The effect of supplementary video lectures on learning in intermediate microeconomics

Melissa Famulari and Zachary A. Goodman*

University of California, San Diego

This version: February 2020

Abstract

In this paper we estimate the effectiveness of a novel video-based textbook replacement for intermediate microeconomics, the Intermediate Microeconomic Video Handbook (IMVH), on learning outcomes. In a field experiment involving nearly 400 students, 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

*mfamulari@ucsd.edu and zgoodman@ucsd.edu. The authors thank the students who took intermediate microeconomics in the fall of 2018 and 2019 who consented to the use of their data for this study. We also thank UC San Diego's Teaching and Learning Commons for providing campus data on the students in this study as well as anonymizing the data for analysis. Finally, we thank the applied microeconomics group at UC San Diego for their help with the experimental design. 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 lectures were provided to the subjects at no charge and neither author has a direct financial interest in the distribution of the IMVH at UC San Diego. As of fall 2020, one of the authors has a financial interest in the distribution of the IMVH outside of UC San Diego.

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 0.15 standard deviations. We rule out large negative spillover effects to other courses taken concurrently, and we observe persistent take-up of the IMVH in the subsequent quarter when direct 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 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, and Siegfried, 2015).

In this paper, we measure the impacts of one such technology, the Intermediate Microeconomic Video Handbook (IMVH), on outcomes in an intermediate microeconomics course. The IMVH was designed to *supplement* lecture as an audiovisual version of a conventional course textbook. Part of the impetus for creating the IMVH was a discussion with a student who described her inability to read the course text, not because of poor reading skills, but because she did not find the text engaging enough to command her attention. We hypothesized that modern students, who have had unprecedented exposure to electronic media, may find videos more engaging and, perhaps, more effective at building human capital.

Besides higher engagement than with conventional texts, the IMVH and video-based learning tools more broadly are of value to university educators and their students for four additional reasons. First, videos carry near-zero marginal cost and are accessible anywhere via internet, helping reduce financial and geographic barriers to high-quality educational materials. Second, video platforms can help students track what content they have already studied and what content they have yet to cover. Third, features embedded into videos,

such as searchable captions and timestamps, can help students locate the information they are seeking faster. Finally, the perceived low cost of watching a brief video may be easier to overcome than perceived higher costs of other studying methods, potentially leading to more frequent studying, which decades of psychological research has demonstrated leads to more long-term learning than does cramming (Kornell, 2009; Cepeda et al., 2006).

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. To estimate the effects of 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, thereby making manifest a need to adjust studying habits, and perhaps standing to gain the most from an intervention that targets studying. We randomly assigned a grade-based incentive to half of the sample to encourage take-up of the IMVH, which was made available to all students in the experiment, 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 IMVH. We observe grades, GPA, and video watching in both the term of the experiment and the subsequent term.

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: 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.

We interpret our results through a theoretical framework in which students, who value grades and leisure, have potentially incorrect priors about the returns of different studying methods. We do so not to test theory, but rather to help educators understand the potential welfare implications of providing students with a learning technology and paternalistic

incentive structure like the one studied. Although we observe that treatment students performed better on course assessments, for welfare analysis one must consider where the time watching videos came from: leisure, work, student organizations, studying for other classes, studying for present class using other methods, etc. If students must reduce time allocated towards leisure or studying for other classes so they can watch more videos, then the welfare impacts of requiring videos could be negative depending on the students' preferences. On the other hand, if the videos are more productive than the next best studying technology, then requiring videos could be utility enhancing.

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 treated 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 dramatically substitute away from studying for other courses.

An important piece to the welfare puzzle is whether treatment students continue to use the IMVH 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 requiring the IMVH is a net positive on underperforming students' academic achievement, both in the quarter of the ex-

periment and beyond. Though formal welfare analysis is beyond the scope of this paper, we present suggestive evidence that requiring the IMVH is unlikely to be substantially utility harming, if not utility enhancing. We believe 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 presents competing models of studying behavior that may explain the observed phenomena. Section 3 provides background on existing related literature. Section 4 describes the study design. Section 5 presents the results of the experiment, and Section 6 discusses those results. Section 7 concludes.

2 Models of Studying Behavior

In this section we consider three models of student studying behavior: a neoclassical model, an imperfect information model, and a behavioral/procrastination model. For all three models, we consider the effects of an instructor's inducement to encourage student use an effective study method. 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 only briefly explore the added issues of inducing students to use a study tool whose usefulness is not known to the instructor.

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. In this model there is no room for an instructor to increase student well-being by intervening in their study decisions. Oettinger (2002) provides some empirical support for the neoclassical model by demonstrating 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 the letter grade cutoffs and student performance on the final exam is higher if the student is just below a grade

threshold.

In addition to teaching specific skills, many would agree that the "raison d'etre" of higher education is to teach students how to learn. There is evidence from psychology that college students do not know the return to various study methods.[1] Universities often fund "Teaching and Learning Centers" or "Academic Skills Centers," part of whose mission is to help undergraduates learn to study more productively.[2] We posit that for many students, a key assumption of the neoclassical model does not hold: that students possess complete information about the returns across studying methods. Instead, we offer the alternative hypothesis 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, [3] Clark et al. (2020) provide empirical evidence that setting tasked-based goals helps improve college student performance.

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, answering homework problems, 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, stu-

[1] See, for example, mccabe2011, prcc2007, drmnw2013

[2] All nine University of California campuses have one. Some others in the US 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

[3] see, for example, Ferrari (1992), P. Chen et al. (2017) and Lavecchia, Liu, and Oreopoulos (2016).

dents gain utility 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 or some outside option that is not directly incentivized (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.

In the neoclassical model, the marginal return to grades of the incentivized input is less than that of the outside option for the 'compliers', or those induced by the incentive to consume at least a fixed level of the incentivized input. 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 strict 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. Compliers update their priors about the returns to watching videos as they work towards hitting the minimum required level. At this cutoff, they make a decision whether to continue watching videos depending on their updated perceptions of the marginal benefit. Hence, bunching at the cutoff is predicted only if the updated marginal benefit at the cutoff is lower than the marginal benefit of the next best studying option or

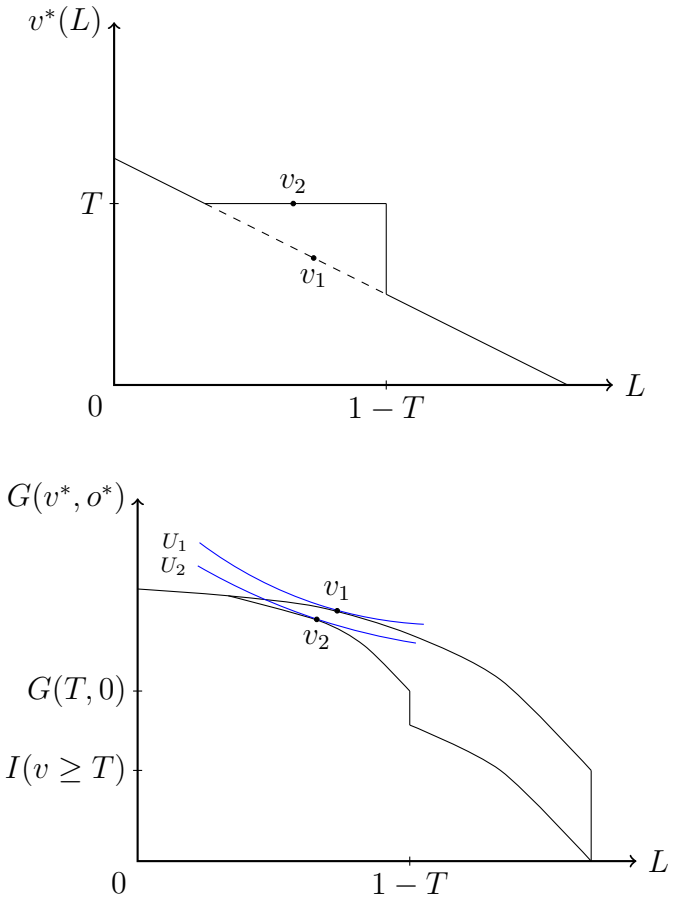


Figure 1: *Above*: Demand curve for video watching as a function of leisure L. At $L = 1-T$, the student maximizes grades $G(v, 0)$ by spending all studying time watching videos, i.e. $v^* = T$. *Below*: Student's utility maximization problem for the neoclassical model. The student maximizes her utility over leisure L and grades G, which is a function of time allocated to video watching v and her next best studying option o. The grade incentive I is given to the student conditional on watching T hours of videos (inner-time budget constraint) or, in the unincentivized case, given regardless of video watching (outer time-budget constraint).

the marginal utility of leisure.

Finally, in the behavioral model, the instructor's inducement helps students stick to study plans. As long as total study time does not fall, the inducement will increase exam performance. In the absence of the inducement, a sharp prediction is that video watching will revert to pre-inducement levels.

In the empirical section, we test for the effects of being induced to watch the IMVH on both exam scores and several other study methods students could use to learn microeconomics (lecture attendance, visits to a class-specific tutoring lab, use of a class discussion board, downloads from the class web page). We also compare grades in other classes taken in the same quarter across the treatment and control group. For a subset of our sample we have survey responses on total study time in the quarter and leisure time. Since the experiment was conducted in the first of a required three-class sequence, we examine video watching in the second class. We test whether the effect of the inducement to watch videos is greater for transfer students (assumed to have more information problems) and non-native English speakers (assumed to benefit the most from closed captioning, a key feature of the IMVH).

3 Related Literature and Contributions

Students have many time-consuming activities to help them learn including attending class, watching recorded lectures, reading the textbook, doing homework, attending office hours, etc. There are several empirical challenges to estimating the causal effects of a learning activity. First, unobserved student characteristics, such as ability and motivation, are likely positively correlated both with the use of a learning activity and class performance. To estimate causal effects, empirical studies must address nonrandom selection into using a study method. Second, most instructors have experience working with motivated students who want to improve their study strategies after a negative exam shock and Oettinger (2002) and Ralph and R. (2008) provide empirical evidence of dynamic selection.[4] Dynamic selection means that including student fixed effects in class performance regressions, a common

[4]Oettinger (2002) finds that students close to a grade threshold before the final exam perform better on the final

empirical approach in this literature, will not uncover causal effects. Third, learning activities are substitutable and inducements to use one study strategy may affect student use of another. In these cases, even randomized experiments will not identify the causal effect of a study method but will rather identify the causal effects of a study policy and all of the changes in behavior caused by the policy. The causal effects of an educational policy, such as requiring homework, are likely useful for educators considering how to design their classes but are less useful for students wanting to know the relative effectiveness of study strategies. Fourth, the existence of the resource may change how the instructor teaches. For example, several papers discuss how lecture capture changes the way instructors lecture. Again, experiments randomly assigning student to classes taught one way versus another will not identify the causal effect of a study method if other aspects of learning transmission are changed. A fifth empirical issue is that study strategies all take time and so most empirical studies jointly test the effectiveness of a particular learning method and devoting more time to the course. It is possible that the primary benefit to students is simply devoting more time to learning the course material, regardless of the study method.

Our study uses a randomized control trial as well as a RD approach which solves issues (1) and (2). Since we use within-class randomization, we solve issue (4). Our study does not solve issues (3) and (5), however, we have empirical evidence on a large number of alternative study methods and test if the inducement to use the video book affected these other study methods to shed light on issue (3) and we have survey data on study time for the course for a selected sample of students which we use to test whether total study time for the course differs across treated and control students to shed light on issue (5).

Attending class: J. Chen and Lin (2008) estimate the average effect of attending lecture for those who attend lecture (average treatment effect on the treated). The same instructor taught two required public finance classes of size 67 students and 47 students. The instructor provided identical PowerPoints to both classes after lecture, but did not lecture on randomly chosen topics to a randomly chosen class. The outcome is whether the student answers a multiple choice exam question correctly. Physically attending lecture involves transportation costs and this costly aspect of lecture is not captured by the authors since the exam score comparison is across students who were all in class. However, this study does identify the

effect of attending a remote lecture versus having a written explanation.

Dobkin, Gil, and Marion (2010) analyze a policy where lecture attendance was voluntary before the midterm, but after the midterm, students scoring below the median were required to attend class. The policy affected 352 students taking three classes, two intermediate micro and one econometrics class. The policy led to a 36 percentage point increase in post-midterm attendance at the threshold. Using a regression discontinuity design, they find that a 10 percentage point increase in overall attendance results in a 0.17 standard deviation increase in the final exam score. They find no effect of the attendance policy on grades in other classes taken the same quarter, attending TA sections, homework scores, and the use of university tutors. Arulampalam, Naylor, and Smith (2012) study section (as opposed to lecture) attendance across intermediate microeconomics, intermediate macroeconomics and econometrics for 444 students. The authors find that absenteeism depends on day of week and time of day and, since students are randomly assigned to sections, use these variables as instruments for absenteeism. They also include student fixed-effects. Surprisingly, they find significant attendance effects only for students in the top quantiles: missing 10 percent of sections results in a 1 percentage point performance loss. The authors have no information about other uses of the student's time use, including attending the main lecture.

Effect of Homework: Trost and Salehi-Isfahani (2012) randomly require two-thirds of students taking Principles of Economics classes to complete a one of three homework assignment for a grade. The other third may complete the homework, but it does not contribute to their grade. The outcome measure is exam performance on questions related to the three homework assignments. The authors use the score on the remaining exam questions as a control variable. They find significant effects pf homework on the first midterm but not the final exam. Grodner and Rupp (2013) use within-class randomization to estimate the effects of required homework for 423 microeconomics principles students. A coin flip determined whether a student was in the treated group, where course points are based on both homework and exams, or in the control group, where all course points are based on exams. Treatment led to a 58 percentage point increase in completing all homework assignments and a 84 percentage point increase in completing the majority of homework assignments. They find that treated students are less likely to drop the class and score higher on the first two but not

the last two exams. The average across the four exams is increased 5-6 percent by treatment and the control group GPA would increase from 2.44 to 2.68 if they had been required to do homework. They find three times larger treatment effects for students who initially fail the first exam (10 to 15 percent vs 4 to 6 percent increase in average test scores). The authors do not examine whether other uses of student time are affected by the homework policy.

Effect of Study time: In a convincing empirical paper on the effects of study time, Ralph and R. (2008) examine 210 Berea College students who were randomly assigned a roommate. Students whose roommates brought a video game to college, earn lower grades and spend less time studying. They authors instrument for study time using presence of a roommate with a video game and find that a one hour increase in study time per day (a .67 standard deviation increase in their sample) has the same effect of first semester GPA as a 5.21 increase in the ACT (an increase of 1.4 standard deviations in their sample).

Other researchers have investigated using technology to improve learning. N. Angrist et al. (2020) conduct an experiment in Botswana during the COVID-19 pandemic and find that text messages and phone calls deployed as low cost, scalable learning technologies improved test scores by 0.16 to 0.29 standard deviations.

Effect of Recorded Lectures: An educational resource closely related to the IMVH is when instructors record their lectures and make them available to students. Recorded lectures have course administration information which the IMVH does not have. Compared to an IMVH video, recorded lectures are typically much longer, less organized, and may include components that do not work well when recorded, such as group work or class discussion. Savage (2009) taught two intermediate micro classes: one 42-student class had "talk and chalk" lectures and the other 45-student class used technology that allowed lecture capture which was then made available to the students. The author finds no significant differences in observed student characteristics across the two sections but exam performance was significantly higher across the classes.

Effect of Setting Goals: Clark et al. (2020), hereafter CGPR, explore the effect of having students set goals on class performance. They find that setting *task-based* goals of completing a specific number of online practice exams improves student performance on exams. Those randomly assigned to set task-based goals completed 0.10 standard deviations more practice

exams and increased total course points by 0.07 standard deviations. The authors found no significant effect of setting *performance-based* goals of achieving specific grades in the course or on exams on total course points. In the present study, we set for the students a *task-based* goal of watching a specific number of videos. Our intervention differs from that of CGPR in two important ways. First, the instructor, not the students, set the value of the goal. Second, we motivate students to watch videos through a quantifiable grade incentive instead of appealing to psychological forces such as internal commitment devices.

This study adds to this body of research by studying the effectiveness of an educational innovation: a video textbook. We randomly assign half the students scoring below the median on the first midterm to a grading scheme which placed 4 percent, or 40 points, of the student's grade on watching 40 videos and down-weighted the first midterm by 4 percent. This experiment allows two empirical strategies to test for causal effects: within class randomization for students scoring below the median on the first exam and a regression discontinuity approach at the median first exam score. We examine a large set of student study behaviors (lecture attendance, homework downloads from course web page, contributions to a discussion board, use of a class-specific tutoring lab) to determine if any of these study methods are substitutes or complements with video watching. We test for spillovers to other classes taken in the same quarter. We test for heterogeneous treatment effects using techniques that are robust to p-hacking. Finally, our research setting allows us to examine video views in the absence of the grade incentive in the next, required intermediate microeconomics class.

4 Study Design

4.1 Description of the sample and institution

We conducted the field experiment in an undergraduates 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,

[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 shared

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*, which was taught by the same instructor during both the winter 2018 and winter 2019 quarters. Both Micro A and B instructors created half of the videos relevant to their course in the IMVH.

The structure is similar across the three courses in the Micro sequence. Students have the option to attend one of two lectures offered back to back twice per week, each lasting about 90 minutes. Two midterm and final exams are held at a common time outside of lecture. In addition to lecture, students have access to 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 TAs and Undergraduate Instructional Assistants (UIAs) 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 sets, and semi-weekly graded online quizzes.

Students were told about the experiment during the first lecture and provided an informed consent form in the syllabus. 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

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, and 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.

[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.

students.

There are two unique demographic features of the class worth noting. First, many non-econ 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, for whom the class is not only their first experience with upper division coursework at a four-year research university, but also typically 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.

4.2 Description of the IMVH

The Intermediate Microeconomics Video Handbook (IMVH) is a collection of 220 short videos that cover the material in a year-long intermediate microeconomics course sequence.[9] The videos, designed to complement or replace a course textbook, include graphical and verbal intuition as well as formal algebraic and calculus-based definitions and proofs.

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. Videos are closed captioned and were checked by graduate students for accuracy.

Given the complexity of the material, a key objective was to keep the web interface clean and simple so as not to distract from the content. The videos are organized by content area (e.g., consumer theory, producer theory, etc.) that help students understand where various topics "live" in intermediate microeconomics. When considering the design of the platform, one goal was to help students find material quickly. Besides a table of contents and an

[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	✓		✓	✓

index, each video contains time stamps of the concepts therein. Another helpful feature is searchable captions, which allow the student to jump to the part of a video containing the searched-for word.

While we do not know of another textbook completely comprised of videos, the IMVH is similar to Khan Academy, lecture podcasts, and textbook websites that incorporate instructional videos. Besides the engaging viewable nature, the IMVH differs from a traditional textbook in that the instructors explain, graph, and derive mathematical results in much the same way one would in a conventional lecture. However, the IMVH differs from lectures in that students control the pace: they can rewatch, speed up, or slow down the videos. Another difference is the ability to read captions and the reduced demand on the instructor's time. We summarize some options to present course material to students in Table 1.[10]

4.3 Experiment Design

The experiment began four weeks into the term following the first midterm exam. All students who scored above the median on the first midterm, the *Above median* arm, and half of students who scored below the median, the *Control* arm, were assigned a conventional grading scheme that places weight only on exams and quizzes. We assigned the remaining half of students below the median to the *Incentive* arm, whose grading scheme allots four

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

Table 2: Grade scheme by treatment arm. *Control* represents same grade scheme as *Above median*. 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%

percentage points conditional on watching at least 40 of 48 eligible videos in the IMVH.[11] These 48 videos review new class content since the first midterm that could be assessed in the second midterm and final exam. Students could still view the remaining 26 course-relevant videos, and as they could have helped 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 points 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 and 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]

[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.

[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

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
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[13]

For expositional ease, we use "watching" to refer to the link-opening behavior as defined above. It is also worth mentioning that although we code video watching as binary, 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 quantity video engagement such as 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,[14] in a few exceptional cases, students opened 40 or more video links within a matter of 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.[15] Though it is doubtful these strategic students gained much from opening so many videos so quickly, to maintain

the first midterm.

[13]We cannot observe the amount of time students spent on each video, so proceed recognizing our time measures likely overestimate true viewing time for most students.

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

[15]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.

interpretability of our results, we do *not* remove these clicks from our video count measures.[16] 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 grades. First, we applied a curve to the *Control* and *Above median* arms as one group to achieve a grade distribution in line with that of previous cohorts. Second, we curved the *Incentive* students' course grades to match the average course grade among *Control* students after curving in the pervious step. Since course grades are by construction *ex post* equal between the *Control* and *Incentive* arms, we use exam scores as our primary outcomes of interest. Our secondary outcomes of interest include term GPA and number of courses passed, which help us understand how treatment may have affected other courses. We examine separately econ and non-econ courses in case effects differ by course content. To better understand mechanisms, we estimate effects of treatment on take-up of other studying tools within Micro A. Finally, we explore video watching and grade outcomes in Micro B, the subsequent course in the intermediate microeconomics sequence, to see if treatment effects persist beyond one term.

4.4 Empirical strategies

In this paper, 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), also referred to as a Complier Average Treatment Effect (CATE) (Imbens and Rubin, 2015).

Below we outline the empirical strategies for estimating both the ITTs and LATEs.

4.4.1 Intention To Treat (ITT)

In this section, we examine the empirical strategy for estimating the causal effect of being assigned to the *Incentive* arm on outcomes of interest, such as exam scores. These are ITT

[16] 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.

estimates and not average treatment effect estimates because of two-sided non-compliance: some students in the treatment arm do not watch videos and some students in the control arm do watch videos. 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 adopting the IMVH or other video-based 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 of the ITT effect (Imbens and Rubin, 2015).[18]

In our baseline estimation of Equation 1, we include in X_i year indicators 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.2. In a third model, to check that our results are robust to potentially nonrandom attrition by treatment arm, we fit Equation 1 including pair fixed effects. These fixed effects subsume the year indicator (since pairs were assigned separately across years), so we drop the year indicator but keep midterm 1 score to control for small differences within pairs along that dimension. As entity fixed effects require at least two

[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 in both years, 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. 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 A1 and A2.

observations within the entity to be estimable, we drop any students whose matched pair attrited.

In our results, we present an additional nonparametric estimate using Neyman's (1923) repeated sampling approach, considering each pair (block) an independent, completely randomized experiment and averaging the results. We estimate the point estimate of the ITT as the mean difference in outcomes across pairs:

$$\widehat{\tau} = \frac{1}{J}\sum_{j=1}^{J}\widehat{\tau_j} = \frac{1}{J}\sum_{j=1}^{J} y_{j,I}^{\text{obs}} - y_{j,C}^{\text{obs}} \tag{2}$$

where $\widehat{\tau}$ is the point estimate of the ITT, J is the number of pairs in the sample, and $\widehat{\tau_j} = y_{j,I}^{obs} - y_{j,C}^{obs}$ is the observed difference in outcome for pair j. The estimated standard error of $\widehat{\tau}$ (Imai, 2008; Imbens and Rubin, 2015; Athey and Imbens, 2017) is:

$$\widehat{SE}(\widehat{\tau}) = \Big(\frac{1}{J}\sum_{j=1}^{J}\widehat{V}(\widehat{\tau_j})\Big)^{\frac{1}{2}} \tag{3}$$

where $\widehat{V}(\widehat{\tau_j})$ is the estimated variance within block (pair) $j \in \{1, ..., J\}$. This within-block variance given one control and one treated unit per block is (Imbens and Rubin, 2015, Athey and Imbens, 2017):

$$\widehat{V}(\widehat{\tau_j}) = s_{j,I}^2 + s_{j,C}^2 \tag{4}$$

where $s_{j,I}$ and $s_{j,C}$ are the *Incentive* and *Control* sample variances within block j, respectively. Unfortunately, these sample variances are not estimable in a matched-pair setting as there is only one unit in each arm per block. As such, we use the following estimator, which is conservative (confidence intervals wider) if there is heterogeneity in the treatment effect (Imai, 2008, Imbens and Rubin, 2015, Athey and Imbens, 2017):

$$\widehat{SE}(\widehat{\tau_j}) = \Big(\frac{1}{J(J-1)}\sum_{j=1}^{J}(\widehat{\tau_j} - \widehat{\tau})^2\Big)^{\frac{1}{2}} \tag{5}$$

Similar to the fixed effect model, the Neyman repeated sampling approach is only estimable if we drop all students whose matched pair attrited. This 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. We present estimates from all four models to demonstrate that the results are generally similar and not sensitive to model specification or choice of control variables.

4.4.2 Local Average Treatment Effect (LATE)

Here we present the empirical strategies for estimating the causal effect of watching videos on outcomes of interest, exam scores. The average causal effect of watching videos can be modeled as:

$$Y_i = \gamma v_i + g(X_i) + u_i \tag{6}$$

where γ is the average causal effect of watching an additional video, Y_i is an outcome of interest (e.g., exam scores) for a student i, $g()$ is a generic function through which X_i, a vector of pretreatment covariates, affects Y_i, and u_i in an unobserved model residual. Because a student's decision to watch videos is likely correlated with unobservable factors (for example, motivation) that are also correlated with outcomes, regressing Y_i on endogenous videos v_i will provide biased estimates of γ. To solve this problem, we rely on variation in v_i induced by an exogenous instrument Z_i. In our setting, Z_i is assignment to the *Incentive* grade scheme. If Z_i is a valid instrument for v_i, then we can estimate γ using two-stage least squares (2SLS):

$$v_i = \alpha Z_i + f(X_i) + e_i \tag{7}$$

$$Y_i = \gamma \widehat{v}_i + g(X_i) + u_i \tag{8}$$

where $f()$ 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 7. 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 and 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 (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 treated students in the same class and, perhaps, studying together.

Although we believe unconfoundedness,[19] excludability,[20] and monotonicity[21] 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

[19] Although we randomly assigned treatment, one concern is nonrandom attrition. We find that the *Incentive* and *Control* arms remain balanced across observables by the end of the experiment. Additionally, we find our results are similar when restricting the sample to students whose matched pair did not attrite.

[20] 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 they would have been doing. This subtle point could matter for external validity as a different population of students with zero leisure time may respond differently to the incentive.

[21] 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 3 shows that the two CDFs do not cross.

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.[22]

Similar to our estimates of Equation 1, we estimate Equation 8 with three sets of controls: only year and first midterm score, controls chosen using PDS, and a fixed-effect model with controls chosen using PDS. We additionally estimate the LATE using Neyman's repeated sampling approach whose estimators we derive in Appendix B.

4.4.3 Treatment Effects at the Cutoff

Here we describe estimation of treatment effects at the first midterm score cutoff. Because the probability of being assigned to the *Incentive* arm changes discontinuously from 0.5 to 0 at the midterm score cutoff, our setting is appropriate for estimating local treatment effects using a regression discontinuity (RD) design (Thistlethwaite and Campbell, 1960; J. D. Angrist and Pischke, 2008; Imbens and Lemieux, 2008). With this method, we compare students in the *Incentive* arm who scored just below the cutoff to those in the *Above median* and *Control* arms who scored just above or below the cutoff, respectively. These two groups are similar across pretreatment characteristics but different in treatment status, thereby providing an estimate of the treatment effect local to those who scored near the cutoff.

Since RD designs require that agents near the cutoff be similar across covariates except treatment status, a threat to validity is manipulation of the forcing variable (in our study, midterm score), which biases treatment effect estimates by nonrandom selection into treatment. This manipulation can occur if agents behave strategically to target a particular side of the cutoff, for example, scoring slightly higher than a published minimum SAT score for college admission. Since students in our experiment do not know the cutoff *ex ante*, it is un-

[22] 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 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.

likely that students would attempt to target a particular side of the midterm score cutoff[23]. Ultimately, we must *assume* continuity of the conditional means of the potential outcomes along the midterm score; however, we do not observe a discontinuity in any observable pre-treatment covariate at the cutoff, which gives us further confidence that this assumption holds.

To estimate local ITT effects using a sharp RD, we return to the potential outcomes framework modeling the treatment effect $\tau(c)$ as the difference in expected outcomes at the cutoff c along the forcing variable x (midterm score):

$$\begin{aligned} \tau(c) &= \lim_{x \uparrow c} \mathrm{E}[Y_i | X_i = x] - \lim_{x \downarrow c} \mathrm{E}[Y_i | X_i = x] \\ &= \mathrm{E}[Y_i(1) | X_i = c] - \mathrm{E}[Y_i(0) | X_i = c] \end{aligned} \tag{9}$$

We estimate $\tau(c)$ using local low-order polynomials, per the advice of Gelman and Imbens (2019).

Sharp RD designs used in the literature frequently do not observe $Y(1)$ and $Y(0)$ for the same values of x. In our setting, however, we observe $Y(0)$ both above and below the cutoff. Hence, we need to assume continuity only for $Y(1)$ as we do not observe any outcomes for treated students scoring above the cutoff but *do* observe outcomes for control students both above and below the cutoff.

5 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

[23]It would be surprising for students who value high grades to target the expected median score since any student capable of doing so would likely earn a higher grade in the course by scoring as high as possible on the midterm exam rather than strategically scoring just below the expected median cutoff.

better understand mechanisms, we examine spillovers to other studying methods and grades in other courses taken during the experiment term. Finally, we 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.

5.1 Attrition and balance

At the university where the experiment took place, Micro A is the first upper-division economics course, often taken by students who have not yet declared a major. As such, it has higher withdraw 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, 8.5% and 13.4% of students who took the first midterm did not take the second midterm and final exam, respectively. Unsurprisingly, the withdrawal rates were greater for students who scored below the median on the first midterm: 14.7% and 24.2% of these students did not take the second midterm and final exam, respectively.[24] In the present study, high attrition is not problematic, other than reducing statistical power, if attrition is independent of potential outcomes. If attrition is influenced by treatment status, which could occur, for example, if treatment students believed they were more likely to pass the course,[25] then the resulting nonrandom selection into our analysis could bias the results.

We assess the presence of nonrandom attrition by comparing rates of attrition by treatment arm and examining balance across observable demographic variables in the analysis sample. Students could have attrited in three ways. First, students under the age of 18 at the start of the experiment were removed from the analysis sample. Because of student privacy considerations, demographic variables (including age) were not observable until the conclusion of the experiment. Second, students who opted out of having their data included in the experiment analysis, an option available to students at any time during the experiment, were removed. The analysis sample was prepared and anonymized by a campus-based

[24]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.

[25]As noted earlier, we informed students that final grades would be curved separately between *Control* and *Incentive* arms to remove any advantage treatment may carry towards passing the course.

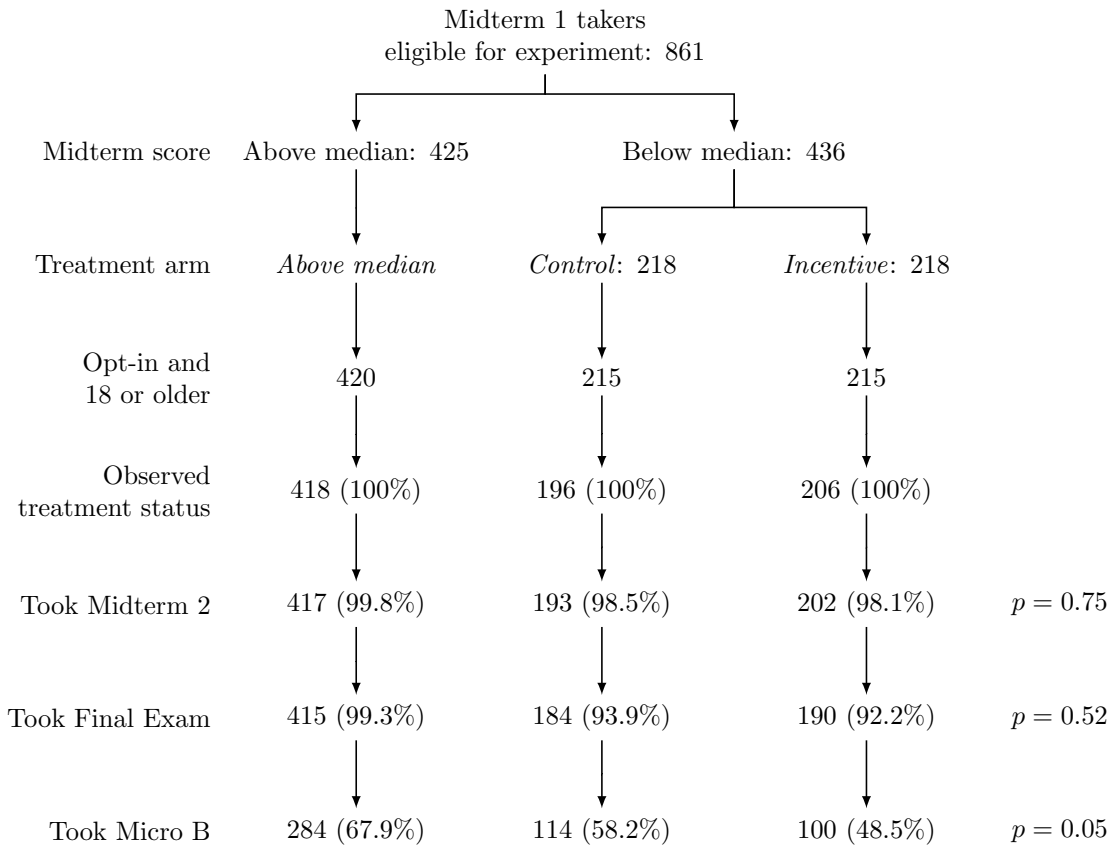
independent education research organization, per IRB requirement, which removed four minors and seven opt-outs from the sample and merged demographic variables before returning the anonymized data to the research team.[26]

The final and largest cause of attrition is withdrawing from the course. At the present university, students may formally drop a course without penalty up to the fourth week of the quarter. Between the fourth and sixth weeks, students may withdraw from a course, but a "W" grade is assigned in lieu of a letter grade, which does not affect GPA. From the seventh week onwards, students may no longer formally withdraw, but some may choose not to take the second midterm or final exams, which almost assuredly results in a failing grade in the course and does factor into the student's GPA. Because the first midterm took place in week four, the same week as the penalty-free drop deadline, many students took the exam and withdrew before finding out their grades, which were posted in week five. However, because of the lag between when a student drops a course and when the instructor is notified, we assigned treatment to several students not knowing they had already dropped the course. In total, 30 students - two *Above median*, 19 *Control* and nine *Incentive* - dropped the course before finding out their first midterm scores and treatment assignments. Among students who waited to find out their first midterm scores and treatment assignments, three *Control* and three *Incentive* students did not take the second midterm, and an additional nine *Control* and 13 *Incentive* students did not take the final exam.

We show attrition at each stage of the experiment in Figure 2. The p-values in the figure are calculated using two-sample t-tests of the equality of attrition rates between the *Control* and *Incentive* arms at each stage. We do not find any statistically significant difference in attrition before the second midterm and final exams that can be attributed to treatment status. However, we do find a significant difference in enrollment rates in Micro B: treatment students were 9.7 percentage points *less* likely to enroll in Micro B than were control students. Conditioning on midterm 1 score and year reduces the gap slightly to 8.9 percentage points ($p = 0.06$). As mentioned in Section 4.4, as a robustness check against potential bias from any nonrandom attrition, we estimate treatment effects using models that only include

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

Figure 2: 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.

matched-pairs.

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 and Imbens, 2017). We check balance on observable characteristics after attrition for both the second midterm and final exam samples. As can be seen in Appendix Tables A1 and A2, 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 4.4, 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).

5.2 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 4.4. We must check that our instrument is both valid and relevant to ensure this method will produce an unbiased estimate of the LATE (Imbens and 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 Appendix Tables A1 and A2, 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 3 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 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, Yogo, et al. (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 3. 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. 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.

5.3 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, 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 8) and nonparametric methods a la the repeated sampling framework of Neyman (1923). We check that our parametric results are robust to model specification by estimating Equations 1 and 8 with and without $f(X_i)$ as a vector of linear control variables chosen via PDS (Belloni, Chernozhukov, and Hansen, 2014b). To rule out nonrandom attrition across treatment arms as a confounder, we fit a fixed effect model that drops any student whose matched pair attrited. These specifications and identification strategies are described in detail in Section 4.4.

Table 4 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 3), 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.

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 4.

5.4 Spillovers during concurrent term

We next estimate spillover effects to other courses taken concurrently during the term of the experiment. Although we find positive effects on outcomes in Micro A, it is important to examine spillover effects to understand the complete picture of how treatment affects student achievement. If the gains in Micro A come at the cost of slowed progress in other courses, then treatment may be counterproductive for some students. On the other hand, since the videos cover concepts and methods that could be covered in other courses, treatment may help students outside Micro A.

Table 5 presents our estimates of Equation 1 where Y_i is GPA, number of classes passed, or portion of classes taken for a letter grade. 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 only economics or non-economics courses.

We additionally estimate the effects of treatment on 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

[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.

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.

Besides estimating spillover effects on other course grades, 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 6, 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 not statistically significantly so. 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.

5.5 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 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 during the subsequent term in Table 7. 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 5.1, treatment students were directionally less likely to take Micro B than were control students.

6 Discussion

Here we discuss the findings reported in Section 5.

Students overwhelmingly report that they find the IMVH useful. But considering psych learning theories, why might videos be effective? The IMVH does not have students use retrieval practice (and example is when students test themselves). However, it is possible the IMVH facilitates interleaving (mixing topics up while studying), spaced practice (attend lecture and watch videos either before or after), learning information from different formats (IMVH presents information both verbally, graphically and algebraically), and using worked examples (IMVH includes worked examples. Again, it is also possible that the inducement to watch the IMVH simply led students to spend more times on the class. Why do the videos not appear effective for students at the median? Not sure unless it is simply consistent with a learning model where students above median already know how to study and the inducement to use the IMVH

6.1 Limitations

The present study has several limitations that should be considered before, for example, creating one's own video handbook and requiring students to use it. 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 univerisity. The extent to which treatment effects vary by course, instructor, university, or along the midterm score distribution is beyond the scope of this paper. 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. We cannot recover the *population* average treatment effect,

though annecdotal evidence and economic theory both suggest that the population average treatment effect is likely greater than the LATE.

Some researchers may wonder why we included only the bottom half of midterm scorers in the experiment instead of the entire class. Though we cannot estimate heterogeneity in treatment effects along the entire midterm distribution with our design, we believe this loss is justified by reduced risk of welfare losses by high performing students. The first midterm provides a signal of which students likely know for themselves how much and what kind of studying they should be doing. Coercing these high-type students to spend time with an alternative studying method is unlikely to be helpful and runs a higher risk of harming utility. On the other hand, students who have made manifest a need for alternative or more time studying stand to benefit the most from instructor-provided guidance.

Another 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 coronavirus 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.

Generalizability of our results. The experiment was conducted in intermediate microeconomics, a required course in all economics programs which typically has high failure rates.[28] It is the first upper division class for many students and, for transfer students, also their first class in the university and under a XX percent quarter system. As such there may be large information problems about how to successfully study for the class.

Future research: most importantly, to see of our results hold up in other educational settings (e.g., different students, types of classes, instructors, and universities). We would like to examine the role of weekly deadlines instead of one final deadline at the end of the term, which may reduce the deleterious effects of binge-watching.

[28]At UC San Diego it is and at LSE it ranges from 12 percent in their least mathematical version to 24 percent in the more mathematical class, see https://www.lse.ac.uk/study-at-lse/The-General-Course/pdf/Choosing-economics-Course-guide-2020-21-HR-1.pdf

7 Conclusion

We examine the effectiveness of an educational innovation, 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 student to score significantly higher on exams. Specifically, we estimate that for students on the margin of watching videos, an additional hour of video watching causes students to score XX to XX 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 cause worse 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, though not statistically significantly different from zero, are positive for most alternative studying methods, suggesting that a potential mechanism of the videos may be helping students realize what they *don't* know, whereas students who selectively study spend too much time on material they already know.

A final concern is one of welfare. In a neoclassical model, instructors cannot make their students better off by forcing on them quantities of studying they would not otherwise have chosen for themselves. In a behavioral model, which we think is more appropriate in our university classroom setting, instructors *can* improve student welfare through intervention when information barriers and myopia lead to suboptimal time allocation decisions. We observe two phenomena that supports the latter model. First, treated students tend not to bunch at the cutoff for the grade incentive. Second, video consumption remains much higher among treated 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 be conducted on the effectiveness of pedagogical technologies.

References

Allgood, S., W. B. Walstad, and J. J. Siegfried (2015). "Research on teaching economics to undergraduates". In: *Journal of Economic Literature* 53.2, pp. 285–325.

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

Angrist, J. D. and G. W. Imbens (1995). "Two-stage least squares estimation of average causal effects in models with variable treatment intensity". In: *Journal of the American statistical Association* 90.430, pp. 431–442.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, N., P. Bergman, C. Brewster, and M. Matsheng (2020). "Stemming Learning Loss During the Pandemic: A Rapid Randomized Trial of a Low-Tech Intervention in Botswana". In: *Available at SSRN 3663098*.

Arulampalam, W., R. A. Naylor, and J. Smith (2012). "Am I missing something? The effects of absence from class on student performance". In: *Economics of Education Review* 31.4, pp. 363–375.

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

Belloni, A., V. Chernozhukov, and C. Hansen (2014a). "High-dimensional methods and inference on structural and treatment effects". In: *Journal of Economic Perspectives* 28.2, pp. 29–50.

Belloni, A., V. Chernozhukov, and C. Hansen (2014b). "Inference on treatment effects after selection among high-dimensional controls". In: *The Review of Economic Studies* 81.2, pp. 608–650.

Bruhn, M. and D. McKenzie (2009). "In pursuit of balance: Randomization in practice in development field experiments". In: *American economic journal: applied economics* 1.4, pp. 200–232.

Cepeda, N. J., H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer (2006). "Distributed practice in verbal recall tasks: A review and quantitative synthesis." In: *Psychological bulletin* 132.3, p. 354.

Chen, J. and T.-F. Lin (2008). "Class attendance and exam performance: A randomized experiment". In: *The Journal of Economic Education* 39.3, pp. 213–227.

Chen, P., O. Chavez, D. C. Ong, and B. Gunderson (2017). "Strategic resource use for learning: A self-administered intervention that guides self-reflection on effective resource use enhances academic performance". In: *Psychological Science* 28.6, pp. 774–785.

Clark, D., D. Gill, V. Prowse, and M. Rush (2020). "Using Goals to Motivate College Students: Theory and Evidence From Field Experiments". In: *The Review of Economics and Statistics* 102.4, pp. 648–663.

Dobkin, C., R. Gil, and J. Marion (2010). "Skipping class in college and exam performance: Evidence from a regression discontinuity classroom experiment". In: *Economics of Education Review* 29.4, pp. 566–575.

Ferrari, J. R. (1992). "Psychometric validation of two procrastination inventories for adults: Arousal and avoidance measures". In: *Journal of Psychopathology and Behavioral Assessment* 14.2, pp. 97–110.

Gelman, A. and G. W. Imbens (2019). "Why high-order polynomials should not be used in regression discontinuity designs". In: *Journal of Business & Economic Statistics* 37.3, pp. 447–456.

Grodner, A. and N. Rupp (2013). "The Role of Homework in Student Learning Outcomes: Evidence from a Field Experiment". In: *The Journal of Economic Education* 44.2, pp. 93–109.

Imai, K. (2008). "Variance identification and efficiency analysis in randomized experiments under the matched-pair design". In: *Statistics in medicine* 27.24, pp. 4857–4873.

Imbens, G. W. and T. Lemieux (2008). "Regression discontinuity designs: A guide to practice". In: *Journal of econometrics* 142.2, pp. 615–635.

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

Kay, R. H. (2012). "Review: Exploring the Use of Video Podcasts in Education: A Comprehensive Review of the Literature". In: 28.3.

Kleibergen, F. and R. Paap (2006). "Generalized reduced rank tests using the singular value decomposition". In: *Journal of econometrics* 133.1, pp. 97–126.

Kornell, N. (2009). "Optimising learning using flashcards: Spacing is more effective than cramming". In: *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 23.9, pp. 1297–1317.

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

Neyman, J. (1923). "On the application of probability theory to agricultural experiments: essay on principles, Section 9". In: *Statistical Science* 5.4, pp. 465–472.

Oettinger, G. S. (Aug. 2002). "The Effect Of Nonlinear Incentives On Performance: Evidence From "Econ 101"" in: *The Review of Economics and Statistics* 84.3, pp. 509–517.

OLEA, J. L. M. and C. PFLUEGER (2013). "A Robust Test for Weak Instruments". In: *Journal of Business & Economic Statistics* 31.3.

Ralph, S. and S. T. R. (June 2008). "The Causal Effect of Studying on Academic Performance". In: *The B.E. Journal of Economic Analysis & Policy* 8.1, pp. 1–55.

Savage, S. J. (2009). "The Effect of Information Technology on Economic Education". In: *The Journal of Economic Education* 40.4, pp. 337–353.

Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011). "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant". In: *Psychological science* 22.11, pp. 1359–1366.

Stock, J. H., M. Yogo, et al. (2005). "Testing for weak instruments in linear IV regression". In: *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* 80.4.2, p. 1.

Thistlethwaite, D. L. and D. T. Campbell (1960). "Regression-discontinuity analysis: An alternative to the ex post facto experiment." In: *Journal of Educational psychology* 51.6, p. 309.

Trost, S. and D. Salehi-Isfahani (2012). "The effect of homework on exam performance: Experimental results from principles of economics". In: *Southern Economic Journal* 79.1, pp. 224–242.

Table 3: 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 4: 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 8 (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 5: 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 6: 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

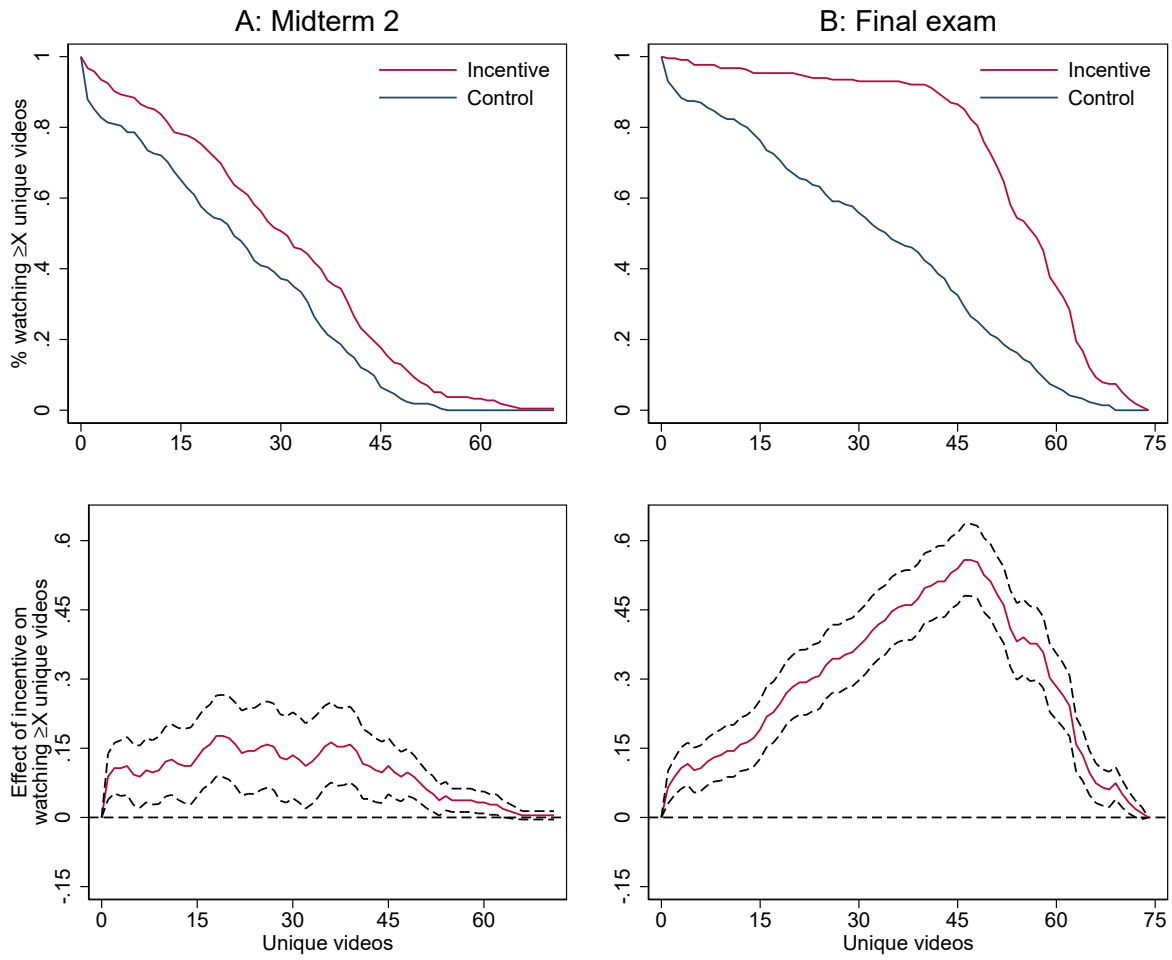
Note: This table reports coefficients on $Incentive_i$ from Equations 1. There were seven *Attendance checks* during the quarter. *Tutoring visits* includes those after the first midterm. Model (1) contains linear controls for midterm 1 score and year; (2) is the difference in means and standard errors 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 7: 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

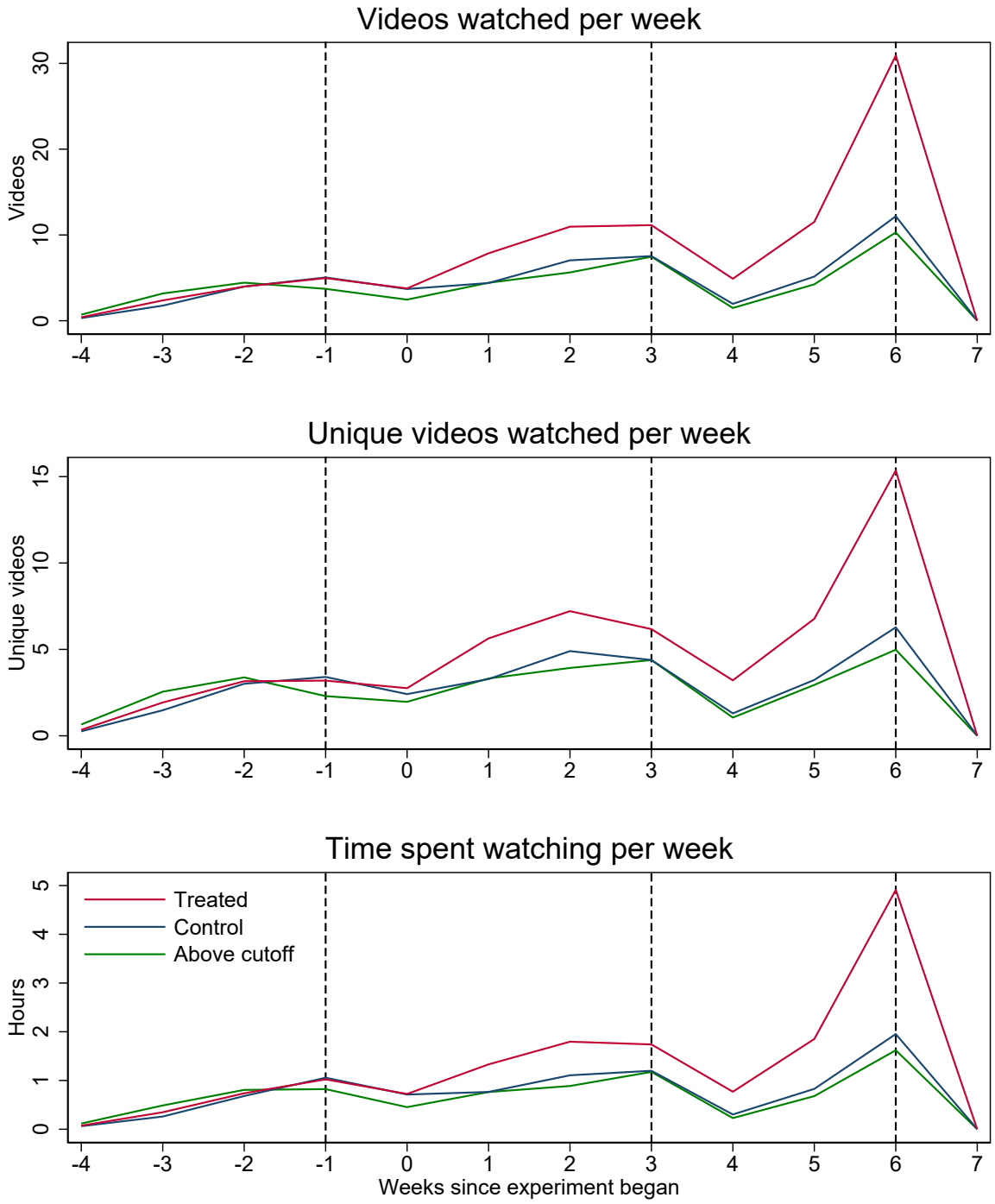
Note: This table reports coefficients on $Incentive_i$ from Equations 1. Panel A restricts the sample to those who completed both the first and second microeconomics courses (Micro A and B). Panel C includes those who completed the first microeconomics course (Micro A). Test scores are measured in standard 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.

Figure 3: 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 $X \in \{0, ..., X_{max}\}$ unique videos (or hours of unique videos) on the student's treatment status.

Figure 4: Weekly video watching by treatment arm



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

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. Imbens and Rubin, 2015, Athey and Imbens, 2017). The

Additionally, 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 assiging 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.

We assigned treatment shortly after issuing midterm exam grades, which occured 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 randomization 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% withdraw rate is in line with the withdraw rates 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 nearly all observable pretreatment covariates, as shown in Tables A2 and A1, which gives us confidence that the *Control* arm is a good counterfactual for the *Incentive* arm.

A.2 Selection of control variables

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

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 stochastic imbalances can occur, which if controlled for can reduce bias of the treatment effect estimator (Athey and 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, and 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 either pair fixed effects or year and 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:

$$\widehat{\gamma} = \frac{\overline{\Delta y}}{\overline{\Delta v}} = \frac{\frac{1}{J}\sum_{j=1}^{J}\Delta y_j}{\frac{1}{J}\sum_{j=1}^{J}\Delta v_j} = \frac{\overline{y_I} - \overline{y_C}}{\overline{v_I} - \overline{v_C}} \tag{10}$$

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:

$$\begin{aligned} Y &= \overline{y_I} - \overline{y_C} \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \\ V &= \overline{v_I} - \overline{v_C} \sim \mathcal{N}(\mu_V, \sigma_V^2) \end{aligned} \tag{11}$$

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

$$\begin{aligned} \mathrm{Var}(g) &= \mathrm{E}[(g - \mathrm{E}(g))^2] \\ &\approx \mathrm{E}[(g(\theta) + (Y - \theta_Y)g'_Y(\theta) + (V - \theta_V)g'_V(\theta) - g(\theta))^2] \\ &= \mathrm{E}[(Y - \theta_Y)^2(g'_Y(\theta))^2 + (V - \theta_V)^2(g'_V(\theta))^2 + 2(Y - \theta_Y)(V - \theta_V)g'_Y(\theta)g'_V(\theta)] \\ &= \mathrm{Var}(Y)(g'_Y(\theta))^2 + \mathrm{Var}(V)(g'_V(\theta))^2 + 2\mathrm{Cov}(Y, V)g'_Y(\theta)g'_V(\theta) \end{aligned} \tag{12}$$

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

$$\begin{aligned} \text{Var}(g) &\approx \frac{1}{\mu_V^2}\text{Var}(Y) + \frac{\mu_Y^2}{\mu_V^4}\text{Var}(V) + 2\frac{-\mu_Y}{\mu_V^{-2}}\text{Cov}(Y,V) \\ &= \frac{\mu_Y^2}{\mu_V^2}(\frac{\sigma_Y^2}{\mu_Y^2} + \frac{\sigma_V^2}{\mu_V^2} - 2\frac{\text{Cov}(Y,V)}{\mu_Y\mu_V}) \end{aligned} \tag{13}$$

We use the following variance estimators of Y and V from Equation 5:

$$\begin{aligned} \widehat{\text{Var}}(Y) = \widehat{\sigma_Y^2} &= \frac{1}{J(J-1)}\sum_{j=1}^{J}(\Delta y_j - \overline{\Delta}y)^2 \\ \widehat{\text{Var}}(V) = \widehat{\sigma_V^2} &= \frac{1}{J(J-1)}\sum_{j=1}^{J}(\Delta v_j - \overline{\Delta}v)^2 \\ \widehat{\text{Cov}}(Y,V) = \widehat{\sigma_{YV}} &= \frac{1}{J(J-1)}\sum_{j=1}^{J}(\Delta y_j - \overline{\Delta}y)(\Delta v_j - \overline{\Delta}v) \end{aligned} \tag{14}$$

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

$$\begin{aligned} \widehat{\mu_Y} &= \text{E}(\mu_Y) = \overline{\Delta}y \\ \widehat{\mu_V} &= \text{E}(\mu_V) = \overline{\Delta}v \end{aligned} \tag{15}$$

Substituting these variance and means estimators into the final step of 13, we arrive at the standard error estimator for $\widehat{\gamma}$:

$$\widehat{\sigma_\gamma} = \frac{\overline{\Delta}y}{\overline{\Delta}v}\sqrt{\frac{\widehat{\sigma_Y^2}}{\overline{\Delta}y^2} + \frac{\widehat{\sigma_V^2}}{\overline{\Delta}v^2} - 2\frac{\widehat{\sigma_{YV}}}{\overline{\Delta}y\overline{\Delta}v}} \tag{16}$$

Table A1: Baseline balance test, Midterm 2 sample

	All students			P-values	Matched pairs		P-values
Variable	Above Median	Control	Incentive	(3) - (2)	Control	Incentive	(5) - (4)
Midterm 1 score	2.048	0.116	0.037	0.398	0.139	0.131	0.933
	(0.025)	(0.063)	(0.068)		(0.065)	(0.066)	
Year = 2019	0.492	0.513	0.500	0.797	0.514	0.514	1.000
	(0.025)	(0.036)	(0.035)		(0.037)	(0.037)	
Cumulative GPA	3.445	2.944	2.948	0.965	2.942	2.992	0.487
	(0.029)	(0.043)	(0.058)		(0.045)	(0.056)	
No cum. GPA	0.230	0.368	0.332	0.452	0.365	0.320	0.377
	(0.021)	(0.035)	(0.033)		(0.036)	(0.035)	
Math quiz score	0.592	0.037	0.106	0.471	0.054	0.137	0.396
	(0.044)	(0.070)	(0.065)		(0.071)	(0.068)	
Tutoring visits	0.269	0.259	0.223	0.655	0.276	0.232	0.612
	(0.042)	(0.059)	(0.056)		(0.062)	(0.061)	
Videos watched	13.228	13.368	13.777	0.750	13.663	13.729	0.961
	(0.681)	(0.886)	(0.931)		(0.929)	(0.986)	
Videos, unique	9.746	9.689	10.188	0.554	9.845	10.116	0.760
	(0.431)	(0.580)	(0.611)		(0.606)	(0.644)	
Hours videos	1.690	1.782	1.825	0.818	1.827	1.804	0.906
	(0.093)	(0.127)	(0.135)		(0.133)	(0.142)	
Hours videos, unique	1.291	1.355	1.387	0.802	1.382	1.364	0.897
	(0.062)	(0.090)	(0.092)		(0.095)	(0.096)	
Asian	0.700	0.694	0.668	0.581	0.713	0.652	0.215
	(0.022)	(0.033)	(0.033)		(0.034)	(0.036)	
Latinx	0.060	0.135	0.158	0.506	0.133	0.166	0.377
	(0.012)	(0.025)	(0.026)		(0.025)	(0.028)	
White	0.151	0.114	0.124	0.765	0.105	0.138	0.336
	(0.018)	(0.023)	(0.023)		(0.023)	(0.026)	
Other ethnicity	0.089	0.057	0.050	0.741	0.050	0.044	0.804
	(0.014)	(0.017)	(0.015)		(0.016)	(0.015)	
Female	0.393	0.342	0.391	0.312	0.343	0.392	0.328
	(0.024)	(0.034)	(0.034)		(0.035)	(0.036)	
Male	0.592	0.653	0.604	0.316	0.652	0.602	0.329
	(0.024)	(0.034)	(0.034)		(0.036)	(0.036)	
Transfer	0.271	0.477	0.455	0.673	0.470	0.436	0.528
	(0.022)	(0.036)	(0.035)		(0.037)	(0.037)	
Observations	417	193	202		181	181	

Note: This table includes all students who completed the second midterm. Descriptions of each variable can be found in Table A4. *Male* and *Female* 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, Final Exam sample

	All students			P-values	Matched pairs		P-values
Variable	Above Median	Control	Incentive	(3) - (2)	Control	Incentive	(5) - (4)
Midterm 1 score	2.049	0.153	0.057	0.291	0.177	0.170	0.938
	(0.025)	(0.061)	(0.069)		(0.064)	(0.065)	
Year = 2019	0.489	0.516	0.500	0.753	0.518	0.518	1.000
	(0.025)	(0.037)	(0.036)		(0.039)	(0.039)	
Cumulative GPA	3.445	2.946	2.959	0.864	2.929	3.001	0.346
	(0.029)	(0.044)	(0.060)		(0.047)	(0.059)	
No cum. GPA	0.231	0.359	0.332	0.583	0.367	0.313	0.299
	(0.021)	(0.035)	(0.034)		(0.038)	(0.036)	
Math quiz score	0.599	0.071	0.152	0.396	0.061	0.157	0.338
	(0.043)	(0.068)	(0.066)		(0.071)	(0.071)	
Tutoring visits	0.270	0.272	0.237	0.684	0.283	0.253	0.746
	(0.043)	(0.061)	(0.060)		(0.066)	(0.066)	
Videos watched	13.292	13.418	13.658	0.856	13.729	13.789	0.966
	(0.682)	(0.909)	(0.953)		(0.978)	(1.023)	
Videos, unique	9.793	9.783	10.111	0.704	9.795	10.181	0.674
	(0.432)	(0.598)	(0.622)		(0.630)	(0.665)	
Hours videos	1.698	1.788	1.805	0.929	1.812	1.803	0.967
	(0.094)	(0.130)	(0.138)		(0.138)	(0.148)	
Hours videos, unique	1.297	1.369	1.372	0.985	1.363	1.366	0.985
	(0.062)	(0.093)	(0.094)		(0.098)	(0.100)	
Asian	0.701	0.696	0.653	0.376	0.711	0.633	0.129
	(0.022)	(0.034)	(0.035)		(0.035)	(0.038)	
Latinx	0.060	0.141	0.158	0.654	0.139	0.169	0.448
	(0.012)	(0.026)	(0.027)		(0.027)	(0.029)	
White	0.149	0.109	0.132	0.497	0.102	0.145	0.244
	(0.018)	(0.023)	(0.025)		(0.024)	(0.027)	
Other ethnicity	0.089	0.054	0.058	0.882	0.048	0.054	0.804
	(0.014)	(0.017)	(0.017)		(0.017)	(0.018)	
Female	0.393	0.348	0.405	0.253	0.337	0.404	0.212
	(0.024)	(0.035)	(0.036)		(0.037)	(0.038)	
Male	0.593	0.647	0.584	0.215	0.657	0.584	0.176
	(0.024)	(0.035)	(0.036)		(0.037)	(0.038)	
Transfer	0.272	0.462	0.447	0.778	0.470	0.416	0.321
	(0.022)	(0.037)	(0.036)		(0.039)	(0.038)	
Observations	415	184	190		166	166	

Note: This table includes all students who completed the final exam. Descriptions of each variable can be found in Table A4. *Male* and *Female* 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 8 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

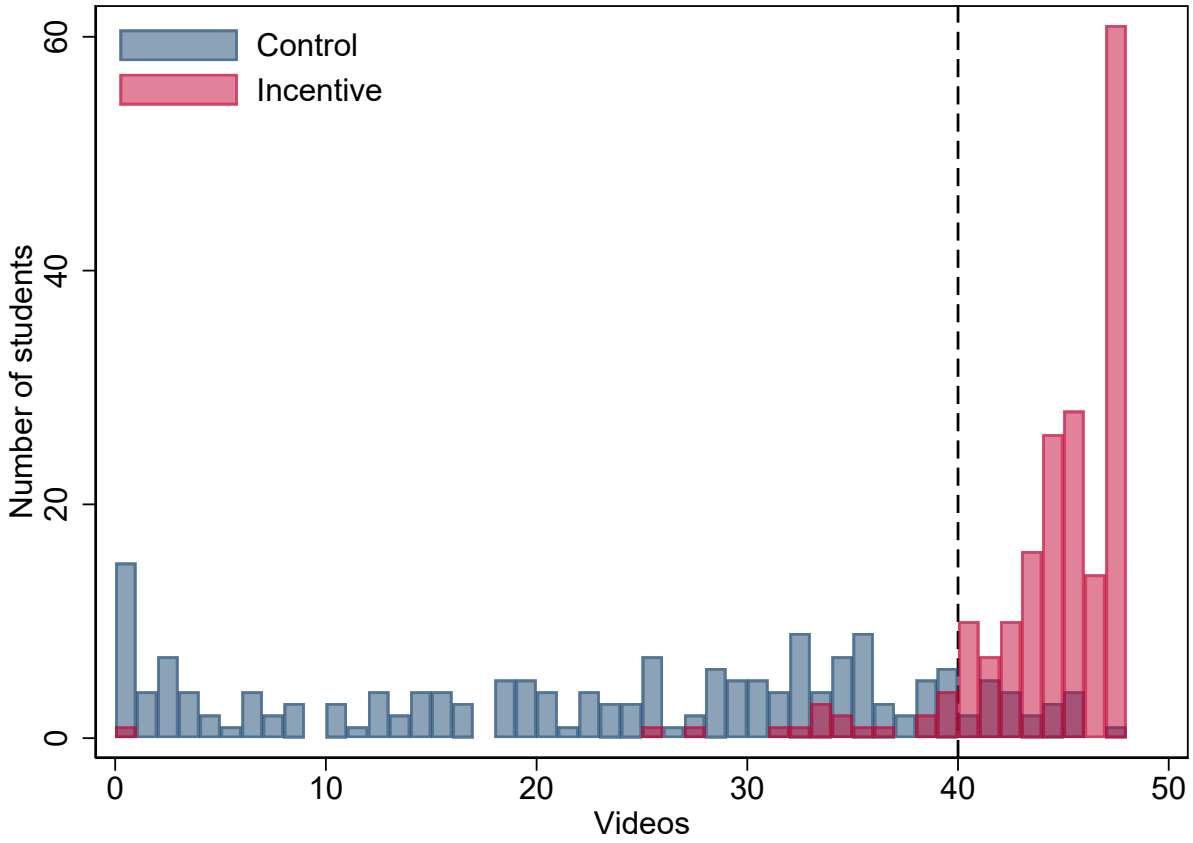
Note: Controls chosen via the PDS procedure of Belloni, Chernozhukov, and Hansen (2014b). In the *All Observations* model, *Midterm 1 score* and *Year = 2019* are additionally included as controls. In the *Fixed Effects* model, pair fixed effects and *Midterm 1 score* are included. All control variables 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

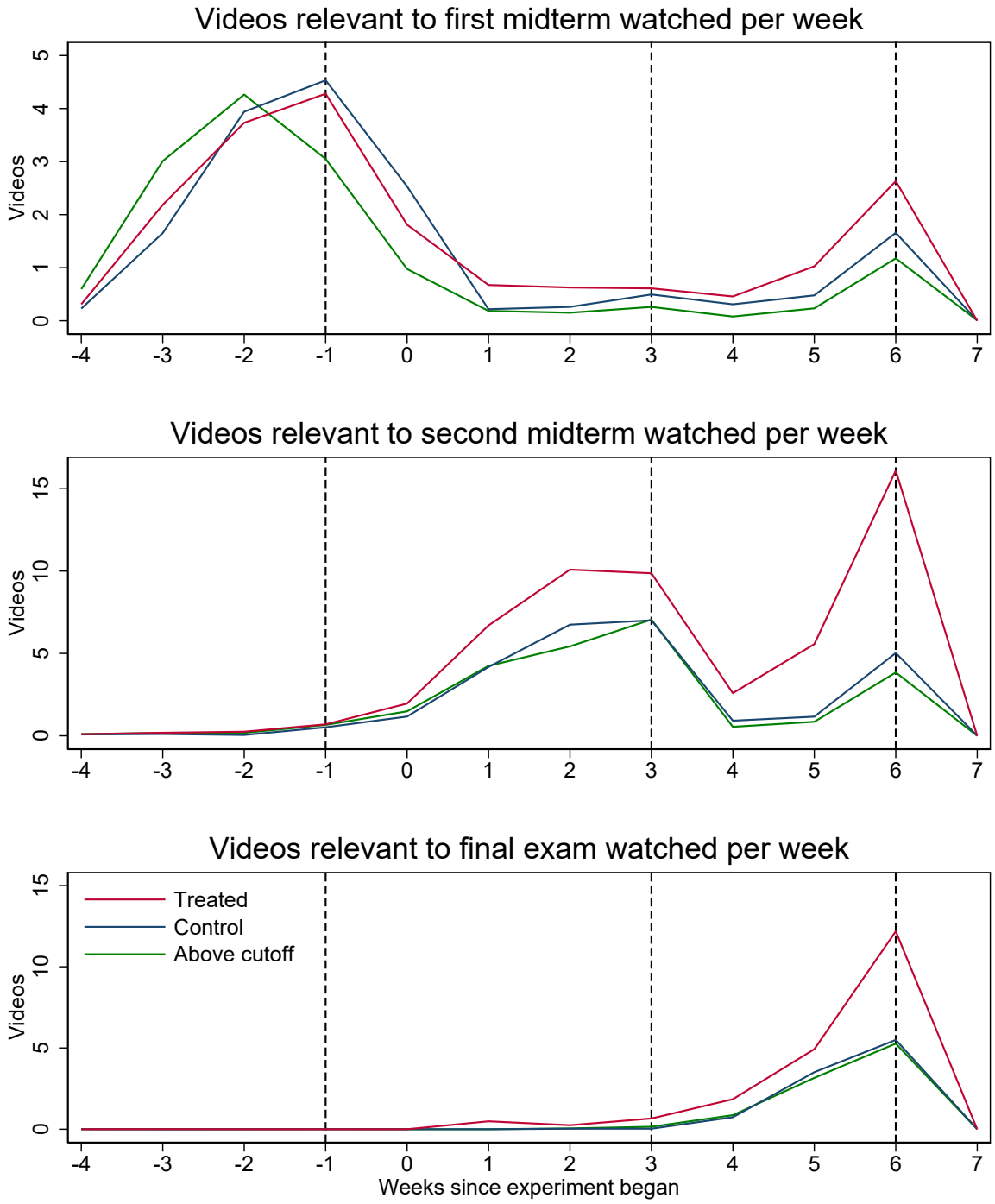
Note: Controls chosen via the PDS procedure of Belloni, Chernozhukov, and Hansen (2014b). In the *All Observations* model, *Midterm 1 score* and *Year = 2019* are additionally included as controls. In the *Fixed Effects* model, pair fixed effects and *Midterm 1 score* are included. All control variables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

Figure A6: Distribution of videos counted towards incentive



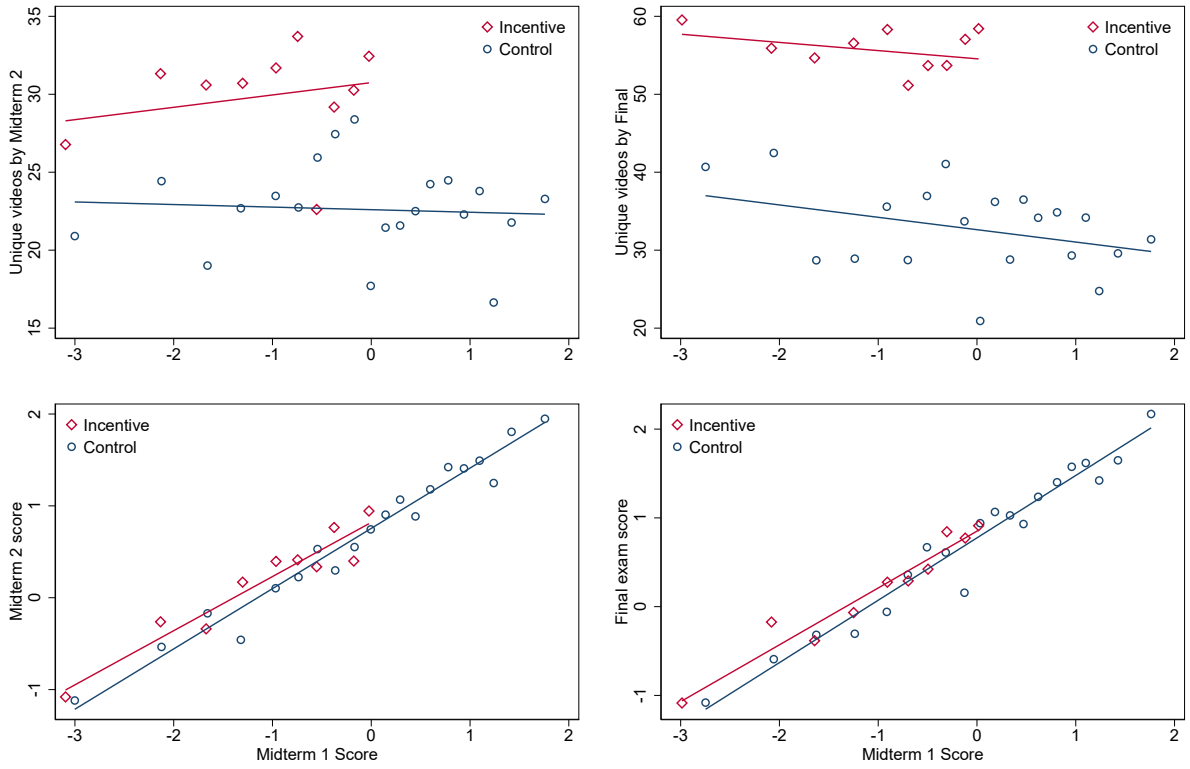
This plot includes only videos that would have counted towards the earning the grade incentive. Students were required to watch 40 unique of 48 eligible videos between the first midterm and final exam to earn the grade incentive. 91% of *Incentive* students met the requirements for the grade incentive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

Figure A6: Weekly video watching by exam topic



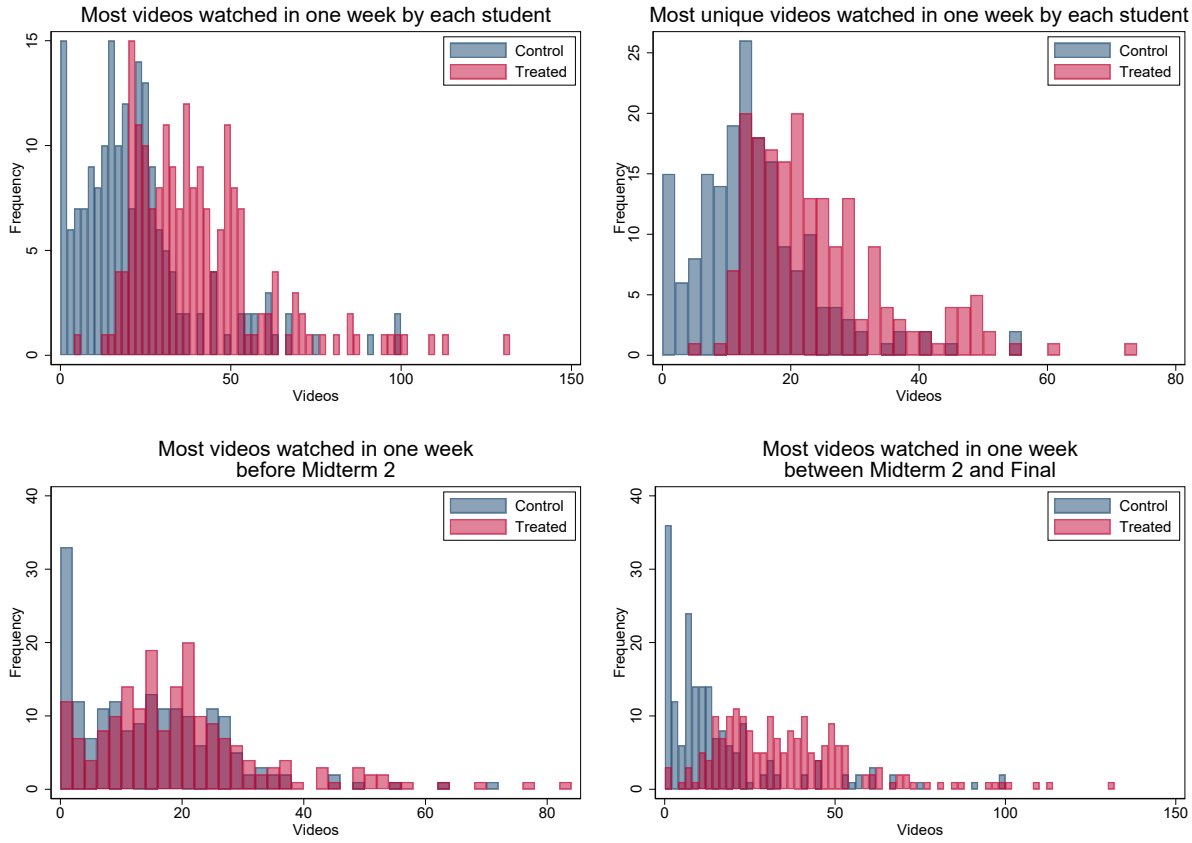
Dashed lines represent Midterm 1, Midterm 2, and Final exams.

Figure A6: Effect of treatment on videos watched and exam scores by Midterm 1 score



Test scores in standard deviation units. Each point comprises 5-percentile bins along the domain. The control points displayed include both *Control* and *Above median* arms.

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.