Evaluating the Flipped Classroom in An Undergraduate History Course

Yiran Zhao and Andrew Ho

Harvard Graduate School of Education

June 16, 2014

Suggested Citation: Zhao, Y., & Ho, A. D. (2014). *Evaluating the flipped classroom in an undergraduate history course* (HarvardX Research Memo).

Abstract
In spring of 2012, Harvard University and MIT announced the launch of edX, an open online learning platform that allowed anyone with an internet connection to register and complete online courses. The edX mission included the improvement of on-campus, residential education as well as advancing educational research. This report documents the impact of an effort to transform an undergraduate history course using a "flipped classroom" model, where online videos and resources from the edX platform supplemented in-class discussions and activities. A baseline administration of the course was in the fall of 2011, prior to the launch of edX. The second administration used the flipped classroom format in the fall of 2013.

By holding the midterm exam fixed from 2011 to 2013, as well as using deidentified institutional data, including achievement test scores and grade point averages, we employed a quasi-experimental design that supports interpretations about the causal impact of the flipped classroom on student learning.

Our main findings include the following:

- Enrollment was smaller for the flipped course, although the evidence that this was caused by the flipped classroom is scant. There are no statistically significant differences in the composition of the enrolled student body in terms of incoming achievement test scores and grade point averages.
- In the 2013 administration, student opinions about the "flipped classroom" model were variable, with 46% of the 37 participants preferring or strongly preferring the flipped model, and 38% of participants preferring or strongly preferring the traditional model.
- There is no compelling evidence of any significant impact of the flipped classroom on midterm examination scores.
- Students found that discussions with a) high percentages of engaged students and b) cross-talk among students to be most beneficial.

Flipped classroom models using edX-like platforms for open online learning are in nascent stages of uptake. The first-year learning gains that we report may represent a lower bound. Consistent with the edX mission, we recommend that instructors and researchers continue quasi-experimental or experimental designs to track the impact on learning over time. At the very least this should include holding one or more assessments fixed over time. We also recommend that classroom assessment technologies be given the same care in development that online lectures are currently receiving.

Table of Contents

Executive Summary

- We used a quasi-experimental design that compared student learning—as measured by a common midterm exam score—from two administrations of the undergraduate Chinese history course (Societies of the World 12: China; abbreviated SW12): in the fall of 2013 (the "flipped" version) and in the fall of 2011 (the "traditional" version). We obtained student learning data from 33 students in 2013 and 65 students in 2011.
- Overall there was no statistically significant difference in the composition of the student body in 2013 in the flipped classroom compared to that of the student body in 2011 in the traditional classroom in terms of incoming achievement test scores, grade point averages and year in school.
- The class size in 2013 was smaller than the class size in 2011. However, there was no evidence that causally linked the change in enrollment size to the flipped classroom model.
- Overall, we found no compelling evidence that suggested that the flipped classroom had any significant impact on student learning as measured by the midterm exam. There were no statistically significant differences between the unadjusted midterm exam scores of the students in the flipped classroom and those of the students in the traditional classroom, controlling statistically for incoming achievement test scores, grade point averages, or year in school.
- The midterm exam required judgmental grading by teaching fellows, and

teaching fellows changed from 2011 to 2013. We had 2011 teaching fellows grade a subset of 2013 responses, and we applied two different adjustments methods, or "corrections," to scores to account for the differences in stringency between 2011 and 2013 graders. After Correction 1, we found that students in the flipped classroom scored, on average, 4.6 points lower (on a 100-point scale with a standard deviation of 6.7) than students in the traditional classroom, after controlling for first-year GPA. After Correction 2, this difference was 6.9 points on average. The magnitude of this average effect is akin to that from a B to a B- after Correction 1 and a B to a C+ after Correction 2.

- The estimated differences between the flipped and traditional classroom were statistically significant under both Correction 1 and Correction 2, however, neither correction incorporated measurement error in the adjustments for rater stringency. In other words, although the average impact is large in magnitude, it is on the order of variation in rater stringency that we observe in practice.
- We administered a survey to the 37 enrolled students in the fall of 2013 for the purpose of understanding their opinions about the flipped classroom model. We found no statistically significant association between student midterm grades and their stated preference for the flipped or traditional model.
- Students were asked whether hearing initially about the "flipped" approach made the course seem more or less appealing. Over half (53%) of the

respondents felt the course was somewhat more or much more appealing when they first heard of this flipped approach, comparing to 14% who felt that the course was somewhat less or much less appealing.

- Out of the 12 students whose interests in this course were not influenced when they first heard of the flipped approach, 5 ended up preferring or strongly preferring the traditional approach at the time of the survey, while 4 switched to preferring the flipped approach.
- In open-ended responses, students' stated reasons for preferring the flipped model included the following: The online modules were convenient and engaging. Modules effectively substituted for the lectures and allowed more time for active student participation in class. The short video segments made it easy to digest the contents. Students could hear various perspectives in discussions to help them learn better. Discussions were effective when students came prepared with the same background knowledge and when a high proportion of students engaged in dialogue.
- In open-ended responses, students' stated reasons for preferring the traditional model included the following: The workload was too heavy. The online assessments had shortcomings. The discussions were of low quality. The cold-calling was stressful and could negatively affect the flow and quality of discussions. Some students preferred that class be used for lectures and not discussions.
- On average, respondents thought the online videos were valuable to their

learning, while the average rating for the online assessments was between somewhat valuable and valuable. Respondents also rated the readings to be between valuable and very valuable on average.

- Analyses of the quantitative data and direct observation of the class discussions lead us to three general recommendations for the design and evaluation of flipped classrooms.
    - First, the transition from a traditional to a flipped classroom is likely to coincide with a high demands on the instructional team to produce videos and support basic user interactions on the digital platform. Assessments and the design of in-class activities are likely to be issues for which instructors have less expertise to address in the transitional stage.
    - Second, consistent with existing research on flipped classrooms, student opinions about flipped classrooms are mixed in this transitional stage, as workload increases without tangible benefits and the norms of classroom behavior change. There is particular need to make expectations explicit to students and model desirable student behavior in transitional periods.
    - Third, to advance the research mission, large classrooms in transition should hold some assessments constant over time, to track potential learning gains due to changing instructional practices. Institutional research should provide incoming student achievement data for

statistical controls, allowing quasi-experimental analyses such as the one presented here.

Section 1: Background

Flipped learning, or the flipped classroom, refers to pedagogical practices that allow students to learn course contents traditionally delivered in classroom lectures prior to class, with the help of technology including but not limited to online videos. Students spend class time engaging in active learning activities and may also receive individually targeted feedback from instructors (Hamdan, McKnight, McKnight, & Afstrom, 2013). Many higher education institutions have flipped their classrooms and evaluated such efforts with the goal of increasing the quality of instruction (Mclaughlin, Roth, Glatt, Gharkholonarehe, Davidson, Griffin, Esserman, & Mumper, 2013). The majority of the studies investigated students' experiences qualitatively, and quasi-experimental or experimental evaluation of the impact of flipped classrooms on student learning outcomes has been limited (Bishop & Verleger, 2013; Zhao & Breslow, 2013).

Zhao and Breslow (2013) reviewed 43 studies on blended learning, defined as a combination of face-to-face learning and online learning (Garrison & Kanuka, 2004; Reasons, Valdares, & Slavkin, 2005; Means, Toyama, Murphy, Bakia, & Jones, 2010; Lack, 2013). These studies all contained quantitative measures of student learning. Among these studies, 22 studies employed a flipped learning model where in-person lectures were fully or partially substituted by online materials, and class time was reduced. Four categories of materials were used to substitute for in-person lectures: online readings/assignments (Maki & Maki, 2002; Riffell & Sibley, 2005); long videos (Lewis & Harrison, 2012); multiple short video segments (Stone, 2012; Mason, Shuman & Cook, 2013); and online tutorials with a combination of videos, embedded

assessments and animations, and images and texts (Christoph, 2001; Broida, 2001; Lovett, Meyer, & Thille, 2008; Holdhusen, 2009; Chen, Stelzer, & Gladding,2010; Upchurch & Hartman, 2011; Bowen, Chingos, Lack & Nygren, 2012; Sadaghiani, 2012; Popyack, 2013).

Overall, results on student learning were mixed (Zhao & Breslow, 2013). 10 studies found greater learning gains for students in the flipped model than students in the traditional model (Broida,2001; Day & Foley, 2006; Chen, Stelzer, & Gladding, 2010; Moravec, Williams, Aguilar-Roca, & O'Dowd, 2010; Upchurch & Hartman,2011; Lewis & Harrison,2012; Pierce& Fox, 2012; Sadaghiani,2012; Mclaughlin et al.,2013; Popyack,2013). 2 studies found significant effects of the flipped model only for a subpopulation of the participants (Maki & Maki, 2012; Riffell & Sibley, 2005). 4 studies reported mixed results on learning outcomes (Lovett, Meyer, & Thille, 2008; Marcey & Brint, 2012; Stone, 2012; Mason, Shuman & Cook, 2013) and 6 studies found no significant difference in student learning when comparing treatment with control groups(Christoph,2001; Reasons, Valdares, & Slavkin,2005; Holdhusen,2009; Bowen, Chingos, Lack & Nygren, 2012; Choi,2013; Davies, Dean, & Ball, 2013).

Bishop and Verleger (2013) conducted a literature review on research of flipped classroom. They defined flipped classroom as "an educational technique that consists of two parts: interactive group learning activities inside the classroom and direct computer-based individual instruction outside the classroom" that must include videos (Bishop & Verleger, 2013, p.5). They found mixed but generally positive student

perceptions of the flipped classroom. Interactive in-class activities were preferred by students compared to lectures and shorter videos were preferred. However, students liked face-to-face lectures better than video lectures. Two studies in their review measured student learning outcomes and both found improvement in student learning. But in one of the studies, the class was only flipped for three times and there were mini-lectures in class. And the result of the other study of a senior level computer interaction course alone was not sufficient to support a generalized claim on the effectiveness of flipped classroom (Bishop & Verleger, 2013).

These evaluations of flipped classroom practices mainly addressed science, technology, engineering, and mathematics courses. In contrast, this study evaluates a humanities course: Societies of the World 12.

Societies of the World12 (SW12) is an undergraduate general education course offered at Harvard College. Traditionally, students attended lectures and small group discussion sections led by teaching fellows every week. In 2013, the instruction team of SW12 created ChinaX, an open online version of SW12, and prepared to offer it to the world on the edX platform. In the fall of 2013, they embarked on a flipped learning initiative by integrating ChinaX into SW12. This study provides quasi-experimental estimates of the impact of this transition on learning as measured by a midterm examination held fixed from the previous course administration in 2011. In addition, we describe differences in the characteristics of the enrolled students from 2011 and 2013, and student perceptions of the flipped classroom from the 2013 administration.

In Section 2, we describe the implementation of the flipped classroom. In Section 3, we describe characteristics of the student body. In section 4, we analyze the impact on student learning. In section 5, we report student perceptions of the flipped classroom. We offer concluding remarks in Section 6.

Section 2: Implementation of the flipped classroom in SW12

In a typical week of SW12 in fall 2013, students were required to complete two to three ChinaX online modules, complete readings of varied length, and attend two one-hour-long class sessions, where they engaged in discussions, debates, or case studies. This differed from the traditional version of the course in 2011, where in-class time was used for lectures. To partially offset the additional work required of students in 2013, sections with teaching fellows were held three times in the semester instead of on a weekly basis.

The online modules were delivered through the ChinaX website on edX "Edge" platform, a version of edX designed for on-campus instructional use. Modules consisted mainly of video segments and assessments. The assessments came in both multiple choice and short-answer formats. Additional materials were available on the ChinaX website and the conventional SW12 course website, including images, Chinese pronunciation guides, and maps. Students were expected to review them before the corresponding class sessions.

In class meetings, every student had a name card in front of him or her. Students either voluntarily participated in in-class activities or were cold-called. All the class meetings were recorded. Students could review the class meeting videos through both websites. Class discussions and debates often centered on images or the reading

materials. Students could access and prepare for the discussion questions on both websites before class meetings. Besides occasional lectures, for example in the first week while the instructional team transitioned students to the new format, in-class time was devoted to discussions, debates, and other non-lecture activities.

Section 3 Analytic Strategy

To investigate the impact of flipped classroom on student learning, we employed a quasi-experimental design by collecting and comparing midterm examination scores from the two administrations of SW12: one in the fall of 2011 and one in the fall of 2013. The instructor team administered the same midterm exam under similar administrative conditions. We also used de-identified institutional data including year in school, achievement test scores taken prior to admission, and grade point averages, in an effort to control for potential confounding factors.

The midterm exam consisted of two types of items, identification (ID) items and essay items. In the ID items, the first 8 items were presented as text, and students needed to choose 4 to identify. The last two ID questions were presented as images and students had to choose 1 to identify. Additionally, there were two essay items from which students selected 1 to write their essay. Responses were graded by teaching fellows, and teaching fellows differed in 2011 and 2013. In 2011, there were three raters who graded the ID items, essay 1, and essay 2, respectively. In 2013, there were two raters: one graded the ID items, and the other graded both essays.

Ideally, the 2013 raters would rate all 2011 midterms or vice versa, to ensure comparable rater stringency across midterm grading procedures. For practical reasons including time constraints for teaching fellows, we had the 2011 raters grade a

random sample of responses from the 2013 midterms (5 essays each for the two 2011 essay raters, and 10 student sets of ID questions for the 2011 ID rater). We use a model where 2011 ratings can be linked to the 2013 ratings by adding a rater stringency difference parameter, κ_r, for each 2011 rater, representing the difference in stringency between the 2013 rater and his or her 2011 counterpart. In addition, we define a judgmentally determined calibration constant for each rater, C_r, that captures interactive calibration between 2011 raters in 2011 that was not possible when they graded independently in 2013:

$$X_{pr}^{2013} = X_p^{2011} + \kappa_r + C_r + \epsilon_{pr}.$$

Here, X_p^{2011} is the ID or essay score for a student p in 2011, κ_r is the rater stringency difference parameter that allows a link to the 2013 scale, C_r is a judgmental calibration constant, and ϵ_{pr} is a random error term capturing rater-student interactions and other unobserved and random sources of error in the link to 2013 scores. We use the difference in average ratings between two raters, across midterms graded in common, as our estimate of the difference in rater stringency:

$$\hat{\kappa}_r = \bar{X}_r^{2013} - \bar{X}_r^{2011}$$

The 2011 scores are thereby linked to the 2013 scale as the sum of the 2011 score and the rater stringency parameter:

$$\tilde{X}_{pr}^{2013} = X_p^{2011} + \kappa_r + C_r.$$

Here, $\tilde{X}_{pr}^{2013}$ can be interpreted as the score that the 2011 student would have received if rated by the 2013 raters. Ignoring the constant, C, this is the rater-based

analog of a "mean-mean equating" in the psychometric literature (Kolen & Brennan, 2002). For raters 1 (ID), 2 (Essay 1) and 3 (Essay 2) in 2011, $\hat{\kappa}$ estimates are 1.3, 2.0, and 10.0 points respectively. The total corrected score, $\tilde{Y}$, is simply the sum of the ID score and the score for whichever essay the student selected.

Because significant time had passed since 2011 raters had graded, we calibrated further by discussing potential differences in rating procedures across years. The third rater in 2011 ($r = 3$) raised concerns that his 2011 grades were a byproduct of interactive calibration with other raters. He acknowledged that in 2011 this resulted in a 5-point downward adjustment that he felt would be appropriate for the 2013 grading as well. For this rater, then, we set $C_3 = -5$.

If the model holds without error, then the 2013 scores Y and the converted scores from 2011, $\tilde{Y}$, are directly comparable, allowing for the quasi-experimental estimation of the impact of the course transformation on midterm examination scores:

$$Y_p = \alpha + \beta\left(Flipped_p\right) + \boldsymbol{\gamma}\mathbf{W_p} + \nu_p$$

Here, Y_p is a vector of total scores that does not distinguish between Y and $\tilde{Y}$, α is a trivial constant, β is the effect of interest on a scale of total midterm points, $Flipped$ is a 0/1 indicator variable for the 2013 version of the course, $\boldsymbol{\gamma}$ is a row vector of regression coefficients for the column vector of covariates $\mathbf{W_p}$, and ν_p is a random error term.

The analytic strategy depends crucially upon this correction process, as it assumes that the conversion from the 2011 to 2013 scale is without error. Interrater reliabilities across the midterm sections graded in common are reasonably high and

lend support for the precision of the estimation of κ. These correlations are 0.82 for the ID grading, .94 for Essay 1 grading, and .78 for Essay 2 grading. However, the only evidence in support of the precision of C estimation is judgmental. As acknowledgment of influence of the correction process on substantive conclusions, we present results from three different approaches as a sensitivity study, the corrected scores under the posited model, the corrected scores without the judgmental calibration constant C (affecting Essay 2), and the uncorrected scores.

$$\text{Corrected: } \tilde{X}_{pr}^{2013} = X_p^{2011} + \kappa_r + C_r$$

$$\text{Corrected 2: } \tilde{X}_{pr}^{2013} = X_p^{2011} + \kappa_r$$

$$\text{Uncorrected: } \tilde{X}_{pr}^{2013} = X_p^{2011}$$

There were 37 students in 2013 and 67 students in 2011. We excluded a student who did not show up for the midterm exam in 2013. Due to the lack of institutional data, we also excluded 2 students from 2011, one who dropped out after the midterm exam and 1 visiting student, along with 2 visiting students and a graduate student from 2013. In total, we have student learning data for 33 students in 2013 and 65 students in 2011.

To understand student perceptions of the flipped classroom model, we administered a survey to all 37 enrolled students in fall 2013. An e-mail with each student's individual access link to the survey was sent out a week before the midterm exam. The survey contained both multiple choice and open-ended questions. We had a partial response rate of 100%. One student did not respond to some of the multiple choice questions, and response rates for the individual open-ended questions were

83.78%, 81.08%, 78.38% and 67.57%, respectively. We conducted a thematic analysis on student answers to open-ended questions. Finally, we also conducted class observations through the fall of 2013 to document the implementation of the flipped classroom approach.

Section 4 Enrollment numbers and student demographics

Table 1 summarizes the enrollment numbers and student demographics for both years. Although the class size in 2013 was smaller than the class size in 2011, there was no evidence that suggested that the smaller enrollment number was caused by the flipped classroom model rather than changes in schedules, competing offerings, and other factors.

The percentages of students in each year in school in 2013 were not significantly different from those in 2011. We also compared the incoming achievement test scores and grade point averages of the enrolled students in 2013 to those of the students in 2011 and found no statistically significant differences. The incoming achievement test scores we examined included the SAT I total score as well as the verbal, quantitative, and writing subscores. We operationalized grade point averages (GPAs) both as first-year GPA and as the GPA up to and including the year the students took the course.

Overall, there was no statistically significant difference between the composition of the student body in the flipped classroom and the student body in the traditional classroom. As a result, the covariates make little interpretive difference over and above the direct comparison of average test scores across years. Table 1 thus foreshadows the results of all upcoming analyses succinctly, with minimal difference

in uncorrected scores, 5-point differences in corrected scores, and 7-point differences under Correction 2.

Table 1. *Enrollment and Student Demographics in 2011 and 2013*

		Traditional(2011)	Flipped(2013)
Sample Size	Enrolled Students	67	37
	Excluded Students*	2	4
	Active Sample	65	33
Year in School	First	10(15.38%)	9(27.27%)
	Second	22(33.85%)	8(24.24%)
	Third	20(30.77%)	5(15.15%)
	Fourth	13(20.00%)	10(30.30%)
	Fifth**	0(0.00%)	1(3.03%)
SAT and GPA	SAT I(Total)	2244.0	2235.9
	SAT I(Verbal)	742.7	738.4
	SAT I(Quantitative)	747.5	747.5
	SAT I(Writing)	753.8	750.0
	GPA (First Year)	3.54	3.46
	GPA (Up to and Including The Year Students Took SW12)	3.59	3.51
Mid-term Exam Score	Corrected Total	90.6	85.5
	Corrected ID***	35.0	32.1
	Corrected Essay	55.6	53.3
	Corrected Total 2	92.9	85.5
	Corrected Essay 2	57.9	53.3
	Uncorrected Total	85.9	85.5
	Uncorrected ID	33.8	32.1
	Uncorrected Essay	52.2	53.3

* We excluded students who were missing institutional data (5), usually because they were visiting students (3), graduate students (1), or students who dropped out after the midterm exam (1). One student also took a make-up exam that was not equivalent to the primary midterm in 2013.

** Because only one student was in the fifth year, in regression analysis, we combined fourth and fifth year students into one category: fourth year and above.

***ID is short for identification questions.

Section 5 Impact of the flipped classroom on student learning

Tables 2 and 3 present unadjusted differences in average scores and proceed through a process of covariate adjustment, controlling for GPA, SAT, year in school, and combinations of covariates thereof. Table 2 presents results for Correction 1, factoring in both rater stringency and the judgmental constant intended to account for the interactive calibration process. Table 3 presents results for Correction 2, excluding the judgmental constant.

The unadjusted mean difference shown in Model 1 in Table 2 is statistically significant and sizable in magnitude in favor of the traditional classroom, at 5.2 points on a 100-point scale with a standard deviation of 6.7. Models 2 and 3 control for total SAT score, as well as first-year and cumulative GPA respectively. The adjusted mean difference is 4.8 and 4.4 points respectively, still in favor of the traditional classroom. Although cumulative GPA is a stronger predictor in Model 3 than first-year GPA is in Model 2, we proceed with first-year GPA. The cumulative GPA includes the year in which the course was taken and thus positively biases the association, and cumulative GPA tends to become less comparable across years in school and substantive concentrations.

Model 4 includes first-year GPA and SAT subscores. The evidence is consistent that first-year GPA is a stronger predictor than SAT score regardless of whether the SAT total score or SAT subscores are used. The restriction of range for SAT scores at Harvard College is a likely factor, as well as the substantive and temporal distance between the SAT exam and college courses (Geiser & Studley, 2001; Atkinson & Geiser, 2009). The substantive interpretation of the estimated causal impact of the

flipped classroom is unchanged.

Table 2. *Midterm Exam Score as A Function of the Flipped Classroom, Incoming Achievement Test Scores, Grade Point Averages and Year in School After the First Correction*

Predictor Categories	Predictors	Model1	Model2	Model3	Model4	Model5	Model6
Flipped Classroom	Flipped	-5.154***	-4.801***	-4.369**	-4.810**	-3.959**	-4.645**
		(1.46)	(1.40)	(1.37)	(1.42)	(1.42)	(1.38)
GPA	First Year GPA		5.761**		5.683**	5.913**	6.893***
			(2.05)		(2.08)	(2.05)	(1.87)
	Cumulative GPA			9.320***			
				(2.50)			
Centered SAT I Scores	Total		0.009	0.007		0.010	
			(0.006)	(0.006)		(0.006)	
	Quantitative				0.009		
					(0.014)		
	Verbal				0.013		
					(0.015)		
	Writing				0.002		
					(0.017)		
Year In School	Second Year					1.803	
						(1.933)	
	Third Year					3.131	
						(2.012)	
	Fourth and Above					-1.553	
						(2.014)	
	_cons	90.64***	70.15***	57.04***	70.43***	68.33***	66.25***
		(0.85)	(7.30)	(9.02)	(7.42)	(7.62)	(6.65)
	N	98	95	95	95	95	98
	r2	0.115	0.249	0.292	0.251	0.306	0.226
	rss	4458.5	3742.6	3528.4	3735.7	3459.5	3899
	mss	581.5	1244.1	1458.4	1251.1	1527.3	1141
	F	12.52	10.08	12.54	5.961	6.475	13.9
	df_r	96	91	91	89	88	95
	df_m	1	3	3	5	6	2
	Standard errors in parentheses						
	="* p<0.05	** p<0.01	*** p<0.001"				

Note: Cumulative GPA is the GPA up to and including the year the student took SW12.

Model 5 includes fixed effects for the year in school, and we can see that scores are higher on average for second- and particularly third-year students and slightly lower for fourth-year students. These differences are not statistically significant.

We conclude with Model 6, our final model for interpretation, that drops all nonsignificant covariates: the estimated effect is 4.7 points in favor of the traditional classroom.

Table 3. *Midterm Exam Score as A Function of the Flipped Classroom, Incoming Achievement Test Scores, Grade Point Averages and Year in School After The Second Correction*

Predictor Categories	Predictors	Model1	Model2	Model3	Model4	Model5	Model6
Flipped Classroom	Flipped	-7.444***	-7.106***	-6.680***	-7.116***	-6.202***	-6.936***
		(1.59)	(1.55)	(1.53)	(1.56)	(1.57)	(1.52)
GPA	First year GPA		5.396*		5.264*	5.511*	6.871**
			(2.27)		(2.30)	(2.27)	(2.07)
	Cumulative GPA			8.958**			
				(2.779)			
Centered SAT I Scores	Total		0.010	0.008		0.011	
			(0.006)	(0.006)		(0.006)	
	Quantitative				0.008		
					(0.016)		
	Verbal				0.020		
					(0.016)		
	Writing				-0.001		
					(0.019)		
Year In School	Second Year					1.987	
						(2.139)	
	Third Year					3.256	
						(2.226)	
	Fourth and Above					-1.831	
						(2.228)	
	_cons	92.93***	73.72***	60.62***	74.19***	71.99***	68.61***
		(0.92)	(8.07)	(10.03)	(8.19)	(8.43)	(7.36)
	N	98	95	95	95	95	98
	r2	0.186	0.291	0.324	0.295	0.344	0.271
	rss	5324.5	4573.3	4360.3	4548.2	4235.1	4768.6
	mss	1212.8	1878.1	2091.1	1903.2	2216.3	1768.7
	F	21.87	12.46	14.55	7.448	7.675	17.62
	df_r	96	91	91	89	88	95
	df_m	1	3	3	5	6	2
	Standard errors in parentheses						
	="* p<0.05	** p<0.01	*** p<0.001"				

Note: Cumulative GPA is the GPA up to and including the year the student took SW12.

Table 3 shows the same progression of models for Correction 2, without the

judgmental calibration factor C_r for rater 3. Without this correction, unadjusted and adjusted mean differences are larger, from 7.4 to 6.9 points in favor of the traditional classroom. Similarly, we find no significant effect of SAT and year in school.

We conducted three additional sets of analyses whose results we do not include for parsimony. First, we tested for interactions among variables in Tables 2 and 3 and found no notable effects, thus we do not include these results. Second, we find no significant effects for uncorrected scores across the model specifications in Tables 2 and 3, thus we do not show these results. This analysis is arguably short-sighted in its neglect of differences in rater stringency. On the other hand, the correction process that estimates differences in stringency relies on decontextualized grading, as acknowledged by the 2011 student raters who graded 2013 midterms. If the calibration process is consistent over years and successfully adjusts for rater stringency, then the unadjusted grades are arguably the most trustworthy.

Third and finally, we decomposed midterm scores by question type (ID and Essay). Table 4 in the Appendix shows one interesting interaction, where the estimated impact of the flipped classroom is negative for first-, second-, and third-years but is significantly different and closer to zero for fourth-year students and above. However, this effect is marginal, particularly in light of the large number of tests that we are running. Tables 5 and 6 show isolated results for the Essay scores and confirm that the effects are not solely due to ID or Essay scores but a combination of the two.

The midterm exam grade was not reported in terms of a letter grade in fall, 2013

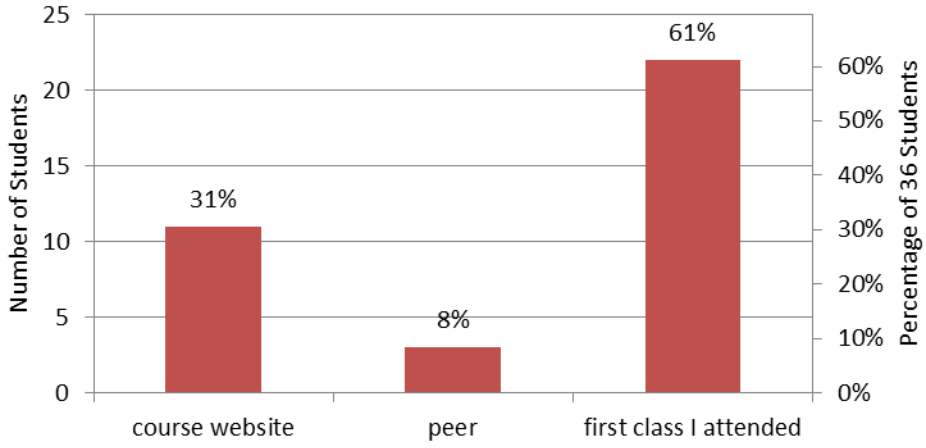
class. However, based on the letter grades assigned to midterm scores in the 2011 version of the course, we can better interpret the magnitude of the effects. The average midterm exam score in 2011 was a B. A 4.7-point average decline on this metric would drop the average score to a B-, and a 6.9-point average decline would drop the average score to a C+. Thus, the estimated average effects for corrected scores are sizable in magnitude.

Section 6 Student Perceptions of the Flipped Classroom

A. Student Preference for Pedagogical Models and Regular Sections

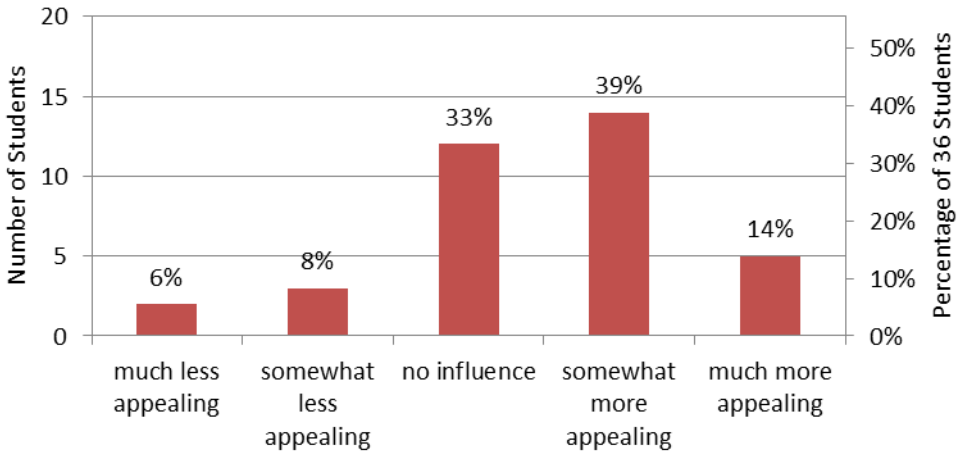
Our survey was administered to students in the 2013 course, approaching the midterm, and was intended to explore the impact of the flipped classroom on enrollment as well as general opinion. Results showed that the majority (61%) of the 36 respondents first learned that the course would use the flipped classroom approach at the first class they attended. Another 31% of the respondents learned this from the SW12 course website. Finally, 8% of the respondents heard about the flipped classroom approach from a peer.

Figure 1. Where did you first learn that the course would use this approach?

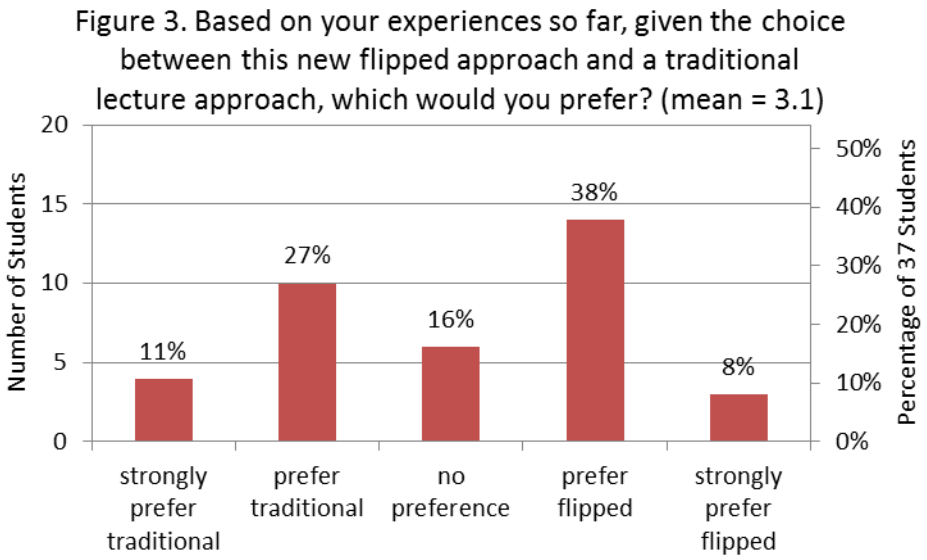


Over half (53%) of the respondents felt the course was somewhat more or much more appealing when they first heard of the flipped approach; 14% felt that the course was somewhat less or much less appealing (Figure 2). One third (12 students) reported not being influenced by the new pedagogical model. Although these data are consistent with the hypothesis that the flipped classroom had a small or, if anything, positive effect on enrollment, it is also quite possible that large numbers of students chose not to take the course because of the flipped classroom. By selecting only enrolled students instead of interested students and shoppers, the results can be expected to be biased away from revealing factors that are negative influences on enrollment.

Figure 2. When you first heard of this "flipped" approach, did it make the course seem more or less appealing to you? (mean = 3.5)



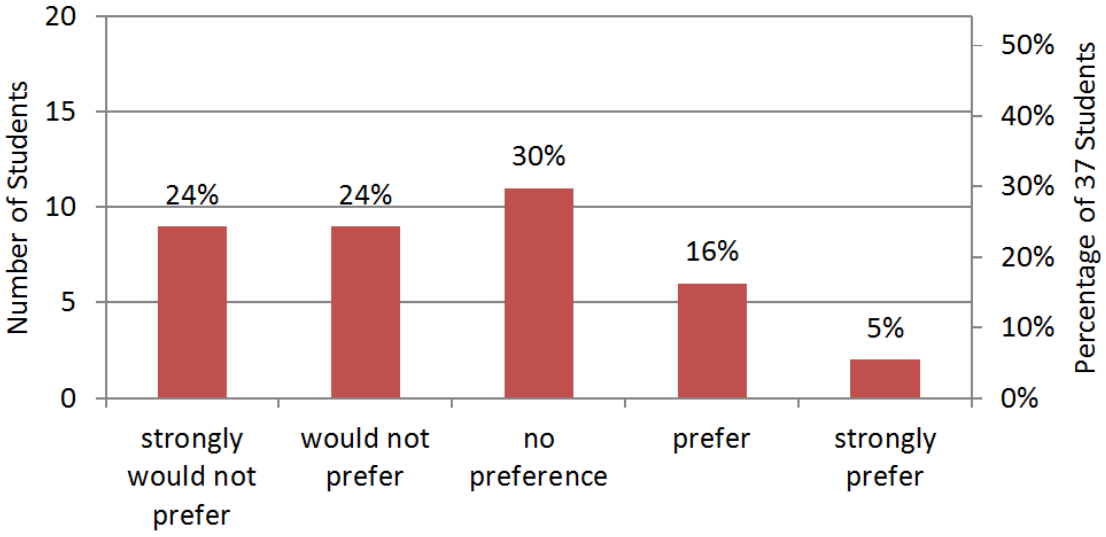
At the time of the survey, near the middle of the semester, preferences for pedagogical models varied. Nearly half (46%) of the respondents preferred or strongly preferred the flipped model, 16% expressed no preference and 38% of the respondents preferred or strongly preferred the traditional model (Figure 3).



Notably, we estimated the correlation between these responses and student midterm grades in 2013 and found no statistically significant association. Out of the 12 students whose interests in this course reported not being influenced when they first heard of the flipped approach, 5 stated that they preferred or strongly preferred the traditional approach based on their experiences so far, whereas 4 stated that they preferred the flipped approach.

In the traditional version, there were weekly small-group discussion sections led by teaching fellows; in the flipped version, there were no regular sections. 48% of the 37 respondents would prefer or strongly preferred not having required weekly sections, while 30% of the respondents had no preference. The rest (21%) of the respondents would prefer or strongly prefer to have a weekly TF section.

Figure 4. General education courses often have a required, weekly 1-hour discussion section with a teaching fellow. This course does not. Based on your experience so far, would you prefer to have required weekly sections with a teaching fellow? (mean = 2.6)



B. Student evaluation of the main components of the flipped classroom

The evaluations for the main components were varied. On average, respondents thought the online videos were very valuable to their learning (Figure 5) while the average rating for the online assessments was between somewhat valuable and valuable (Figure 6). Respondents rated the readings to be between valuable and very valuable on average (Figure 7).

Figure 5. Based on your experiences so far, how valuable are the online videos to your learning in this class? (mean = 4.0)

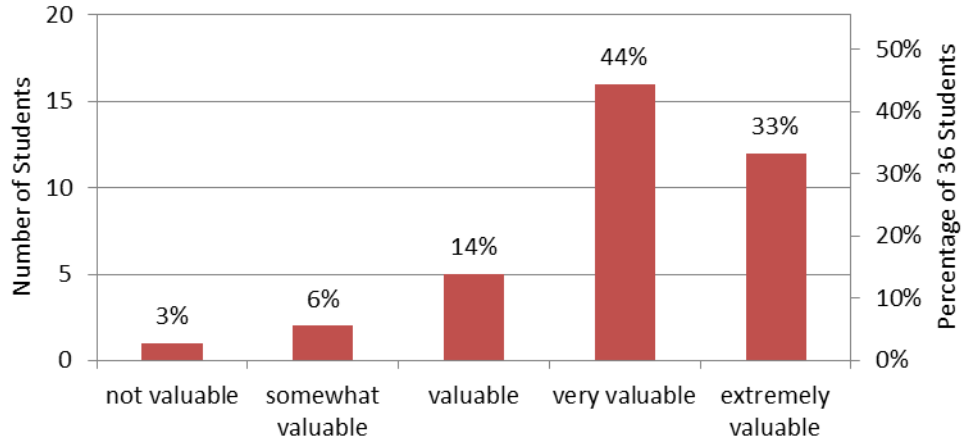


Figure 6. Based on your experiences so far, how valuable are the online assessments to your learning in this class? (mean=2.7)

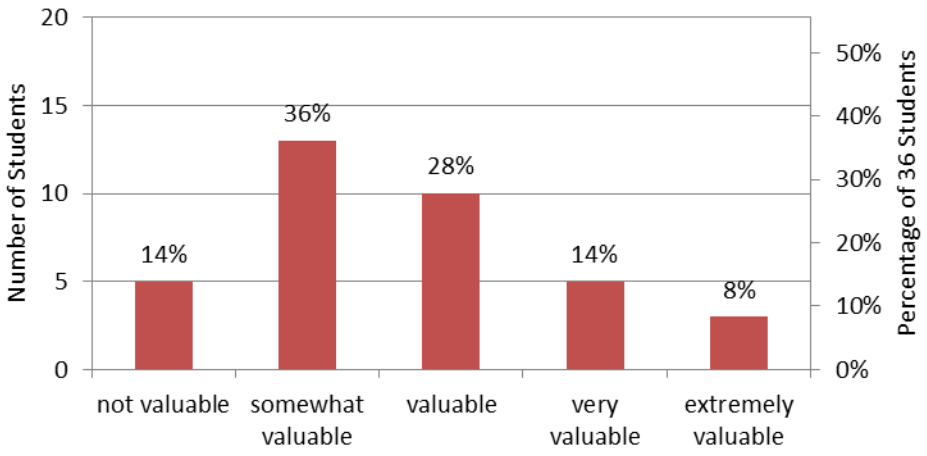
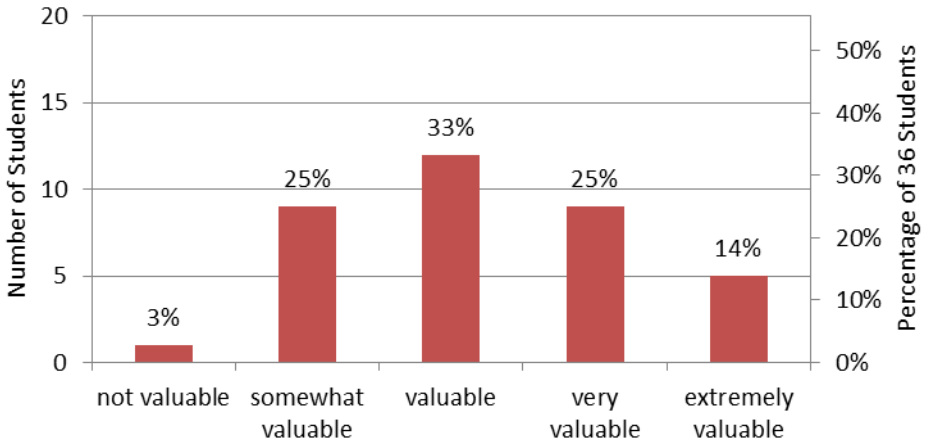


Figure 7. Based on your experiences so far, how valuable are the readings to your learning in this class? (mean = 3.2)



C. Thematic Analysis of Students' Answers to Open-ended Questions

We asked the fall 2013 students to list the ways in which the online modules or in-class activities were effective and the ways in which they could be improved. Among the 37 students, 2 students did not respond to these questions at all. 34 students commented on the online modules while 35 students commented on the in-class activities. Themes extracted from these student responses follow.

Students thought that the online modules were convenient and engaging. They were convenient because students could self-pace their learning, learn at their own

time of convenience, and break down the work because the video was in short segments. It was easy to rewind and review the videos. Videos also facilitated note-taking by allowing students to pause and having bullet points outlining important concepts. Students felt “effective and efficient” about their learning.

The modules were engaging because there were a mixture of material types and presentation styles. Sometimes an instructor lectured alone, and sometimes the instructors had conversations. The integration of multimedia and the “quasi-movie style” of the videos made them “visually stimulating”. Materials otherwise inaccessible were available through the videos. For example, scholars who studied China around Harvard were featured in the videos, and there were visits to museums and even to historical buildings in China.

Students also thought that the online modules served its practical purposes very well. They were an effective “substitute for lectures” because they were “thorough”. Modules provided a “good overview”, helped students “understand main points”, provided a variety of information and different perspectives, “deepened topics” and “summarized information”. Because modules were divided into sections, students found it easier to focus on each small topic than an hour-long lecture. One student found them fairly concise, while another student liked that they allowed students to try to interpret the contents themselves and then provide a correct version. Students generally felt that the videos provided a basis for discussion and allowed more participation in class.

Students also felt there was room for improvement. Aside from technical

problems with editing, some would like more texts, diagrams or transcripts for the videos. Students wanted video lengths, both total and for each segment, to be clearly indicated so that they could manage time better. While the purpose of dividing a module into sections is to allow students to divide up the work, there was one student who wanted to "block off" a chunk of time for the modules. Some asked for even shorter videos while others asked to provide a "double-speed" playing function. Because there were two websites for this course and the contents on these websites were not always the same, a student asked to put all materials including reading articles and photos that were "otherwise hidden" in the conventional course website into the edX modules.

In terms of the contents of the modules, two students felt that the modules were not organized and asked for greater connections among the modules. A student felt that the modules "lacked proper conclusions" to the topics in each segment. And a student felt that the interviews with other scholars were not as effective as other videos. One student noted that the modules were not effective because they were not organized while another student felt that an online module "lost the interaction and dynamism" of a face-to-face session but it was "as good as it could get." Another student mentioned that the amount of information could be "overwhelming" which made it difficult to identify critical concepts.

Students reported greatest dissatisfaction with the online assessments in the modules. Although one student said that the assessments in the modules helped students "get key points" and "solidify knowledge," multiple students commented that

assessments should allow multiple attempts because single attempt was too stressful. Some found the assessment questions were worded ambiguously purposefully. They said that sometimes the assessments failed to provide explanation to the correct answer choices. A student suggested that the assessments be put under the corresponding video clip instead of on another page. Two students went to the extreme to ask for removal of the assessments completely as they did not "help understanding" the contents.

Several students felt that the assessments focused on "minute details" of facts instead of "larger themes" or analysis. And a student asked for more written response because it could stimulate thinking more. One student thought the assessments should cover all important points from the previous videos.

A possible useful clustering of student responses is based on responses that indicated a desire to learn by reading vs. a desire to learn by videos. The former thought readings provided more detail and expressed a desire for the transcripts for the videos (these are now available). A student said videos were "a much bigger pain" than readings because when you wanted to review a small part in the video, you might need to watch a long part of or an entire video to locate it. In contrast, the latter found the videos to be more engaging than the background reading or were a great "complement to the readings".

Students offered several criteria for an effective discussion. First, discussions were effective when the majority of the class participated in the discussions. Second, effective discussions required dialogue among students. Third, effective discussions

exposed participants to a variety of perspectives that facilitated the comparison of ideas and the formation of new ideas. Fourth, when students were prepared and had the same background knowledge, discussions were effective. Fifth, effective discussions should have depth. Because debates encompassed such characteristics, several students found them effective and stated a preference for more debates.

A student mentioned that discussions were effective when discussing "thoroughly laid out issues." Another student thought the discussions were useful when they were about "overarching topics and main ideas." Two students pointed out that discussions could help students learn different ways of thinking. Students liked the discussions because they provided an opportunity for active engagement.

Some students felt that large-class discussions were beneficial, while others expressed a desire for more small-group discussions. Some believed that weekly sections could suffice for discussions because they preferred that class be used for lectures. Others felt the class size was too big for everyone to have a chance to speak. The pressure to speak was also heightened by the fact that a portion of their final grades depended on "meaningful participation" in class. One student stated that participating for the sake of grades could lead to repetitive comments from different students. Another student mentioned that he or she was not a confident public speaker. A student suggested that class time should be increased to allow more participation opportunities.

Students also had varied opinions about professor involvement in the discussions. Those who were not so keen on professor involvement asked for more dialogue

among the students and more room for students to choose the discussion questions they use. Even though the discussions were useful for emphasizing what the professor wanted the students to learn, it could also feel like the professor was fishing for correct answers from the students. There were also students who preferred professor involvement. Students felt that the professor could effectively direct and move the conversation forward and help students understand some larger themes. Students who wanted more professor involvement wanted the professor to answer more student questions, do more lecturing, and talk about his own interpretations of the history. Two students wanted the professor to cover the overview in class. One of them suggested shortening the discussion time on some materials to make room for the overview.

Other aspects of the in-class activities that could be improved include better pacing so that the class did not have to rush or skip contents towards the end of the hour. There could be more structure to the discussions so it did not "jump from point to point". One student felt it was not beneficial to spend an entire class discussing one story.

The instructors utilized cold-calling. Students recognized the merits of doing so as it "kept everyone in their game" and could induce wider participation from more students. It also could be useful at the beginning to start the discussions. However, many students disliked cold-calling because it was stressful and intimidating, and they believed it slowed down the conversation when there were volunteers. It could also stimulate "not-well-thought-out" responses which would confuse the students. A

student pointed out that people may only have comments on certain topics, and cold-calling could prevent those people to speak when they had something meaningful to contribute.

A student felt that cold-calling was stressful especially when students do not fully understand certain topics while another student felt that students were more likely to give wrong answers in discussions that centered on the readings. These suggested that some students thought students' answers were judged and they had to be correct. The discussion did not felt like an open discussion where it was perfectly acceptable to show lack of understanding or negotiate the merits of different interpretations. A student suggested that the instructors could use "positive reinforcement for participation".

We found that students' prior experience with similar pedagogy would affect how they perceive this flipped classroom. A student felt unsure about what she could improve for the in-class activities because of the lack of experience with such courses, while another student who had a case-study based course before was very sure about how important in-depth passionate, full dialogues among students were.

Students felt that discussions were effective when they involved the module contents or readings. Discussions could provide a "refined understanding" of the modules and a review of important materials or main points from the modules and readings. Discussions could also "clear up misconceptions" and deepen students' understandings of the concepts in the modules or the readings. Moreover, discussions could expand learning from the modules and readings, particularly the interesting

topics. However, two students felt that the modules should be better incorporated into the discussions. Four students explicitly stated that they liked the discussions as they were.

Several students felt that the in-class discussions were mildly effective or not effective. One student believed that class time should be used for lectures; two students felt that the discussions were "empty" and were only a time for students to fish correct answers from the professor; and one student stated that the stress created by cold calling distracted the student from focusing on the contents of the discussions.

One common complaint that students had with the flipped classroom was the heavy workload. They found the class to be time-consuming. Students felt that it took a long time to finish the modules and readings. One student described it as "2-4 hours" of homework per class. And attendance was also mandatory. One student felt that the total time spent on this class was actually "doubled or tripled." Another student felt it was more stressful than what a general education course should be.

Section 7 Discussion and Conclusion

This paper evaluates the impact of the flipped classroom in an undergraduate humanities course on the composition of the enrolled student body, midterm examination scores, and student perceptions of the flipped classroom pedagogy. Here in this section, we summarize and discuss our main findings and provide limited recommendations. Our intent is for this research to inform the future design, implementation, and research for on-campus teaching that uses open online content.

There were no statistically significant differences in the enrolled student

body in terms of incoming achievement test scores, grade point averages and the year in school. Although the enrollment size was smaller in 2013 compared to 2011, no evidence linked this to the flipped classroom model. In addition, because in-class activities were designed to be discussions and debates, the instructors may have been more stringent and selective in their marketing of the class, to ensure maximum opportunities for students to participate in learning activities. In student responses, a number of students did indicate that the class size of 37 might be too big for every student to have sufficient meaningful participation in the discussions.

Overall, we found no compelling evidence that suggested that the flipped classroom had any significant impact on midterm examination scores. **Moreover, because flipped classroom practices in a nascent stage of development, any impact on learning that we are seeing now may be a lower bound**. We used midterm exam scores to measure student learning by keeping the midterm exam fixed from 2011 to 2013. There were two types of questions on the exam: identification (ID) questions and essay questions. There was no statistically significant difference in the unadjusted midterm exam scores of students in 2013 and students in 2011 controlling for incoming achievement test scores, grade point averages and year in school.

To account for difference in rater stringency between the two years, we applied two corrections to the scores. The estimated average impact following this correction amounted to a drop from a B to a B- under Correction 1, which utilized a stringency correction and a calibration correction, and from a B to a C+ under Correction 2,

which utilized only a stringency correction. Due to the variation in the correction process, and the importance of calibration procedures in real grading contexts that represent an argument for uncorrected scores, we conclude that there is no compelling evidence in support of an average impact on midterm examination scores.

Several structural factors in the design and implementation of the flipped classroom are possible explanatory factors for student opinion and differential student learning in this transitional course administration. First, in the traditional version, there were regular small group discussion sections, led by teaching fellows, in which students were actively engaged. If these contributed significantly to learning, the flipped classroom model may not have added significantly to the number of active learning activities in which students participate. If students benefit from discussing the material with teaching fellows and not only the professor, and if the full-class discussions effectively dilute the opportunity for close conversation, then the time for active learning opportunities may not have increased across administrations.

Second, in the flipped classroom model, the online modules were used to deliver lecture content. Student opinion was varied, with average appreciation for online videos and relatively low satisfaction with online assessments. Many core technological capabilities that are now available were not available during the course administration, including video transcription and navigation functions.

Third, midterm scores may have been affected by misalignment between what was taught and what was tested. As important as it is to keep a common assessment

over time to track learning, this practice also risks decreased alignment between instruction and assessment, as instructional priorities and learning goals shift. This is a particularly important issue in a course that is transitioning to emphasize discussion-based proficiency as well as written proficiency. Teaching fellows reported relatively less coverage of content measured by some ID items, simply due to the shifting priorities of the course. From this perspective, the midterm examination score does not offer a complete measure of desired learning, because it represents an outdated and biased sample of the learning goals that currently drive instruction in the course.

Fourth, in 2011, students were required to submit a response paper about the readings before sections. Teaching fellows might ask students to address certain questions in their response papers. In the sections, teaching fellows might base discussions partly on students' response papers. In addition, they would provide feedback for these response papers. Such intensive writing exercises in 2011 might have deepened their understanding, enhanced their historical thinking abilities, and prepared them better for the essay questions.

Finally, in 2011, teaching fellows would integrate lecture content into discussions. They would lead students to think about the implicit implications of the professor's instruction in class. Sometimes they would even point these out directly if it was very difficult for students to understand on their own. In this way, teaching fellows helped students summarize and clarify their learning in the lectures. In open-ended survey responses, several students mentioned a desire for clearer

exposition of key points. Clearly, it is difficult, in a humanities course, to distinguish between a list of key points as a means to an end and the ability to list key points as an end in itself. From this perspective, if teaching fellows were making it too easy in 2011, the scores in that year may have been inflated. Distinguishing among these hypotheses requires more data and a longer-term data collection horizon. Greater attention to the assessment process is necessary to address this, including the specification of desirable outcome variables, including both traditional short-term outcomes like midterm and final exams, and longer term outcomes like course taking patterns, deciding to write a thesis, career choices, income, and participation in alumni activities.

Around midway through the semester, students' opinions about the flipped classroom model varied with 46% of the 37 participants preferring or strongly preferring the flipped model, and 38% of the participants preferring or strongly preferring the traditional model.

We have summarized the reasons why students might prefer the flipped or traditional model. For the flipped model, the online modules were convenient and engaging, more so than lectures and, for some students, readings. At the same time, it effectively substituted the lectures and allowed more time for students' active participation in class. The short video segments made it easy to digest the contents. The in-class activities provided active engagement for the students. Students could hear various perspectives to help them learn better. Discussions were effective when students came prepared with the same background knowledge and when a high

proportion of students engaged in deep dialogues. The main reasons why students preferred the traditional model could be that the workload was too heavy; the online assessments were limiting; the discussions were too shallow; and the cold-calling was stressful and could negatively affect the flow and quality of discussions. There were also students who felt that class should be used for lectures and history courses should not be discussion-based.

For professors who want to flip their classrooms with open online courses and scholars who want to study such practices in the future, we have the following recommendations.

First, equal if not more attention currently paid to the production of videos should be paid to developing assessment tools, both online and for use in the flipped classroom. Assessments should not only help students stay focused on the online modules, but also help students develop analytical ability in classroom activities. This benefits the evaluation and research enterprise as well, particularly as longitudinal and cross-sectional data collection opportunities increase, to track the progress of a course over time while charting the paths of its students after each cohort completes the course.

Second, for humanities courses that traditionally use discussion sections, it is important to carefully design the in-class active learning activities to make sure that they add more value to student learning than the regular sections did. This value is difficult to estimate, however, it seems clear that smaller, TF-led sections can be more responsive to individual student needs than a large discussion focused more on

cold-calling and the tracking and evaluation of student contributions. It might also be a choice to continue having small group sections in addition to the class meetings, but in that case, the instruction team should carefully control the workload.

Fourth, in-class activities should create an open environment for every student to participate and learn from each other. For this purpose, the instructional team should have means to incentivize and check that students come to class fully prepared.

Finally, because flipped classrooms practices with open online content are still in an early stage of adoption, we believe that the impact on learning that we estimate are likely to be a lower bound. We therefore recommend more quasi-experimental or experimental studies to investigate the impact of flipped classrooms on student learning, particularly in this time of rapid and varied adoption by instructors on campus. Over this transitional period, it is important to manage the expectations of students appropriately, both in terms of their workload and in terms of the criteria by which they will be evaluated and the norms that they are expected to follow. This is a time of rapid change for teaching practices that must align with a similar change for learning practices. Throughout, we recommend that instructors remain nimble in their ability to respond to evidence, both anecdotal and systematic, while providing a clear framework within which students can have their high expectations for learning managed and met.

References

Atkinson, R.C. & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher, 38* (9), 665–676. DOI: 10.3102/0013189X09351981

Bishop, J. L., & Verleger, M. A. (2013). The flipped classroom: A survey of research. *120th ASEE annual conference and exposition.* Retrieved April 2, 2013 from http://www.asee.org/public/conferences/20/papers/6219/view

Bowen, W. G., Chingos, M. M., Lack, K. A., & Nygren, T. I. (2012). *Interactive learning online at public universities: Evidence from randomized trials.* Retrieved May 12, 2013, from http://sr.ithaka.org/research---publications/interactive---learning---online---public---universities---evidence---randomized---trials.

Broida, J. (2001). Introductory psychology course redesign. National Center for Academic Transformation (NCAT). Retrieved August 25, 2013, from http://www.thencat.org/PCR/R1/USMe/USMe_Overview.htm.

Chen, Z. Z., Stelzer, T., & Gladding, G. (2010). Using multimedia modules to better prepare students for introductory physics lectures. *Physics Education Research, 6*(1), 101018-010108-5.

Choi, E. M. (2013). Applying inverted classroom to software engineering education. *International Journal of e-Education, e-Business, e-Management and e-Learning, 3*(2), 121---125.

Christoph, K. (2001). General chemistry course redesign. National Center for Academic Transformation (NCAT). Retrieved August 25, 2013, from

http://www.thencat.org/PCR/R1/UWM/UWM_Overview.htm

Davis, J., Odell, M., Abbitt, J., & Amos, D. (1999). Developing online courses: a comparison of web---based instruction with traditional instruction. *Society for Information Technology & Teacher Education International Conference (SITE).* Retrieved August 25, 2013, from http://www.editlib.org/index.cfm?fuseaction=Reader.ViewAbstract&paper_id=7520&from=NEWDL

Day, J. A., & Foley, J. D. (2006). Evaluating a web lecture intervention in a human---computer interaction course. *IEEE Transactions on Education, 49*(4), 420-431.

Garrison, D. R., & Kanuka, H. (2004). Blended learning: uncovering its transformational potential in higher education. *The Internet and Higher Education, 7*(2), 95-105.

Geiser, S., & Studley, R. (2001). *UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California*. Retrieved from Stanford University website: http://www.stanford.edu/~rag/ed351B/sat_study.pdf

Hamdan, N., McKnight, P., McKnight, K., & Afstrom, K. M. (2013). *A review of flipped learning*. Retrieved from Flipped Learning Network website: http://www.flippedlearning.org/cms/lib07/VA01923112/Centricity/Domain/41/LitReview_FlippedLearning.pdf

Holdhusen, M. (2009, June 14-17). Comparison of engineering graphics courses delivered via face-to-face, online, synchronous distance education, and hybrid formats. Proceedings of the Annual Meeting of the American Society for Engineering

Education. Retrieved May 18, 2013, from http://search.asee.org/search/fetch?url=file%3A%2F%2Flocalhost%2FE%3A%2Fsearch%2Fconference%2F19%2FAC%25202009Full1065.pdf&index=conference_papers&space=129746797203605791716676178&type=application%2Fpdf&charset

Lack, K. A. (2013). *Current status of research on online learning in postsecondary education.* Retrieved May 12, 2013, from http://sr.ithaka.org/research---publications/current---status---research---online---learning---postsecondary---education

Lewis, J. S., & Harrison, M. A. (2012). Online delivery as a course adjunct promotes active learning and student success. *Teaching of Psychology, 39*(1), 72-76.

Lovett, M., Meyer, O., & Thille, C. (2008). The Open Learning Initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media in Education.* Retrieved May 25, 2013, from http://jime.open.ac.uk/2008/14

Maki, W. S., & Maki, R. H. (2002). Multimedia comprehension skill predicts differential outcomes of web---based and lecture courses. *Journal of Experimental Psychology Applied, 8*(2), 85-98.

Marcey, D. J., & Brint, M. E. (2012). Transforming an undergraduate introductory biology course through cinematic lectures and inverted classes: a preliminary assessment of the CLIC model of the flipped classroom. *Biology Education Research Symposium at the meeting of the National Association of Biology Teachers.* Retrieved August 25, 2013, from

https://www.nabt.org/websites/institution/File/docs/Four%20Year%20Section/2012%20Proceedings/Marcey%20&%20Brint.pdf

Mason, G. S., Shuman, T. R., & Cook, K. E. (2013). Comparing the effectiveness of an inverted classroom to a traditional classroom in an upper---division engineering course. *IEEE Transactions on Education.* Retrieved August 25, 2013, from http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6481483&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6481483

McLaughlin, J. E., Roth, M. T., Glatt, D. M., Gharkholonarehe, N. , Davidson, C. A., Griffin, L. M., Esserman, D. A., & Mumper, R. J. (2013). The flipped classroom: A course redesign to foster learning and engagement in a health professions school. *Academic Medicine, 89(2),* 236-243.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2010). Evaluation of evidence---based practices in online learning: a meta analysis and review of online learning studies. *U. S. Department of Education Office of Planning, Evaluation, and Policy Development Policy and Program Services* S*tudy*. Retrieved July 20, 2013, from http://www.2.ed.gov/rschstat/eval/tech/evidence---based---practice/finalreport.pdf.

Moravec, M., Williams, A., Aguilar---Roca, N. & O'Dowd, D. K. (2010). Learn before lecture: a strategy that improves learning outcomes in a large introductory biology class. *CBE—Life Sciences Education, 9*, 473–481.

Pierce, R. & Fox, J. (2012). Vodcasts and active-learning exercises in a "flipped classroom" model of a renal pharmacotheraphy module. *American Journal of*

Pharmaceutical Education, 76(10): 196.

Popyack, J. (2003). Computer programming course redesign. National Center for Academic Transformation (NCAT). Retrieved August 25, 2013, from http://www.thencat.org/PCR/R3/DU/DU_Overview.htm

Reasons, S. G., Valadares, K., & Slavkin, M. (2005). Questioning the hybrid model: student outcomes in different course formats. *Journal of Asynchronous Learning Networks*, *9*(1), 85-98.

Riffell, S., & Sibley, D. (2005). Using web---based instruction to improve large undergraduate biology courses: an evaluation of a hybrid course format. *Computers & Education*, *44*(3), 217-235.

Sadaghiani, H. R. (2012). Online prelectures: an alternative to textbook reading assignments. *Physics Teacher, 50*(5), 301-303.

Stone, B. B. (2012). Flip your classroom to increase active learning and student engagement. *28th annual conference on distance teaching and learning.* Retrieved August 25, 2013, from http://www.uwex.edu/disted/conference/Resource_library/proceedings/56511_2012.pdf

Upchurch, R. H., & Hartman, J (2001). American national government course redesign. National Center for Academic Transformation (NCAT). Retrieved August 25, 2013, from http://www.thecat.org/PCR/R1/UCF/UCF_Overview.htm

Zhao, Y.R., & Breslow, L.(2013). *Literature review on hybrid/blended learning*. Unpublished manuscript, HarvardX, Harvard University, Cambridge, MA.

Appendix

Table 4. *ID Question Score as A Function of the Pedagogical Model, SAT I scores, First year GPA and Year in School After Correction*

Predictor Categories	Predictors	Model1	Model2	Model3	Model4	Model5	Model6
Flipped Classroom	Flipped	-2.903***	-2.658**	-2.427**	-2.659**	-1.982*	-3.907*
		(0.84)	(0.83)	(0.82)	(0.84)	(0.81)	(1.66)
GPA	First Year GPA		2.752*		2.730*	2.781*	3.006*
			(1.22)		(1.24)	(1.17)	(1.17)
	Cumulative GPA			4.711**			
				(1.492)			
Centered SAT I Scores	Total		0.004	0.003		0.005	0.006
			(0.003)	(0.003)		(0.003)	(0.003)
	Quantitative				0.004		
					(0.008)		
	Verbal				0.006		
					(0.009)		
	Writing				0.002		
					(0.010)		
Year In School	Second					1.643	0.476
						(1.102)	(1.432)
	Third					2.329*	1.871
						(1.147)	(1.420)
	Fourth and Above					-1.453	-3.679*
						(1.148)	(1.553)
Interactions Between Year In School And Flipped Classroom	Secxflip						2.416
							(2.252)
	Thirdxflip						-0.448
							(2.410)
	Fourxflip						4.746*
							(2.230)
	_cons	35.04***	25.22***	18.03**	25.30***	24.13***	24.32***
		(0.49)	(4.34)	(5.39)	(4.41)	(4.34)	(4.37)
	N	98	95	95	95	95	95
	r2	0.111	0.198	0.237	0.199	0.318	0.368
	rss	1482.7	1320.7	1257.1	1319.9	1124.2	1041.1
	mss	184.5	326.6	390.2	327.4	523.1	606.2
	F	11.94	7.501	9.416	4.415	6.825	5.499
	df_r	96	91	91	89	88	85
	df_m	1	3	3	5	6	9
	Standard errors in parentheses						
	="* p<0.05	** p<0.01	*** p<0.001"				

Note: Cumulative GPA is the GPA up to and including the year the student took SW12.

Table 5. *Corrected Essay Score as A Function of the Pedagogical Model, SAT I score, GPA and Year in School*

Predictor Categories	Predictors	Model1	Model2	Model3	Model4	Model5	Model6
Flipped Classroom	Flipped	-2.252**	-2.143*	-1.941*	-2.152*	-1.978*	-1.985*
		(0.84)	(0.82)	(0.81)	(0.83)	(0.86)	(0.81)
GPA	First Year GPA		3.010*		2.953*	3.132*	3.610**
			(1.20)		(1.22)	(1.24)	(1.09)
	Cumulative GPA			4.609**			
				(1.481)			
Centered SAT I Scores	Total		0.004	0.004		0.005	
			(0.003)	(0.003)		(0.003)	
	Quantitative				0.005		
					(0.008)		
	Verbal				0.007		
					(0.009)		
	Writing				0.000		
					(0.010)		
Year In School	Second Year					0.16	
						(1.171)	
	Third Year					0.802	
						(1.219)	
	Fourth and Above					-0.100	
						(1.220)	
	_cons	55.60***	44.92***	39.01***	45.13***	44.21***	42.82***
		(0.49)	(4.27)	(5.34)	(4.34)	(4.62)	(3.89)
	N	98	95	95	95	95	98
	r2	0.0694	0.189	0.216	0.191	0.196	0.165
	rss	1487.8	1281.2	1238	1278	1270.3	1334.4
	mss	111	298.6	341.8	301.8	309.5	264.4
	rmse	3.937	3.752	3.688	3.789	3.799	3.748
	F	7.159	7.07	8.375	4.204	3.574	9.413
	df_r	96	91	91	89	88	95
	df_m	1	3	3	5	6	2
	Standard errors in parentheses						
	="* p<0.05	** p<0.01	*** p<0.001"				

Note: Cumulative GPA is the GPA up to and including the year the student took SW12.

Table 6. *Essay Score as A Function of the Pedagogical Model, SAT I score, GPA and Year in School After The Second Correction*

Predictor Categories	Predictors	Model1	Model2	Model3	Model4	Model5	Model6
Flipped Classroom	Flipped	-4.541***	-4.448***	-4.253***	-4.457***	-4.220***	-4.276***
		(1.02)	(1.01)	(1.01)	(1.02)	(1.06)	(0.99)
GPA	First Year GPA		2.644		2.534	2.729	3.588**
			(1.48)		(1.50)	(1.53)	(1.34)
	Cumulative GPA			4.247*			
				(1.84)			
Centered SAT I Scores	Total		0.006	0.005		0.006	
			(0.004)	(0.004)		(0.004)	
	Quantitative				0.004		
					(0.010)		
	Verbal				0.014		
					(0.011)		
	Writing				-0.003		
					(0.012)		
Year In School	Second Year					0.343	
						(1.442)	
	Third Year					0.928	
						(1.501)	
	Fourth and Above					-0.377	
						(1.502)	
	_cons	57.89***	48.50***	42.60***	48.89***	47.87***	45.19***
		(0.59)	(5.27)	(6.63)	(5.33)	(5.68)	(4.79)
	N	98	95	95	95	95	98
	r2	0.172	0.25	0.267	0.256	0.258	0.23
	rss	2172.5	1946.3	1902.8	1929.3	1925.8	2020.9
	mss	451.3	648.6	692.1	665.6	669.1	602.9
	rmse	4.757	4.625	4.573	4.656	4.678	4.612
	F	19.94	10.11	11.03	6.14	5.096	14.17
	df_r	96	91	91	89	88	95
	df_m	1	3	3	5	6	2
	Standard errors in parentheses						
	="* p<0.05	** p<0.01	*** p<0.001"				

Note: Cumulative GPA is the GPA up to and including the year the student took SW12.