The Journal of Economic Education

ISSN: 0022-0485 (Print) 2152-4068 (Online) Journal homepage: http://www.tandfonline.com/loi/vece20

# Evaluating the flipped classroom: A randomized controlled trial

Nathan Wozny, Cary Balser & Drew Ives

**To cite this article:** Nathan Wozny, Cary Balser & Drew Ives (2018): Evaluating the flipped classroom: A randomized controlled trial, The Journal of Economic Education, DOI: 10.1080/00220485.2018.1438860

**To link to this article:** https://doi.org/10.1080/00220485.2018.1438860

Published online: 14 Mar 2018.

THE JOURNAL OF ECONOMIC EDUCATION
2018, VOL. 0, NO. 0, 1–15
https://doi.org/10.1080/00220485.2018.1438860

Routledge
Taylor & Francis Group

RESEARCH IN ECONOMIC EDUCATION



# Evaluating the flipped classroom: A randomized controlled trial

Nathan Wozny[a], Cary Balser[b], and Drew Ives[a]

[a]Department of Economics and Geosciences, United States Air Force Academy, USAF Academy, CO, USA; [b]Department of Economics, University of Notre Dame, Notre Dame, IN, USA

**ABSTRACT**

Despite recent interest in flipped classrooms, rigorous research evaluating their effectiveness is sparse. In this study, the authors implement a randomized controlled trial to evaluate the effect of a flipped classroom technique relative to a traditional lecture in an introductory undergraduate econometrics course. Random assignment enables the analysis to eliminate other potential explanations of performance differences between the flipped and traditional classrooms, while assignment of experimental condition by section and lesson enables improved statistical precision. The authors find that the flipped classroom increases scores on medium-term, high-stakes assessments by 0.16 standard deviation, with similar long-term effects for high-performing students. Estimated impacts are robust to alternative specifications accounting for possible spillover effects arising from the experimental design.

**KEYWORDS**
econometrics; flipped classroom; randomized controlled trial; video lecture

***JEL CODES:***
*A22*; *C93*

Educators have recently expressed increased interest in "flipping the classroom." While there are many definitions of the flipped classroom, our working definition is any method of delivering content primarily outside the classroom so that time in the classroom can be devoted to a variety of other activities that promote active learning (Lage, Platt, and Treglia 2000). For example, instructors may use class time to increase individual or small-group interaction with the instructor (Prober and Heath 2012) or to correct common student misconceptions through "microlectures" (McLaughlin et al. 2014). Proponents of the flipped classroom highlight numerous potential benefits; Goodwin and Miller (2013) provide a broad overview.

Despite substantial interest from researchers and educators, empirical research on the effects of flipped classrooms[1] faces some major but understandably pervasive limitations. Observational or quasi-experimental studies, often the most feasible to implement, may confound teaching method with other factors outside of the researcher's control. Randomized trials eliminate these systematic biases, but the clustered designs typically required in educational interventions limit statistical precision in all but large-scale trials.

In observational or quasi-experimental studies comparing flipped and traditional classrooms, differences in student performance across the two groups may reflect either the effect of flipping the classroom or other differences between the groups. Swoboda and Fieler (2016) found greater improvement in *Test of Understanding in College Economics* (TUCE) scores across the semester for introductory microeconomics students in flipped (blended) classrooms than for students in traditional courses using nonrandom treatment at the instructor level. Caviglia-Harris (2016) assigned courses to teaching methods nonrandomly by semester and found higher final exam scores in partially and fully flipped courses. However, in such settings, the effect of the teaching method cannot be isolated from the effect of student composition, instructor quality and experience, or other confounding factors.

**CONTACT** Nathan Wozny nathan.wozny@usafa.edu Associate Professor, Department of Economics and Geosciences, United States Air Force Academy, 2354 Fairchild Drive, Suite 6K-144, USAF Academy, CO 80840, USA.

Some studies of the flipped classroom employ randomization of teaching methods to classes to avoid systematic differences between study groups. Calimeris and Sauer (2015) randomly assigned one of two sections of an introductory microeconomics course to participate in a flipped classroom, with the other section participating in a traditional classroom. They reported standardized exam scores two thirds to one full letter grade higher for students in the flipped classroom. However, cluster-randomized experiments face the well-known limitation that common shocks affecting all students in a section (such as time of day or class disruptions) lead to incorrect statistical inference when observations are assumed to be independent across students within the same section. In comparisons involving only one flipped section and one lecture section, the effect of the flipped class cannot be distinguished from these other (potentially random) correlated effects across students in the same section. These correlated effects may be nonrandom in studies where instructors are selected purposefully to implement a given teaching methodology, as appears to be the case in a study evaluating the efficacy of "deliberate practice" in a physics classroom (Deslauriers, Schelew, and Wieman 2011).

Randomized trials with multiple sections in each condition enable correct statistical inference using clustered standard errors. Lape et al. (2014) evaluated the effect of the flipped classroom in engineering and mathematics courses by comparing four flipped sections and four control sections. Yamarik (2007) evaluated the effect of a different teaching method (cooperative learning) in an intermediate macroeconomics course by comparing two treated sections and two control sections. However, reported statistical precision in these studies does not appear to account for clustering within sections. Properly accounting for correlated outcomes within each section (through clustered standard errors) in studies like these generally results in imprecise estimates of the teaching method for small- to moderate-scale studies that implement experimental conditions coursewide.[2]

In contrast, large-scale studies are capable of producing precise estimates even using clustered designs. A number of large-scale studies have evaluated educational interventions such as online learning. These studies achieved meaningful statistical precision with appropriate clustering of standard errors by having multiple treatment and control groups randomly assigned across multiple semesters (Alpert, Couch, and Harmon 2016), multiple institutions (Bowen et al. 2013), or large samples of students (Joyce et al. 2015). Nonetheless, large-scale studies adequate to produce meaningful statistical precision are often impractical. Given the breadth of educational interventions and settings of interest to researchers, widespread use of these studies would be extremely costly. Even when feasible, critics of large-scale randomized trials of educational interventions highlight their high costs and limited ability to generalize to other populations and settings (Thomas 2016).

This study addresses the key limitations of current research on flipped classrooms without the need for a large-scale study, also enhancing the possibility for replication. First, random assignment of the teaching method to sections and lessons eliminates unobservable differences between the two groups (aside from random statistical error). Any statistically significant differences in student performance measured after flipped classrooms versus traditional classrooms can therefore be attributed solely to the effect of the teaching technique. Second, the method of assigning the treatment greatly increases the precision of estimated effects, even after accounting for correlated outcomes. Specifically, combinations of section and lesson were randomly assigned a teaching method, so that each student, each instructor, and each section are observed under both the flipped and lecture class.[3] Consequently, we are able to detect impacts of the flipped classroom with precision similar to large-scale randomized trials with far fewer resources. Although this approach faces the same limitations of many educational studies in its ability to generalize to other populations and settings, the method of random assignment enables replication in other settings at far lower cost than the more traditional approach of section-level assignment. We are aware of one other study of an educational intervention with a comparable design: Prunuske et al. (2016) randomly assigned four groups of medical students to sequences of four modules, each using one of two online learning methods.[4] In addition to employing this uncommon design, ours is the first study we are aware of that examines the flipped classroom in an upper-level undergraduate economics course.

We estimate a positive and statistically significant impact (0.16 standard deviation) of the flipped classroom on medium-term, high-stakes assessments. The flipped classroom appears to benefit only higher-performing students on long-term high-stakes assessments, and we find no impacts on

short-term, low-stakes assessments. These findings prove robust to a variety of assumptions about possible spillover effects that one lesson's teaching method may have on performance in a subsequent lesson and to other concerns regarding internal validity.

## Experimental setting

This study took place in a junior-level undergraduate introductory econometrics course at the United States Air Force Academy. Each of the 137 students enrolled in the course was assigned to one of seven sections (each meeting at a different time) based on scheduling constraints. Each section had 14 to 23 students, and the three authors each taught two to three sections of the course. The course was highly standardized across instructors and sections, with the same exams, assignments, textbook, and practice exercises. Furthermore, two of the instructors consistently audited the third instructor to increase standardization. Instructors practiced blind, parallel grading, meaning that one instructor scored responses for a particular problem for students across all sections without knowing the student's identity while scoring. The course is required for economics and operations research majors, and attendance is mandatory for all students at the Academy.

We implemented a randomized controlled trial of the flipped classroom in the course. Each section had a total of 25 lessons covering new course content (excluding days dedicated to reviewing previous material or testing). We identified 10 lessons (referred to as "experimental lessons" below) appropriate for either a flipped class or a traditional lecture methodology. After students and instructors were assigned to sections of the course, we randomly assigned each combination of section and experimental lesson to a treatment condition (flipped classroom) or a control condition (traditional lecture), conditional on each section having exactly five flipped lessons and each experimental lesson having three or four flipped sections.[5]

We assigned students a video lecture and graded comprehension questions in advance of each flipped lesson. For each flipped lesson, the instructor reviewed comprehension questions, using student responses as a basis for discussion. Next, the instructor facilitated independent or small group work on exercises and provided mini-lectures as appropriate for topic and student needs. Students did not have any assignment in advance of each lesson selected as a traditional lecture but listened to the instructor lecture, during class, on the same material covered in the video for the flipped lesson. Students in lecture lessons had access to the same exercises offered in the flipped classes, but the lecture group generally did not have available class time to complete the exercises. The key difference in the two conditions is therefore timing rather than the primary learning resources provided: both groups received a lecture (before class for the flipped group and during class for the lecture group) and exercises (during class for the flipped group and after class for the lecture group).[6] The appendix describes the experiment and each condition in more detail. We taught the 15 nonexperimental lessons using methods more similar to the lecture classes, although we generally devoted some class time to facilitating student-centered exercises.[7]

Assessments were administered to all students to test comprehension of material in the short-, medium-, and long-term. Six classes ended with an online, unannounced, ungraded formative assessment testing comprehension of content covered in approximately the three lessons preceding the assessment. Four announced, written graded exams administered throughout the semester measured medium-term comprehension on content covered in approximately the eight lessons preceding the exam.[8] A comprehensive written final exam administered at the end of the semester measured long-term comprehension. Two ungraded assessments also included an optional 10-question online survey about students' experiences preparing for and learning from that day's lesson. The subjective questions asked about pre-class preparation time, the helpfulness of classroom and preparation activities, attention, instructor feedback, learning style, pace, and self-reported comprehension.

The military setting of the Academy might raise concerns about the applicability of findings to other institutions. However, the Academy's student body and academic curriculum are similar in many respects to other liberal arts colleges (USAFA 2015). Students complete a fully accredited academic program with 31 majors, and all graduates earn a Bachelor's of Science along with a commission in the U.S. Air Force. The average SAT math and verbal scores are 672 and 642, respectively, and the admission

rate is 13 percent. Despite a regimented daily schedule, eight to nine hours of each weekday are devoted to academics, where students are free to complete academic work on their own schedule when they are not attending class. Mandatory classes are perhaps the most unique characteristic of the Academy for the purpose of our study, so that our findings do not reflect the effect of teaching method on attendance. Burton et al. (2007) found no statistical difference in the behavior of students at the Academy and at Queens University in Belfast in an experimental study.

## Analysis and data

The analysis is divided into three components. First, the randomized controlled trial enables us to estimate the impact of the flipped classroom on student comprehension as measured on assessments. Next, we examine the possibility that impact estimates are biased by spillover effects that one lesson's teaching method may have on comprehension of a subsequent lesson's material. Third, qualitative data provide context for the impact analysis. At the end of this section, we also summarize the data used in these analyses.

### *Impact analysis*

Our preferred specification of the impact analysis assumes that student performance for a given lesson is independent of the types of other lessons, an assumption we relax in a variety of robustness checks. This assumption enables us to compare student assessment scores for concepts covered in experimental flipped lessons to student scores for the same concepts covered in experimental lecture lessons. Due to the random assignment of the flipped versus lecture condition, the raw difference in average scores is a consistent estimate of the impact of the flipped classroom on student performance. However, we increase the precision of our estimates by controlling for student fixed effects (which capture differences in student performance that are consistent across topics) and lesson fixed effects (which capture differences in difficulty level or grading standards of questions across topic areas). Furthermore, we include data on both experimental and nonexperimental lessons, coding nonexperimental lessons as lectures.[9] Specifically, we estimate the following ordinary least squares regression model:

$$Y_{isl} = \beta F_{sl} + \alpha_i + \gamma_l + u_{isl}, \tag{1}$$

where $Y_{isl}$ is the normalized score for student $i$ in section $s$ on questions covering material from lesson $l$, $F_{sl}$ is a binary indicator equal to 1 if section $s$, lesson $l$ was assigned to the flipped condition, $\alpha_i$ is a student fixed effect, $\gamma_l$ is a lesson fixed effect and $u_{isl}$ is an idiosyncratic error term. The parameter $\beta$ is the impact of the flipped classroom on normalized student scores. We also estimate impacts on subgroups of students defined by academic performance prior to the start of the course.

### *Robustness checks*

Violations of our assumption that lesson impacts are independent may lead to biased impact estimates. Although each experimental lesson is designed to introduce a new topic, any impact of the instructional method may also affect comprehension of a subsequent topic. For instance, if a flipped lesson has a positive impact on assessment scores, that benefit may "spillover" into a subsequent lecture lesson. The constrained random assignment procedure dictates that a flipped lesson is more likely to be followed by a lecture lesson than another flipped lesson, so those spillover effects are more likely to be attributed to a lesson of the opposite type. This leads to biasing the estimated impacts towards zero.

In addition to treating our primary impact estimates as likely lower bounds on the true effect, we consider alternative specifications that explicitly account for this potential bias. Any robustness check requires assumptions about the dynamics of possible spillover effects, and we consider two alternatives. First, we suppose that the teaching method for the previous lesson (or previous several lessons) directly affects assessment scores for the current lesson. Second, we suppose that the spillover effects are proportional to the total number of previous flipped lessons.

We estimate a dynamic model to account for spillover effects of recent lessons:

$$Y_{isl} = \beta F_{sl} + \sum_{j=1}^{J} \delta_j F_{s(l-j)} + \alpha_i + \gamma_l + u_{isl}, \tag{2}$$

where $F_{is(l-j)}$ is an indicator variable for a flipped class $j$ lessons prior. We again include both experimental and nonexperimental lessons in this model, reasoning that recent lessons of any type may influence comprehension of the topic currently being covered. If spillover effects of this form bias the estimated impact, then models (1) and (2) will yield different estimates $\hat{\beta}$ on the same sample. As an alternative, we suppose that the effects of previous classes are cumulative and persistent over time. Accordingly, we estimate a model that controls for the number of flipped classes prior to a given lesson,

$$Y_{isl} = \beta F_{sl} + \psi \sum_{j=1}^{l-1} F_{sj} + \alpha_i + \gamma_l + u_{isl}. \tag{3}$$

Once again, if spillover effects of this form bias the estimated impact, models (1) and (3) will yield different estimates $\hat{\beta}$.

As an additional robustness check, we estimate model (1) excluding any experimental lessons that were closely preceded by another experimental lesson, thus avoiding potential bias from spillover effects that are limited to recent lessons. An extreme version of this specification test is to estimate the impact using only the first experimental lesson, because no spillover effects are possible at that point. In addition to decreasing precision, specification checks that exclude data have the downside that any changes in the $\hat{\beta}$ estimate may also reflect changes in the sample.

## Student experience

Students' subjective reports of their experience in the classroom provide additional context for the study. We compare student responses to subjective survey questions at the end of two experimental lessons between flipped and lecture lessons. Fisher's exact test determines whether the distributions of discrete responses were statistically different between teaching methods. We pool responses for the two lessons in which the subjective assessments were administered. Although we compare the responses of students across the randomly assigned flipped versus lecture classes, we consider this comparison suggestive as we offered surveys after only two classes.[10]

## Data

Table 1 summarizes content scores for each student-lesson combination. We present statistics separately for the 10 experimental lessons, the focus of this study, and the 15 nonexperimental lessons that also covered new material and were tested on at least one assessment. Accordingly, 137 students enrolled in the course results in a total of 1,370 possible student-lesson observations for each assessment type in

**Table 1.** Summary statistics.

| Lesson Type: | (1) | (2) Experimental | (3) | (4) | (5) Nonexperimental | (6) |
|---|---|---|---|---|---|---|
| Assessment: | Short-term | Medium-term | Long-term | Short-term | Medium-term | Long-term |
| Flipped (mean) | 49.5% | 50.0% | 50.0% | 0.0% | 0.0% | 0.0% |
| Percent score | | | | | | |
| Mean | 56.5% | 76.6% | 74.3% | 49.8% | 72.8% | 73.0% |
| Standard deviation | 34.3% | 20.9% | 28.1% | 37.7% | 26.7% | 36.3% |
| Observations | 1,239 | 1,370 | 1,310 | 622 | 2,055 | 1,703 |

*Notes:* Percent score is the percent of possible points received by a student on all questions related to a lesson's topic. Medium-term assessments have observations for all 137 students for 10 experimental lessons and 15 nonexperimental lessons. See text for explanations of missing short-term and long-term scores.

experimental lessons and 2,055 possible observations for nonexperimental lessons. There are no missing data for the medium-term assessments, which consisted of graded exams throughout the semester. The long-term assessment (final exam) was similar, except that six top performers were excused from the exam. Short-term assessments were administered at the end of selected classes and focused on material covered in experimental classes, and makeups were not offered to absent students. In analysis available from us, we show that missing an assessment score is uncorrelated to the experimental lesson type.

Excluding missing data, half of experimental observations are of the flipped status by design, as each student experienced five out of the 10 experimental lessons in the flipped condition. Percent scores are the percent of possible points received by a student on all assessment questions related to a lesson's topic. The lower scores on short-term assessments may be explained by the fact that they were ungraded and unannounced, minimizing student preparation. However, response rates that exceed 90 percent and average scores well above those expected from random guessing suggest that most students took them seriously. Subsequent analyses use scores normalized to have mean zero and standard deviation one for each lesson and assessment type (short-term, medium-term, or long-term).

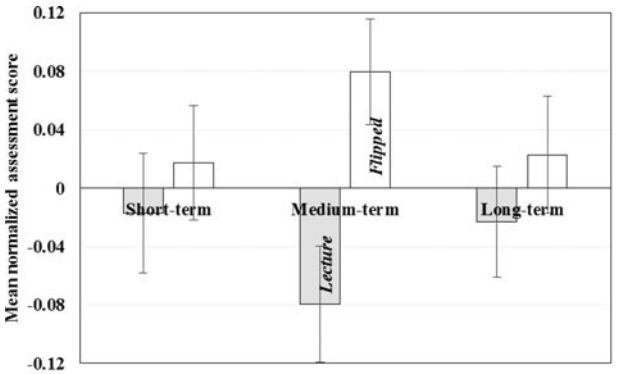
## Results and discussion

Results indicate positive impacts of the flipped classroom on medium-term assessment scores. These impacts are larger for students with above-median grade point averages, and we find positive effects of the flipped classroom on long-term assessments only for above-median students. We find no impacts on short-term, low-stakes assessments. Finally, students reported similar experiences in the flipped and lecture classes with a limited number of exceptions. This section details the estimated effects and explores possible mechanisms for the patterns of effects.

### *Impact analysis*

Students scored higher on assessments when material was covered during a flipped class as compared to a lecture class (figure 1). This difference, measured only for the 10 experimental classes, was greatest for medium-term assessments, where scores were 0.08 standard deviation above the mean for flipped classes and 0.08 standard deviation below the mean for lecture classes. The overall difference of 0.16 standard deviation is statistically significant ($p < 0.01$). Differences in scores for short-term and long-term outcomes are much smaller and not statistically significant ($p > 0.05$).

Regression analysis supports the conclusions drawn from the raw comparison. Although the random assignment of the lesson type ensures that the two groups differ on average only in their lesson type, regression analysis enables us to control for systematic differences across students (such as study habits) and across lessons (such as a topic's difficulty level) using fixed effects.[11] Regression analysis also enables




*Notes:* Data includes only experimental classes. Vertical bars show one standard error above and below mean.

**Figure 1.** Higher assessment scores for flipped classes.

**Table 2.** Impact of flipped classes on assessment scores.

| Assessment: | (1) Short-term | (2) Medium-term | (3) Long-term |
|---|---|---|---|
| =1 if flipped | 0.033 | 0.164*** | 0.047 |
| | (0.055) | (0.046) | (0.058) |
| Student fixed effects | Y | Y | Y |
| Lesson fixed effects | Y | Y | Y |
| Observations | 1,861 | 3,425 | 3,013 |
| Within R-squared | 0.000 | 0.003 | 0.000 |

*Notes*: Dependent variable is normalized mean assessment score for each student and lesson. Robust standard errors clustered at student level in parentheses.
*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$

us to include data on nonexperimental lessons because lesson fixed effects ensure that the impact of the flipped classroom is still identified on the random assignment of treatment in experimental lessons. Consistent with the differences in means, table 2 shows a positive impact of the flipped classroom on assessments. That impact is statistically significant ($p < 0.01$) for medium-term assessments (column 2) but is insignificant ($p > 0.10$) for short-term and long-term assessments (columns 1 and 3).[12] The estimated effect size for medium-term assessments, 0.16, matches that of the comparison of means. For all specifications, we present standard errors clustered at the student level. Estimated standard errors are smaller when clustering by section or by both student and section using the multi-way clustering algorithm developed by Cameron, Gelbach, and Miller (2011).[13]

We examine how estimated impacts differ by students' prior academic performance. Table 3 shows estimates of equation (1) separately for students with below or above median grade point average (GPA) prior to the start of the course. While we continue to not detect any impacts of the flipped classroom on short-term assessments (columns 1 and 2), we note that the point estimate is slightly larger for above-median students than for below-median students. Medium-term impacts remain positive and statistically significant for both groups, with the point estimates in equation (3) and (4) again indicating a slightly larger impact of the flipped classroom for above-median students compared to below-median students (0.183 vs. 0.145 standard deviation). Equation (5) shows no significant effect of the flipped classroom on long-term assessment scores for below-median students, but equation (6) shows a positive and significant effect for above-median students comparable to medium-term assessments (0.163 standard deviation). Estimates of interaction models (not shown) confirm a statistically significant difference in estimated impacts between the below-median and above-median students for the long-term assessment ($p < 0.05$), but not for short-term or medium-term assessments. We do not detect any differences in impacts across experimental lessons.[14]

We propose possible mechanisms for the pattern of results. The lack of short-term impacts is surprising given the flipped classroom's incentives to prepare for class and the additional in-class practice. The delayed impact suggests that a key part of the learning process occurs after the flipped class, and

**Table 3.** Impact of flipped classes by prior performance.

| Assessment: | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Short-term | | Medium-term | | Long-term | |
| Subgroup: | < median | ≥ median | < median | ≥ median | < median | ≥ median |
| =1 if flipped | 0.028 | 0.049 | 0.145* | 0.183*** | −0.092 | 0.163** |
| | (0.082) | (0.077) | (0.075) | (0.057) | (0.095) | (0.064) |
| Student fixed effects | Y | Y | Y | Y | Y | Y |
| Lesson fixed effects | Y | Y | Y | Y | Y | Y |
| Observations | 887 | 974 | 1,650 | 1,775 | 1,518 | 1,495 |
| Within R-squared | 0.014 | 0.012 | 0.012 | 0.015 | 0.013 | 0.019 |

*Notes*: Dependent variable is normalized mean assessment score for each student and lesson. Subgroups are above or below median grade point average (GPA) prior to start of course. Robust standard errors clustered at student level in parentheses.
*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$

**Table 4.** Models of spillover effects.

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| Sample Excluded: | First Lesson | | First 2 Lessons | | First 3 Lessons | | None |
| =1 if flipped | 0.164*** | 0.164*** | 0.159*** | 0.167*** | 0.139*** | 0.157*** | 0.186*** |
| | (0.046) | (0.046) | (0.046) | (0.047) | (0.050) | (0.051) | (0.053) |
| =1 if 1 lesson prior flipped | | 0.016 | | 0.019 | | 0.035 | |
| | | (0.050) | | (0.050) | | (0.053) | |
| =1 if 2 lessons prior flipped | | | | 0.092* | | 0.096* | |
| | | | | (0.054) | | (0.054) | |
| =1 if 3 lessons prior flipped | | | | | | 0.123** | |
| | | | | | | (0.056) | |
| Number prior flipped lessons | | | | | | | 0.044 |
| | | | | | | | (0.046) |
| Student fixed effects | Y | Y | Y | Y | Y | Y | Y |
| Lesson fixed effects | Y | Y | Y | Y | Y | Y | Y |
| Observations | 3,288 | 3,288 | 3,151 | 3,151 | 3,014 | 3,014 | 3,425 |
| Within R-squared | 0.003 | 0.003 | 0.003 | 0.004 | 0.002 | 0.005 | 0.004 |

*Notes*: Dependent variable is normalized mean medium-term assessment score for each student and lesson. Robust standard errors clustered at student level in parentheses.
$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

one possibility is that students again use videos as a study tool for high stakes assessments.[15] The lack of long-term impacts for below-median students is consistent with those students using feedback from medium-term assessments to identify deficiencies, which they then correct prior to the long-term assessment. The greater effects for above-median students is consistent with the fact that flipped classrooms put more responsibility on the students to prepare for class, so that above-median students may be obtaining the greatest benefit through better preparation. However, it is less clear why the long-term effects did not fade for above-median students through the same feedback mechanism that we propose for below-median students. Unfortunately we do not have the data to test these particular mechanisms, although the analysis of student experience data lends further support for the proposed mechanism explaining the delayed impact of the flipped classroom.

## *Robustness checks*

The estimated impacts are robust to alternative specifications presented in the previous section that account for possible spillover effects. We present robustness checks only for the full-sample medium-term assessments, but other specifications are available from us upon request.

The first six columns of Table 4 compare lagged independent variable models to fixed effects models estimated on the same sample. Column (1) repeats the primary specification (Equation 1) for medium-term assessments but excludes data from the first lesson with new content, matching the sample for a model with one lag. Column (2) estimates equation (2), which controls for the first lag of the independent variable, an indicator for the previous lesson being flipped. The columns that follow add the second and third lags of the independent variable (columns 4 and 6, respectively) and show the primary specification estimated on respective matched samples (columns 3 and 5). Although the estimated impact of the flipped classroom varies with the sample, controlling for previous flipped classes either changes the estimated impact little or increases it by up to 0.02 standardized unit. These changes are consistent with a slight negative bias in equation (1) caused by a positive spillover effect of previous flipped classrooms, also evidenced by positive coefficients on the lagged independent variables.

Column (7) of table 4 uses the full sample but estimates equation (3) to control explicitly for cumulative effects of previous flipped classes. We do not detect a significant effect of the number of previous flipped classes, although the estimated impact increases slightly to 0.186, once again consistent with the baseline model slightly understating the impact due to positive spillover effects.

**Table 5.** Spillover effect robustness checks.

| Sample Excluded: | (1) 1 Lesson after Experimental | (2) 2 Lessons after Experimental | (3) 3 Lessons after Experimental | (4) All Except First Lesson |
|---|---|---|---|---|
| =1 if flipped | 0.107* | 0.176** | 0.236** | 0.179 |
| | (0.061) | (0.079) | (0.091) | (0.109) |
| Student fixed effects | Y | Y | Y | N |
| Lesson fixed effects | Y | Y | Y | N |
| Observations | 2,055 | 1,233 | 685 | 137 |
| Within R-squared | 0.002 | 0.005 | 0.014 | 0.008 |

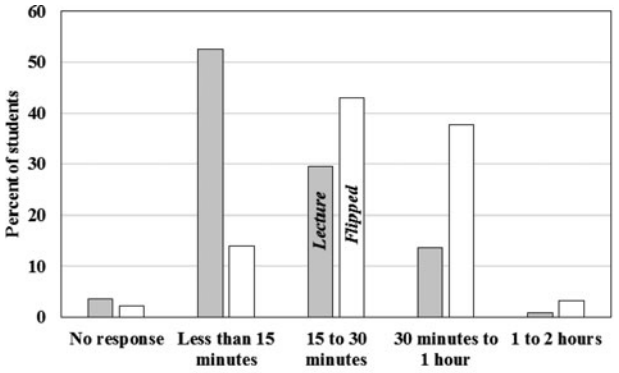
*Notes*: Dependent variable is normalized mean medium-term assessment score for each student and lesson. Robust standard errors clustered at student level (columns 1–3) or section level (column 4) in parentheses.
$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 5 estimates equation (1) on samples that limit the bias caused by short-term spillover effects. Columns (1) through (3) exclude data from lessons that follow an experimental lesson within one, two, or three lessons, respectively. Although the estimated impact varies with the sample, they remain positive and are statistically different from zero at least at the 10 percent level of significance. These specifications support the robustness of the sign of the flipped classroom's impact, although the changing sample in these specifications provides limited insight into the magnitude of any potential bias. Column (4) estimates the impact of the flipped classroom on only the first experimental lesson, perhaps the most conservative robustness test for spillover effects.[16] Although the point estimate of the impact is similar to the full sample, it is too imprecise to distinguish from zero.

The estimated effects are robust to a number of additional threats to internal validity. Attrition and nonresponse were minimal and unrelated to the lesson's assigned teaching method, highlighting a further advantage of randomization of teaching method by section and lesson. We prevented any potential instructor bias from affecting assessment scoring by implementing blind, parallel grading of exams. Psychological effects such as the Hawthorne effect are also unlikely because each student participated in both lecture and flipped classes, and assessments contained a mix of questions generally covering both flipped and lecture lessons with no apparent distinction between lesson types. Limitations to the study's internal validity, such as imperfect compliance with the assigned lesson type, will generally bias our analysis against detecting impacts. For instance, if some students assigned to a lecture class accessed videos intended for the flipped classes, their outcomes will be more similar to the flipped class than in a study with full compliance.[17] While making the instructor blind to the treatment status is virtually impossible in any study in an educational setting, the standardized procedures for each lesson type, described in the appendix, minimize the risk of instructor bias in delivery of the lessons.

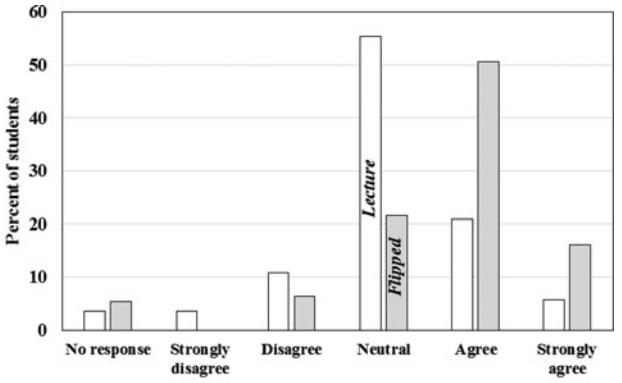
## *Student experience*

Students reported very similar experiences in flipped versus lecture classrooms. Students in flipped classes reported spending more time preparing for class (figure 2) and finding that pre-class preparation more helpful (figure 3). Given that students in lecture classes did not have any graded assignment due before the start of class, the results are consistent with most students preparing for class only when short-term incentives encourage them to do so. Students reported no difference in the helpfulness of the class lecture component; the helpfulness of the class non-lecture component; difficulty paying attention; helpfulness of instructor feedback; consistency of the class with student's learning style; appropriateness of the pace; and self-rated understanding of the day's topic (detailed results available upon request). The similarity of reported student experiences in the classroom perhaps lends further support to the hypothesis that the use of videos as a study tool is a larger part of the flipped classroom's impact than has previously been recognized.




*Note:* Prompt to students: "The amount of time on task I spent preparing for today's class was:"

**Figure 2.** Students prepare more for flipped classes.




*Note:* Prompt to students: "The preparation I completed before today's class helped me to understand hypothesis testing/fixed effects." (Topic name depended on class.)

**Figure 3.** Preparation helpful in flipped classes.

## Conclusion

This study uses a randomized controlled trial to estimate the impact of the flipped classroom. We find a statistically significant positive impact of the flipped classroom for medium-term assessments of 0.16 standard deviation. Impacts are slightly larger and also persist to long-term assessments for high-performing students, but the relatively similar effects for above-median and below-median students support the generalizability of the results to a wide spectrum of students. We also find patterns of results that we argue are consistent with students in flipped classrooms benefitting from access to videos as study tools and not solely from the pre-class and in-class activities.

The diversity of methods for implementing flipped classrooms limits the external validity of any one study. While the flipped and lecture classes in this study are plausibly similar to many other college-level classrooms, impacts may vary with the topic area, specific pre-class and in-class activities, and institutional setting. Understanding the conditions under which the flipped classroom is effective inevitably requires carefully designed studies in a variety of settings. Furthermore, future research could isolate the effects of specific components of the flipped classroom by comparing flipped classrooms that vary those components in a controlled manner.

This study creates a model by which researchers can evaluate educational interventions rigorously while using only modest resources. Random assignment of each lesson and section to a flipped or lecture class ensures that the teaching method is uncorrelated with student characteristics, and observation of students under both conditions improves the precision of the estimates. The more common approach of comparing classrooms with and without an intervention risks confounding the efficacy of the teaching methods with differences in the instructors, student body, or other factors that are difficult to measure.

By contrast, this study's design ensures that experimental comparisons isolate the effect of the intervention while improving statistical power. The study design does require that the educational intervention can plausibly be implemented in a single lesson, and that the material covered in an experimental lesson can be assessed independently of other lessons' material. While this limitation makes the design inappropriate for some interventions, small-scale randomized controlled trials have the potential to bring rigor to evaluating the efficacy of promising teaching practices when large-scale randomized trials are infeasible.

## Acknowledgments

The authors thank Victoria Bhavsar, Ashley Miller, Amy Munson, Gregor Novak, Sarah Robinson, Lauren Scharff, participants at the 2017 Conference on Teaching and Research in Economic Education and the 2016 International Society for the Scholarship of Teaching and Learning Conference, two anonymous reviewers at the U.S. Air Force Academy, and two anonymous referees for their comments and suggestions on study design, implementation, and analysis. The authors also thank the Department of Economics and Geosciences and the students in Econometrics I for their support in implementing the research. The views expressed in this article are those of the authors and not necessarily those of the U.S. Air Force Academy, the U.S. Air Force, the Department of Defense, or the U.S. Government.

## Notes

1. See Bishop and Verleger (2013) for an overview of research on flipped classrooms.
2. Schochet (2008) provides a thorough treatment of statistical precision in clustered randomized experiments.
3. Randomized trials with assignment at the student by lesson level are possible for certain types of interventions, such as technological interventions, providing similar benefits when feasible.
4. So-called "crossover" designs are more common in clinical trials (Wellek and Blettner 2012). However, the analytic methods used by Prunuske et al. (2016) appear to treat students' learning gains as independent despite the clustered nature of the design. Single-case experimental design, another technique for estimating the impact of an intervention with a modest sample size, requires a similar assumption that impacts of an intervention do not carry over after the intervention is stopped by the researcher (Dallery, Cassidy, and Raiff 2013).
5. We used a computer program with a random number generator to determine the classroom assignments. As suggested by Moulton (2004), the assignment program looped until it identified an outcome that met the criteria for the number of flipped lessons in each section and number of sections with flipped lessons. This method ensures the randomness but not independence of random draws across lessons or sections.
6. This study cannot distinguish between the effects of the various components of the flipped classroom condition, such as more individualized instructor feedback or the availability of the videos.
7. We were interested in flipping more lessons but were limited by a lack of videos appropriate for the course content and the time-consuming nature of creating new multimedia content.
8. Because an exam question may draw on material from multiple lessons, we each made independent determinations of the lesson material primarily associated with each question. The pattern of results is not sensitive to which author's judgments are used.
9. Lesson type remains uncorrelated with the model's error term under our assumption because the lesson fixed effects absorb any systematic differences in performance in the nonexperimental lessons. Intuitively, the additional data improve identification of the individual fixed effects but not the treatment effect.
10. In particular, our analysis of qualitative assessment data does not control for student effects, as this would limit the sample to sections of the course that were assigned opposite styles on the two classes with these surveys.
11. Because each student participated in equal numbers of flipped and lecture classes by design, however, student fixed effects are mechanically uncorrelated with lesson type, except where data are missing.
12. Estimated impacts are nearly identical when limiting the sample to experimental lessons.
13. An exception is for the short-term effects, where the standard error of the impact estimate is slightly larger when clustering by section (0.057) or individual and section (0.059).
14. Specifically, we estimate a model that interacts flipped status with dummies for each experimental lesson, and we fail to reject that the impact of the flipped classroom is the same across all lessons ($p > 0.10$).
15. Despite the low stakes of the short-term assessments, average scores were well above levels consistent with random guessing, suggesting that lack of student effort does not explain the results. In results available from us, we also found that the estimated impact does not vary with the time elapsed from the lesson's coverage to the assessment, suggesting that study patterns rather than time elapsed explain the difference between short- and medium-term assessments.
16. Out of necessity, this specification excludes individual and lesson fixed effects, and the standard error is clustered by section.
17. We did not have the ability to track student access to videos, so we cannot rule out that students accessed videos assigned to other sections. However, we provided students with links to videos assigned only to them, and did not advertise which lessons had videos assigned to another section.

## References


Alpert, W. T., K. A. Couch, and O. R. Harmon. 2016. A randomized assessment of online learning. *American Economic Review* 106 (5): 378–82.

Bishop, J. L., and M. A. Verleger. 2013. The flipped classroom: A survey of the research. Paper ID #6219. 120th ASEE Annual Conference and Exposition, Atlanta, GA, June 23–26.

Bowen, W. G., M. M. Chingos, K. A. Lack, and T. I. Nygren. 2013. Interactive learning online at public universities: Evidence from a six-campus randomized trial. *Journal of Policy Analysis and Management* 33 (1): 94–111.

Burton, A. C., K. S. Carson, S. M. Chilton, and W. G. Hutchinson. 2007. Resolving questions about bias in real and hypothetical referenda. *Environmental and Resource Economics* 38 (4): 513–25.

Calimeris, L., and K. M. Sauer. 2015. Flipping out about the flip: All hype or is there hope? *International Review of Economics Education* 20:13–28.

Cameron, C. A., J. B. Gelbach, and D. L. Miller. 2011. Robust inference with multiway clustering. *Journal of Business and Economic Statistics* 29 (2): 238–49.

Caviglia-Harris, J. 2016. Flipping the undergraduate economics classroom: Using online videos to enhance teaching and learning. *Southern Economic Journal* 83 (1): 321–31.

Dallery, J., R. N. Cassidy, and B. R. Raiff. 2013. Single-case experimental designs to evaluate novel technology-based health interventions. *Journal of Medical Internet Research* 15 (2): e22. doi:10.2196/jmir.2227

Deslauriers, L., E. Schelew, and C. Wieman. 2011. Improved learning in a large-enrollment physics class. *Science* 332 (6031): 862–64.

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

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

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

Lape, N. K., R. Levy, D. H. Yong, K. A. Haushalter, R. Eddy, and N. Hankel. 2014. Probing the inverted classroom: A controlled study of teaching and learning outcomes in undergraduate engineering and mathematics. Paper ID #9475. 121st ASEE Annual Conference and Exposition, Indianapolis, IN, June 15–18.

McLaughlin, J. E., M. T. Roth, D. M. Glatt, N. Gharkholonarehe, C. A. Davidson, L. M. Griffin, D. A. Esserman, and R. J. Mumper. 2014. The flipped classroom: A course redesign to foster learning and engagement in a health professions school. *Academic Medicine* 89 (2): 236–43.

Moulton, L. H. 2004. Covariate-based constrained randomization of group-randomized trials. *Clinical Trials* 1 (3): 297–305.

Prober, C. G., and C. Heath. 2012. Lecture halls without lectures—A proposal for medical education. *New England Journal of Medicine* 366 (18): 1657–59.

Prunuske, A. J., L. Henn, A. M. Brearley, and J. Prunuske. 2016. A randomized crossover design to assess learning impact and student preference for active and passive online learning modules. *Medical Science Educator* 26: 135–41.

Schochet, P. Z. 2008. Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics* 33 (1): 62–87.

Swoboda, A., and L. Feiler. 2016. Measuring the effect of blended learning: Evidence from a selective liberal arts college. *American Economic Review* 106 (5): 368–72.

Thomas, G. 2016. After the gold rush: Questioning the "gold standard" and reappraising the status of experiment and randomized controlled trials in education. *Harvard Educational Review* 86 (3): 390–411.

United States Air Force Academy (USAFA). 2015. *USAFA quick facts*. http://www.usafa.af.mil/AboutUs.aspx (accessed December 21, 2016).

Wellek, S. and M. Blettner. 2012. On the proper use of the crossover design in clinical trials. *Deutsches Ärzteblatt International* [*German Physician Sheet International*] 109 (15): 276–81.

Yamarik, S. 2007. Does cooperative learning improve student learning outcomes? *Journal of Economic Education* 38 (3): 259–77.


## Appendix: Implementation notes

Any study evaluating the effect of a flipped classroom depends on how the flipped and comparison classes are implemented. In this study, we attempted to follow commonly accepted best practices in the flipped classes while using as a comparison a lecture style common in many postsecondary educational settings today. This appendix describes the implementation of the experiment followed by more detail about the flipped and lecture classes.

**Table A1.** Course Outline.

| Lesson | New content? | Experimental? | Topic |
|---|---|---|---|
| 1 | X | | Course introduction |
| 2 | | | Statistics review |
| 3 | X | X | OLS mechanics |
| 4 | X | X | Regression interpretation |
| 5 | X | X | Sampling distributions |
| 6 | X | | Gauss-Markov Theorem |
| 7 | X | X | *t*-tests |
| 8 | X | | Graded assessment / *p*-values |
| 9 | X | | Confidence intervals |
| 10 | X | | Multiple regression |
| 11 | X | X | Omitted variable bias |
| 12 | X | | Simple vs. multiple regression |
| 13 | | | Review day |
| 14 | | | Graded assessment |
| 15 | X | | Multicollinearity |
| 16 | X | X | Binary independent variables |
| 17 | X | X | *F*-tests for linear restrictions |
| 18 | X | | Interaction variables |
| 19 | X | X | Intro to nonlinear models |
| 20 | X | | Nonlinear models, continued |
| 21 | X | | Model specification |
| 22 | | | Enrichment topics |
| 23 | | | Review day |
| 24 | | | Graded assessment |
| 25 | X | | Internal vs. external validity |
| 26 | X | X | Measurement error |
| 27 | X | | Heteroskedasticity and correlated errors |
| 28 | X | | Multiple regression in practice |
| 29 | X | | Panel data |
| 30 | X | X | Fixed effects regression |
| 31 | X | | Time effects |
| 32 | | | Enrichment topics |
| 33 | | | Review day |
| 34 | | | Graded assessment |

*Notes:* Lessons marked as experimental were randomly assigned to a flipped or lecture condition for each section. Nonexperimental lessons that introduce new content use methods more similar to the lecture condition.

### *Experiment implementation*

The experiment was implemented with the intent of minimizing disruption to the learning environment. As instructors and authors, we identified 10 experimental lessons appropriate for either teaching method. Even in the absence of the experiment, the instructors would have alternated a flipped class on these lessons with other techniques on the 15 other instructional days covering new material, so that differences in instructional type across lessons were not unusual to students. After randomly assigning 5 of the 10 experimental lessons to the flipped class for each of the 7 sections, we created separate course outlines for each section with the flipped classes (or at least the associated assignments) marked. Instructors notified students of the study at the start of the semester and told students that other sections may have different assignments on their course outlines. Although a student could have accessed outlines for other sections to identify the experimental classes selected as lecture, a casual observer would not distinguish between an experimental lesson selected as a lecture and a lesson not included in the experiment. A list of topics covered in each lesson similar to the course outline is shown in table A1. The table shows that experimental lessons are mixed with nonexperimental lessons and also with lessons not covering new material, such as review of previous material or assessments.

The lead instructor for the course developed common course materials for all lessons. In particular, he developed separate materials for each flipped class and the corresponding lecture version. Both sets of notes and corresponding classroom activities were developed based on experience in teaching the course for several years before the experiment was implemented. However, notes for the lecture

were revised to follow the examples and explanations in the video closely. The other two instructors followed those notes closely and normally audited the lead instructor's class before teaching their own lessons.

As in any educational study implemented by the researchers, the instructors must put forth their full effort into teaching both types of classes to avoid biasing the study's results in favor of a "preferred" technique. Although we cannot guarantee that we were free from all subconscious bias, instructor favoritism was unlikely to drive the findings. In particular, the random assignment of instructional method made it difficult to favor one type of technique: an instructor may have taught a lecture class followed by a flipped class on one day and a flipped class followed by a lecture class on another day. Furthermore, preparation for lecture and flipped classes overlapped significantly, as teaching either lesson type required being prepared to address student questions on content and assist in the completion of exercises.

### *Flipped classroom procedures*

The flipped classes required students to watch videos before each class. We created these videos before the start of the semester and posted them on a public video-sharing Web site. Each video was approximately 10 minutes in length and used graphics and text in a presentation with narrated explanations, providing the students with a lecture-like overview of the lesson's topics. We prepared the slides for the videos along with a script in advance, making the pace of the videos significantly faster than the same content would be delivered in a typical classroom lecture. Accordingly, students were advised that pausing or re-watching parts of the video may be necessary. We specifically selected topics for videos that were expected to require significant lecture with little room for student interaction. The topics, also shown in table A1, were: ordinary least squares mechanics; regression interpretation; sampling distributions; *t*-tests; omitted variable bias; dummy variables; *F*-tests for linear restrictions; introduction to nonlinear models; measurement error; and fixed effects regression.

Each video was accompanied by a Web-based comprehension assessment. After reviewing the video and before the lesson's start, students in the flipped class were required to complete a series of multiple choice, quantitative calculation, and short-answer questions. Students also were afforded the opportunity to provide comments on the video or questions on the material. Students were permitted to view the video multiple times in an effort to answer the questions. All instructors briefly reviewed student responses before the start of each class to identify potential problem areas that informed a discussion at the beginning of each class. We also graded student responses in parallel and provided individualized feedback after the lesson's completion. Each assessment was worth one percent of the course grade for the semester.

Class time in each flipped lesson was generally divided into two parts. Approximately the first 20 minutes of each flipped lesson were allotted to review and discussion of the comprehension assessment questions with the students. The instructor presented summaries of answers to quantitative and multiple choice questions, discussing the correct answers with a focus on problems most commonly missed. When appropriate, we probed students for explanations of their answers to ensure comprehension. The instructor also presented examples of student answers to the short-answer questions (without identifying respondents), generally asking students to explain the strengths and shortcomings of each answer. The remainder of each 53-minute class was primarily devoted to student time to work on practice exercises. The instructor walked around the classroom to gauge student progress and answer questions. In most classes, the instructor also provided mini-lectures and clarifications throughout this time when appropriate for the material or to address frequently asked questions.

### *Lecture class procedures*

Lecture classes covered the same material as the videos. We delivered lectures covering the same objectives and using similar examples and explanations as the video. Content was delivered primarily

on a whiteboard, with some slides projected onto a screen as supplements. Although the lectures were heavily instructor-led, the instructor did engage students by asking questions about comprehension and also responding to student questions. Due to this additional engagement and the time required to write on the board, the lecture typically required at least 30 minutes to cover the same topics as in the video. When time permitted, students worked on the same exercises provided to the flipped classes, although lecture classes consistently had considerably less time to work on exercises compared to flipped classes. Students were encouraged to complete exercises out of class and before the next lesson, although the exercises were not collected and graded.