The effect of supplementary video lectures on learning in intermediate microeconomics

Melissa Famulari and Zachary A. Goodman*

University of California, San Diego

This version: November 2020

Abstract

The abstract goes here eventually.

*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 the University of California. As of fall 2020, one of the authors has a financial interest in the distribution of the IMVH outside of the University of California.

1 Introduction

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

— Anonymous student

University students spend tens of thousands of dollars annually on tuition and hundreds of hours in lecture and completing assignments, in large part, to learn. Instructors can improve how well students learn by employing pedagogical tools that have the greatest returns per unit time and financial cost. Despite the importance of comparing the effectiveness of different teaching technologies, little empirical work exists that estimates . In this paper we examine the impact of low marginal cost, video-based learning materials on exam scores in a large, intermediate microeconomic theory course.

The Intermediate Microeconomic Video Handbook (IMVH) at UC San Diego was designed to *supplement* lecture, not replace it, 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 current students, who have had unprecedented exposure to electronic media, would find video materials more engaging and ultimately study more effectively or for more time than they would have if provided only conventional studying materials. Though students may use the IMVH more than the textbook, it is an empirical question whether the videos ultimately improve learning outcomes.

We answer this question using a field experiment involving over 400 undergraduates enrolled in the same microeconomics course over two years. Only students who scored below the median on the first midterm were eligible for the experiment, since previous work and institutional knowledge suggests that students in the top half of the distribution would not benefit (and may even be harmed) from being induced to watch the videos. While the optimal experimental design for identifying average treatment effects would involve restricting access to the IMVH to only treated students, ethical considerations required that all students have access to the IMVH. Hence, we opted for an encouragement design in which treated students are induced to watch more videos than their control group peers through a grade-based

incentive, which more than doubled the number of videos watched by treated students. This experimental design permits identification of treatment effects local to those students induced by the encouragement to watch more videos.

Among students who performed poorly on the first midterm, we find the treated students watched XX more videos than the control group, an X percent increase in video watching. We find that the incentive increased midterm and final exam scores by 0.18 and 0.17 standard deviations, respectively, and that the marginal hour of video watched increased exam scores (LATE) by 0.08 standard deviations. Although the confidence intervals are, admittedly, wide, the point estimates are statistically and economically significant: a student could increase their course letter grade by one step (e.g. from a B+ to A-) by watching XX hours of videos. Our estimates suggest that XX percent of students in the control group who failed the course would have earned passing grades had they watched as many videos as their treated counterparts.Our experiment also allows us to use a regression discontinuity approach to estimate the effect of treatment for students near the median on the first midterm. We find no effect of treatment for students near the median on the first exam.

Although treated students performed better on course assessments, for determining welfare it is important to identify where the time watching videos came from: leisure time, working, student organizations, studying for other classes, studying for the current class using other methods, etc. On one hand, if watching videos is more productive than the next best studying method, then the utility of requiring videos is unambiguously positive as students can substitute studying time towards the more productive option. On the other hand, if students must reduce time allocated towards leisure or studying for other classes so they can watch more videos, then the welfare implications are less clear and could be negative depending on the students' preferences.

We attempt to disentangle whether treated students spent more time studying or used their time more effectively by examining proxies for time use including class attendance, visits to a tutoring center (specific to this course), downloading materials from the course website, posting on the class discussion board, and reported time use from an in-class survey. Although our estimates are noisy, we find no statistically significant differences between treatment and control, and we can reject large decreases in take-up of other study methods

by treated students. Surprisingly, in nearly all cases, the point estimates suggest that the treatment group used study methods beyond the videos at *greater* rates than did their control peers. Though estimates are noisey, we find no significant differences in reported leisure time across treatment and control. Finally, we investigate spillovers to other courses taken during the same academic term as the experiment and similarly find that treated students perform *better* than their control peers, which suggests that watching the videos likely did not dramatically reduce time spent studying for other classes.

Finally, we attempt to distinguish between two models of student learning which could explain why the video watching inducement improved student exam performance: an incomplete information model where students do not know how to study effectively versus a two-selves model where students like good grades but dislike studying. One important (and testable) difference between the incomplete information model and the two-selves model is what happens after exogenous incentives to watch videos are removed. While the former predicts that students exposed to treatment will continue watching videos given their new knowledge of a relatively productive studying technology, the latter predicts that the students will return to their lower baseline levels of video watching as their doer selves no longer have a commitment device reducing the temptation of immediately gratifying leisure. We examine video watching behavior during the term following the experiment in the subsequent microeconomics course and find that treated students watch significantly more videos than their control classmates, consistent with the incomplete information model.

Collectively, we interpret our findings as strong evidence that requiring studying tools known by the instructor to be effective is utility enhancing for students who manifest their limited knowledge of how best to study, perhaps through poor performance on an early stage assessment. Finally, we provide suggestive evidence that students found the IMVH to be a relatively effective study method by examining video watching across the treatment and controls in the next class in the sequence. The rest of the paper is organized as follows. Section 3 provides background on existing related literature. Section 4 describes the experimental 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 of 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 do not 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. Since college instructors have little knowledge about student utility functions, they do not know student preferences over performance in other classes and other aspects of their lives that may have large payoffs in the labor and marriage markets and so there is no room for an instructor to increase student well-being by intervening in their study decisions. **oettinger2002** provides 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 letter grade cutoffs and student performance on the final exam is higher if the student is just below a grade threshold. **kow2020** also finds support for the neoclassical model for students at the margin of poor performance. The authors use a regression discontinuity approach to examine the effects of a university policy that required students who performed poorly in their first year to attend at least 70 percent of the tutorials for each class in their second year. In courses where tutorial attendance was not required for all students, the policy increased both tutorial and lecture attendance by over 50 percent, did not increase total study time and reduced grades by 0.16-0.26 standard deviations. The policy had its biggest impact on students who lived far from campus and those who were most likely to miss section in their first year. The authors conclude that the

university policy prevented students from using their optimal mix of study methods and so reduced grades.

A key assumption of the neoclassical model is that students possess complete information about the returns across studying methods. However, there is evidence from psychology that college students do not know the return to various study methods[1] and many universities fund "Teaching and Learning Centers" or "Academic Skills Centers," part of whose mission is to help undergraduates learn to study more productively.[2] Further, the "raison d'etre" of higher education is not only to teach students specific skills but to teach students how to learn. As an alternative to the neoclassical model, we hypothesize that students supply a quantity of study time that is optimal given their information constraints. In this 'imperfect information' model, students choose study methods and quantities that are suboptimal relative to those they would have picked in a full information setting. Hence, an intervention by an entity that has more information about returns to studying across various methods (i.e. an instructor) can enhance student utility.

A third model is a behavioral one in which students plan to study more than they end up studying when the time comes. This phenomenon is consistent with two-self models in which a person's "planner" self, the one who desires high grades at the expense of leisure, is at odds with her "doer" self who must choose between immediately gratifying leisure and delayed gratification from higher grades. Indeed, survey and experimental data suggest that many students study less than they report they "should" and finish the term with grades lower than what they had anticipated they would earn at the start of the term, [3] **cgpr2020** provide empirical support for this model by finding that setting tasked-based goals helps improve college student performance. Also, **blmo2019** find that students that do much worse than expected in college are those who say they have poor time management or procrastination issues, including a tendancy to cram and spending very little time studying.

We consider the testable implications of the three models applied to a setting where students are incentivized to use a time-consuming educational input, say, a set of instructional

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

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

[3] see, for example, **ferrari1992**, **ccog2017** and **llo2016**.

videos (or attending class, reading the textbook, working on homework, etc.). The incentive is structured such that students who consume the educational input receive a higher grade in the course by consuming a set level of the input. In this simple setting, students gain utility 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 incentived 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 incentived 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 a 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

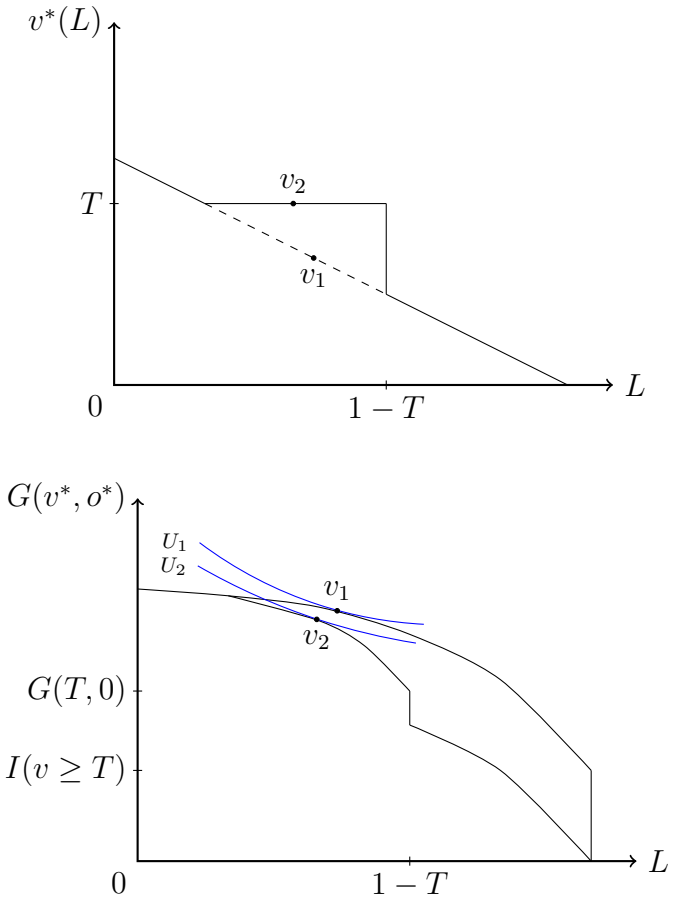

Figure 1: *Above*: Demand curve for video watching as a function of leisure L. At $L = 1-T$, the student maximizes grades $G(v, 0)$ by spending all studying time watching videos, i.e. $v^* = T$. *Below*: Student's utility maximization problem for the neoclassical model. The student maximizes her utility 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).

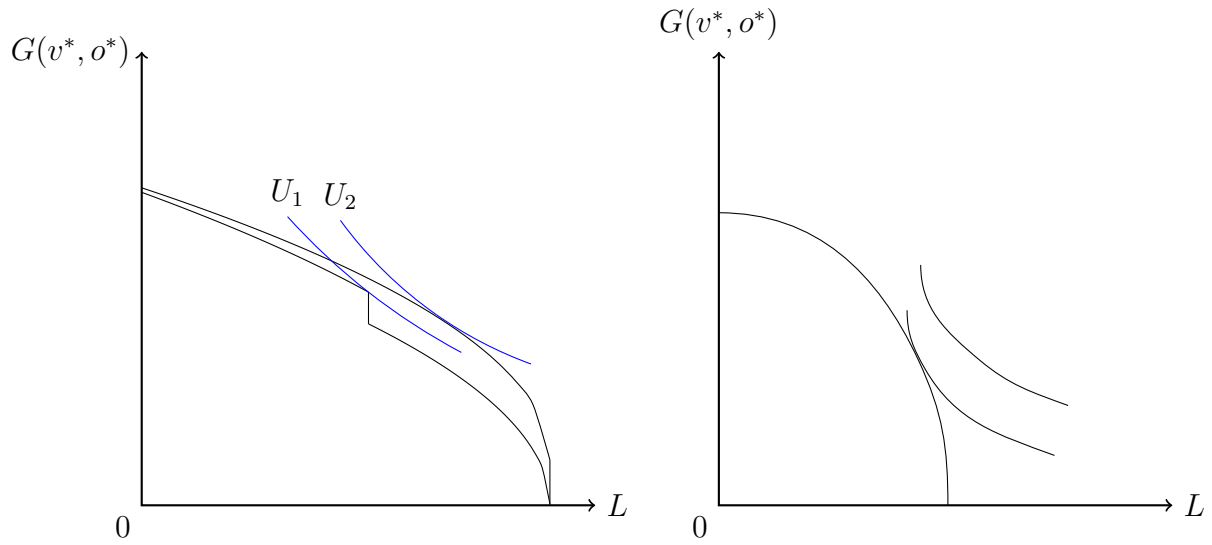

Figure 2: Student's utility maximization problem for the neoclassical model. The student maximizes her utility over leisure L and grades G, which is a function of time allocated to video watching v and her next best studying option o.

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

Finally, in the behavioral model, the instructor's inducement helps students stick to study plans up until the cutoff. As long as total study time does not fall, the inducment will increase exam performance. This model also predicts bunching at the incentivised level cutoff as long as using the incentivised input does not change the student's "planner" and "doer" selves. In the absence of the inducement, i.e., in future classes, a sharp prediction is that video watching will revert to pre-inducment levels.

In the empirical section, we test for the effects of being induced to watch the IMVH on

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 examine the effects of treatment on grades in other classes taken in the same quarter. We compare bunching of video watching across treatment status. 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 sequence, we test for differences in video watching across treated and control students in the next class.

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 or practice exams, or attending tutoring labs, and there are many empirical challenges to estimating the causal effects of a learning activity. First, a student's decision to use a study method is likely influenced by unobservable student characteristics, such as motivation or ability, that are also likely to affect student exam performance. To estimate causal effects, researchers must address selection into using the study method. Second, most instructors have experience with students who want to improve their study strategies after a negative exam shock which suggests "dynamic selection" into the use of a study method and has been found empirically by **oettinger2002**, **ko2005**, **ss2008**, **bo2012** and **bo2015**.[4] Dynamic selection means including student fixed effects in class performance regressions will not uncover the causal effect of a study method. Third, study methods may be substitutes or complements in student learning and experimental inducements to use one study strategy may affect student use of another. In these cases, even randomized experiments will not identify the causal effects of a particular study method but will instead identify the causal effects of a study

[4] **oettinger2002** finds that students close to a grade threshold before the final exam perform better on the final. **ko2005** show that students reduce the number of hours they study after getting higher midterm scores. **ss2008** find that IV estimates of studying on grades are much larger than OLS and using two years of data, provide suggestive evidence that students increase effort in semesters when semester-specific elements of grades are low. Finally, **bo2012** and **bo2015**

policy and all of the changes in student behavior caused by that policy.[5] Finally, experimental inducements to use a study method may change the total time devoted the course. In this case, experiments jointly test the effectiveness of a particular learning method and devoting more (or less) time to the course.

We focus our review on studies that use experiments or quasi-experiments to explore the effects of attending lecture, discussion sections/tutorials, or tutoring labs, doing homework, studying, and watching recorded lectures. **km2003** use student-reported travel time to campus as an instrument for lecture attendance and find travel time increases student attendence. Since they find no evidence of selection bias, they report OLS results and find the usual postitive relationship between lecture attendance and exam grades. **dgm2010** analyze a policy where lecture attendance was voluntary before the midterm but after the midterm was graded, required for students who scored below the median. At the threshold there was a 36 percentage point increase in post-midterm attendance which allows the authors to use a regression discontinuity approach to estimate the effects of attendance on the final exam. They find a 10 percentage point increase in student attendance led to a .17 standard deviation increase in final exam score. **tlad2020** randomly assign students to either weekly or bi-weekly grading of clicker questions (answered in lecgture). Weekly grading increased student lecture attendance by 11 percent but not student-reported self study hours. Course grades were 6.31 percent higher for students randomly assigned to weekly grading and the effects of weekly grading were much stronger if the student preferred bi-weekly grading, had lower prior GPAs and lower self-control scores.[6]

ans2012 study discussion section (as opposed to lecture) attendance. Since students are randomly assigned to sections and section attendance depends on day of week and time of

[5]While the causal effects of an educational policy are useful for educators considering how to design their classes they are less useful for students wanting to know the most productive use of their study time. We should also point out that instuctors may find learning transmission methods substitutable or complementary. As an example, in **msc2019** many instructors report that lecture capture, where lectures are recorded and made available to students, changed the way they lectured in the classroom. Experiments randomly assigning students to classes taught one way versus another will not identify the causal effect of a study method if other aspects of learning transmission are simultaneously changed.

[6]**cl2008** randomly omit material from lecture from a randomly chosen class and find students who heard the material did better on related multiple choice questions. However since both treatment and controls are in lecture, this study abstracts from the time costs of attending lecture and we focus on empirical studies exploring study methods that take time.

day, they use these variables as instruments. They find no effects of section attendance for most students but for students in the top quantiles missing 10 percent of sections resulted in a 1 percentage point performance loss. **kow2020** also examine the effects of section but like **dgm2010**, study an attendance policy focused on students who reveal themselves as poor performers. For students whose first year GPA was below a threshold, a European university required 70 percent section attendance in the second year. Nearly all affected students met the attendance threshold since there was a substantial penalty for not meeting it. They find students at the policy threshold attended 50 percent more sections and lectures but reported no significant difference in total study hours (lectures plus section plus self study). Students at the policy threshold earned grades that were 0.16 to 0.24 *lower*.

bs2013 have data on attendance for lecture and discussion sections combined. Their attendance measure is highly correlated with student reported self study hours and including self study hours significantly reduces the estimated effects of attendance. Using student residence as an instrument and controlling for self study hours, the authors find significantly positive attendance effects on grades only for quantitative courses. Finally, **mgm2010** find athletes are significantly more likely to attend peer tutoring labs which they attribute to frequent remindeers from coaches. Using being an athlete as an instrument for tutorial hours, they find significantly positive effects of peer tutoring: to increase a student's letter grade, they would need to spend about one hour per week being tutored over a 14 week semester.

Turning to the effectiveness of a student's own study time, we consider resesarch examining student-reported hours of self study as well research examineing what students could be doing in those self-study hours such as doing homework or practice exams. **ss2008** examine 210 Berea College students who were randomly assigned a roommate. Students whose roomates 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). **mbp2011** However, **oppp2019** randomly assign first year economics students at three different institutions to a variety of low

touch interventions to increase study hours (help planning study schedules at the start of the semester, information about the value of studying, and weekly reminders about study plans) which increased student's self-reported study hours by two hours per week at two of the three institutions but has no effect on grades, credit accumulation, or retention. **cgpr2020** also explore a low touch intervention, having students set goals on the number of practice exams they will complete, and find those randomly assigned to set goals completed 0.102 of a standard deviation more practice exams and increased total course points by .068 of a standard deviation. **ts2012** examine the causal effects of homework by randomly requiring two-thirds of 682 students in Principles of Economics classes to complete one of three homework assignments for a grade while the other third may complete the homework, but it does not contribute to their grade. The outcome is exam scores on questions related to the three homework assignments. They find significant effects of homework on the first midterm but not the final exam. **gr2013** use within-class randomization to assign 423 students in four principles of microeconomics courses to homework required (10 percent of grade) and not required groups (the homework grade incentive was spread across four exams). They find that 90 percent of the homework-required students completed 7 or more homework assignments compared to 6 percent of the controls. Treated students did better on the first two of four exams and the average across all four exams was higher.

Finally, we review the research on flipped classrooms and lecture capture, both of which have aspect similar to the IMVH. In the flipped classroom, the instructor has students watch recorded lectures outside of the classroom, take a quiz on basic concepts before lecture (to help ensure the videos are watched) and then do problem solving, group work and other activities during lecture. With lecture capture, lectures are recorded and then made available to students. In yet another interesting RCT using Air Force Acadamy students, **wbi2018** assign lecture type (flipped vs traditional) to randomly chosen topics and randomly chosen instructors. The course, Introduction to Econometrics, is highly standardized (common exams, textbook, assignments and practice exercises). They give students six short term exams (based on preceeding three lessons), four medium term exams (based on preceeding eight lessons) and one long term exam (a comprehensive final). They find positive effects of the flipped classroom, .16 standard deviations, for the medium-term assessments. Students

report spending more time preparing for class and find that pre-class preparation more hlepful in flipped classes but report no differences in the usefulness of lecture, the helpfulness of the non-lecture components of class, difficulty paying attention, helpfulness of instructor feedback, self-rated understanding of the day's topic, and appropriateness of the pace. Given how similar student reports of the classroom are, and the delayed effects of the flipped classroom on performance, the authors hypothesize that the students may be using the videos as an effective study tool before high stakes assessments.

Other researchers have investigated using technology to improve learning. **abbm2020** 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.

The Intermediate Microeconomics Video Handbook (IMVH) is a collection of 220 short-ish videos that cover the material for a year-long intermediate microeconomics class. Most topics were covered by two videos: one video introduces the concepts more intuitively with verbal explanations and graphs and the other video has the more formal, calculus-based definitions. The videos were created by six UC San Diego faculty members with professional videographer and production support. Many videos were created using an innovative presentation technology, the "learning glass," where the instructor uses neon markers to write on a large sheet of glass that has lights embeded along the glass edge to make the colors pop. The remaining videos are PowerPoint presentations that faculty edit as they talk and the instructor's image is superimposed next to the PowerPoint. Videos are closed captioned and were checked by graduate student for accuracy. Given the complexity of the material, a key objective of the faculty was to keep the web interface clean and simple so as not to distract from the content. A second objective was to help students find what they want quickly and so the IMVH has both a table of contents and an index, there are time stamps for each video on the IMVH web site to let students know what is in each video, and the video captions are searchable so, while watching the video, students can seach for a keyword and jump to the part of a video with that word. The last objective was to help students know where various topics "live" in interemediate microeconomics (in consumer theory? in producer theory? in welfare theory?) and so we keep the table of contents on the left panel displayed at all times

Table 1: Information Transmission Formats.

	Lecture	Book	Lecture Capture	IMVH
Instructor's time used	x			
Instructor-Learner Interaction	x			
Learner-Learner Interaction	x			
Readable		x	?	x
Scalable	?	x	x	x
Searchable		x		x
Skimmable		x		x
Stoppable	?	x	x	x
Watchable	x		x	x

except when the student is watching a video. While we do not know of another textbook completely comprised of videos, the IMVH is similar both to the Khan Academy web site, to lecture podcasts, and to textbook web sites that incorporate instructional videos.

The syllabus for all classes include a week-by-week list of topics to be covered, the textbook chapters that cover those topics and the videos in the IMVH that cover the topic. Thus students knew what the instructor expected them to read and watch each week.

Table 1 presents a classification of some options to present course material to students.[7] The IMVH differs from a traditional textbook because intructors verbally explain, graph and derive mathematical results in much the same way one would in a lecture. Like a textbook, the IMVHThe primary benefit of lecture over the IMVH is that students can stop the instructor and get an answer as soon as they are confused, or wonder about a connection of the material with their lives or other courses, or want to know if they are right about an extension of the material, etc. Further, student questions may have import externalities for the learning of other students in the class. There is also an important social aspect of lectures as students can interact with each other before, during and after lecture. Students cannot ask questions or interact with other students during an IMVH lecture. Some of the advantages of the IMVH over lecture are first, students control the IMVH in that they can rewatch, speed up or slow down an IMVH lecture. Second, compared to a large lecture hall, all students can clearly see and hear the IMVH presentation. Finally, the IMVH is closed-captioned which may be particularly useful when English is not the first language of the

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

instructor and/or the student. The IMVH differs from lecture capture because the IMVH videos are much shorter, averaging under ten minutes. The IMVH web site is well organized so students can see where the topic "lives" at all times. A useful feature of recorded lectures is they have course administration information which the IMVH does not have. However recorded lectures may also include components that do not work well when recorded, such as group work or class discussion.

4 Experimental Design

We conducted the field experiment in four intermediate microeconomics courses, two in fall 2018 and two in fall 2019. The university is a large, diverse and selective public research four-year university. At this institution, intermediate microeconomics is a three quarter sequence required for students majoring in Economics. The experiment was conducted in the first course of the sequence, Micro A. For students who take Micro B the quarter after Micro A, we have exam grades and video viewing from Micro B.

Students were told about the experiment in the first lecture and the syllabus. At any time during the quarter, students could opt out of having their data included in the analysis sample.[8] Students below the age of 18 at the start of the course as well as students enrolled via Extension were removed from the analysis dataset.[9]

The course is ten weeks long and the experiment began four weeks into the term following the grading of the first midterm exam, and as such, only students who took the first midterm were included in the experiment. Following the first midterm, students were assigned to one of three treatment arms: *Incentive*, *Control*, or *Above median*. Students who scored below the median on the first midterm were randomly assigned to *Incentive* or *Control* whereas *Above median* includes all other students in the experiment. We used pairwise random assignment stratified on midterm score and year cohort (details on treatment assignment can be found in the Appendix). Students in the *Incentive* arm received a grading scheme

[8]Students opt out via an online form owned by the Teaching + Learning Commons (T+LC) so that neither the instructor nor research team could observe which students decided to opt out.

[9]Students under the age of 18 were excluded per IRB protocol. We exlude Extension students because of their small number, their unknown and potentially very different preparation for the course compared to UC San Diego students, and we are missing all demographic information for these students

Table 2: Grade scheme by treatment arm. *Control* represents same grade scheme as *Above median*. Differences between the two grade schemes in red.

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%

that encouraged watching elegible videos, those related to the material covered after the first midterm, whereas *Control* and *Above the median* received the standard grading scheme that did not directly reward watching these videos. The two different grading schemes are outlined in Table 2. Specifically, the *Incentive* grading scheme requires that at least 40 of 48 eligible videos be watched to earn 4 percentage points towards the students' final grade and there was no partial credit[10]. Notably, the 4 percentage points comes at the expense of the first midterm score, which had already occurred at the time of treatment assignment. Hence, we isolate the video incentive as the sole difference between treatment arms, giving us more confidence that the exclusivity assumption of our encouragement design holds.

Many non-majors take the class typically to either satisfy general education requirements or to explore majoring in Economics. Thus there are many students at the margin of majoring in economics in the class and so an important outcome is the likelihood the student takes the second class in the sequence. The other unique feature of this class at this university is the large fraction of transfer students, for whom the class is not only their first experience with upper division coursework, but also the first time taking a class under the quarter system (community colleges in the state are on the semester system), and their first class at a 4-year research university. Thus, we expect transfer students to have greater information problems.

While the experiment was conducted in the first class of the sequence (Micro A) which was taught in the fall quarter, we also have data on exam grades and video watching from treatment and control students who took the second class in the sequence (Micro B) in the

[10]Watched in standard speed, 40 videos 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 incentivized videos in standard speed would require just under eight hours.

winter quarter. Importantly, in MicroB, the IMVH was made available to all students but video watching was not incentivised. Fortunately, one instructor taught all four of the Micro A classes (one of the authors) and another instructor taught all four of the Micro B classes. Both instructors created half of the videos lectures for the course they taught.

Enrollments are high enough that the Micro A and Micro B instructors both taught two classes back-to-back each quarter but offered all exams at a common time out-of-class. Given the same exams were given at the same time to both classes in a quarter, we treat the two classes in a quarter the same but account for different years in the empirical analysis.

Neither author graded any exams for this course.

Micro A had an identical structure across both classes and years. In addition to a live lecture on T-Th, either 11-12:30 or 12:30-2:00, students had access to weekly one-hour discussion sections run by Econ PhD students (including one of the authors), a tutoring lab staffed by both the graduate TAs and top undergraduates, weekly supplemental sessions taught by top undergraduates majoring in Economics and trained by the university in supplemental instruction, a discussion board (monitored by the instructor as well as the grad and undergrad TAs), four years of the instructor's old exams (no answers provided), weekly ungraded problem sets (with detailed answers provided one week after the problem set was posted), graded online quizzes on most Mondays, and free acces to the IMVH, a video handbook containing 220 short-ish videos covering the entire Micro A-B-C sequence. This video handbook was created by the two instructors teaching Micro A and Micro B as well as four other faculty members who also regularly teach in the intermediate microeconomics sequence.

As part of the experiment, the Micro A instructor committed to having the course grade distribution identical across treatment and controls. Thus the two key experimental outcomes are scores on the second midterm and the final exam.

Note we do not actually know if students watched the vidoes. Video "watching" was based on opening the video, leaving it open for at least half the video length, and clicking a link at the end of the video which took students to a Qualtrics survey where they had to record their e-mail address. After random assignment, we surveyed students to make sure they knew which group they were in. Twice during the quarter we let students in the experiment know how many videos they had watched and how many more they needed to

watch to earn the grade incentive.

We are primarily interested in the causal effect of watching videos on exam grades. The average causal effect of watching videos can be modeled using the potential outcome framework or Rubin Causal Model (**ir2015**):

$$\tau = E[Y_{it}(1) - Y_{it}(0)] \tag{1}$$

where τ is the average causal effect of watching videos and $Y_{it}(1)$ and $Y_{it}(0)$ are the potential outcomes (e.g., exam scores) for a student i in year t who does and does not watch videos, respectively. We can observe treatment assignment for each student, $Z_{it} \in \{0, 1\}$, as well as the observed outcome $Y_{it}(Z = z_{it})$ and a vector of pretreatment covariates X_{it}.

A student's decision to watch videos is endogenous, so we must rely on an exogenous instrument to calculate an unbiased estimate of τ. By randomly assigning treatment that induces significantly greater video watching, the treatment indicator Z_{it} satisfies the instrument validity and relevancy conditions required to estimate τ using an instrumental variable approach (citation). We calculate an estimate of τ using a two-stage least squares approach:

$$v_{it} = \alpha Z_{it} + f(X_{it}) + e_{it} \tag{2}$$

$$y_{it} = \tau \hat{v}_{it} + g(X_{it}) + u_{it} \tag{3}$$

where $\hat{v}_{it}$ is instrumented videos watched estimated by Equation 2, $f()$ and $g()$ are generic functions through which X_{it} affects v_{it} and y_{it}, respectively, and e_{it} and u_{it} are model residuals assumed to be mean-zero conditional on observables. $\hat{\tau}$ is the estimate of the local average treatment effect (LATE) of watching videos local to those students induced by treatment to watch videos.

Under the assumptions of independence, monotonicity, and non-interference, $\hat{\tau}$ is an unbiased estimate of the LATE (**ai1995**). Independence assumes that outcomes (e.g. grades) are only impacted by treatment 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 refered to as the "no de-

fiers assumption", is required because of two-sided noncompliance and requires that students assigned treatment are weakly more likely to watch videos than 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 excludability[11] and monotonicity[12] are reasonable assumptions, we have more pause about the non-interference assumption because of the potential for spillovers between students in the same class. If we had unlimited resources, a more 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 large effect sizes. Hence, we proceed acknowledging the potential for spillovers between students. We hypothesize that spillovers likely bias our estimates of the treatment effect *downwards* as we believe control students are more likely to benefit from having well-studied peers than they are to lose from, for example, having peers too busy watching videos to join a study group.

We also estimate the average causal effects of being assigned to the *Incentive* arm, or average Intention To Treat (ITT) effects. These are ITT estimates and not average treatment effect estimates because of non-compliance: some students in the treatment arm do not watch videos and some students in the control arm do watch videos. However, since the incentive itself in our setting is representative of how future instructors may require their students to use the IMVH, the ITT estimates are relevant for an instructor deciding whether to use a grade incentivize students to watch videos or simply make the videos available as a supplemental resource with no grade incentive.

[11]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

[12]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, in Appendix Figure **??** shows that the two CDFs do not cross.

Following the methods proposed by **robinson1988** and more recently **wa2018**, our primary estimating equation is the partially linear model:

$$Y_{it} = \beta Z_{it} + f(X_{it}) + \epsilon_{it} \tag{4}$$

where Y_{it} is an outcome of interest (e.g. videos watched or test scores) for student i in year t, $Z_{it} \in \{0, 1\}$ is a treatment indicator, $f()$ is a generic function through which X_{it}, a vector of controls, affects Y_{it}, and ϵ_{it} is an unobserved residual. Following the Neyman-Rubin potential outcomes framework, $Y(Z = 1|X) - Y(Z = 0|X)$

To identify the causal effect parameter of interest, β, we make three assumptions: 1) *treatment status unconfounded given controls*: $Z_{it} \perp\!\!\!\perp \Big(Y_{it}(Z = 1), Y_{it}(Z = 0)\Big)\Big|X_{it}$; 2)*Excludability*: 3) *Non-interference/Stable Unit Treatment Value Assumption (SUTVA)*:

4.1 ITT

We are interested in knowing the average effect of being assigned to the *Incentive* arm on exam scores, which we will refer to as the Intent to Treat Effect (ITT). We will test the null hypothesis that being treated has zero effect against the alternative that being treated has a nonzero effect on exam scores.

We will estimate the ITT using Neyman's (**neyman1923**) repeated sampling approach, considering each pair (block) a completely randomized experiment and averaging the results. We begin by estimating the point estimate of the ITT as the mean difference in outcomes across pairs:

$$\hat{\tau} = \frac{1}{J}\sum_{j=1}^{J} \hat{\tau_j} = \frac{1}{J}\sum_{j=1}^{J} y_{j,I}^{\text{obs}} - y_{j,C}^{\text{obs}} \tag{5}$$

where $\hat{\tau}$ is the point estimate of the ITT, J is the number of pairs in the sample, and $\hat{\tau_j} = y_{j,I}^{obs} - y_{j,C}^{obs}$ is the observed difference in outcome for pair j.

The point estimate $\hat{\tau}$ is appropriate if treatment is unconfounded by observable or unobservable variables. While this assumption is true in expectation given random assignment of treatment, as a robustness check we estimate specifications that control for differences in

observed characteristics, described later in this paper.

Following Neyman's repeated sampling approach, the estimated standard error of $\hat{\tau}$ (**imai2008**; **ir2015**; **ai2017**) is:

$$\hat{SE}(\hat{\tau}) = \Big(\frac{1}{J}\sum_{j=1}^{J}\hat{V}(\hat{\tau_j})\Big)^{\frac{1}{2}} \tag{6}$$

where $\hat{V}(\hat{\tau_j})$ 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 (**ir2015**, **ai2017**):

$$\hat{V}(\hat{\tau_j}) = s^2_{j,I} + s^2_{j,C} \tag{7}$$

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 with 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 (**imai2008**, **ir2015**, **ai2017**):

$$\hat{SE}(\hat{\tau_j}) = \Big(\frac{1}{J(J-1)}\sum_{j=1}^{J}(\hat{\tau_j} - \hat{\tau})^2\Big)^{\frac{1}{2}} \tag{8}$$

4.2 Treatment Effects at the 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 (**tc1960**; **ap2008**; **il2008**). With this method, we compare students who scored just below the cutoff to those who scored just above the cutoff, two groups 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, the first 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 unlikely that students attempted to target a particular side of the midterm score cutoff[13]. 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 pretreatment covariate at the cutoff, which gives us further confidence that this assumption holds.

To estimate local treatment 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:

$$\tau(c) = \lim_{x \uparrow c} \mathbb{E}[Y_{it}|X_{it} = x] - \lim_{x \downarrow c} \mathbb{E}[Y_{it}|X_{it} = x] = \mathbb{E}[Y_{it}(1)|X_{it} = c] - \mathbb{E}[Y_{it}(0)|X_{it} = c] \quad (9)$$

We estimate $\tau(c)$ using local low-order polynomials, per the advice of **gi2019**.

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 establish that *Incentive* and *Control* arms are balanced on observable characteristics both when students were assigned treatment and when they took midterm and final exams. Second, we show that the grade-based encouragement worked, i.e. that students in the *Incentive* arm watched significantly more videos than did their *Control* peers. Third, we present results from the LATE and ITT specifications as described in the previous

[13]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.

section. Finally, we estimate spillover effects in other courses during the experiment term as well as we the term following.

5.1 Balance betweeen treatment arms

5.2 Relevancy of the encouragement instrument

As described in the previous section, we use a Two-Stage Least Squares approach to estimate the LATE of watching videos on exam performance. We must check that our instrument is both valid and relevant to ensure this method will produce an unbiased estimate of the LATE (**ir2015**). The validity condition is met by assigning the treatment arm at random conditional on observable characteristics, namely exam score and year of instruction. Additionally, Table 3 gives us further confidence that treatment status is uncorrelated with demographics. Next we check relevancy, i.e. whether treatment status generates significantly more video watching. Below in Table 4 we present estimates from Equation 2 and find that being assigned to the *Incentive* arm induces students to watch XX more videos than does being assigned to the *Control* arm. The F-statistic is XX, far greater than 10, the generally accepted minimum value required to consider an instrument relevant.

5.3 Estimation of causal effects

First, we estimate the causal effect 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. Second, 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 5 and **??**) and nonparametric methods a la the repeated sampling framework of **neyman1923**. We check that our parametric results are robust to model specification by estimating Equations 5 and **??** with and without $f(X_{it})$ as a vector of linear control variables chosen via the Post-Double-Selection procedure of

bch2014a. To rule out differential attrition across treatment arms as a confounder, we fit the aforementioned models dropping any student whose matched pair attrited before the conclusion of the experiment.

5.3.1 Effect of grade incentive on video watching

Table **??**

5.3.2 Effect of treatment on exam scores

Table 3

5.3.3 Substituion from other forms of studying

Table 7

5.4 Spillover effects

Here we estimate spillover effects to other other courses taken concurrently during the term of the experiment. We also estimate spillover effects to Micro B, the subsequent course in the intermediate microeconomics sequence. It is important to examine spillover effects

5.4.1 Concurrent courses

Table 6

5.4.2 Subsequent intermediate microeconomics course

Table 8

6 Discussion

Contributions: First, we find that a small grade incentive is effective in motivating poorly performing college students to take-up video watching. We also reduced the weight on an early assessment and allowed students to earn back the lost points fully by meeting the video

watching requirement, which may also be an important motivator. This result adds to the literature on what motivates college students to use educational inputs including financial incentives (see papers cited in **gmr2011**), requiring students to attend class if they perform poorly on an early assessment but with no penalty for not attending as in **dgm2010**), or by having students set goals on the use of an practice quizzes and, again no penalty for failing to acheive the goal(**cgpr2020**).[14] Grade incentives have the unappealing feature that grades are directly a function of input use. The grade incentive used in this study was small, at most four percent of the student's grade, which may help mitigate this concern.

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

Third, while not statistically significant, it seems worth pointing out that we found positive point estimates on treated student's use of other educational inputs (attending class, downloading material from the course web page, and attending a tutoring lab) and grades in other courses taken in the same quarter compared to the control group. This surprising result was also found by **dgm2010** who required poorly performing students to attend class. The authors posit that the statistically insignificant but positive spillovers they find may be due to fixed effects of coming to campus. A fixed effects argument is less compelling in our context.

Finally, we find support for a "poor information" model of learning because treated stu-

[14]Note, in contrast to **cgpr2020**, **oppp2019** find that setting a weekly schedule ahead of time and weekly reminders via a text message had only a small effect on study time and no effect on output as measured by grades, retention or credit accumulation. In contrast to the effectiveness of incentives on college student use of educational inputs, several papers find incentives to directly increase output, such as grades or performance on exams, are less effective. See for example, **fryer2011** shows that paying students in K-12 to increase inputs shows greater effects than paying for increased output, **gmr2011** also point out that incentives in education appear to work better for inputs than output

dents frequently watched more than the 40 videos they were required to watch to get the grade incentive and because treated students continued watching videos at a significantly higher rate in the following class when videos were not incentivised. **lo2009** also find continued higher use of academic support services after the incentivised year for women. A poor information model of learning can also account for why incentivising educational inputs has been found to be more effetive than incentivising grades or exam performance directly (see studies cited in **gmr2011**)

aws2015 review the literature on using technology to provide supplemental aids for students in traditional classrooms and conclude that there is little causal evidence that student acheivement is improved. The current study stands in contrast to this prior literature. In particular, they conclude that online instruction appears to have a negative effect on course grades and persistence. A more recent paper by **bflt2017** confirms these results, particularly for low performing students. The IMVH is clearly the backbone for an online class as it includes all the videos one would require students to watch as part of an online course. Perhaps the main effect of the video-watching requirement in this study was to increase student study time. Would encouraging at-risk students to watch instructional videos in an online class be as effective as we find for a in-person class? We consider this an important area for future research.

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.

Though we followed the approach of **dgm2010**, 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.[15] It is the first upper division class for many students and, for transfer students, also their first class in the university and under the quarter system. As such there may be large information problems about how to successfully study for the class under study and that there would be smaller effects of such an intervention in later classes.

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 (though we should point out many top students also "binge watch" before midterm 2 and the final exam and

[15]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

annecdoatally we have been told that they find this approach the most useful for reviewing the material.)

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 **kay2012**. 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 myopia or information barriers lead to suboptimal time allocation decisions. We observe two phenomena that supports the information barriers 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.

Table 3: 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)	
PSET 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 A2. *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 4: 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***	9.09***	9.53***
		(2.85)	(3.12)	(2.04)	(2.18)
Unique videos	23.13	6.63***	6.79***	5.97***	6.11***
		(1.54)	(1.70)	(0.98)	(1.11)
Hours of videos	4.08	1.19***	1.19***	1.13***	1.20***
		(0.37)	(0.39)	(0.24)	(0.26)
Hours of unique videos	2.97	0.79***	0.79***	0.75***	0.79***
		(0.23)	(0.23)	(0.13)	(0.15)
Observations		395	362	395	362
Panel B: By Final Exam					
Videos	53.09	39.25***	39.07***	38.77***	38.42***
		(4.06)	(4.37)	(3.42)	(3.79)
Unique videos	33.95	21.55***	21.08***	21.34***	20.49***
		(1.55)	(1.66)	(1.22)	(1.27)
Hours of videos	6.32	4.02***	4.01***	4.04***	3.98***
		(0.52)	(0.55)	(0.40)	(0.44)
Hours of unique videos	4.35	2.35***	2.34***	2.36***	2.32***
		(0.25)	(0.27)	(0.18)	(0.20)
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 **bch2014a** to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A3. 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 5: 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*	0.295*	0.286*
	(0.146)	(0.154)	(0.151)	(0.146)
2SLS: 1 hour of videos	0.224*	0.233*	0.238**	0.222*
	(0.128)	(0.138)	(0.121)	(0.124)
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*	0.083**	0.088*
	(0.041)	(0.049)	(0.041)	(0.046)
2SLS: 1 hour of videos	0.074**	0.074*	0.074**	0.058
	(0.038)	(0.045)	(0.037)	(0.044)
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 5 (Reduced Form, *RF*) and $\hat{Video_i}$ from Equation **??** (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 **bch2014a** to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A3. 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 6: 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)	(0.11)
		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 5. 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 **bch2014a** to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A3. 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 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)
Num. Piazza views	49.81	10.64	8.51	10.64	3.69
		(7.64)	(8.25)	(7.60)	(8.05)
Num. Piazza days online	10.40	1.43	1.89	1.43	1.67
		(1.55)	(1.59)	(1.54)	(1.65)
Num. Piazza questions asked	0.53	0.32	0.30	0.32	0.30
		(0.25)	(0.30)	(0.25)	(0.31)
Num. Piazza answers	0.47	0.08	0.01	0.08	-0.02
		(0.26)	(0.28)	(0.26)	(0.28)
Num. of PSET 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 5. There were seven *Attendance checks* during the quarter. *PSET 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 **bch2014a** to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A3. Models (2) and (4) include only students whose matched-pair did not attrite from the experiment. *Control Mean* is the mean for the Control students included in models (1) and (3), which is nearly identical to the mean for the Control students included in models (2) and (4). Standard errors in parentheses are robust to heteroskedasticity. ***, **, and * indicate significance at the 1, 5, and 10 percent critical levels, respectively.

Table 8: Spillover Effects during Subsequent Quarter

	Control Mean	(1)	(2)	(3)	(4)
Panel A: Videos during subsequent quarter					
Num. of videos	25.46	14.00***	12.78*	12.32***	12.25*
		(4.45)	(6.74)	(4.24)	(6.47)
Num. unique videos	19.77	9.87***	8.85**	8.65***	8.29**
		(3.03)	(4.04)	(2.89)	(3.87)
Hours of videos	2.87	1.47***	1.30*	1.29***	1.34*
		(0.52)	(0.77)	(0.48)	(0.78)
Hours unique videos	2.31	1.10***	0.93*	0.94***	0.92*
		(0.38)	(0.52)	(0.36)	(0.51)
Observations		211	108	211	108
Panel B: Effects on classes passed					
Midterm 1 score		-0.04	-0.24	-0.04	-0.31
		(0.13)	(0.18)	(0.13)	(0.19)
		213	112	213	112
Midterm 2 score		0.00	-0.04	0.00	0.04
		(0.13)	(0.20)	(0.13)	(0.22)
		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 5. 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 **bch2014a** to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A3. 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.

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. **ir2015**, **ai2017**). 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 (the math quiz was given on the second class day and so before some students enrolled in the class).

We assigned treatment shortly after issuing the first midterm exam grades, which 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 sample by XX students before the second midterm and XX students before the final exam. A XX% rate of withdraw 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 XX% of students attriting before the second midterm and XX% before the final exam, the two treatment arms below the median are balanced on nearly all observable pretreatment covariates, as shown in Table 3, 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 5 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 (**ai2017**). 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 **bm2009** 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 (**sns2011**). 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 **bch2014a**.

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 A2 below, we describe all covariates observable in our study. In Table A3, 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 **bch2014b**.

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)	
PSET 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 A2. *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: 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
PSET visits	Number of PSET 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 A3: 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	Hours videos
	Hours unique videos by Mid. 2	Hours videos	Hours videos
	Hours videos by Final	Hours videos	Hours videos
	Hours videos by Mid. 2	Hours videos	Hours videos
			PSET visits
			Videos
	Num. unique videos before Final	Hours videos	Videos
		Videos	
	Num. unique videos before Mid. 2	Videos	Videos
	Num. videos before Final	Hours videos	Hours videos
		Videos	Videos
	Num. videos before Mid. 2	Videos	PSET 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	Transfer
	Excluding Micro A	Cumulative GPA	Transfer
	Excluding econ classes	None	None
	Letter grade in Micro A	Cumulative GPA	Cumulative GPA
		Latinx	
		Transfer	
	Num. classes not passed	None	None
	Num. classes passed	Cumulative GPA	Cumulative GPA
		Transfer	Transfer
	Num. classes taken P/NP	Latinx	Latinx
	Num. classes taken for letter	Cumulative GPA	Cumulative GPA
		No cum. GPA	
	Num. classes withdrawn	None	None
	Num. units taken P/NP	Latinx	Latinx
	Num. units taken for letter grade	Cumulative GPA	Cumulative GPA
		No cum. GPA	
	Num. units withdrawn	None	None
	% classes taken P/NP	None	Latinx
	% classes taken for letter	None	Latinx
Table 4	Attendance checks	Female	PSET visits
		Math quiz score	
		PSET visits	
	Num. Piazza answers	None	None
	Num. Piazza days online	None	None
	Num. Piazza questions asked	None	None
	Num. Piazza views	None	Asian

Continued on next page

Table A3 (continued)

	Num. of PSET visits	PSET visits	PSET visits
Table 5	Hours of videos	Hours videos	Latinx
		PSET visits	Math quiz score
			PSET visits
			Videos
	Midterm 1 score	None	Latinx
			Math quiz score
	Midterm 2 score	None	Asian
			Latinx
			Math quiz score
	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
			PSET 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
			PSET visits
	Term GPA, econ courses ex. Micro B, winter	None	Math quiz score
	Term GPA, ex. Micro B	Cumulative GPA	Cumulative GPA
			PSET visits
	Term GPA, ex. econ courses	None	PSET 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
	Hours unique videos	Hours videos	Latinx
			Math quiz score
			PSET visits
			Videos
	Num. unique videos	Hours videos	Latinx
			Math quiz score
			PSET visits
			Videos
	Pass Micro B	None	Latinx
			Math quiz score
			Videos

Note: Controls chosen via the PDS procedure of **bch2014a**. 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 A4: 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	Hours videos
Final exam score	Videos, unique	Hours videos Videos	Hours videos Videos
Midterm 2 score	Hours videos, unique	Hours videos Math quiz score PSET visits	Hours videos
Midterm 2 score	Videos, unique	Hours videos Videos	Hours videos Videos

Note: Controls chosen via the PDS procedure of **bch2014a**. 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.