EFFECTS OF FLIPPED CLASSROOM INSTRUCTION: EVIDENCE FROM A RANDOMIZED TRIAL

Elizabeth Setren
(corresponding author)
Department of Economics
Tufts University
Medford, MA 02155
Elizabeth.Setren@tufts.edu

Kyle Greenberg
Department of Social Sciences
United States Military Academy
West Point, NY 10996
kyle.greenberg@westpoint.edu

Oliver Moore
Office of the Deputy Chief of Staff for Programming
Army G-8, Headquarters
Department of the Army
Washington, DC 20310
oliver.c.moore1@gmail.com

Michael Yankovich
Department of Mathematical Sciences
United States Military Academy
West Point, NY 10996
michael.yankovich@westpoint.edu

Abstract

In a flipped classroom, an increasingly popular pedagogical model, students view a video lecture at home and work on exercises with the instructor during class time. Advocates of the flipped classroom claim the practice not only improves student achievement but also ameliorates the achievement gap. We conduct a randomized controlled trial at West Point and find the flipped classroom produced short-term gains in math and no effect in economics. The flipped model broadened the achievement gap: Effects are driven by white, male, and higher-achieving students. We find no long-term average effects on student learning but the widened achievement gap persists. Our findings demonstrate feasibility for the flipped classroom to induce short-term gains in student learning; however, the exacerbation of the achievement gap, the effect fade-out, and the null effects in economics, suggest that educators should exercise caution when considering the model.

https://doi.org/10.1162/edfp_a_00314

1. INTRODUCTION

Technology plays an increasing role in education and opens up a myriad of possibilities for educators to innovate on the traditional lecture format. One option, called the "flipped classroom," involves students learning the material by watching video lectures prior to class. This frees up class time for more in-depth discussion and application of the concepts through practice problems, group work, and increased interaction with the instructor (Brame 2013). Industry surveys estimate that over half of U.S. colleges use the flipped classroom and its popularity is growing (Schaffhuaser 2016; Schaffhauser and Kelly 2016). A range of education nonprofits, textbook publishers, and technology companies have capitalized on increasing interest in the format by providing videos and other educational tools.

Despite the proliferation of the flipped classroom, little well-identified evidence exists on its impact on student learning. Proponents claim this model not only boosts student achievement but also ameliorates the achievement gap through increased student–teacher interaction (Supiano 2018). The increased contact with students may make instructors more responsive to students' needs, which could be particularly beneficial for lower-achieving students who might otherwise not seek out assistance (Bergmann and Sams 2012; Goodwin and Miller 2013).

Opponents of the flipped classroom worry that it requires extensive work by instructors to create engaging videos and interactive classroom activities and hinges on students' engagement with the material outside of class (Lo and Hew 2017). The flipped classroom may take more time for students and will leave them without a foundational overview of the concepts if they do not watch the videos. Furthermore, if lower-income students have less reliable Internet access, it could exacerbate inequalities.

This study presents causal estimates of the flipped classroom's impact on student learning using a randomized controlled trial at West Point. We conducted the study during one unit in two mandatory core-curriculum courses, Introduction to Calculus and Principals of Economics, allowing us to explore the impact of the flipped classroom in two different subjects. Aspects of West Point and these two classes make it an ideal setting for this randomized controlled trial. Both courses require extensive problem-solving, a common attribute of flipped classroom courses (Berrett 2012). They also lend themselves well to consistent grading to provide an objective measurement of student learning. West Point standardizes the curriculum, teaching, and exams of these two high-enrollment courses across the eighty course sections. Additionally, the registrar randomly assigns students to course sections so that the sections have similar sets of students.

We randomly assigned course sections to flipped classrooms or standard lectures. To remove individual instructor effects on student learning outcomes, we assigned each instructor to at least one section in the control group and one section in the treatment group. The flipped classroom treatment consisted of a standardized video lecture that instructors told students to watch before class, and interactive problem-solving during class time. Students in the control group received a standardized lecture in class, with identical content to the video lecture. They were also given the same problems to work on as the treatment group, but to solve outside of class. To isolate the impact of the flipped format, we held the lecture material, instructor, and practice problems constant, and altered the format and time in which students engage with the practice problems

and lectures. If students complete all of the assigned work, we expect students to spend a similar amount of time preparing for class: completing the readings and watching the videos for the flipped classroom and completing the readings and working on the practice problems for the standard lectures.

We find that the flipped classroom does not reduce the achievement gap as proponents suggest. The flipped classroom produced a strong, positive short-term effect in math and null effect in economics. Students in the flipped math classrooms scored 0.3 standard deviations above the mean on the unit quiz relative to their peers in the standard classroom. However, female, black, and Hispanic students, and students with lower baseline academic performance (measured by their ACT scores), do not experience gains from the math flipped classroom; the math effects are driven by white, male, and higher-achieving students. The flipped classroom has a 69 percent larger white/nonwhite (black or Hispanic) achievement gap relative to the standard lecture and it exacerbates the difference by 23 percent between students who scored in the top and bottom ACT quartile. Although the average effects fade by the course final, the achievement gaps persist. We observe lower levels of student engagement (both in and out of the classroom) and instructor interest in the less-effective flipped classrooms. Our findings demonstrate that it is feasible for the flipped classroom to induce large gains in student learning in a short period of time, but that the effects vary by subject, student characteristics, and teacher motivation for the flipped classroom technique. The exacerbation of the achievement gap, the fade-out of effects, and the different effects by subject suggest that educators should exercise caution when considering the flipped classroom.

This study contributes to a growing literature on technology and education (see Escueta et al. 2017 for a summary). Within the education technology literature, work on the impact of online courses is particularly relevant. Randomized studies find small negative effects of fully online courses compared to in-person lectures (Figlio, Rush, and Yin 2013) and similar effects of standard lectures relative to reduced lecture time with access to online course materials or machine-guided instruction (Bowen et al. 2014; Joyce et al. 2015; Alpert, Couch, and Harmon 2016). Bettinger et al. (2017) find that negative effects of online course-taking are particularly large for students with lower prior grade point averages.

Descriptive flipped classroom research finds mixed results.[1] However, because of these studies' nonrandom designs, differences between student outcomes in the flipped and standard classrooms could be due to differences in course material, instructor quality, student preparation or characteristics, or other factors. Our study is most akin to Wozny, Balser, and Ives (2018), whose randomized controlled trial finds that the flipped classroom boosted students' scores in econometrics. One drawback is that the authors' seven course sections serve as the sample for the study. Because of the small number of sections, they randomize the flipped or standard teaching methods across lectures within the same sections, which could produce spillover effects.[2]

1. See Lage, Platt, and Treglia (2000); Bergmann and Sams (2009); McLaughlin et al. (2014); Schultz et al. (2014); Findlay-Thompson and Mombourquette (2014); Davies, Dean, and Ball (2013); Overmyer (2014); Swoboda and Feiler (2016).
2. For example, if students observe increased retention after flipped classrooms, they might ask more follow-up questions or focus more of their study efforts on material from standard lectures to compensate. Wozny, Balser,

Our study contributes to this relatively understudied topic by running a randomized controlled trial with a large number of class sections, students, and instructors, and by holding all aspects other than the flipped classroom constant, including the course materials, course content, and instructors. All twenty-nine instructors of the two courses participated in the experiment, allowing us to understand the effects of flipped classrooms for a range of instructor types, not just those most motivated to pedagogically innovate. We provide the first causal flipped classroom evidence from classroom level randomization and the first flipped classroom estimates for two separate subjects.

The next section provides background on West Point and the flipped classroom. Section 3 describes the design of the experiment and section 4 details the characteristics of the students, classrooms, and instructors in the study. Section 5 lays out the empirical framework. Section 6 presents the results, provides evidence to explain the differential effects in math and economics, and explores the equity implications. Section 7 offers concluding thoughts.

2. INSTITUTIONAL DETAILS

The United States Military Academy at West Point is a four-year undergraduate institution that prepares students to become military officers. In addition, West Point shares many characteristics with small, liberal arts schools. Each cohort has approximately 1,100 cadets and they complete a twenty-seven-course liberal arts curriculum. West Point caps class sizes at eighteen cadets per instructor and the average class has sixteen cadets per instructor. West Point also has some unique characteristics that distinguish it from other postsecondary institutions. Cadets must receive permission to miss class, there is a high level of discipline in the classroom, the cadet population is predominately male, and, considering the propensity for cadets to serve in combat after graduation, cadets are probably less risk-averse than the average U.S. college student.[3]

Admissions

West Point has a competitive admissions process. Applicants must receive a nomination from one of their Congressional representatives and must demonstrate physical fitness. As a result, West Point students are more athletic and geographically diverse than typical universities. The U.S. News and World Report ranks West Point as number eighteen in their list of National Liberal Arts Colleges.[4] West Point students have a mean SAT score of 627 in Reading and 645 in Math (out of a possible score of 800 in each), comparable to similarly ranked liberal arts colleges (West Point 2019).

Faculty

West Point has three types of faculty members: senior military faculty, civilian professors, and junior rotating military faculty. The senior military faculty serve permanently at West Point and most hold a doctorate degree. Civilian professors typically serve on

and Ives (2018) produce a variety of robustness checks that suggest the teaching methods of previous lessons do not impact test scores from current lessons.

3. See table 1 of Carter, Greenberg, and Walker (2017) for a comparison of West Point to similarly ranked liberal arts colleges and to all four-year postsecondary institutions.
4. See https://www.usnews.com/best-colleges/west-point-2893/overall-rankings.

Table 1. Class Size and Teaching Load

	Full Sample			Math			Economics		
	All (1)	Treatment (2)	Control (3)	All (4)	Treatment (5)	Control (6)	All (7)	Treatment (8)	Control (9)
Number of instructors	29	28	27	20	19	18	9	9	9
Number of sections	80	40	40	51	26	25	29	14	15
Sections taught by instructor	2.8 (0.9)	1.4 (0.5)	1.5 (0.5)	2.5 (0.8)	1.4 (0.5)	1.4 (0.5)	3.2 (1.0)	1.6 (0.5)	1.7 (0.5)
Number of students	1328	661	667	852	435	417	476	226	250
Average class size	16.6 (1.0)	16.5 (1.0)	16.7 (0.9)	16.7 (0.6)	16.7 (0.6)	16.7 (0.7)	16.4 (1.4)	16.1 (1.5)	16.7 (1.2)

Notes: This table describes the number of instructors, sections, and students in the treatment and control groups. Standard deviations are reported in parentheses.

the faculty for a prolonged period and have PhDs. The junior military faculty spend three years teaching and hold a master's degree in their area of instruction. The junior military faculty teach lower level electives and the more basic core curriculum courses, while civilian and senior military faculty teach more advanced courses.

Curriculum

The first two years at West Point are almost exclusively core curriculum courses. In order to accommodate the large enrollment, West Point offers many sections of these courses. Students are highly incentivized to do well because grades determine job placement after graduation. Because of the importance of course performance, West Point prioritizes standardizing courses and course grading. Course directors set the syllabus, lesson objectives, assignments, and exams so that they are consistent across all instructors teaching the course. New instructors receive training from the course director to further ensure standardization across class sections. The course standardization and randomization of students to course sections makes West Point an ideal place to study the flipped classroom.

3. EXPERIMENTAL DESIGN

The experiment took place in the 2016 Fall semester in two required core courses: Introduction to Calculus and Principles of Economics. We selected these courses because their quantitative nature lends themselves to interactive problem-solving. This method links well with the active teaching style the flipped classroom strategy utilizes.

Cadets take the math class in their first semester at West Point and take Principles, their first economics course, in their sophomore year. Students with stronger math backgrounds and students interested in majoring in Economics can take more advanced and in-depth versions of these classes. Because of the small number of advanced classes, we excluded these class sections from the study.

We also chose these two courses because they have a large number of courses sections, students, and faculty. All faculty who taught at least two sections of these courses participated in the study. Twenty-nine faculty members, 80 class sections, and 1,328 students participated in the experiment (see table 1). Math constituted a larger portion

of the sample with 20 instructors and 852 students, while 9 instructors and 476 students in economics participated. Forty class sections were randomly assigned to the flipped classroom treatment, 26 in math and 14 in economics. Another 40 sections were randomly assigned to the standard classroom control (25 in math and 15 in economics). Each instructor taught at least 1 flipped and 1 control classroom. On average, instructors taught 2.8 classes in the study. While random assignment happened at the classroom level, the registrar randomly assigned students to class sections, balancing on baseline academic ability.

We selected a three-lesson unit from both the math and economics courses to conduct the experiment. This discrete group of lessons enabled strict adherence to the experimental design. We think our experiment gives a lower bound on the impact of the flipped classroom for several reasons. First, the faculty involved are all new to the flipped classroom format and we would expect their effectiveness to improve with more experience. Second, our experiment occurs in the middle of the courses: in the lessons 14 through 16 of the math course and in lessons 22 through 24 for economics (after the microeconomics units and before the macroeconomics units). It may be challenging for faculty to switch and be disruptive for students. Lastly, we observe instructors in both the standard lecture and the flipped classroom, so instructors cannot focus their preparation time on one type of class.

We chose the vector and personal finance[5] units for the study because neither required nor built upon prior knowledge of the subject area. The math unit covered dot products and parametric equations. The personal finance unit in economics covered budgeting, present discounted value, and retirement and investment calculations. Some young people are exposed to the basics of personal finance through self-study, interactions with parents, and previous employment. In contrast, students have little to no prior knowledge of the vector math covered in the unit.

The course directors created and lectured in the videos. We chose to have one set of videos for each course (instead of having each instructor create their own video series) to ensure that each treated section had access to the same quality video with an experienced lecturer. We vetted the videos for consistent formatting between math and economics and piloted them in the summer prior to the experiment. We posted the videos to an internal West Point Web site that required students to log in to view the videos. This allowed us to monitor each student's video watching. The Web site allowed any student with the link to watch the video as many times as they wanted.[6]

Class Structure

Consistency across course sections is highly valued at West Point because grades factor into students' job placement after graduation. This standardization across sections strengthens our design and allows us to hold all components of the course constant, except for the "flipped" aspects. Figure 1 shows the timing of the different course

5. The primary purpose of the financial literacy unit is to prepare students for managing a large (upwards of $30,000) loan they receive in the middle of their junior year. The loan gives students upfront capital to purchase life necessities (including vehicles, uniforms, and furniture) to begin their Army career.
6. As a backup, the videos were also loaded to YouTube and the instructors were informed of this alternate capability to help them troubleshoot viewing problems during execution of the experiment. We cannot track YouTube video watching except for student reported surveys.

	Prior to Class	In Class	After Class
Flipped classroom (treatment)	Readings Video lectures	Problem set Q&A session	
Standard lecture (control)	Readings	Standard lecture Q&A session	Problem set

Figure 1. Class Format for Treatment and Control Groups

components. Both the flipped and standard classrooms were assigned identical readings to complete before lecture.

In addition, students in the flipped classroom were tasked with watching a 20-minute video lesson before each of the three lessons in the unit. Students in the treatment group were e-mailed instructions to watch videos before class and given reminders during lecture. They were informed that instructors would track whether they watched the videos. The e-mail emphasized that watching the video was important to their learning and they would lose participation points for not watching. Instructors were directed not to inform students about their participation in the experiment.

Upon arriving to class, instructors took attendance and made class announcements for both the treatment and control groups. Then, the treatment classrooms proceeded with a question-and-answer session with the instructor about the video for that lesson. The instructors were provided clear guidance to avoid lengthy lectures during this period, but to use the opportunity to clarify specific questions.

Then the flipped classrooms worked on a problem set with ten to fifteen practice problems that linked directly to the lesson objectives. The video covered the material in the problem sets and modeled how to solve similar problems. Instructors were given strict guidelines to not conduct a traditional lecture, but instead guide the students through the worksheet and answer any questions they might have. Implementation of the problem set varied from instructor to instructor. Some instructors would have the cadets complete one problem at a time and then have the class discuss the solution. Other instructors would let the cadets work the entire way through the worksheet uninterrupted and circulated the room to answer individual questions.

The control group experienced little change to the normal class procedures. After class administrative tasks, the instructors delivered a standardized lecture to the cadets that mirrored the content in the videos. The math course director provided a detailed script for instructors to follow in delivering the lesson. The economics control group class delivered the lecture following the same slides built into the economics flipped classroom video. During the course of the lecture, both math and economics instructors worked through quantitative problems on the board. Instructors could take questions throughout the lecture.

As the control group students departed at the end of the class period, instructors handed out practice problem worksheets that the flipped classroom students worked on during class. Instructors encouraged students to complete the worksheets prior to the next class but instructors did not collect or grade the worksheets. Therefore, the

control group students had the same opportunity to practice problems as the flipped classroom. Course directors circulated the classrooms during the experiment to check for proper implementation of both the flipped and control classrooms.

There were no strong incentives to watch the video or to complete the practice problems outside of class. Because of the military nature of West Point, students may be more likely to follow instructions than the typical college student. However, survey data find (see descriptive statistics below) that students spend less time preparing for class than recommended and do not all watch the videos in full. If compliance for class instructions is stronger among West Point students relative to other college settings, it affects both the treatment and control students similarly, since they are each instructed to do work outside of the classroom (e.g., watching videos or completing practice problems). West Point has a strict attendance policy, so we are able to control the amount of time that students spend in class with instructors. If the flipped classroom changes student attendance in less strict settings, then we will not see the effects of this in our study.

Outcomes

After the completion of the three-lesson block, students in both the flipped and standard classrooms took an in-class quiz that covered the material of the experiment's lessons. The in-class quiz accounted for 3 percent of the math course and 3.5 percent of the economics course grades. Both the treatment and control groups received identical quizzes.

To test for fade-out of knowledge or whether students improved their knowledge of the material before the final exam, we analyze performance on the unit-specific questions on the final exams and the overall final exam grade. For math, the exam was administered in May 2017 and for economics the exam took place in December 2016. For both classes, the exam accounted for 25 percent of the course grade.

4. DESCRIPTIVE STATISTICS AND IMPLEMENTATION DETAILS

Student Characteristics and Covariate Balance

The random assignment of students and instructors effectively balanced the demographic composition and baseline academic ability of the standard and flipped classrooms. Columns 1 and 2 of table 2 show that students assigned to the control group, with classes taught in a traditional lecture format, and students assigned to the treatment group, with classes taught in a flipped classroom format, have similar average baseline characteristics. Women constitute about 21 percent of students in both groups. Roughly 60 percent of students in the experiment are white, African American students account for approximately 17 percent of the sample, and Hispanic students account for just over 10 percent. Nearly one in six students has prior military service as enlisted soldiers and the average composite ACT score for both groups exceeds 28.[7] The flipped and standard classroom students also scored similarly on West Point's College Entrance Exam Rank (CEER), an admissions tool that factors in high school class ranking, SAT or ACT scores, physical fitness, West Point faculty evaluations, and extracurricular activities.

7. For students who did not take the ACT, we map SAT scores to comparable ACT scores. The highest possible ACT score is 36, with 21 being the average score.

Table 2. Student Characteristics and Covariate Balance

	Mean		Difference between Treatment and Control		
Baseline Characteristic	**Control (1)**	**Treatment (2)**	**Full Sample (3)**	**Math Classes (4)**	**Economics Classes (5)**
Female	0.205	0.212	−0.002	−0.002	−0.003
	(0.404)	(0.409)	(0.014)	(0.019)	(0.019)
White	0.600	0.611	0.002	−0.008	0.018
	(0.490)	(0.488)	(0.024)	(0.036)	(0.025)
Black	0.175	0.169	−0.006	0.000	−0.016
	(0.381)	(0.375)	(0.022)	(0.028)	(0.034)
Hispanic	0.111	0.100	−0.003	−0.007	0.002
	(0.314)	(0.300)	(0.012)	(0.017)	(0.016)
Age	19.461	19.390	−0.014	−0.043	0.033
	(1.155)	(1.231)	(0.044)	(0.055)	(0.074)
Prior military service	0.165	0.159	0.006	0.005	0.008
	(0.371)	(0.366)	(0.016)	(0.020)	(0.026)
Composite ACT	28.085	28.324	0.127	0.150	0.091
	(3.448)	(3.350)	(0.141)	(0.187)	(0.214)
College Entrance Exam Rank score	6.017	6.053	0.017	0.031	−0.004
	(0.682)	(0.645)	(0.030)	(0.044)	(0.034)
p-value (joint χ^2 test)			0.966	0.890	0.951
Observations	667	661	1,328	852	476

Notes: This table reports descriptive statistics of students in the experiment. Column 1 reports mean characteristics of the control group (students in classrooms with the standard lecture format) and column 2 reports means for students in the treatment group (flipped classrooms). Standard deviations are reported in parentheses. Columns 3, 4, and 5 report coefficient estimates from a regression of the baseline characteristics on an indicator variable that equals one if a student is assigned to a flipped classroom. The regressions used to construct estimates in columns 3, 4, and 5 include (course) x (instructor) and (course) x (hour) fixed effects. Standard errors, clustered on classrooms (each instructor-hour combination), are reported in parentheses. The reported *p*-values come from a joint test of the null hypothesis that all coefficients are equal to zero.

Column 3 of table 2 reports the regression-adjusted differences between students assigned to treatment classrooms and students assigned to control group classrooms. We construct these differences from regressions that include instructor fixed effects and class schedule block fixed effects. The differences between treatment and control classroom characteristics are all small and statistically insignificant, suggesting assignment to treatment or control groups was as good as random. A test of the joint-hypothesis that all differences in baseline characteristics equal 0 yields a *p*-value of 0.966 (bottom row of table 2), further suggesting that the randomization was effective. These similarities would indicate that any difference in the performance of the flipped and standard classrooms can be attributed to the treatment of the flipped classroom.

Columns 4 and 5 of table 2 report the same covariate balance checks after restricting the sample to math and economics classes, respectively. As with the estimates reported in column 3, there are no noticeable differences between the treatment and control groups for either class type.

Table A.1, which is available in a separate online appendix that can be accessed on *Education Finance and Policy*'s Web site at https://doi.org/10.1162/edfp_a_00314, explores attrition rates for taking the unit quiz and final exam, the key outcomes of interest. Approximately 97 percent of the sample takes the quiz and the treatment and control groups have similar quiz-taking rates. The final exam attrition rate is twice as large as the quiz attrition. While attrition is not statistically significantly different in the

economics treatment and control groups, the treatment group in math has a marginally significantly higher attrition rate relative to the control group. We estimate Lee (2009) treatment-effect bounds and find that attrition does not bias our findings.

Classroom Characteristics

Instructors and Class Size

Consistent with West Point's small class sizes, the average class in the study has 16.6 students (see table 1). Flipped and standard math classes both had student–teacher ratios of 16.7. Economics classrooms had similar class sizes of 16.1 for flipped and 16.7 for standard lectures.

A total of twenty-nine instructors participated in the experiment, with twenty from math and nine from economics. The majority of instructors were military officers with three or fewer years of teaching experience. The rest of the instructors were senior military officers (three in math, one in economics) and civilian faculty (also three in math, one in economics). These faculty each had at least five years of teaching experience.[8]

Before teaching any classes, all U.S. Military Academy instructors must pass a rigorous six-week training course where they learn best teaching practices, observe experienced instructors teach summer classes, and present practice lectures to a panel of senior military and civilian faculty members. The experiment took place at roughly the mid-point of the fall semester, so all instructors had a minimum of two months of teaching experience, plus the six-week training course, before the experiment began.

Class Time Allocation

We administered an instructor survey at the conclusion of the course to gather descriptive information about how the flipped and standard classrooms functioned in practice. Over 86 percent of instructors completed the survey. Table A.1 in the online appendix shows no differential attrition across whether treatment or control students had an instructor who responded to the survey. We also administered student surveys at the end of the class and collected video-watching data that we discuss below.

Table 3 displays the results of the survey and confirms that faculty carried out the flipped classroom and standard classroom models as instructed. Consistent with classroom observations, faculty reported using the class time as instructed: on average, instructors reported spending 85 percent of the class time lecturing in the standard classroom, relative to 8 percent of the time in the flipped classroom.

The flipped classroom also involved more in-class group and independent work than the standard lecture: Instructors said that students worked in groups 76 percent of the time in flipped classrooms and 5 percent of the time in standard classrooms. Math instructors reported that students worked in groups for only 1.5 percent of class time

8. After randomizing classrooms to treatment and control groups, but before the start of the school year, the Math Department added an additional instructor. The Math Department assigned the new instructor to teach three sections that had been assigned to three separate math instructors in our experiment. These sections included two treatment and one control. We confirmed with the Math Department that our experiment had no bearing on the decision to bring in an additional instructor, including the decision about which sections the new instructor would teach. As a result of this swap, the three instructors who lost a section taught only one section and therefore had no variation in treatment and control sections during the experiment. We did not adjust the random assignment after this section swap occurred and we did not permit instructors to self-select their sections into treatment or control groups.

Table 3. Classroom Characteristics

	Full Sample		Math		Economics	
	Flipped (1)	Standard (2)	Flipped (3)	Standard (4)	Flipped (5)	Standard (6)
Panel A: Class Time Allocation (in Percent)						
Lecturing	8.0	85.0	8.8	88.2	6.3	78.1
	(11.9)	(16.1)	(12.3)	(15.6)	(11.6)	(16.0)
Students working in group	76.0	5.0	73.5	1.5	81.3	12.5
	(30.2)	(10.2)	(32.4)	(6.1)	(25.9)	(13.4)
Students working alone	30.0	6.0	36.8	4.4	15.6	9.4
	(38.9)	(13.1)	(42.5)	(13.2)	(26.5)	(12.9)
Answering individual questions	42.0	6.0	44.1	5.9	37.5	6.3
	(26.7)	(13.1)	(30.0)	(14.1)	(18.9)	(11.6)
Answering questions for class	36.0	53.0	44.1	50.0	18.8	59.4
	(22.9)	(34.9)	(18.8)	(33.1)	(22.2)	(39.9)
Review old material	5.0	6.0	7.4	8.8	0.0	0.0
	(10.2)	(10.9)	(11.7)	(12.3)	(0.0)	(0.0)
Other/administration	10.0	9.0	8.8	7.4	12.5	12.5
	(12.5)	(12.2)	(12.3)	(11.7)	(13.4)	(13.4)
Panel B: Percent of Student Who Do the Following at Least Once during Class Time						
Work in group	76.0	11.0	77.9	11.8	71.9	9.4
	(25.5)	(24.0)	(24.8)	(26.7)	(28.1)	(18.6)
Work alone	23.0	18.0	25.0	11.8	18.8	31.3
	(27.9)	(30.2)	(28.0)	(26.7)	(29.1)	(34.7)
Asking questions	56.0	45.0	57.4	41.2	53.1	53.1
	(19.5)	(20.4)	(21.2)	(19.6)	(16.0)	(20.9)
Answering questions	48.0	45.0	44.1	39.7	56.3	56.3
	(24.9)	(19.1)	(24.3)	(19.9)	(25.9)	(11.6)
Taking notes	47.0	75.0	35.3	73.5	71.9	78.1
	(28.2)	(21.7)	(21.8)	(22.5)	(24.8)	(20.9)
Paying attention	74.0	65.0	77.9	61.8	65.6	71.9
	(18.4)	(17.7)	(15.0)	(17.9)	(22.9)	(16.0)
N	25.0	25.0	17.0	17.0	8.0	8.0

Notes: This table describes the flipped and standard classrooms using data from a post-study instructor survey. Panel A displays the average percent of time instructors reported spending on different activities. Items did not need to add up to 100 percent because some activities could happen simultaneously (e.g., answering individual questions and having students work in a group). Panel B shows the percent of time the instructor observed students doing an activity in class. Standard deviations are reported in parentheses.

in the standard lecture, while group work was more common in the economics standard lectures: constituting 12.5 percent of time. Students also worked alone on practice problems more frequently in the flipped relative to the standard classrooms, with the starkest difference in math (36.8 percent versus 4.4 percent) and a smaller difference in economics (15.6 percent versus 9.4 percent).

Math faculty spent more time answering questions for the whole class in the standard classroom than the flipped classrooms, but more time circulating around the classroom to answer individual student questions in the flipped classroom. Economics instructors reported a similar pattern but with a larger difference between the amount of time answering questions in front of the class.

Instructors spent small and similar amounts of time reviewing old material and other tasks in the flipped and standard lectures.

Student Behaviors

We also asked instructors to report behaviors of students during class time (see panel B in table 3). Some student behaviors reflect clear differences between flipped and standard classrooms: Faculty reported 76 percent of students working in groups in a typical flipped classroom, compared with 11 percent of students in the standard lecture.

Other survey responses suggest different implementation in the math sections compared to the economics sections. In math, working alone was more than twice as common in the flipped versus the standard classroom. However, in economics, the relationship was flipped. Additionally, while asking and answering questions were more common in the math flipped classrooms relative to the standard lectures (consistent with the flipped classroom model), students asked and answered questions at similar rates in the flipped and standard economics classrooms. Also, math instructors reported higher note-taking rates in the standard lecture than the flipped classroom, while economics instructors reported only a slightly higher rate in the standard classrooms. Together, these survey results suggest that the math classes implemented the flipped classroom model more fully than economics.

Lastly, instructors perceived that more students paid attention in math in the flipped classroom (77.9 percent versus 61.8 percent), while attention in economics was greater in the standard lecture (71.9 percent versus 65.6 percent in the flipped classroom).[9] This shows that math instructors found the flipped classroom to be more engaging, while economics faculty found the standard lecture preferable for students.

Time and Activities Outside of Class

Video Watching

Using the unique logins to the West Point Web site that hosted the videos, we tracked student streaming of the class videos. Panel A of table 4 shows that almost 80 percent of math and 73 percent of economics students in the flipped classrooms watched at least some of a video. On average, students watched roughly two out of three of the videos. As expected, the control group rarely accessed the video: Only 2 percent of math and 3 percent of economics standard lecture students ever logged in. To log in, they needed to get the link from someone in a flipped classroom.

We track the proportion of the video data that streamed to students' computers and find that on average students watch roughly 50 percent of the video content in math and economics.[10] In a survey at the end of the course, we asked students about how they watched the videos. The responses in panel B of table 4 show that over three fourths of the math flipped classroom students report repeating sections when they watch the video compared with only 37 percent of economics students. Students reported low rates of multitasking while watching the videos in math (6 percent), but higher rates (31 percent) in economics.

9. The columns in table 3, panel A, do not need to add up to 100 percent because multiple activities can occur at the same time. For example, instructors can answer individual questions while students work in groups.
10. We measure percent of video watched by dividing the number of bytes downloaded by the total number of bytes for each lesson. If students streamed a lesson's video more than once, we take the session where they watched the largest proportion of the video. This conservatively measures student video watching by undercounting the proportion of the video watched if students do not restart from the beginning.

Table 4. Video Watching

	Math		Economics	
	Flipped (1)	Standard (2)	Flipped (3)	Standard (4)
Panel A: Video Watching				
Ever watched video	0.798	0.022	0.730	0.036
	(0.402)	(0.145)	(0.445)	(0.187)
Proportion of videos students watched	0.559	0.008	0.544	0.015
	(0.370)	(0.056)	(0.398)	(0.080)
Proportion of video length watched	0.453	0.006	0.543	0.015
	(0.304)	(0.050)	(0.399)	(0.080)
N	435	417	226	250
Panel B: Video Watching Style				
Repeated sections	0.770	–	0.371	–
	(0.421)	–	(0.485)	–
Multitasked while watching	0.066	–	0.308	–
	(0.249)	–	(0.463)	–
N	318	–	143	–

Notes: This table reports average video watching behaviors of students in flipped and standard classrooms. Panel A data come from log-in and streaming data to the Web site that hosted the video lectures. Panel B data come from a post-study student survey. Standard deviations are reported in parentheses.

Class Preparation

The random assignment of classrooms to flipped or standard lecture format ensures that we estimate the causal impact of the flipped classroom. However, the flipped classroom format could increase the time spent on the class or time spent on practice problems. If these changes influence student outcomes, then it is possible the effects are driven by changes in the amount of time students spend on the class and not from the flipped classroom format.

We find no evidence that students in the treatment group spend more time preparing for the class than students in the control group. Panel A of table 5 shows that students in the math standard lecture spend approximately 12 minutes more preparing for lecture compared with the math flipped classroom students (1.35 hours compared with 1.15 hours). This 12-minute difference is statistically significant, though a relatively small amount of time. Because the flipped classroom students spend slightly less time preparing for lecture, the treatment effects cannot be driven by students spending more time preparing. Economics students report spending less time preparing for lecture overall (0.87 and 0.73 hours for flipped and standard lectures, respectively).[11] This suggests that students spend similar amounts of time preparing for lecture by watching videos or doing the assigned readings.

11. While self-reported data are not ideal, students knew their responses would not affect their grade or be viewed by their instructors. The average student responded that they spent 30 to 50 percent less time preparing for class than West Point suggests (2 hours per class, see panel A of table 5). This signals students' willingness to give non-favorable answers. Furthermore, the flipped and standard classrooms did not have different incentives to over- or underreport their course preparation, so any measurement error should be consistent across the two groups.

Table 5. Student Preparation Outside of Class

	Math			Economics		
	Mean		*P*-value of Flipped = Standard	Mean		*P*-value of Flipped = Standard
	Flipped (1)	Standard (2)	(3)	Flipped (4)	Standard (5)	(6)
Panel A: Time Spent on Course						
Hours spent preparing for lecture (watching videos or reading combined)	1.148 (0.485)	1.346 (0.481)	0.000	0.867 (0.384)	0.729 (0.476)	0.007
N	318	325		143	144	
Hours spent on practice problems outside of class	0.910 (0.545)	0.909 (0.631)	0.980	0.661 (0.516)	0.865 (0.598)	0.002
N	318	325		143	144	
Panel B: Course Reading						
Complete all of the assigned reading for every class	0.233 (0.423)	0.252 (0.435)	0.563	0.105 (0.307)	0.125 (0.332)	0.595
Complete part or all of the assigned reading for every class	0.830 (0.376)	0.855 (0.352)	0.381	0.664 (0.474)	0.590 (0.493)	0.196
Never read for class	0.082 (0.274)	0.037 (0.189)	0.016	0.119 (0.325)	0.229 (0.422)	0.014
N	318	325		143	144	
Panel C: Rating for Helpfulness of Learning Tool (0 = Not helpful, 1 = Very helpful)						
Reading (for those who read)	0.410 (0.230)	0.429 (0.229)	0.316	0.388 (0.233)	0.350 (0.193)	0.171
N	296	304		123	115	
Video (just treatment)	0.615 (0.262)			0.353 (0.231)		
N	301			129		
Reading (just treatment)	0.410 (0.230)			0.388 (0.233)		
N	296			123		
Readings (just control)		0.225 (0.325)			0.036 (0.126)	
N		417			250	

Notes: This table reports end-of-course survey responses from students in flipped and standard classrooms. Columns 1, 2, 4, and 5 show the mean responses and columns 3 and 6 show the *p*-value of the test of whether the flipped and standard classroom means are equal. Standard deviations are reported in parentheses.

Math flipped classroom and standard classroom students report spending similar amounts of time outside of class on practice problems (see panel A of table 5). Because the treatment group worked on the problems in class, this signals they have overall more time to work on practice problems relative to the control group, which could contribute to the higher average scores on the quiz. In contrast, economics students in standard classrooms spend significantly more time on practice problems outside of class relative to their flipped classroom peers.

Flipped classroom students are also similarly likely to complete some or all the readings for class (see panel B of table 5). Approximately 24 percent of math students report completing the readings for every class and over 83 percent complete some of the readings for each class. Reading completion is less common in economics: Over 10 percent

of students report completing all of the readings and over 59 percent complete some of the readings.

5. EMPIRICAL FRAMEWORK

We estimate the equation below to compare outcomes between students in the flipped (treatment) classrooms to students in the standard lecture (control) classrooms:

$$Y_{ijh} = \alpha + \beta T_{jh} + \kappa_j + \lambda_h + \gamma' X_i + \epsilon_{ijh}, \quad (1)$$

where Y_{ijh} is the exam score of student i with instructor j during class-hour (schedule block) h. T_{jh} is a binary variable that equals 1 for students in the flipped classrooms and 0 for students in the standard lecture classrooms. X_i is a vector of individual controls, including race, gender, age, prior military service, and composite ACT and CEER scores. Equation 1 also includes instructor fixed effects, κ_j, to control for instructor ability that is constant across the type of class taught. Class-hour fixed effects, λ_h, capture unobserved mean differences in academic performance across class hours. Due to the random assignment of flipped-classroom status, estimates of β capture the causal effect of learning in a flipped classroom environment. We cluster standard errors at the classroom level.

6. RESULTS

We find the flipped classroom has a short-term positive effect that fades by the time students take the final exam. Table 6 reports the estimates of equation 1 for the quiz at the end of the unit, the questions specific to that unit in the final exam, and the overall final exam score. All test scores are standardized to have a mean of 0 and standard deviation of 1. In our study, flipped classrooms increase quiz scores by 0.165 standard deviations relative to standard lecture classrooms (column 1 of table 6). The results remain similar after controlling for student-level baseline covariates in column 2.[12]

Subsequent columns of table 6 reveal that only math classes experience the short-term positive effect of flipped classrooms. Columns 3 and 4 indicate that the flipped classroom environment improved test scores of students in math classes by roughly 0.3 standard deviations. We see positive effects for both rote questions that involve memorization and more advanced questions that require problem-solving.

In contrast to the math results, the average test scores among economics students in flipped classrooms were roughly 0.07 standard deviations lower than the average test scores of economics students in standard classrooms, though this estimate is only marginally significant and indistinguishable from zero when we correct for the number of clusters.[13,14]

12. Results with and without instructor fixed effects are also similar.
13. With only twenty-nine economics sections, our clustered standard errors for economics classes are potentially biased downwards. To investigate this further, table A.2 in the online appendix reports conventional standard errors, robust standard errors, and standard errors constructed from section-level unit-quiz means. The results of this investigation suggest that the marginally significant negative estimate among economics classrooms is indistinguishable from zero when we correct for the small number of clusters (column 4). Our positive estimates for math classrooms, however, are statistically significant regardless of how we estimate standard errors.
14. Data were not available to analyze rote versus problem-solving effects for economics.

Table 6. Effect of Treatment on Academic Outcomes

	Full Sample		Math		Economics	
	(1)	(2)	(3)	(4)	(5)	(6)
Unit quiz	0.165***	0.148***	0.304***	0.282***	−0.055	−0.071*
	(0.048)	(0.044)	(0.062)	(0.058)	(0.043)	(0.037)
N	1,281	1,281	818	818	463	463
Final: Unit-specific questions	0.035	0.025	0.057	0.039	0.000	−0.004
	(0.057)	(0.053)	(0.071)	(0.061)	(0.094)	(0.096)
N	1,254	1,254	801	801	453	453
Final: Overall score	−0.020	−0.035	0.027	0.001	−0.093*	−0.100**
	(0.042)	(0.033)	(0.060)	(0.048)	(0.050)	(0.039)
N	1,262	1,262	801	801	461	461
Rote questions on quiz			0.250***	0.238***		
			(0.058)	(0.060)		
N			848	848		
Problem-solving questions on quiz			0.223***	0.205***		
			(0.076)	(0.068)		
N			848	848		
Instructor and time block controls	X	X	X	X	X	X
Demographic controls		X		X		X
Clusters (classroom)	80	80	51	51	29	29

Notes: This table reports estimates from regressions of exam scores on an indicator for being assigned to a flipped classroom. All scores are standardized to have a mean of 0 and a standard deviation of 1 for each subject. Baseline controls include instructor fixed effects and class hour (i.e., time block) fixed effects. Demographic controls include indicators for female, white, black, Hispanic, and for having prior military service, plus linear terms for age, ACT score, and West Point's College Entrance Exam Rank score. Standard errors, clustered on classroom, are reported in parentheses.

***Significant at the 1% level; **significant at the 5% level; *significant at the 10% level.

To investigate whether the positive impact of flipped classrooms on math classes persists, we test the effect of the flipped classroom on the final exam. We estimate the effect for both the overall score and for the questions specific to the experimental unit.[15] We ensured that the difficulty and content of the final exam questions for the experimental unit material was comparable to that of the experimental unit quiz. Students in the math-flipped classrooms perform similarly to those in the standard lecture for both the questions on the final—specific to the experimental unit—and the final exam overall. The point estimates are positive but small, ranging from 0.039 to 0.057 standard deviations, and are indistinguishable from zero. Although we cannot rule out positive effects on the order of one fifth of a standard deviation, these estimates do suggest the positive impact of the flipped-classroom environment on math comprehension likely faded with time. The overall final exam scores between students in flipped classrooms and students in standard lecture classrooms did not vary, which is unsurprising because the flipped classroom experiment did not extend beyond the specific set of lectures described above.

In online table A.3, we investigate the faded-out effects by reporting the mean test scores of treatment and control students for the quiz, and the questions specific to the experimental unit on the final. We find that flipped classroom students increase their knowledge of the experimental unit's content following the quiz: They score

15. The units that followed the experimental unit did not build upon knowledge from the experimental unit. Instructors did not spend time on the experimental unit's material after the quiz.

22 percent higher on the unit-specific final exam questions relative to the quiz (see column 1 of table A.3). Because the quiz and final questions cover the same content with highly comparable questions, this denotes that average student knowledge of the subject grew over time. The control group also increased their average score from the quiz to the final and caught up to the flipped classroom students: Both groups scored similarly, on average, on the final questions specific to the experimental unit. This means the null effects on the final exam stem from the control group's catching up and not that the flipped classroom students' knowledge faded.

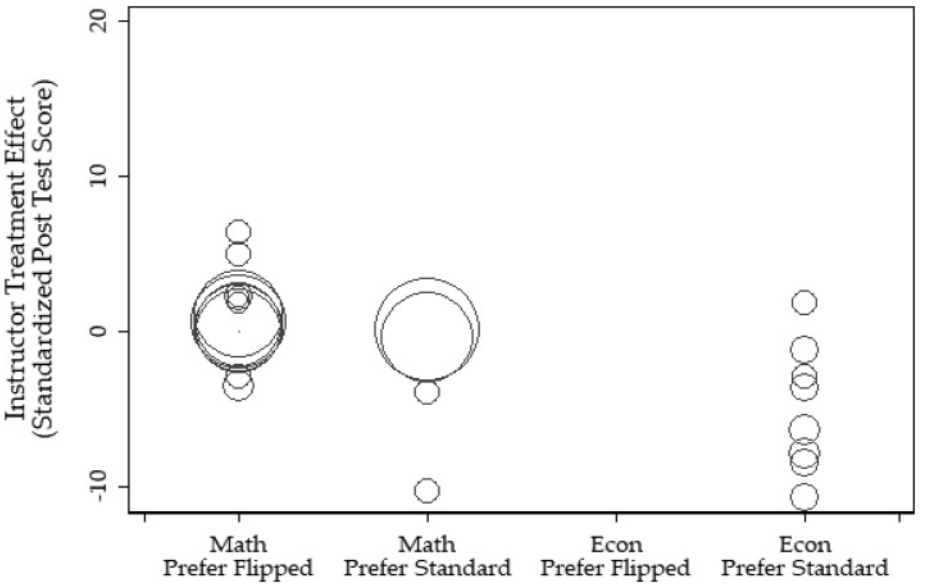
Table 6 reports no long-term test effects for students in economics classes for the unit-specific final exam questions, suggesting that if the flipped classroom had any initially deleterious effects on economics students, they likely faded over time. Puzzlingly, students in the flipped classroom scored about 0.1 standard deviations lower on the final exam overall relative to students in the control group. However, after we account for the small number of clusters among economics students, this effect is not statistically significant (see the *p*-values in online table A.4). As a result, we think the small difference in point estimates is due to noise and the small number of clusters.

Differential Effects Across Subjects

What explains the short-term positive effects of the flipped classroom in math and the null effects in economics? We chose these subjects because both are quantitative and involve problem-solving, common features of flipped classroom subjects. However, while both courses have problem-solving aspects, the math lessons included a higher proportion of problem-solving questions than economics, which included relatively more rote memorization. Because of this, perhaps the material in the math lecture lent itself better to the flipped classroom.

In addition, prior to the study, the math department employed a more interactive teaching style relative to economics. The math faculty could have been more confident in implementing flipped classrooms due to the teaching styles of the department. Additionally, instructor preferences could impact their effort and teaching effectiveness. Figure 2 supports this hypothesis: It plots individual instructor effects on student quiz scores by whether instructors prefer flipped or standard lectures. Math instructors who preferred flipped classrooms generate similar or larger learning gains for their flipped classrooms relative to their standard lectures. In contrast, math instructors who prefer standard classrooms are either equally effective in flipped or standard classrooms, or more effective in standard lectures. Having a math instructor who prefers teaching the flipped classroom is linked to an approximate doubling of the flipped classroom treatment effect, although the point estimate is noisy (see table 7). All economics faculty prefer standard lectures and all but one instructor has a stronger impact on their standard lecture students relative to their flipped classroom students. This suggests that instructor preferences could play a role in the impact of the flipped classroom.

There are two issues with these survey data. Because we surveyed instructors at the conclusion of the experiment, their preferences could have been influenced by how effective they thought they were in the flipped versus standard classroom. Second, it is possible that instructors exhibit the Hawthorne effect: Instructors who do not like the flipped classroom may choose to reduce their effort, which could then explain the



Notes: This figure plots the instructor-specific quiz effects by whether they preferred the flipped or standard classroom in a post-study survey. Estimates come from regressions of exam scores on an indicator for being assigned to a flipped classroom that include baseline demographic controls and class hour fixed effects. Larger circles reflect more precise estimates: circle size is weighted by the inverse variance of the effects.

Figure 2. Instructor Treatment Effects by Subject and Preference

null or negative outcomes for instructors who prefer the standard lecture. Together, these complications suggest the need for instructors to actively want to implement the flipped classroom, along with training and support instructors.

We find evidence of lower levels of engagement among economics flipped classroom students, both in and out of the classroom. This could contribute to the differential effects between the subjects. Economics students rated the video less useful relative to math students in our end of semester survey (see panel C of table 5). On average, the math treatment group rated the video 50 percent more helpful than the reading. In contrast, economics treatment group students rated the videos and readings similarly useful. Also, the economics students found the videos less helpful than the math students: Economics students rated the videos an average of 0.35 out of 1 (with 0 denoting not helpful and 1 denoting very helpful), compared with the math students' rating of 0.62 out of 1. Economics flipped classroom students also report lower rates than math students of repeating sections of the video and higher rates of multitasking while watching the video. These findings suggest higher levels of engagement and interest in the videos (a key component of the flipped classroom model) among math students compared with economics students.

Instructor survey responses also reveal higher levels of student engagement for the flipped classroom relative to the standard classroom for math. Table 3 shows that math instructors observe higher rates of students paying attention, asking questions, working in groups, and working independently in their flipped classrooms relative to their standard lectures. In contrast, economics faculty report lower rates of paying attention

Table 7. Effect of Math Instructor Preferences on Unit Quiz Score

	(1)	(2)	(3)
In flipped classroom	0.304***	0.151	0.152*
	(0.062)	(0.097)	(0.092)
Instructor prefers teaching flipped classroom	−0.006	−0.112	−0.135
	(0.182)	(0.228)	(0.154)
Instructor prefers teaching flipped classroom × In flipped classroom		0.198*	0.168
		(0.116)	(0.108)
Instructor and time block controls	X	X	X
Demographic controls			X
R^2	0.083	0.085	0.206
N (Number of Students)	785	785	785
Clusters (classroom)	49	49	49

Notes: This table reports estimates from a regression of unit-quiz exam scores on an indicator for being assigned to a flipped classroom for students in the math section. Columns 2 and 3 report estimates from regressions that interact the flipped classroom indicator with an indicator for whether the instructor preferred teaching a flipped classroom in a post-study instructor survey. We do not report analogous estimates for economics classrooms because all economics instructors preferred the standard lecture format over the flipped classroom format. All scores have been standardized to have a mean of 0 and a standard deviation of 1 and the controls are the same as those described in table 5. Standard errors, clustered on classroom, are reported in parentheses.

***Significant at the 1% level; *significant at the 10% level.

and similar rates of asking questions in the flipped classrooms relative to their standard classrooms—signaling lower levels of engagement. Lastly, math instructors report increased student–teacher and student–peer interaction in the flipped classroom through more question-asking and group work (see table 3). In contrast, economics instructors only report increased group work in their flipped classrooms but similar amounts of student questions. In summary, the math flipped classrooms had more student engagement and student-teacher interactions than the math standard lectures, but economics treatment and control classrooms had fewer differences for these classroom characteristics.

Lastly, students in economics standard classrooms spend significantly more time on practice problems outside of class relative to the flipped classroom economics students. In contrast, math students report spending similar amounts of time on practice problems outside of class in the flipped and standard classrooms. This additional practice time for the control group in economics may play a role in the similar test scores of the treated and control groups in economics.

Equity Implications: Subgroup Effects

Subgroup analysis suggests the short-term positive impact of flipped classrooms on math performance is driven by men, white students, and higher-achieving students. This can be seen in column 4 of table 8, which reports the flipped classroom effects on quiz scores by student characteristics. The flipped classroom has a large positive effect for men's quiz scores (0.32 standard deviations gains) and a smaller and statistically insignificant effect for women. White students experience gains of 0.385 standard deviations higher on short-term learning, while black and Hispanic students have statistically insignificant effects close to zero. The difference in treatment effects between white students and black or Hispanic students is statistically significant (*p*-value

Table 8. Subgroup Analysis

	Full Sample		Math		Economics	
	(1)	(2)	(3)	(4)	(5)	(6)
Female	−0.015	0.025	0.100	0.084	−0.191	−0.146
	(0.103)	(0.094)	(0.131)	(0.112)	(0.156)	(0.153)
N	269	269	173	173	96	96
Male	0.207***	0.175***	0.358***	0.320***	−0.035	−0.057*
	(0.058)	(0.052)	(0.079)	(0.072)	(0.043)	(0.032)
	1012	1012	645	645	367	367
p-value of female = male	*0.077*	*0.167*	*0.148*	*0.089*	*0.331*	*0.544*
White	0.223***	0.208***	0.405***	0.385***	−0.047	−0.068
	(0.066)	(0.060)	(0.088)	(0.077)	(0.071)	(0.075)
N	784	784	494	494	290	290
Black or Hispanic	0.054	0.034	0.129	0.065	−0.091	−0.012
	(0.077)	(0.079)	(0.087)	(0.093)	(0.147)	(0.148)
N	346	346	229	229	117	117
p-value of white = non-white	*0.070*	*0.045*	*0.012*	*0.005*	*0.738*	*0.785*
ACT bottom quartile	0.017	−0.006	0.091	0.068	−0.140	−0.165
	(0.075)	(0.086)	(0.089)	(0.098)	(0.137)	(0.165)
N	274	274	176	176	98	98
ACT 2nd quartile	0.259**	0.293**	0.407***	0.435***	0.022	0.116
	(0.112)	(0.117)	(0.155)	(0.169)	(0.122)	(0.125)
N	251	251	167	167	84	84
ACT 3rd quartile	0.207***	0.208***	0.296***	0.280***	0.059	0.061
	(0.055)	(0.059)	(0.073)	(0.077)	(0.068)	(0.090)
N	405	405	269	269	136	136
ACT top quartile	0.123	0.160*	0.403***	0.419***	−0.251***	−0.225**
	(0.087)	(0.091)	(0.114)	(0.123)	(0.078)	(0.098)
N	351	351	206	206	145	145
p-value of ACT bottom quartile = Not bottom quartile	*0.075*	*0.060*	*0.011*	*0.012*	*0.775*	*0.627*
Instructor and time block controls	X	X	X	X	X	X
Demographic controls		X		X		X

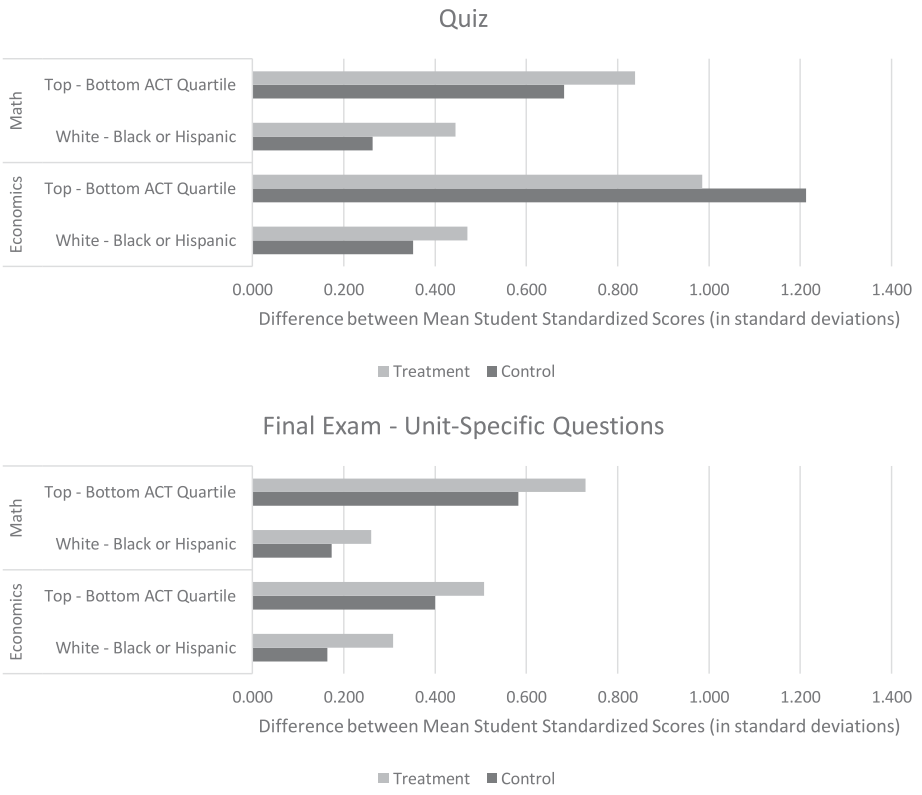
Notes: This table reports estimates from a regression of unit-quiz exam scores on an indicator for being assigned to a flipped classroom for the subgroups identified in each row. All scores have been standardized to have a mean of 0 and a standard deviation of 1 for each subject. All estimates include the controls described in table 5. Standard errors, clustered on classroom, are reported in parentheses.

***Significant at the 1% level; **significant at the 5% level; *significant at the 10% level.

= 0.005). Next, we use ACT scores to proxy for students' baseline math ability and interest. Students who scored in the bottom quartile of the ACT (relative to their classmates) experienced no significant effects from the flipped classroom, while the higher scoring students experienced significant positive gains.

Together, these subgroup effects show that the flipped classroom has the opposite equity implications as proponents claim. By having a null effect on the bottom of the math ability distribution, the flipped classroom exacerbated the achievement gap while not serving women, black students, and Hispanic students.[16,17]

16. We also used the method suggested by Abadie, Chingos, and West (2018) to investigate the impact of the flipped classroom by how we predict students will score on the math quiz in the standard classroom. The results of this investigation are similar to our estimated treatment effects by ACT quartile: The flipped classroom boosts student performance for students whom we predict will score in the top of the distribution, but there are no gains for those in the bottom quartile. Results are available from the authors upon request.
17. We find no substantial differences in the student survey responses on class preparation or students' views on the usefulness of the videos and readings that might explain these differential effects.



Notes: This figure displays the racial and baseline academic ability achievement gaps for the treatment and control groups for both the quiz and the final exam questions specific to the experimental unit's content. Achievement gaps are calculated by differencing the mean standardized scores of the top and bottom ACT quartiles and the white and Black or Hispanic students.

Figure 3. Racial and Baseline Academic Ability Achievement Gaps

Figure 3 illustrates how the flipped classroom impacted the achievement gap. The gap in math quiz performance between white students and black or Hispanic students is 0.263 standard deviations in the control group. The flipped classroom achievement gap is 69 percent larger, at 0.444 standard deviations. Similarly, the difference in math quiz means between the top and bottom quartile of ACT scorers is larger in the treatment group by 23 percent (0.838 standard deviations difference in the flipped classroom compared with 0.682 in the standard classroom). Although we find no aggregate effects of the flipped classroom on the final exam questions related to the experimental unit, the achievement gap differences persist at similar rates. We find a 51 percent larger racial achievement gap difference and a 25 percent larger ACT achievement gap in the flipped classroom relative to the standard lecture.[18]

The subgroup analysis among Economics classrooms reveals few noticeable differences, although it does appear that the flipped classroom is least effective for the

18. It is worth noting that this exacerbates the achievement gap among a relatively high-performing group of students who went through the selective West Point admissions process. In the 2017–18 admissions cycle, 25 percent of admitted applicants scored below 550 on the English section and 590 on the Math section of the SAT while the top quartile of admitted applicants scored above 660 and 690 on English and Math, respectively (West Point 2019).

highest ACT quartile (even though there is no statistically significant difference between the top and bottom of the ACT distribution as there is with math). It appears that the racial achievement gap is larger in the flipped classroom relative to the standard lecture for both the quiz and final exam questions (see figure 3), although we find mixed results for the ACT quartiles achievement gap.

7. CONCLUSION

Several design features of the experiment mute the potential effectiveness of the flipped classroom model. We conducted the experiment for three class sessions. This could lead to an underestimation of the model's effect because it may have been disruptive for the treatment group to switch the class format without enough time to establish strong classroom norms. In addition, we expect that instructors would improve over time as they get used to implementing the new pedagogy. While it is a strength of our study design that we observe instructors in both the standard lecture and the flipped classroom, it means that instructors have to prepare two types of classes. This extra preparation may lead to lower-quality classes than if they focused on one type of lecture. Additionally, we did not allow faculty or students to select into the study. Those who prefer the new pedagogy might be better instructors and students in this model, as suggested by our survey results. Lastly, additional instructor training and support could improve the quality of implementation. Together, these constraints suggest that our findings could be a lower bound for the potential impacts of the flipped classroom. Our finding that the flipped classroom generates substantial gains in math—but widens the achievement gap—might also serve as a lower bound for the potential effects for longer implementations without these limitations.

Although the standardized nature of course content, teaching, and grading offers an ideal setting for a randomized control trial, West Point is a unique academic setting and our results might not generalize to other postsecondary institutions. On one hand, West Point's selective admissions criteria and small class sizes are comparable to selective liberal arts colleges. On the other hand, cadets at West Point and the predominantly military faculty who teach them might differ on unobservable dimensions from students and faculty at other colleges. In particular, West Point's emphasis on class attendance and classroom discipline could make flipped classroom instruction at West Point more effective (e.g., through increased participation) or less effective (e.g., if marginal cadets negatively influence their peers) than flipped classroom instruction in other settings. Considering our results, additional research on the impact of the flipped classroom model in more traditional postsecondary settings is clearly warranted.

We implement a clean flipped classroom experiment in a setting with real stakes and glean insights about this popular pedagogical model. The results of our experiment show that the flipped classroom can generate large learning gains in a short period of time and that implementation quality, instructor preferences, and student engagement likely play key roles in its effectiveness. We find substantial short-term effects in math and null effects for economics. Suggestive evidence points to a few potential explanations. First, instructors who preferred teaching the flipped classroom generated larger effects, suggesting that instructor interest contributes to the success of new pedagogical models. Second, we find higher levels of student engagement in math relative to

economics: Math instructors report higher rates of students paying attention, asking questions, and working in groups and independently in class relative to the math control group and the economics treatment group. Also, math students rate the video more useful than economics students. Survey data also show that the math classrooms increased student engagement and student-teacher interactions more than the economics classrooms (perhaps those aspects are important for an effective flipped classroom implementation).

Despite the short-term effects in math, we find no longer-term gains in learning and the flipped classroom exacerbates the achievement gap instead of reducing it. Short-term gains in math are concentrated among male, white, and high-achieving students. The flipped classroom has a 69 percent larger racial achievement gap and a 23 percent larger baseline, academic ability achievement gap than the standard lecture, and these differences persist through the final exam. Combined, these findings suggest educators should exercise caution when implementing the flipped classroom.

Even with null long-term effects, educational institutions may still choose the flipped classroom model if it maintains average levels of learning, but at lower costs. Schools can reduce costs by using the flipped classroom model by hiring lower-skilled instructors, such as teaching assistants or tutors to facilitate the flipped classroom, and paying a one-time cost to produce high-quality videos by a skilled lecturer.

ACKNOWLEDGMENTS

Special thanks go to the Math and Economics departments of the United States Military Academy for participating in the study. We also thank Sandra Black, Susan Dynarski, David Figlio, Joshua Goodman, Sarah Komisarow, Jonah Rockoff, and seminar participants at Tufts University, the United States Military Academy, the Northeast Economics of Education Workshop, the Association for Public Policy Analysis & Management, and the Western Economic Association for helpful comments. Setren was supported by a National Science Foundation Graduate Research Fellowship. The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense.

REFERENCES

Abadie, Alberto, Matthew Chingos, and Martin West. 2018. Endogenous stratification in randomized experiments. *Review of Economics and Statistics* 100(4): 567–580.

Alpert, William T., Kenneth A. Couch, and Oskar R. Harmon. 2016. A randomized assessment of online learning. *American Economic Review* 106(5): 378–382.

Bergmann, Jonathan, and Aaron Sams. 2009. Remixing chemistry class: Two Colorado teachers make vodcasts of their lectures to free up class time for hands-on activities. *Learning & Leading with Technology* 36(4): 22–27.

Bergmann, Jonathan, and Aaron Sams. 2012. *Flip your classroom: Reach every student in every class every day*. Washington, DC: International Society for Technology in Education.

Berrett, Dan. 2012. How "flipping" the classroom can improve the traditional lecture. *Chronicle of Higher Education*, 19 February.

Bettinger, Eric P., Lindsay Fox, Susanna Loeb, and Eric S. Taylor. 2017. Virtual classrooms: How online college courses affect student success. *American Economic Review* 107(9): 2855–2875.

Bowen, William G., Matthew M. Chingos, Kelly A. Lack, and Thomas I. Nygren. 2014. Interactive learning online at public universities: Evidence from a six-campus randomized trial. *Journal of Policy Analysis and Management* 33(1): 94–111.

Brame, Cynthia. 2013. *Flipping the classroom*. Available http://cft.vanderbilt.edu/guides-sub-pages/flipping-the-classroom. Accessed 10 January 2020.

Carter, Susan Payne, Kyle Greenberg, and Michael S. Walker. 2017. The impact of computer usage on academic performance: Evidence from a randomized trial at the United States Military Academy. *Economics of Education Review* 56:118–132.

Davies, Randall S., Douglas L. Dean, and Nick Ball. 2013. Flipping the classroom and instructional technology integration in a college-level information systems spreadsheet course. *Educational Technology Research and Development* 61(4): 563–580.

Escueta, Maya, Vincent Quan, Andrew Joshua Nickow, and Philip Oreopoulos. 2017. Education technology: An evidence-based review. NBER Working Paper No. 23744.

Figlio, David, Mark Rush, and Lu Yin. 2013. Is it live or is it internet? Experimental estimates of the effects of online instruction on student learning. *Journal of Labor Economics* 31(4): 763–784.

Findlay-Thompson, Sandi, and Peter Mombourquette. 2014. Evaluation of a flipped classroom in an undergraduate business course. *Business Education & Accreditation* 6(1): 63–71.

Goodwin, Bryan, and Kirsten Miller. 2013. Evidence on flipped classrooms is still coming in. *Educational Leadership* 70(6): 78–80.

Joyce, Ted, Sean Crockett, David A. Jaeger, Onur Altindag, and Stephen D. O'Connell. 2015. Does classroom time matter? *Economics of Education Review* 46:64–77.

Lage, Maureen J., Glenn J. Platt, and Michael Treglia. 2000. Inverting the classroom: A gateway to creating an inclusive learning environment. *Journal of Economic Education* 31(1): 30–43.

Lee, David S. 2009. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies* 76(3): 1071–1102.

Lo, Chung Kwan, and Khe Foon Hew. 2017. A critical review of flipped classroom challenges in K-12 education: Possible solutions and recommendations for future research. *Research and Practice in Technology Enhanced Learning* 12(4): 1–22.

McLaughlin, Jacqueline E., Mary T. Roth, Dylan M. Glatt, Nastaran Gharkholonarehe, Christopher A. Davidson, LaToya M. Griffin, Denise A. Esserman, and Russell J. Mumper. 2014. The flipped classroom: A course redesign to foster learning and engagement in a health professions school. *Academic Medicine* 89(2): 236–243.

Overmyer, Gerald Robert. 2014. The flipped classroom model for college algebra: Effects on student achievement. Doctoral dissertation, Colorado State University, Fort Collins, CO.

Schaffhauser, Dian, and Rhea Kelly. 2016. 55 percent of faculty are flipping the classroom. *Campus Technology*, 12 October.

Schaffhuaser, Dian. 2016. Research: Video usage in ed continues ramp-up. *Campus Technology*, 18 July.

Schultz, David, Stacy Duffield, Seth C. Rasmussen, and Justin Wageman. 2014. Effects of the flipped classroom model on student performance for advanced placement high school chemistry students. *Journal of Chemical Education* 91(9): 1334–1339.

Supiano, Beckie. 2018. Traditional teaching may deepen inequality. Can a different approach fix it? *The Chronicle of Higher Education*, 6 May.

Swoboda, Aaron, and Lauren Feiler. 2016. Measuring the effect of blended learning: Evidence from a selective liberal arts college. *American Economic Review* 106(5): 368–372.

West Point. 2019. *Class profile*. Available https://westpoint.edu/admissions/class-profile. Accessed 10 January 2020.

Wozny, Nathan, Cary Balser, and Drew Ives. 2018. Evaluating the flipped classroom: A randomized controlled trial. *Journal of Economic Education* 49(2): 115–129.