 0 otherwise
Cumulative GPA	Cumulative GPA from prior term, 0 if not observed
No cum. GPA	1 if Cumulative GPA unobserved, 0 otherwise
Math quiz score	Score on a quiz assessing prerequisite math skills
Tutoring visits	Number of group tutoring lab visits as of the first midterm
Videos watched	Number unique videos watched as of the first midterm
Hours videos	Hours of unique videos watched as of the first midterm
Asian	1 if ethnicity is Asian, 0 otherwise
Latinx	1 if ethnicity is Latinx, 0 otherwise
White	1 if ethnicity is White, 0 otherwise
Female	1 if female, 0 otherwise
Transfer	1 if transfer student, 0 otherwise

Note: *Midterm 1 score* and *Math quiz score* are measured in control standard deviations. *Cumulative GPA* is measured on a 4.0 scale. Videos included in *Videos watched* and *Hours videos* are unique course-relevant videos. The ethnicity variables are coded by university records: *Asian* includes "Chinese/Chinese American", "Vietnamese", "East Indian/Pakistani", "Japanese/Japanese American", "Korean/Korean American", and "All other Asian/Asian American"; *Latinx* includes "Mexican/Mexican American", "Chicano", and "All other Spanish-American/Latino"; *White* includes "White/Caucasian"; and the omitted category inludes "African American/Black", "Pacific Islander", and "Not give/declined to state".

Table A5: ITT model controls selected via post-double-selection

Table	Dependent Variable	Controls, All Observations	Controls, Fixed Effects
Table 1	Hours unique videos by Final	Hours videos Videos	Hours videos Videos
	Hours unique videos by Mid. 2	Hours videos	Hours videos Videos
	Hours videos by Final	Hours videos	Hours videos
	Hours videos by Mid. 2	Hours videos	Hours videos Tutoring visits Videos
	Num. unique videos before Final	Hours videos Videos	Videos
	Num. unique videos before Mid. 2	Hours videos Videos	Videos
	Num. videos before Final	Hours videos Videos	Hours videos Videos
	Num. videos before Mid. 2	Hours videos Videos	Hours videos Tutoring visits Videos
Table 2	Final exam score	None	Math quiz score Transfer
	Midterm 2 score	None	Math quiz score
Table 3	All classes	Cumulative GPA	Cumulative GPA Math quiz score Transfer
	Econ classes ex. Micro A	None	Cumulative GPA Transfer
	Excluding Micro A	Cumulative GPA	Transfer
	Excluding econ classes	None	None
	Letter grade in Micro A	Cumulative GPA Latinx Transfer	Cumulative GPA
	Num. classes not passed	None	None
	Num. classes passed	Cumulative GPA Transfer	Cumulative GPA Transfer
	Num. classes taken P/NP	Latinx	Latinx
	Num. classes taken for letter	Cumulative GPA No cum. GPA	Cumulative GPA
	Num. classes withdrawn	None	None
	Num. units taken P/NP	Latinx	Latinx
	Num. units taken for letter grade	Cumulative GPA No cum. GPA	Cumulative GPA
	Num. units withdrawn	None	None
	% classes taken P/NP	None	Latinx
	% classes taken for letter	None	Latinx

Continued on next page

Table A5 (continued)

Table 4	Attendance checks	Female Math quiz score Tutoring visits	Tutoring visits
	Discussion board answers	None	None
	Discussion board days online	None	None
	Discussion board questions asked	None	None
	Discussion board views	None	Asian
	Tutoring visits	Tutoring visits	Tutoring visits
Table 5	Hours of videos	Hours videos	Hours videos Latinx Math quiz score Tutoring visits Videos
	Midterm 1 score	None	Latinx Math quiz score Videos
	Midterm 2 score	None	Asian Latinx Math quiz score Videos
	Num. classes not passed	None	None
	Num. classes passed	None	None
	Num. classes taken P/NP	None	Transfer
	Num. classes taken for letter	None	No cum. GPA
	Num. classes withdrawn	None	None
	Num. of videos	Hours videos	Hours videos Latinx Math quiz score Tutoring visits Videos
	Num. units taken P/NP	None	Transfer
	Num. units taken for letter grade	None	None
	Num. units withdrawn	None	None
	Term GPA	Cumulative GPA	Cumulative GPA Tutoring visits
	Term GPA, econ courses ex. Micro B, winter	None	Math quiz score
	Term GPA, ex. Micro B	Cumulative GPA	Cumulative GPA Tutoring visits
	Term GPA, ex. econ courses	None	Tutoring visits
	Took Micro B	None	Math quiz score
	% classes taken P/NP	None	No cum. GPA Transfer
	% classes taken for letter	None	No cum. GPA Transfer
Table None	Final exam score	None	Latinx Math quiz score Videos

Continued on next page

Table A5 (continued)

Hours unique videos	Hours videos	Hours videos
		Latinx
		Math quiz score
		Tutoring visits
		Videos
Num. unique videos	Hours videos	Hours videos
		Latinx
		Math quiz score
		Tutoring visits
		Videos
Pass Micro B	None	Latinx
		Math quiz score
		Videos

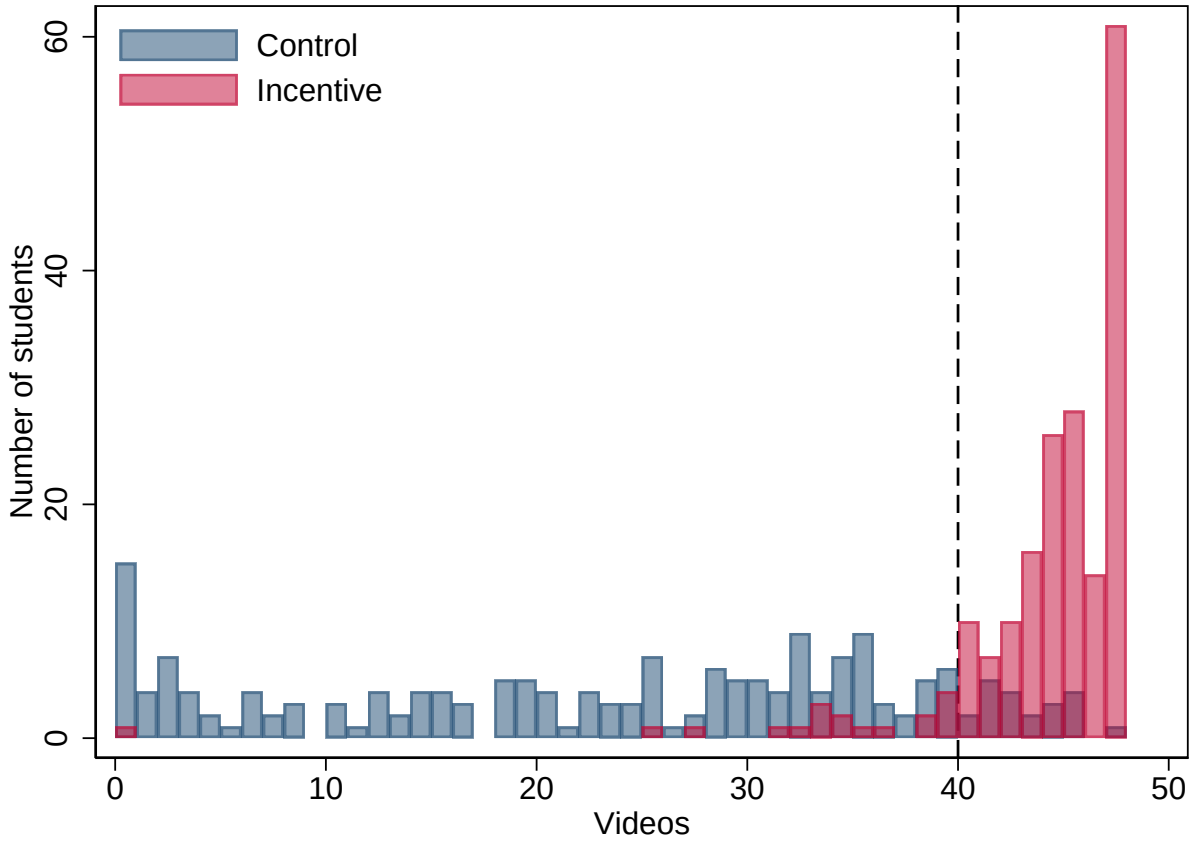
Note: Controls chosen via the PDS procedure of Belloni, Chernozhukov, and Hansen (2014b). In the *All Observations* model, *Midterm 1 score* and *Year = 2019* are additionally included as controls. In the *Fixed Effects* model, pair fixed effects and *Midterm 1 score* are included. All control variables are measured before the start of the experiment, e.g. *Hours videos* is the hours of videos watched as of the first midterm.

Table A6: LATE model controls selected via post-double-selection

Dependent Variable	Instrumented	Controls, All Observations	Controls, Fixed Effects
Final exam score	Hours videos, unique	Hours videos Math quiz score Transfer Videos	Hours videos Videos
Final exam score	Videos, unique	Hours videos Videos	Hours videos Videos
Midterm 2 score	Hours videos, unique	Hours videos Math quiz score Tutoring visits Videos	Hours videos
Midterm 2 score	Videos, unique	Hours videos Videos	Hours videos Videos

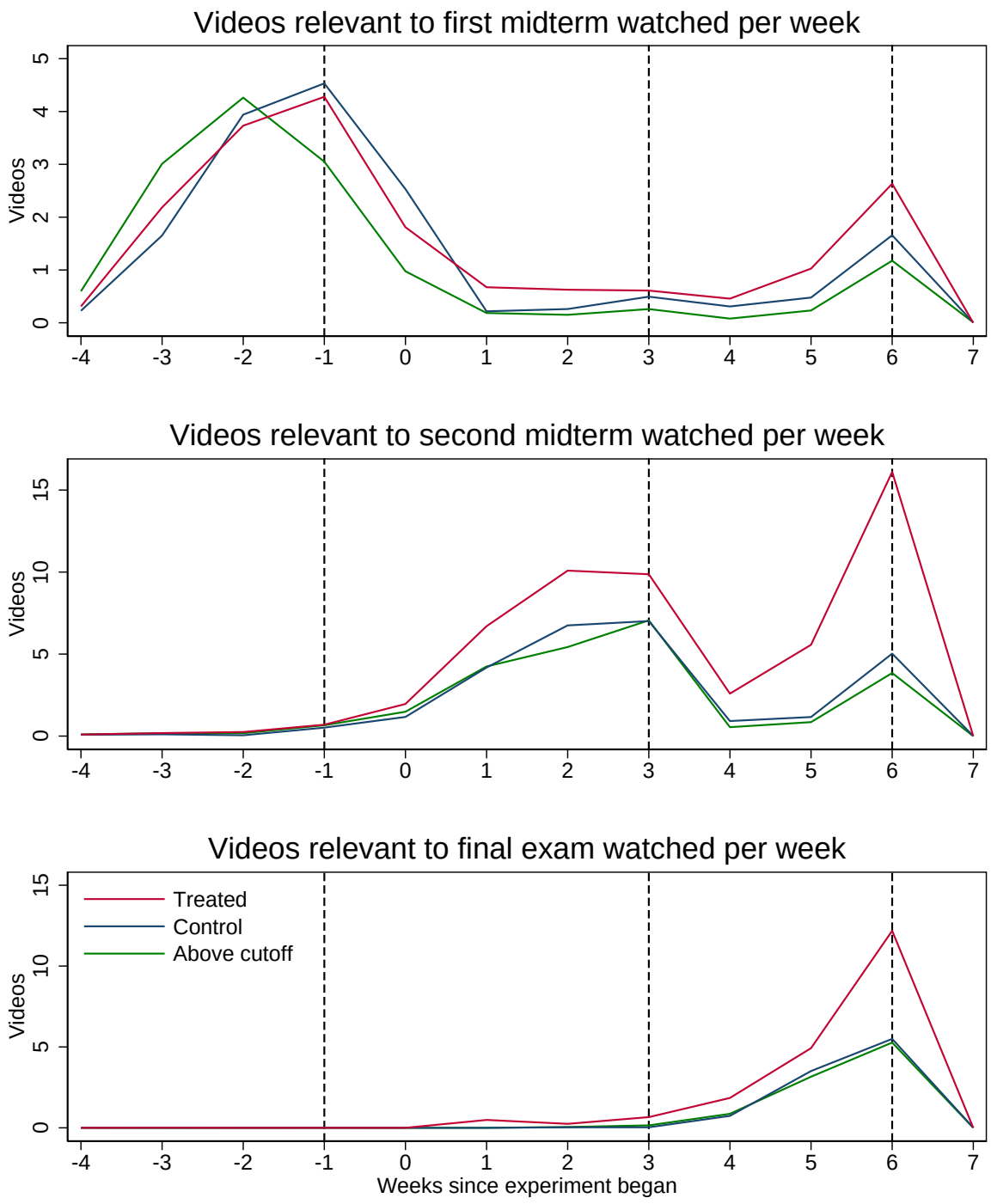
Note: Controls chosen via the PDS procedure of Belloni, Chernozhukov, and Hansen (2014b). In the *All Observations* model, *Midterm 1 score* and *Year = 2019* are additionally included as controls. In the *Fixed Effects* model, pair fixed effects and *Midterm 1 score* are included. All control variables are measured before the start of the experiment, e.g. *Hours videos* is the hours of videos watched as of the first midterm.

Figure A2: Distribution of videos counted towards incentive



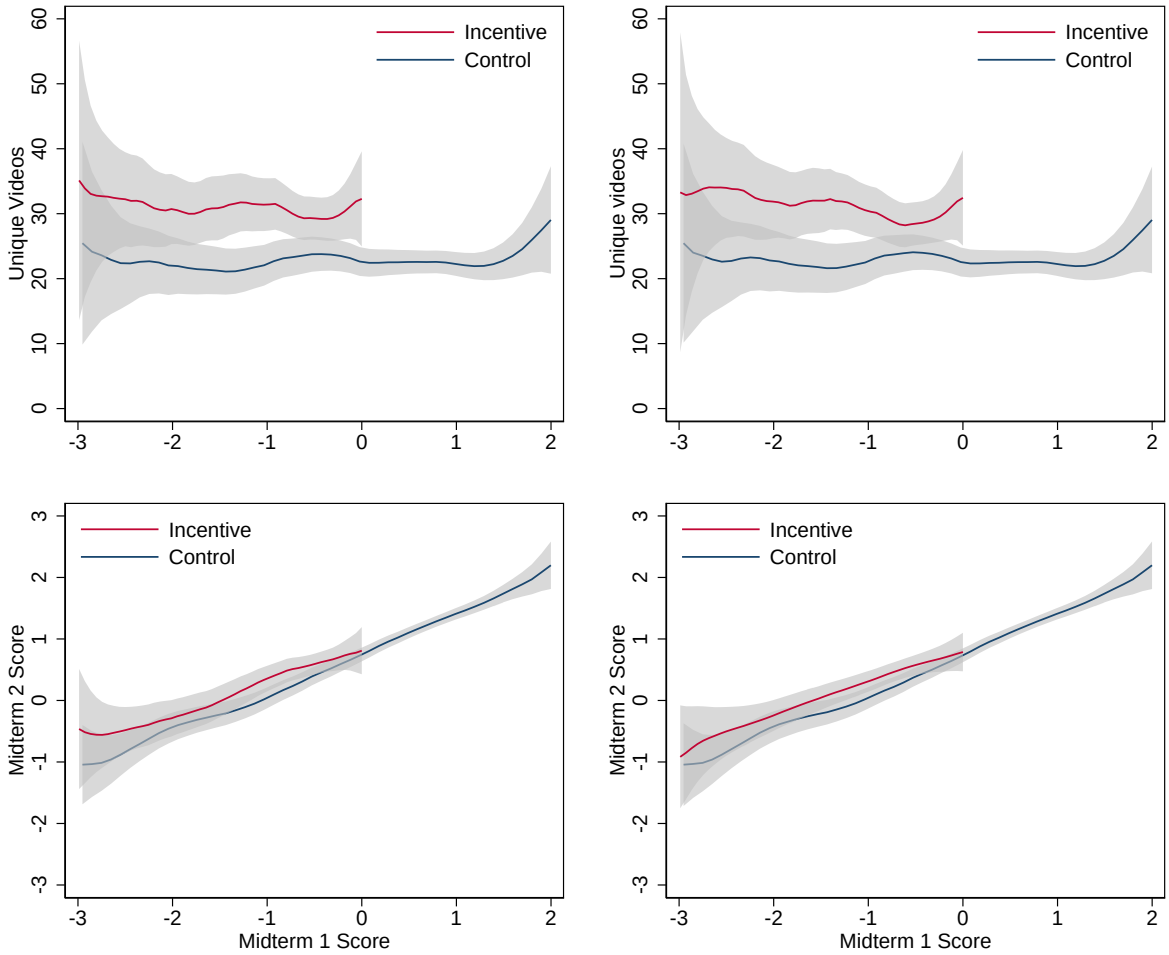
This plot includes only videos that would have counted towards the earning the grade incentive. Students were required to watch 40 unique of 48 eligible videos between the first midterm and final exam to earn the grade incentive. 91% of *Incentive* students met the requirements for the grade incentive versus 11% of *Control* students.

Figure A3: Weekly video watching by exam topic



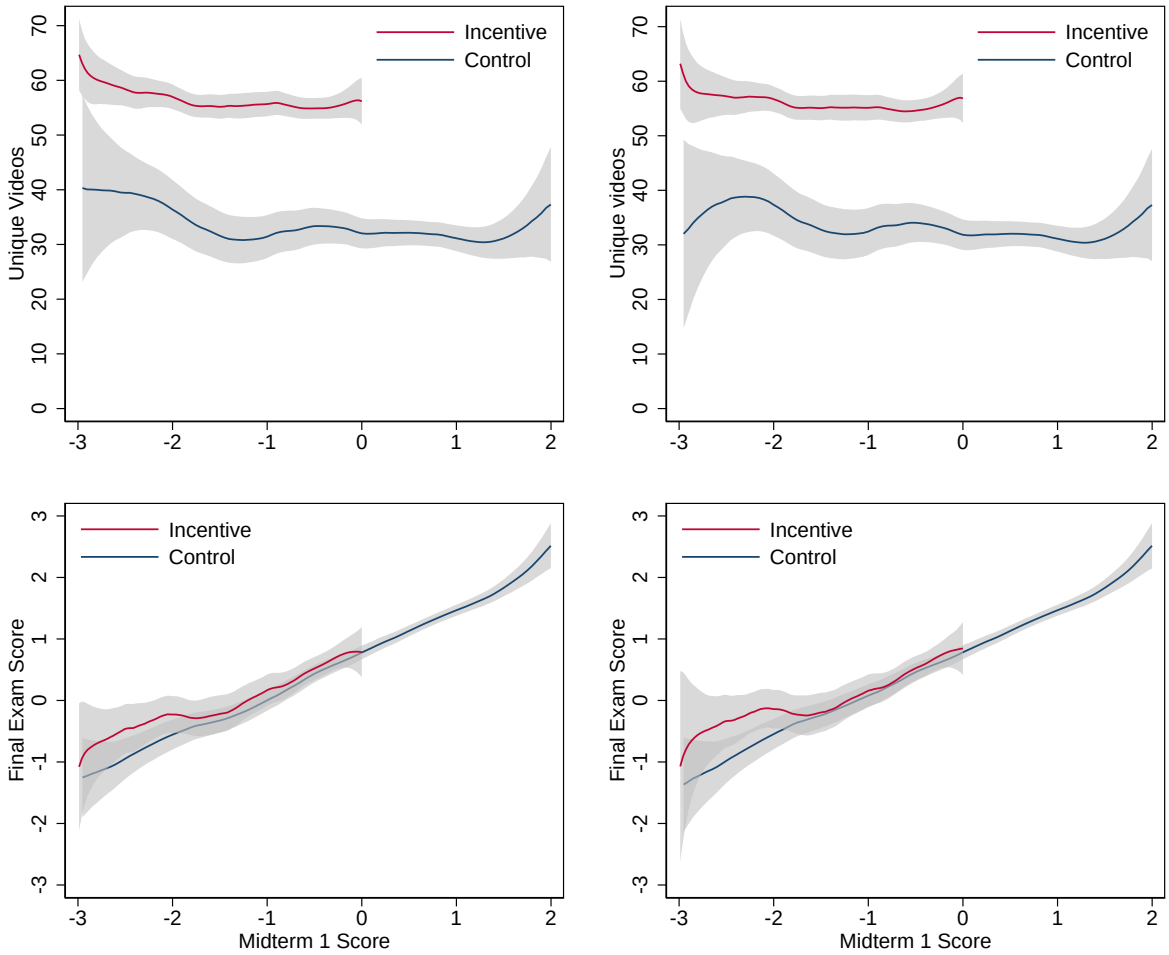
Dashed lines represent Midterm 1, Midterm 2, and Final exams.

Figure A4: Effects of treatment along first midterm score, by midterm 2



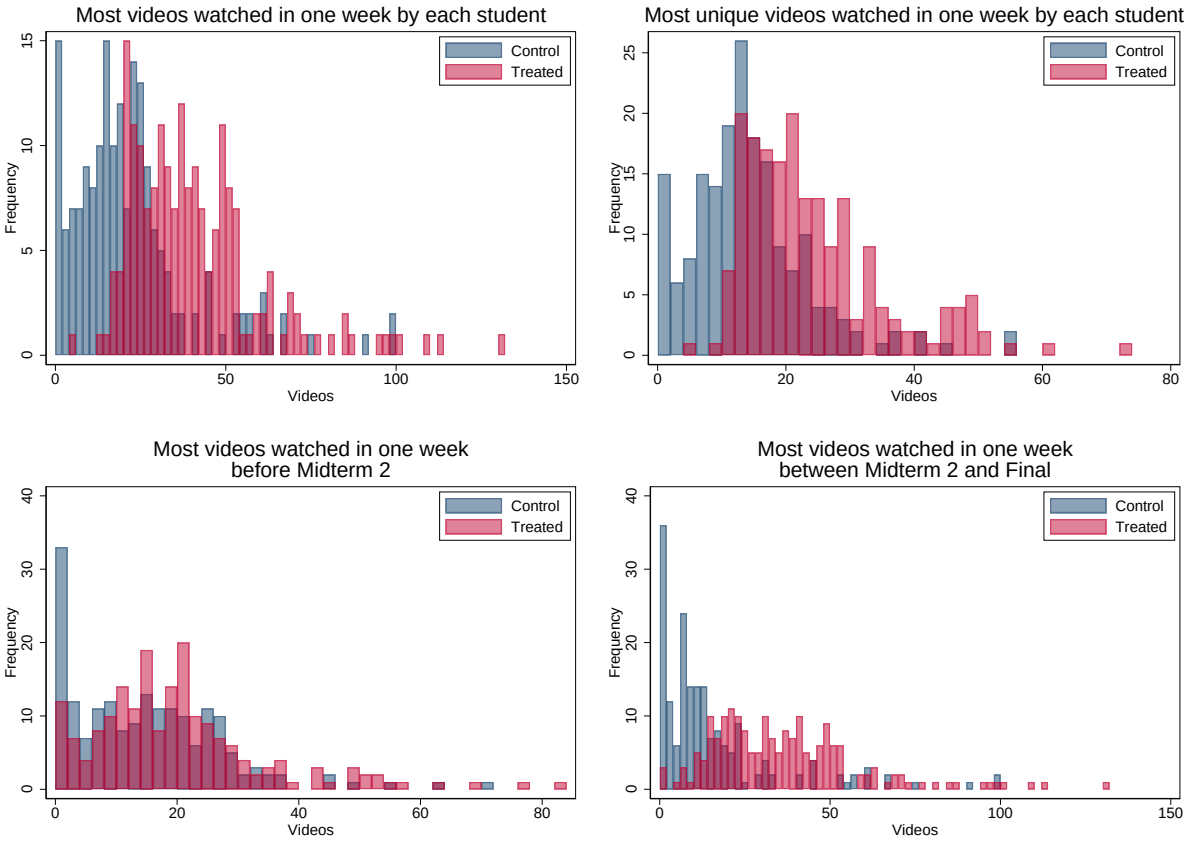
Videos (top) includes unique videos watched before the second midterm exam. Exam scores (bottom) are measured in control standard deviations. Confidence bands represent 95% confidence intervals of the conditional mean outcome. The left plots includes all students who took the second midterm while the right plots exclude any students whose matched pair attrited.

Figure A5: Effects of treatment along first midterm score, by final



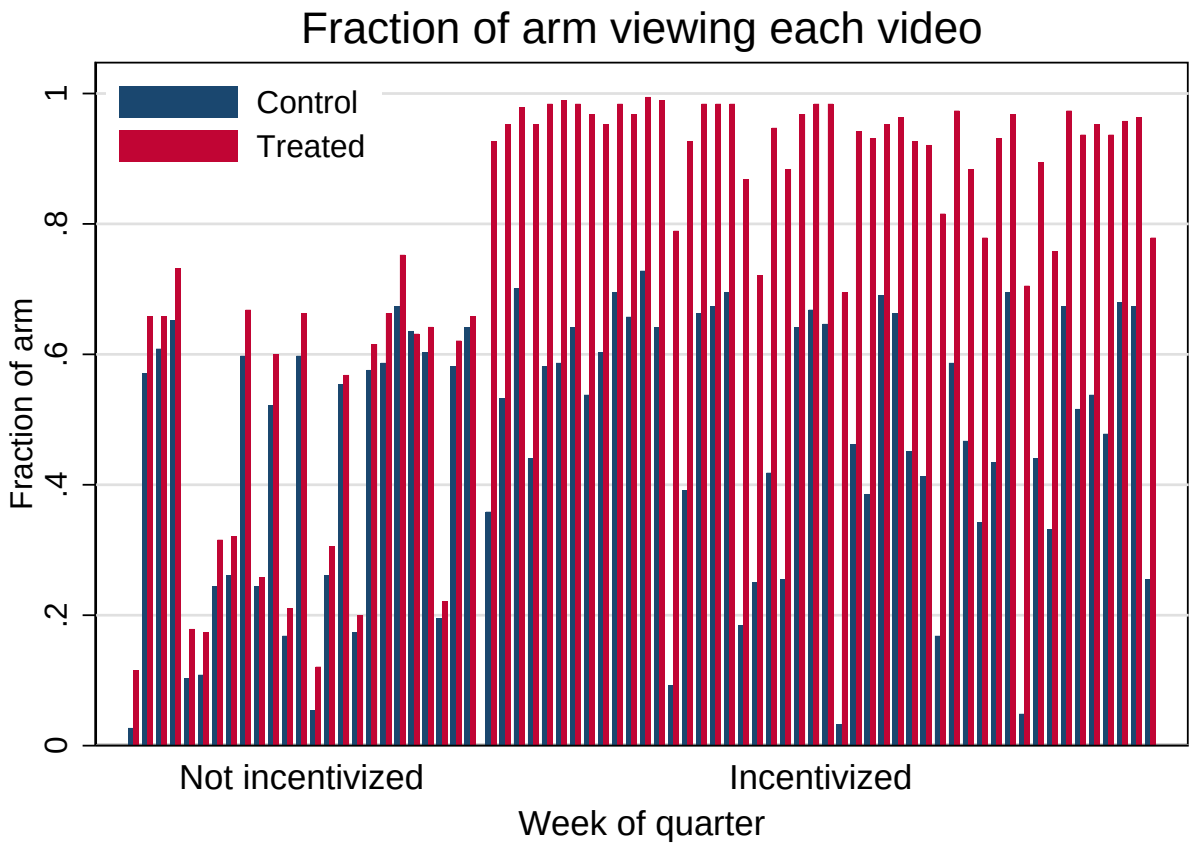
Videos (top) includes unique videos watched before the final exam. Exam scores (bottom) are measured in control standard deviations. Confidence bands represent 95% confidence intervals of the conditional mean outcome. The left plots includes all students who took the final exam while the right plots exclude any students whose matched pair attrited.

Figure A6: Distribution of max videos watched in one week



These plots help illustrate potential "binge watching" behavior. Compared to the *Control* students, *Incentive* students are more likely to watch 40 or more unique videos in a week, which occurs in the weeks preceding the final and not the second midterm.

Figure A7: Video watch rates by video and treatment arm, grouped by incentive, ordered by video duration



Each bar represents the fraction of the treatment arm that watched a particular video. Bars are in order of video duration separately for incentivized and non-incentivized videos.