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 UC San Diego. As of fall 2020, one of the authors has a financial interest in the distribution of the IMVH outside of UC San Diego.

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.

We find that being assigned treatment (ITT) 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.

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 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 lesire 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 increase the use of an effective study method. We do not address the issue that the IMVH is a realtively unique study tool in that, to our knowledge, it is the first instructional book to be created entirely of videos. However, given the availability of close substitutes to the IMVH (lecture capture for example) we only briefly explore the added issues of inducing students to use a study tool whose usefulness is not known to the instructor.

Neoclassical models of studying behavior assume that rational agents know their returns to studying using the methods available to them and allocate the optimal amount of study time to each method given their utility function, which is increasing in leisure and grades and decreasing in time spent studying. In this model there is no room for an instructor to increase student well-being by intervening in their study decisions. Oettinger (2002) provides some empirical support for the neoclassical model by demonstrating that student effort responds rationally to nonlinear grade incentives. Across 1200 students in a principles of economics class with absolute grading standards, he finds evidence of bunching just above the letter grade cutoffs - student performance on the final exam is higher if the student is just below a grade threshold.

However, in addition to teaching specific skills, many would agree that the "raison d'etre" of higher education is to teach students how to learn. There is evidence from psychology that college students do not know how to learn effectively.[1] Universities often fund "Teaching and Learning Centers" or "Academic Skills Centers," part of whose mission is to help undergraduates learn to study more efficiently.[2] We posit that for many students, a key assumption of the neoclassical model does not hold: that students possess complete information about the returns across studying methods. Instead, we offer the alternative hypothesis that students supply a quantity of study time that is optimal given their information constraints. In

[1]See, for example, Pashler, Rohrer, Cepeda, Carpenter (2007). "Enhancing learning and retarding forgetting: Choices and consequences," Psychonomic Bulletin and Review 2007, 14, 2, 187-193; Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M.J., and Willingham, D. T. (2013). "Improving students' learning with effective learning techniques promising directions from cognitive and educational psychology." Psychological Science in the Public Interest. Afton Kirk-Johnson, Brian M. Galla, Scott H. Fraundorf (2019). "Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice" Cognitive Psychology, Volume 115

[2]All nine University of California campuses have one. Some others in the US include Dartmouth's Academic Skills Center, Michigan's Center for Research on Teaching and Learning, UNC's Learning Center, and Yale's Teaching and Learning Center

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 be utility enhancing.

A third model is a behavioral one in which students plan to study more than they end up studying when the time comes. 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] Recent empirical evidence suggests that setting tasked-based goals helps improve college student performance.[4] 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.[5]

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

[3]Ferrari, J.R. "Psychometric validation of two Procrastination inventories for adults: Arousal and avoidance measures." J Psychopathol Behav Assess 14, 97–110 (1992). Patricia Chen, Omar Chavez, Desmond C. Ong, Brenda Gunderson (2017) "Strategic Resource Use for Learning: A Self-Administered Intervention That Guides Self-Reflection on Effective Resource Use Enhances Academic Performance" Psychological Science, Vol. 28, Issue 6, 774-785. Stinebrinckner and Stinebrickner (2008) show large grade declines if one is randomly assigned a roomate who brings a video game to college

[4]Clark, Damon, David Gill, Victoria Prowse, and Mark Rush (2020) "Using Goals to Motivate College Students: Theory and Evidence From Field Experiments" The Review of Economics and Statistics 2020 102:4, 648-663

[5]see review paper, Adam M. Lavecchia, Heidi Liu, and Philip Oreopoulos (2016), "Behavioral Economics of Education: Progress and Possibilities" Handbook of the Economics of Education, Volume 5, Pages 1-74)

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 weak decrease in other studying and leisure consumption. It is ambiguous whether cumulative study time increases or decreases as this depends on relative utility benefits of leisure and grades and the returns to studying by each method. However, if cumulative study time remains constant or decreases, then exam performance should strictly decrease since students are now suboptimally allocating study time versus their first-best allocation when considering only marginal returns to studying. On the other hand, if cumulative study time increases, students may earn greater exam grades but achieve lower utility compared to baseline. Importantly, this model predicts that in subsequent quarters students return to their pre-incentive levels of studying.

In the imperfect information model, students' *ex ante* allocations to each studying method are not necessarily first-best. Compliers update their priors about the returns to watching videos as they work towards hitting the minimum required level. At this cutoff, they make a decision whether to continue watching videos depending on their updated perceptions of the marginal benefit. Hence, bunching at the cutoff is predicted only if the updated marginal benefit at the cutoff is lower than the marginal benefit of the next best studying option or the marginal utility of leisure.

that the marginal benefit at the cutoff is greater (lower) than the marginal benefit of their next best option

the incentive to watch videos will increase exam performance as long as total study time does not fall. A sharp prediction is that video watching will continue at the incentivized level

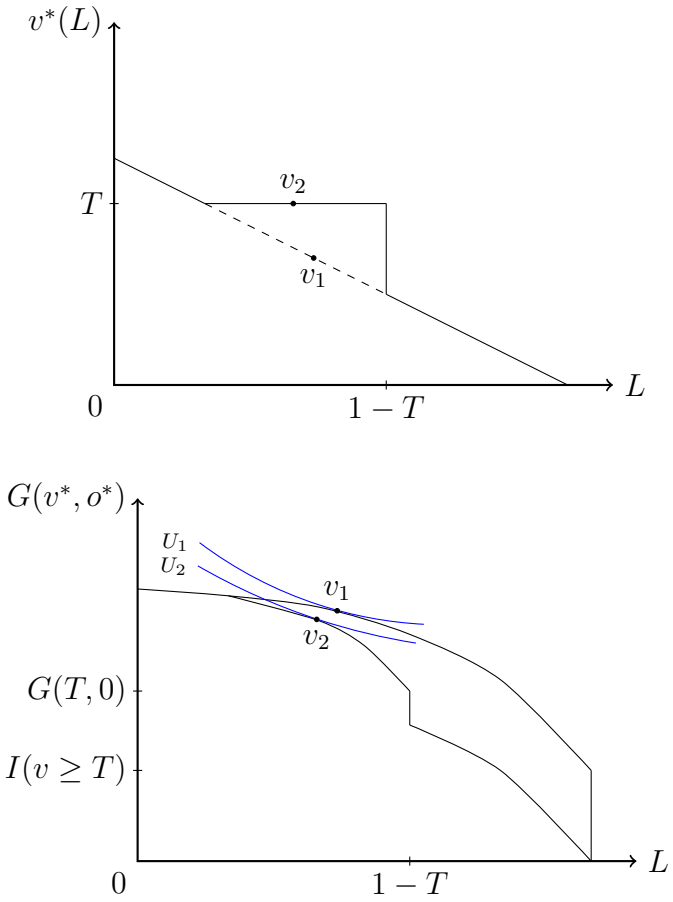

Figure 1: *Above*: Demand curve for video watching as a function of leisure L. At $L = 1-T$, the student maximizes grades $G(v, 0)$ by spending all studying time watching videos, i.e. $v^* = T$. *Below*: Student's utility maximization problem for the neoclassical model. The student maximizes her utility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

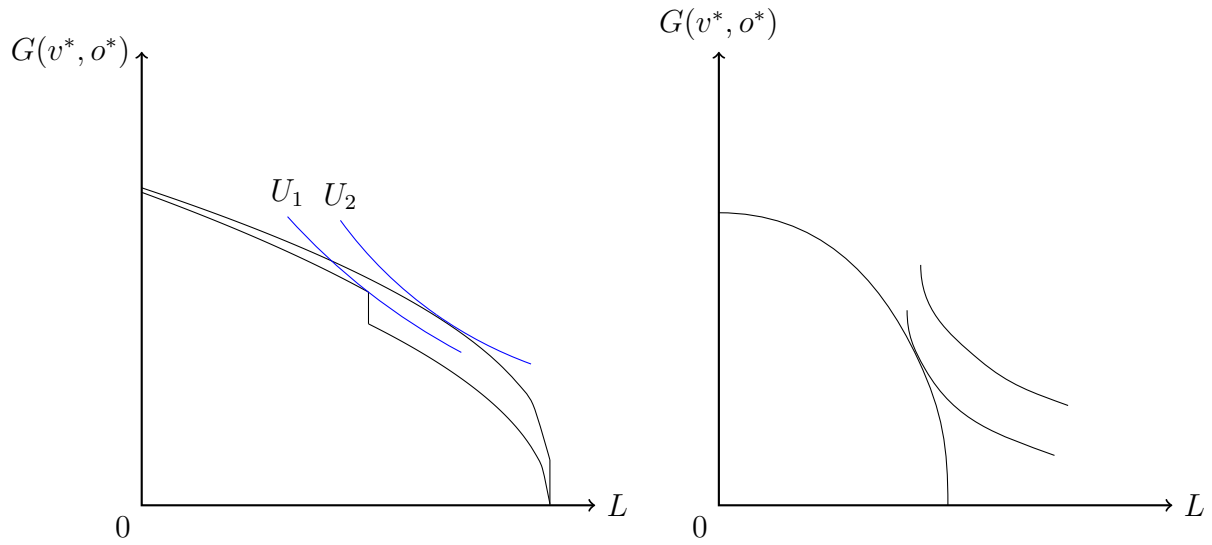

Figure 2: Student's utility maximization problem for the neoclassical model. The student maximizes her utility over leisure L and grades G, which is a function of time allocated to video watching v and her next best studying option o.

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 and one plausible group is transfer students who are taking their first class at UC San Diego, their first upper division class and, typically, their first class under the quarter system (vs semester).

Finally, in the behavioral model, the instructor's inducement helps students stick to study plans. As long as total study time does not fall, the inducment will increase exam performance. In the absence of the inducement, 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 both exam scores and several other study methods student use to learn microeconomics (lecture attendance, visits to a class-specific tutoring lab, use of a class discussion board, downloads from the class web page). We also compare grades in other classes taken in the same quarter across the treatment and control group. For a subset of our sample we have survey responses on total study time in the quarter and leisure time. Since the experiment was conducted in the first of a required three-class sequence, we examine video watching in

the second class. We test whether the effect of the inducemenmt to watch videos is greater for transfer students (assumed to have more information problems) and non-native English speakers (assumed to benefit the most from closed captioning, a key feature of the IMVH).

3 Related Literature and Contributions

Students have many time-consuming activities to help them learn including attending class, watching recorded lectures, reading the textbook, doing homework, attending office hours, etc. There are several empirical challenges to estimating the causal effects of a learning activity. First, unobserved student characteristics, such as ability and motivation, are likely positively correlated both with the use of a learning activity and class performance. To estimate causal effects, empirical studies must address the selection into using a study method. Second, most instructors have experience working with motivated students who want to improve their study strategies after a negative exam shock. Several papers have found empirical evidence of such dynamic changes in study behavior including Oettinger (2002), Ralph and R. (2008) and Cannon (2011) and dynamic selection means that fixed effects estimates will not uncover causal effects and difference and difference estimation may also be subject to selection bias. Third, learning activities are substitutable and inducements to use one study strategy may affect student's use of another. In these cases, even randomized experiments will not identify the causal effect of a study method but will rather identify the causal effects of a study policy and all of the changes in behavior caused by the policy. The causal effects of an educational policy, such as requiring homework, may be the most useful for educators considering how to design their classes. However, such estimates are of limitted use to help students identify the most effective study strategies. A final empirical issue in identifying causal effects is that study strategiess all take time and so most empirical studies jointly test the effectiveness of a particular study strategy and devoting more time to the course. It is possible that the primary benefit to students is simply devoting more time to learning the course material–however they do it.

Attending class: Chen and Lin (2008)[6] collect attendance and randomly leave out exam

[6]Chen, J. and Lin, T-F. (2008). Class attendance and exam performance: A randomized experiment.

material in one but not the other lecture of the same course. By comparing performance of students who attended with those who did not,

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

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

Journal of Economic Education, 39, 213– 227.

and a 84 percentage point increase in completing the majority of homework assignments. They find that treated students are less likely to drop the class and score higher on the first two but not the last two exams. The average across the four exams is increased 5-6 percent by treatment and the control group GPA would increase from 2.44 to 2.68 if they had been required to do homework. They find three times larger treatment effects for students who initially fail the first exam (10 to 15 percent vs 4 to 6 percent increase in average test scores). The authors do not examine whether other uses of student time are affected by the homework policy.

Effect of Study time: In a convincing empirical paper on the effects of study time, Ralph and R. (2008) examine 210 Berea College students who were randomly assigned a roommate. Students whose 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).

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

Effect of Recorded Lectures: Perhaps the educational resource most closely related to the IMVH is when the instructor's recorded lecture is made available to students. Recorded lectures have course administration information which the IMVH does not have. Recorded lectures are much longer than a typical IMVH video, less organized, and it may not be clear what topics are covered in the lecture video and/or where in the video it is covered. Savage (2009) taught two intermediate micro classes: one 42-student class had "talk and chalk" lectures and the other 45-student class used technology that allowed lecture capture which was then made available to the students. The author finds no significant differences in exam performance across the classes.

Clark et al. (2020) explore the effect of having students set goals on class performance. They find that task-based goals, completing a specific number of online practice exams,

but not performance-based goals, course and exam grades, improve student performance on exams. Those randomly assigned to complete task-based goals completed more, 0.102 of a standard deviation, practice exams and increased total course points by .068 of a standard deviation. They caution that it is important that the tasks are based on doing a productive study method. Our intervention seems quite similar but the goal was set by the instructor as opposed to the student and there was a direct grade penalty to not acheiveing the goal.

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

The Intermediate Microeconomics Video Handbook (IMVH) is a collection of 220 shortish 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 in 2014 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

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

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" and so we keep the table of conntents on the left panel displayed at all times excpet when the student is watching a video. While we do not know of another textbook coimpletely comprised of videos, the IMVH is similar both to the Khan Academy web site and to textbooks that incorporate innstructional videos on the textbook web site.

Table 1 presents a classification of some options to present course material to students.[7] The IMVH differs from a traditional textbook because intructors explain, graph and derive mathematical results in much the same way one would in a lecture. The primary benefit of lecture is that students can stop the innstructor, ask questions and get answers in real time. There is also an important social aspect of lectures as students can interact with each other before, during and after lecture. The IMVH differs from lectures because students control the IMVH lecture: they can rewatch, speed up or slow down the lecture. Compared to a large lecture hall, all students can clearly see and hear the IMVH lecture. Finally, the IMVH is closed-captioned which may be particularly useful when English is not the first language of the instructor and/or the student. Compared to lecture capture, the IMVH presents the material differently than lecture

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

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-term sequence required for students majoring in Economics. The experiment was conducted in the first course of the sequence. We also have grades and video viewing in the second course of the sequence.

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 experiment began four weeks into the term following 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. Random assignment was issued using 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 that encouraged watching videos during the rest of the term whereas *Control* and *Above the median* received the standard grading scheme that does not directly reward watching 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 in the IMVH be watched to earn 4 percentage points towards the students' final grades[10]. Notably, the 4 percentage points comes at the expense

[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 most do not have information are so different from we do not know if they met the course prerequisites nor and

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

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%

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.

which and there are two unique features of this class. First, 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 is the large fraction of transfer students, for whom the class is not only their first experience with upper division coursework, but typically 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 (100A), we also have data from the second class in the sequence (100B). Fortunately, one instructor taught all four of the 100A classes (one of the authors) and another instructor taught all four of the 100B classes. Both instructors created half of the videos lectures for their course.

Enrollments are high enough that the 100A and 100B instructors both taught two classes back-to-back each quarter but offered all exams at a common time out-of-class. So we treat the two classes in a quarter the same but account for different years in the empirical analysis.

The 100A course had an identical structure across all quarters and years. In addition to the textbook and the option to attend a live lecture at either 11-12:30 or 12:30-2:00, students had access to weekly one-hour discussion sections run by graduate TAs who were

all Economics PhD candidates (including one of the authors), a tutoring lab staffed by both the graduate TAs (in lieu of office hours) and top undergraduates (some earning course credit for learning how to teach economics and others hired by the Department to help cover the tutoring lab hours, M-Th 5:30-8:30 and Sunday 4-8pm), weekly supplemental instruction sessions offered by an undergraduate 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), graded online quizzes on Mondays of most weeks that did not include an exam, and a video handbook containing 220 short-ish videos covering the entire 100ABC sequence. This video handbook was created by the two instructors teaching 100A and 100B as well as four other faculty members who also regularly teach in the intermediate microeconomics sequence.

Students scoring below the median on the median on midterm 1 were randomly assigned to a "Videos Required" group where their midterm 1 score was down-weighted to 180 points and 40 points (all or nothing, so no partial credit) was put on watching 40 of the 47 remaining videos. Students were told that final letter grades would not be affected by being in the experiment, so the two key experimental outcomes are the scores on the second midterm and on the final exam. Following [cite], we randomized the students into the experiment by sorting the students by the first midterm score, creating ordered student pairs and then choosing one of students in the pair randomly to be in the experiment.

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 on link at end of the video. Right 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

In particular, we use an assessment (exam) to identify students most likely to be struggling with meeting the learning objectives of the course. We require a random sample of students who score below the median on the first midterm to partake in a required (for one's grade) study strategy consisting of watching at least 40 course-relevant supplementary

videos. The videos were freely available to all students in the class and the instructor committed to having the course grade distribution identical across treatment and controls. To keep the weights on the second midterm and final exam, our primary outcome measures, the same across the treated and control groups, the first midterm was down-weighted for the treated group and the weight was put on watching 40 videos. Nearly all students in the treated group watched the 40 videos and treatment led to a X percentage point increase in video views relative to the control group. We find the grading policy led to an

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 (G. W. Imbens and Rubin, 2015):

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

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 (J. D. Angrist and G. W. Imbens, 1995). 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 defiers 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

[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.

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 interested in deciding whether to require her students watch videos or simply make them available with no grade incentive.

Following the methods proposed by Robinson (1988) and more recently Wager and Athey (2018), 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 (1923; translated in 1990) repeated sampling approach, considering each pair (block) a completely randomized experiment and combining 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 $\hat{\tau}$ (Imai, 2008; G. W. Imbens and Rubin, 2015; Athey and G. W. Imbens, 2017) 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 (G. W. Imbens and Rubin, 2015, Athey and G. W. Imbens, 2017):

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

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 (Imai, 2008, G. W. Imbens and Rubin, 2015, Athey and G. W. Imbens, 2017):

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

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 (Thistlethwaite and Campbell, 1960; J. D. Angrist and Pischke, 2008; G. W. Imbens and Lemieux, 2008). 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, 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 Gelman and G. Imbens (2019).

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

[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.

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 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 (G. W. Imbens and Rubin, 2015). 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 3) and nonparametric methods a la the repeated sampling framework of Neyman (1923; translated in 1990) and Generalized Random Forests proposed by Wager and Athey (2018). We check that our parametric results are robust to model specification by estimating Equation 2 with and without $f(X_{it})$ as a vector of linear control variables and including $\hat{v}_{it}$ in Equation 3 with both linear and quadratic components. 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 receiving encouragement

5.3.2 Effect of watching videos

5.4 Spillover effects

Here we estimate spillover effects in courses taken concurrently during the term of the experiment. We also estimate spillover effects to the subsequent course in the intermediate microeconomics sequence.

5.4.1 Concurrent courses

5.4.2 Subsequent intermediate microeconomics course

6 Discussion

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, 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. the 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.

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

students more comfortable with electronic media

7 Conclusion

We examine the effectiveness of an educational innovation, a video handbook, where 220 (mostly) short instructional videos were organized into a book. Instructors may have concerns about making a new resource available if they believe students will substitute away from more productive study methods. Our experiment focused on students who demonstrated that they may benefit from an educational intervention by scoring below the median on an early assessment. We randomly assigned half of those scoring below the median on midterm 1 to a new grading scheme: the first midterm was down-weighted 4 percent and the points were fully awarded if the studnet watched 40 videos. Most treated students watched the 40 videos and the grade inducement led to a doubling of the number of videos watched relative to the control group. Both the second midterm and the final exam scores were significantly higher for the treated group with effect sizes of .18 of a standard deviation.

Our field experiment is unable to determine whether the videos uniquely helped the treated students or whether any intervention that induced additional study time would have been effective. However, we have suggestive evidence that the videos were uniquely effective: treated students in the next couse in the sequence, where there was no grade inducement

to watch videos, continued to watch the videos at a higher rate than the control students. This result is consistent with a model of student learning with imperfect information and not consistent with the neoclassical model or procrastination modle of student learning.

Generalizability of our results. Experiment was conducted in large classes with hisorically high failure rates. It is the first upper division class for many students and, for transfer students, also their first class in the university. As such there may be large information problems about how to successfully study for the class.

Future research: most importantly, to see of our results hold up in other educational settings e.g., different student, types of classes, instructors and universities. We would not allow bing-watching by requiring a certain number of videos each week as opposed to our experiment which simly required the student to watch 40 videos beore the end of the quarter.

Though randomized controlled trials estimating causal effects of teaching technologies are limited in the literature,

References

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

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

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

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

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

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

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

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

Cannon, E. (2011). "Comment on Chen and Lin 'Does Downloading Power-Point Slides Before the Lecture Lead to Better Student Achievement?" In: *International Review of Economics Education* 10.1, pp. 83–89.

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

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

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

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

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

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

Neyman, J. (1923; translated in 1990). "On the application of probability theory to agricultural experiments: essay on principles, Section 9". In: *Statistical Science* 5.4, pp. 465–472.

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

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

Robinson, P. M. (1988). “Root-N-consistent semiparametric regression”. In: *Econometrica: Journal of the Econometric Society*, pp. 931–954.

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

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

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

Wager, S. and S. Athey (2018). “Estimation and inference of heterogeneous treatment effects using random forests”. In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242.

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* do not include 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: This table reports coefficients on $Incentive_i$ from Equations 5 and TBD. Model (1) contains linear controls midterm 1 score and year; (2) is the difference in means and standard errors calculated using the repeated sampling framework of Neyman (1923); (3) and (4) use the post-double-selection (PDS) procedure of Belloni, Chernozhukov, and Hansen (2014b) to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A3. Models (2) and (4) contain 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 **??** and $Video_i$ from Equation TBD. Test scores are measured in standard deviation units. Model (1) contains linear controls midterm 1 score and year; (2) is the difference in means and standard errors calculated using the repeated sampling framework of Neyman (1923); (3) and (4) use the post-double-selection (PDS) procedure of Belloni, Chernozhukov, and Hansen (2014b) to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A3. Models (2) and (4) contain 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 and TBD. Model (1) contains linear controls midterm 1 score and year; (2) is the difference in means and standard errors calculated using the repeated sampling framework of Neyman (1923); (3) and (4) use the post-double-selection (PDS) procedure of Belloni, Chernozhukov, and Hansen (2014b) to select control variables then estimate treatment effects and standard errors. The control variables selected using PDS are listed in Table A3. Models (2) and (4) contain only students whose matched-pair did not attrite from the experiment. 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. *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.

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

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

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

Although our treatment assignment method provides a better chance of balance than does simple random sampling, by random chance and through non-random attrition, it is possible that the two treatment arms vary on *ex post* observable and unobservable covariates that are correlated with the outcomes of interest, thereby confounding our treatment effect estimates. The primary cause of attrition was withdrawing from the course, which reduced our 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 **??** includes a vector of control variables related linearly to the outcomes of interest. Although d_i, the treatment indicator is randomly assigned and in expectation d_i is orthogonal to all observed and unobserved pretreatment covariates, in small samples stochastic imbalances can occur, which if controlled for can reduce bias of the treatment effect estimator (Athey and G. W. Imbens, 2017). Even if perfect balance is achieved, controlling for orthogonal covariates can improve precision of the treatment effect estimator if the covariates can predict unexplained variance in the outcome.

By definition it is not possible to guarantee balance on unobserved covariates. As dis-

cussed in Appendix A.1, we mechanically balanced the treatment arms on first midterm score, one of the few observables at the time of treatment assignment, with our knowledge from previous cohorts' data that the first midterm score explains a significant amount of variance in final exam score. Hence, in our estimation strategies including controls, we always include the first midterm score and year, following the recommendations of Bruhn and McKenzie (2009) to control for all covariates used to seek balance when assigning treatment.

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

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

In Table 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 Belloni, Chernozhukov, and Hansen (2014a).

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* do not include 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	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	Final exam score, winter	None	Latinx
			Math quiz score
			Videos
	Hours unique videos, winter	Hours videos	Latinx
			Math quiz score
			Videos
	Midterm 1 score, Micro B	None	Latinx
			Math quiz score
	Midterm 2 score, Micro B	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. unique videos, winter	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, 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
	Term GPA, winter	Cumulative GPA	Cumulative GPA
			PSET visits
	Took 100B for a letter grade	None	Math quiz score
	% classes taken P/NP	None	No cum. GPA
			Transfer
	% classes taken for letter	None	No cum. GPA
			Transfer

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

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