The impact of the flipped classroom in a principles of microeconomics course: evidence from a quasi-experiment with two flipped classroom designs

Chiara Lombardini[a],* , Minna Lakkala[b] , Hanni Muukkonen[c]

[a]Department of Economics and Management, Faculty of Agriculture and Forestry, P.O. Box 27 (Latokartanonkaari 5), 00014 University of Helsinki, Finland

[b] Faculty of Educational Sciences, P.O. Box 9 (Siltavuorenpenger 1A), 00014 University of Helsinki, Finland

[c] Faculty of Education, University of Oulu, P.O. Box 2000, 90014, Finland

∗ Corresponding author at: Department of Economics and Management, Faculty of Agriculture and Forestry, P.O. Box 27 (Latokartanonkaari 5), 00014 University of Helsinki, Finland. E-mail addresses: chiara.lombardini@helsinki.fi (C. Lombardini), minna.lakkala@helsinki.fi (M. Lakkala), hanni.muukkonen@oulu.fi (H. Muukkonen)

Abstract

This study uses a quasi-experimental, non-equivalent group design to analyze the outcomes in terms of students' learning and satisfaction of the redesign of a first-year, principles of microeconomics course from a lecture-based course using active learning techniques in 2013 to a partial flipped classroom in 2014 and a full flipped classroom in 2015.

Students perceived a higher degree of achievement of the learning goals in both flipped courses compared to the non-flipped active learning course. Moreover, participating in the partial or full flipped classroom decreased the odds of a D or F grade or of withdraw. However, only the partial flip was associated with overall better learning outcomes in the final exam, while there was no statistically significant difference between the non-flipped active learning course and the full flip. Age was negatively associated with learning outcomes and increased the odds of a D or F grade or of withdraw. Gender had no statistically significant impact on learning outcomes. Students were least satisfied with the full flip and equally satisfied with the non-flipped active learning course and the partial flip. Lower satisfaction appears to be due to increased workload, which students evaluated to be highest in the full flip, as well as to elements of group work design. In the flipped classroom design, the pre-class multiple choice tests on Moodle emerged as a clear favorite in students' teaching evaluations.

Keywords: flipped classroom, flipped teaching, inverted classroom, regression analysis, binary logit, non-parametric statistics

JEL codes: A22

1. Introduction

In economics teaching, traditional lecturing still takes up the largest share of class time, an estimated 60 % (Goffe and Kauper 2014) to 83 % (Watts and Shaur 2011). This central role of lecturing is under increasing scrutiny as empirical evidence suggests that active learning techniques are more effective than lecturing in promoting learning (see e.g. Freeman et al. 2014).

A pedagogical approach that decreases lecturing, thus freeing class time for active learning, is the inverted classroom (Lage et al. 2000) or the classroom flip (Baker 2000), whereby first exposure to the material is moved outside the classroom usually in the form of lecture videos (Abeysekera and Dawson 2014). The classroom flip appears to improve learning outcomes as reported in three major reviews of the existing literature (Bishop and Verleger 2013, Giannakos et al. 2014, O'Flaherty and Phillips 2015, O'Flaherty et al. 2015) even though reviewers express some concern for the lack of a "*robust scientific approach*" in evaluating these learning outcomes (O'Flaherty and Phillips 2015, 89). Recent studies using more robust methods, seem to confirm the existence of improvement in learning outcomes from flipping the classroom, albeit moderate (Anderson and Brennan 2015, Calimeris and Sauer 2015).

Moving beyond the analysis of how flipping the classroom affects learning outcomes overall, some researchers have begun to investigate more fine-grained questions. Ryan and Reid (2016) asked how flipping the classroom affects the outcomes of weaker students. They found a 56% reduction in Ds and Fs grades and in the withdrawal rate when flipping the classroom even though no improvements in learning emerged at the aggregate level. Touchton (2015) focused on what type of learning flipping affects the most. He found a larger improvement in learning outcomes in the sections which students generally find most challenging even though at the aggregate level the improvements were very small. Olitsky and Cosgrove (2016) examined whether gains in learning outcomes become larger as students adapted to the flip and showed that this was the case: the gains in learning increased as the flipped course progressed and students became better acquainted with the approach. Jensen et al. (2015) raised the question of how much the impact of the flipped classroom on learning outcomes depends on the choice of the control against which the learning outcomes of the flip are evaluated. They found

insignificant learning benefits of the flipped classroom compared with a non-flipped, active learning course.

On the costs of flipping the classroom, little is said in the literature. Olitsky and Cosgrove (2016) suggested that blended classroom flips can help save resources with no negative impacts on learning compared to a moderately blended class because they allow to decrease face-to-face class time. However, they did not take into account the extra time needed to develop the course online materials such as video lectures. McPherson and Bacow (2015), on the other hand, argued that flipping the classroom is most unlikely to help reduce costs as in current "traditional courses" lectures are relatively cheap while the major costs come from the staff and physical space needed for the discussion sections and laboratories. They suggested that one could obtain real savings if face-to-face discussion sessions, not lectures, could be substituted with interactive sessions run by technology

In this study, we further explore the questions raised by Jensen et al. (2015) and Ryan and Reid (2016) and compare the outcomes a lecture-based microeconomics principles course which makes use of active learning techniques with two flipped course designs. We examine the following main research questions:

1. How did the two flipped course designs impact the learning outcomes and likelihood of a D or F grade or withdraw compared to the non-flipped active learning course?
2. How did the two flipped course designs impact students' perceived learning and teaching evaluations (satisfaction, workload, perceived difficulty) compared to the non-flipped active learning course?

We use linear multiple regression and binary logistic regression models as well a non-parametric statistics to examine these impacts. The costs of inverting the classroom is also briefly discussed. The courses' re-design as a flipped classroom and the related learning outcomes are analyzed following the research-based design approach (Edelson 2002). Design-based research integrates

empirical educational research with theory-driven design of learning settings. If focuses on how to implement pedagogical practices in authentic educational contexts, and simultaneously develop new theoretical insights (Design-Based Research Collective 2003). Typical for design-based research is to include successive and iterative phases of research and design: the design of educational settings is based on prior models and theories and results are used to develop both theories and successive implementations of the pedagogical methods (Design-Based Research Collective 2003; Cobb et al. 2003, Cobb et al. 2015). In practice, the course objectives, implementation, and assessment are developed in an iterative fashion through the refinement of pedagogical design and the collection of empirical evidence on learning outcomes. The remainder of this article is organized as follows. Section two describes the materials and methods. Section three presents the results while section four discusses the results and concludes.

2. Materials and methods

2.1 The non-flipped active learning and flipped classroom course designs

The flipped classroom was developed from the non-flipped, active learning microeconomics section of a principles of economics course with 157 students enrolled taught in 2013. The non-flipped, active learning course included both micro- and macroeconomics and lasted for a semester. The microeconomics section of the course ended with a midterm exam. It included thirteen 90-minutes' classes bi-weekly in the first period of the fall term for a total of 13 class meetings. Approximately two thirds of class time was devoted to lecturing and one third to active learning such as think-pair-share and clicker questions. Continuous exposition by the lecturer was interspersed with active learning tasks so that uninterrupted lecturing did not exceed segments of 20-25 minutes. Students had to hand in three exercise sets as post-class assignments. There were no pre-class assignments. Students were provided with PowerPoint lecture notes which closely followed the textbook, Mankiw's and Taylor's Economics.

In 2014, the principle of economics course (N=146) was split into two separate courses: principles of microeconomics and principles of macroeconomics. The former was redesigned as a flipped classroom without any changes to the amount of hours of in-class instruction, the learning objectives, the schedule, the topics and their order of presentation, the textbook, and

PowerPoint slides. Lectures were moved outside class and offered on Moodle as lecture videos of length varying from 4 to 12 minutes. The lecture videos, produced by the instructor, followed closely the lecturing done in class in previous year. Students were required to prepare for class by watching the videos and/or reading the corresponding chapters in the textbook and then answering a multiple-choice test on Moodle. For a passing grade, students had to answer at least seven out of ten Moodle tests with a score of at least 80% right answers. The Moodle test could be repeated up to four times before the deadline. Students were strongly encouraged to post any queries about the video-lecturers and the Moodle tests on a discussion board to be then addressed in class.

Class time was devoted to pair and group activities, to lecturing tailored to answering students' questions posted on the discussion board, and to different kinds of exercises and discussions. Tailored lecturing did not take more than one third of class time. In terms of Bloom's taxonomy of cognitive learning (Bloom 1956), the pre-class activities focused on remembering, understanding and to some degree applying. The class activities instead focused more on applying and analyzing, although, especially at the beginning of class, remembering and understanding were tested by clicker questions. In the first, partial flip, students could choose to participate in group-work, which took place for the most part outside class with presentations in class. The group-work consisted of two assignments: one on sin taxes and one on market power in the retail food sector, two topics chosen for their relevance to the focus of the faculty of Agriculture and Forestry, where the course was taught. The group-work assignments required students to make connections between a wide range of concepts and models presented in the course and to use them to analyze and evaluate economic instruments and policies thus focusing on the highest levels of Bloom's taxonomy. Students were also asked to evaluate the group work reports of two other groups for the first assignment. Of the 108 students who took the final exam, 61 (56%) chose to participate in the group work.

In the flipped course in 2015 (N = 117), the full flip, group work was made compulsory for all students but more time was dedicated in class to group work, and the number of group work tasks was reduced from two to one: the same assignment on sin taxes as the previous year. These latter two changes were made based on the students' teaching evaluations of the first flip. In-

class lecturing was further reduced compared to the previous flip from one third to about one fifth of class time.

Table 1 presents the building elements of the three courses and how each element contributed to the final grade. In the flip courses, students could gain some points for the final grade by completing satisfactorily the pre-class tests on Moodle. This gave student an incentive to come to class prepared and thus take the most advantage of in-class activities.

Table 1. Assessment of the 2013 non-flipped active learning course and the 2014 and 2015 flips

	2013	2014		2015
		with group work	without group work	
Entry-test	No entry test	Did not give points to the final grade	Did not give points to the final grade	4.5 points for taking the test
Video-lectures	No video lectures	Video lectures available.	Video lectures available	Video lectures available
Moodle tests (10 tests)	No Moodle tests	Max 30 points (3 p/test), a minimum of 7 tests required	Max 15 points (1.5 p/test), a minimum of 7 tests required	Max 20 points (2 p /test), a minimum of 7 tests required
Group-work	No group work	Max. 30 points (15p/group-work, 2 group works)		Max 19 points, one group work
Peer-evaluation of group works	No peer evaluation	Max. 8 points		Required, did not give points to the final grade
Post-lecture exercise sets	Three compulsory exercise sets, half of the exercises had to be done satisfactorily, the sets gave no points to the final grade	No exercise sets	No exercise sets	No exercise sets
Final exam	Max. 100 points, a pass required at least 50 points.	Max. 32 points, a pass requires at least 8 points in the final exam	Max 85 points	Max 50 points, a pass required at least 25 points in the final exam
Attendance	No attendance requirements	A minimum of 8 classes out of 13, did not give points to the final grade	A minimum of 8 classes out of 13, did not give points to the final grade	Max 6.5 points (0.5 p /class)

Other elements of the flipped classroom were also incentivized. Given the different way the course grades were formed, in this study learning outcomes were compared based on the using the final exam scores and not the final grade.

2.2 Data description

2.2.1 Demographic data

We collected data on students' age, major, gender, and enrollment year. A chi-square test showed that the students' populations did not differ significantly in gender (χ^2 (2) = 5.19, p = .075), in the major being an economics or non-economics one (χ^2 (2) = .772, p = .680), or in freshmen status (χ^2 (2) = 1.306, p = .521). Table 2 summarizes the key features of the three student populations. Moreover, a Kruskall-Wallis test lead to retain the null hypothesis that the distribution of students' age (KW = .123, p = .940) and of the number of years students had been enrolled at the university at the beginning of the course (KW =.469, p = .791) was the same in the three student populations.

Table 2. The students' populations by gender, freshman status, major, age and years of enrollment at the university in the three courses.

	2013	2014	2015	**p**
N	157	146	117	
Females	56 %	58.9%	69.2 %	.075
Freshman	52.8 %	47.3 %	59 %	.521
Economics major	74. 2%	78.1 %	77.8 %	.680
Average age (Mdn)	23.67 (23)	24 (22)	23.66 (22)	
Average number of years enrolled (Mdn)	1.53 (0)	1.41 (0)	1.44 (0)	

2.2.2 Students' teaching evaluations

Students' teaching evaluations were collected anonymously from students at the end of each course in conjunction with the final exam. Thus students did not know their final grade but could evaluate whether the exam was aligned with the learning objectives and the implementation of the course. In order to guarantee full anonymity, no information about the gender, age, freshmen status, year of enrollment or major of the respondents was asked. Most items in the students' teaching evaluation survey asked for the degree of agreement with different statements using the Likert scale 1 = 'totally disagree', 2 = 'disagree', 3 = 'neither agree nor disagree', 4 = 'agree', 5 = 'totally agree'. The survey also included some open questions as well as questions in which students were asked to evaluate how large a percentage of class meetings they had attended, how large a percentage of lecture videos they had watched and to give the course a grade.

2.2.3 Entry test, pre-class tests and the final exam

In 2013, no entry test data was collected. In 2014, the first 15 questions of the standardized test of microeconomics knowledge TUCE (Walstad et al. 2007) were administered as an entry test on the first class. Overall 81 students out of the 146 enrolled (55%) took the test. In 2015, students were required to take the entry test in the course Moodle page before the second class of the course. The test included all 30 questions of the standardized test of microeconomics knowledge TUCE. Students could answer the test only once and were given the same amount of time per question as the previous year. We chose to have the entry test taken on Moodle rather than in class as this ensured that more students enrolled took the entry test. This came, however, at the cost of less perfect monitoring. Thus it is difficult to assess whether the higher average score in the TUCE in 2015 compared to 2014 may be due to students getting help from others or consulting learning materials when answering the TUCE online. On the other hand, as the number of right answers on the TUCE entry test did not affect the final grade there was no grade-incentive to cheat on the Moodle TUCE test. In 2015, 97 out of 117 students took the TUCE on Moodle (83%).

For both flipped courses, we also collected data on the results of the multiple choice, pre-class tests on Moodle. We evaluated the learning outcomes using the final exam scores rather than the course grade to ensure better comparability across the three courses, given the different way the course grades were formed (see Table 1). Note that although the final exams were designed to examine the development of the same learning objectives, they were not identical. This is due to the fact that exam questions and answer keys are made available to students after the exam, which limits the possibility of using the same exams multiple times. Table 3 presents the summary statistics for performance in the TUCE entry test and in the final exam

Table 3. Summary statistics for performance in the entry test and in the final exam

	2013	2014	2015
N enrolled	157	146	117
TUCE **%** score			
N (% enrolled)	No test for 2013	81 (55)	97 (83)
Mean (SE)		50.30 (1.83)	58.85 (1.47)
Median		50	60
Std. Deviation		15.35	12.29
Final exam **%** score			
N (% enrolled)	122 (77.7)	108 (74)	82 (70)
Mean (SE)	69.16 (1.46)	74.94 (1.45)	65.95 (1.77)
Median	71	74	68.5
Std. Deviation	16.17	15.04	16.04

We examined the correlation between participation to group work and score on the TUCE entry test as well as between participation to group work and score in the final exam using Kendall's tau-b. Point biserial correlation could not be calculated as the assumption of normality and homogeneity of variance were not met by the data. There was a weak, negative association between TUCE score and participation in group work (τ_b = -.047, p = .506) as well as between percentage score in the final exam and participation in group work (τ_b = -.041, p = .558). However, these correlations were not statistically significant. No significant correlation was

found between score on the TUCE entry test and score in the final exam with Spearman rho being $r_s(140) = .106, p = .211$.

2.3 Data analysis methods

All the statistical analyses were conducted with the SPSS Statistics 23 software. Using linear regression analysis, we estimated the effect of course design on learning outcomes. The first linear model was used to predict student *i*'s performance in the final exam measured as a percentage of right answers, $PERSCORE_i$ using data from all three courses

$$PERSCORE_i = \beta_0 + \beta' X_i + \varepsilon_i \qquad (1)$$

where $X \in \{age, economics\ major, female, flipped14, flipped15, freshman, years\ enrolled\}$ and with Flipped14 and Flipped15 being two dummy variables, which took value 1 for the 2014 and 2015 flipped courses respectively and zero otherwise.

Using the richer data we had for the flipped courses, we estimated a second linear model to better assess the impact of the two flipped course designs on learning. In the second model, student *i*'s performance in the final exam of the flipped courses of 2014 and 2015 measured as a percentage of right answers, $PERSCORE_{flipped,i}$ is

$$PERSCORE_{flipped,i} = \beta_0 + \beta' X_i + \varepsilon_i \qquad (2)$$

where $X \in \{age, economics\ major, female, flipped14, freshman, Moodle\ test\ average\ percentage\ score, participation\ to\ group\ work, TUCE\ entry\ test\ percentage\ score, years\ enrolled\}$ with significance levels: * = p < 0.1, ** = p < 0.05, *** = p < 0.01.

We also analyzed the impact of course design on weaker students by estimating a binary logit model to help identify which factors affected the likelihood of a student getting a D or F grade or withdraw, hereafter DFW. For computing the D and F grades we adapted the evaluation scale used in our courses to the one used in the US. In US higher education, a D grade generally corresponds to a percentage of 60-69 % right answers and a fail to any percentage between 0 and 59%. In our curriculum, a failing grade corresponds to any percentage below 50%. For percentages between 50 and 59, students get the lowest grade, that is, grade one, and for

percentage between 60 and 69, grade two. For better comparability of the scores across years, we used the percentage score in the final exam or first retake rather than overall course percentage score. In the binary logit regression, the dependent variable was coded 1 if the student got a percentage score in the final exam or retake below 70, that is, a D, F or no score, this latter meaning that he had dropped out, and 0 otherwise. The results of the binary logit regressions were used to predict the probability of DFW, given specific student's characteristics, namely gender, age, being an economic major, being a freshman, years enrolled as well as of course design, using the dummy variables Flipped14 and flipped15.

In order to compare the differences in perceived learning, satisfaction with the course as well as in the evaluation of how well the different elements of course design supported students learning across the three courses, we analyzed the students' teaching evaluations. For all answers in the students' teaching evaluations using a Likert scale, we first tested the null hypothesis of normal distribution of each answer. As this was rejected for all Likert-items based on the Shapiro-Wilk test, we tested the null hypothesis of the equality of distribution of the answers to each Likert-item across the three years using the Kruskal-Wallis test for nonparametric data. When the main p value indicated that within our data set at least two years differed from each other, we performed pairwise comparisons. There is some disagreement as to whether one should do the pairwise comparisons by using the Dunn's procedure or by running multiple Mann-Whitney U tests – one for each pairwise comparison – using Bonferroni's correction for multiple comparisons. Since Dunn's (1964) procedure uses data from all three groups when making each pairwise comparison while the Mann-Whitney U tests uses only the data from the two groups being compared, these two methods can lead to different results. To check the robustness of our results we used both Dunn's and Mann-Whitney U tests with and without the Bonferroni correction.

3 Results

3.1 Learning outcomes and teaching method: performance in the final exam

Table 4 shows the regression results from model 1 with data from all three courses: The partial flip (Flipped2014) has a statistically significant, positive impact on performance in the final

exam, while both age and being an economics major have a significant, negative impact. No impact of gender, being a freshman, years of enrollment or of the full flip was found.

Table 4 Summary of linear regression results for model 1 with the dependent variable percentage score in the final exam (All three courses, N=310)

	Unstandardized coefficents	Std. Error	t	p	VIF
(Constant)	91.317	5.581	16.363	.000	
age	-.660***	.222	-2.979	.003	1.181
yearsenrolled	-.723	.607	-1.190	.235	1.980
1= female	-1.997	1.819	-1.098	.273	1.031
1 = econ major	-5.002*	2.712	-1.844	.066	1.612
1=freshman	-.859	2.708	-.317	.751	2.299
1=flipped 15	-3.196	2.239	-1.428	.154	1.275
1=flipped14	5.767**	2.046	2.818	.005	1.244

Model's 1 unstandardized coefficients suggest that an increase in age by one year is associated with a score in the final exam lower by 0.66 percentage points. Being an economics major is associated with a score lower by 5 percentage points while being in the partial flip is associated with a score higher by 5.8 percentage points.

In model 2, the sample size is reduced to N=73 from the N= 263 of students enrolled in the partial and full flip together. This is due to the fact that we included only those students who had both participated in the TUCE entry exam and in the final exam. This subset is not a random selection of the wider group of students participating in the course. In fact, among the excluded from the sample, less motivated or weaker students may be over-represented as these tend to be those who dropped out of the course, were not in class when the TUCE was administered in 2014, or skipped the TUCE Moodle test in 2015. Thus caution should be exercised when comparing the results of models 1 and 2 due to this possible selection bias.

The regression results for model 2 are presented in Table 5. As in the first model, the flipped course design of 2014, the partial flip, is still associated with better learning outcomes compared to the full flip even when we control for entry level. Model 2 suggests that attending the course with the partial flip (1 = flipped14) was associated with a score higher by 23 percentage points compared to the total flip and that an increase of one percentage point in the TUCE score was associated with a .243 percentage point increase in the final exam score. Instead, participation to group work, the average score on the Moodle test, gender, years of enrollment or freshman status had no statistically significant impact on learning outcomes. Unlike model 1, age too was not significant anymore as p = .106 is just above the 10 % cutoff value.

Table 5 Summary of linear regression results for model 2 with dependent variable percentage score in the final exam (Flipped courses only, N=73)

	Unstandardized coefficents	Std. Error	t	p	VIF
(Constant)	34.06	44.441	.766	.446	
age	-.62	.380	-1.639	.106	1.300
yearsenrolled	-1.52	1.349	-1.130	.263	2.462
1= female	-.065	3.322	-.020	.984	1.083
1 = econ major	-3.14	5.216	-.603	.549	1.632
1 = freshman	-3.30	4.980	-.663	.509	2.461
1 = flipped14	23.05**	10.326	2.232	.029	10.928
TUCE_perc	.243*	.126	1.922	.059	1.342
Moodle_av_score	2.929	3.238	.905	.369	10.170
1= groupwork	3.192	6.295	.507	.614	1.226

Adjusted R^2=.208, F = 3.100 (.004)

3.2 Likelihood of getting a D or F grade or withdraw

We examined which factors affect the likelihood of a D or F grade or withdraw (DFW) by running a binary logistic regression using the same variables as for the first linear regression. The results are summarized in Table 6. The Hosmer and Lemeshow test [$\chi^2(8) = 9.384$, p = .311>.05]

suggests that there is no evidence to suggest that the model is misspecified. According to the Wald chi square statistics, holding all other variables in the regression at a fixed value, a freshmen was 1.848 times more likely to get a D or F or withdraw compared to non-freshmen with significance p =.038. This is the only predictor with p-value below 5 %. Age and the dummy variables for participation in the partial and total flip are significant at the 10 % level while the other variables are not significant. One additional year of age increased the odds of a DFW by 4 % (odds ratio 1.041) while participation to the flipped courses decreased the likelihood of DFW. In fact, the coefficient for the flip dummy variables flipped14 and flipped15 are negative and the odds ratios are less than one.

Table 6 Summary of binary logistic regression analysis for predicting the odds of getting a grade D or F or of withdraw (N = 420)

Predictors	**β**	S.E.	Wald	df	p	Odds Ratio	Inverse OR
1= female	.343	.209	2.689	1	.101	1.409	
age	.040	.023	2.939	1	.086	1.041*	
1 = freshman	.614	.295	4.324	1	.038	1.848**	
yearsenrolled	.050	.053	.858	1	.354	1.051	
1 = econ major	.240	.298	.650	1	.420	1.271	
1 = flipped14	-.424	.238	3.165	1	.075	.654*	1.529
1 = flipped15	-.476	.253	3.530	1	.060	.621*	1.610
Constant	-1.730	.600	8.323	1	.004	.177	
Model summary			χ^2	df	p		
Hosmer & Lemeshow			9.384	8	.311		
Nagelkerke R^2 =.060							
-2 Log likelihood = 557.912							
Correctly classified = 58.8 % (cut value = .5)							

To interpret more easily these odd ratios, we calculated the inverse odds ratio by applying the formula 1/odds ratio. A student who did not enroll in the partial flip was 1.529 times more likely to get a DFW and 1.610 times more likely if he did not enroll in the full flip. Equivalently, enrollment in the partial flip decreased the odds of DFW by (1-0.654)*100= 34.6 % and enrollment in the full flip by (1-0.621)*100= 39 %. The Nagelkerke R^2 statistics indicates that the model accounts for 6 % of the variation in the dependent variable. The overall rate of correct classification of the model was 58.8 %, for a cut value of .50 against 55.5% with a model including only the constant.

3.3 Comparison between the non-flipped active learning course and the flipped course designs based on students' teaching evaluations

3.3.1 Students' perceived learning

As a proxy of students' perceived learning we used the item *I achieved the learning objectives for this course* with scale 1 = 'totally disagree', 2 = 'disagree', 3 = 'neither agree nor disagree', 4 = 'agree', 5 = 'totally agree'. Since the null hypothesis of equal distribution of the answers to this item across the three courses was rejected (Kruskal-Wallis H χ^2 (2) =11.848, p = .003), we performed pairwise comparisons. Students in the flipped courses in 2014 and 2015 showed more agreement with the statement *I achieved the learning objectives for this course* compared with the non-flipped active learning course of 2013 as shown in Figure 1 and Table 7.

Figure 1 Students' perceived learning

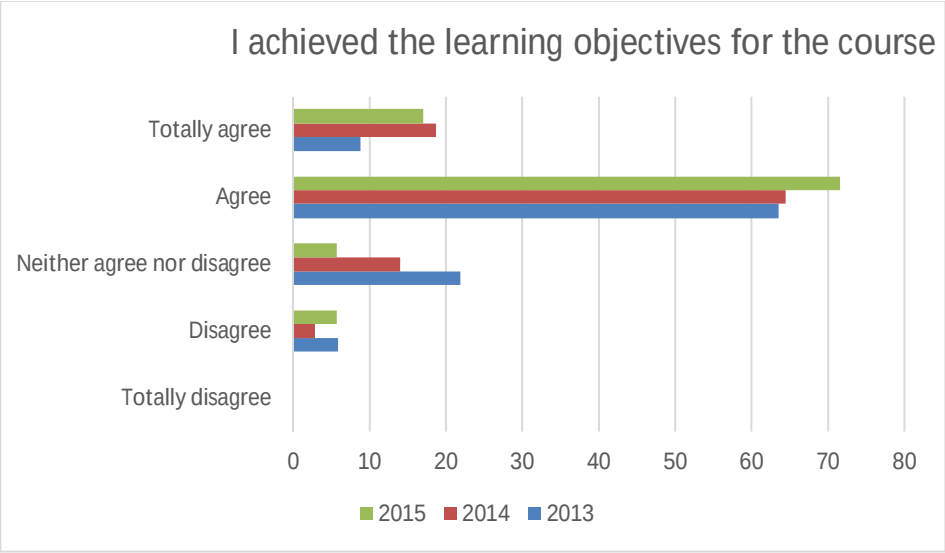


Table 7. Students' perceived learning according to the students teaching evaluations

Item	Year	N	Mdn	Interquartile range	Kruskall-Wallis $\chi^2(2)$	p
I achieved the learning objectives for the course					11.848	.003
	2015	88	4	0		
	2014	107	4	0		
	2013	137	4	1		
	Pairwise comparisons	Dunn	Adj-p	Mann Whitney u	p	Adj p 0.0167 Bonferroni
I achieved the learning objectives for the course						
	2013-2014	-28.892	.017	6058.500	.007	*
	2014-2015	-4.203	1.000	4593.500	.724	
	2013-2015	-33.095	.008	4821.500	.002	*

3.3.2 Students' satisfaction with the course

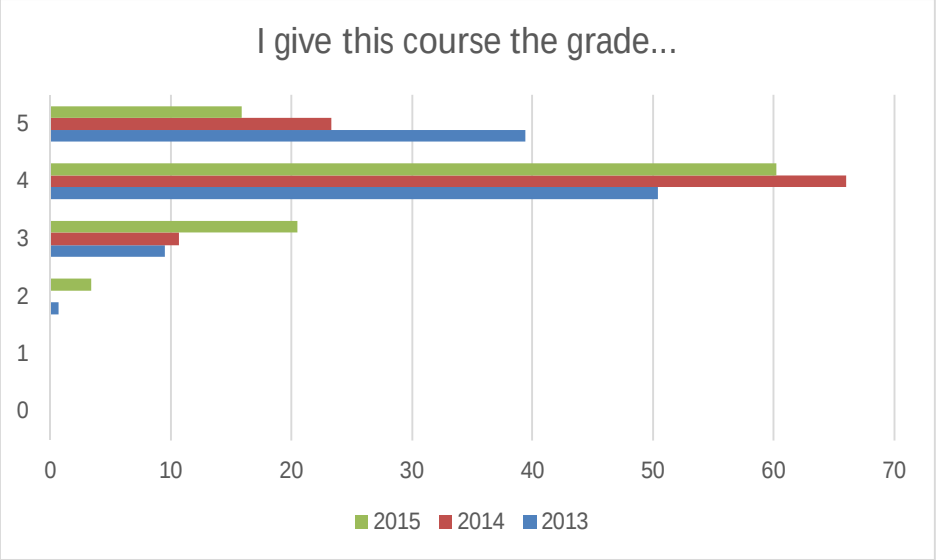
Students' satisfaction with the course was measured by the grade students gave the course in the student evaluation form using the scale 5 = excellent; 4 = very good; 3 = good; 2 = satisfactory; 1 = poor; 0 = fail (see Figure 2 and Table 8). A Kruskal-Wallis H test leads to reject the null hypothesis of equal distribution of the grades given by the students (χ^2 (2) = 18.887, p = .000 with a mean rank of 185.05 for 2013, 161.97 for 2014 and 135.47 for 2015). Pairwise analysis in Table 8 shows statistically different distribution of the grades as a measure of satisfaction between the 2013 non-flipped active learning course (mean grade 4.3, median 4) and the 2015 flip (mean grade 3.9, median 4) based on Dunn's test using 5 % significance level. The Mann-Whitney U-test however suggests differences in satisfaction also between the two flips, with the partial flip of 2014 (mean grade = 4.1, median = 4).

Regardless of which test we use, it emerges clearly, that students were least satisfied with the 2015 implementation of the flipped classroom compared with the non-flipped active learning course and that there was no statistically significant difference in satisfaction between the non-flipped, active learning course and the first, partial flip.

Table 8. Course satisfaction

Item	Pairwise comparison	Dunn	Adj-p	Mann Whitney U	p	Adj p = .0167 (Bonferroni)
I give the course the grade						
	2013-2014	23.089	0.104	6020.500	0.028	
	2014-2015	26.500	0.088	3758.000	0.018	*
	2013-2015	49.589	0.000	4247.000	0.000	*

Figure 2. Students' satisfaction with the course: Item: "I give this course the grade…" with scale 5 = excellent; 4 = very good; 3 = good; 2 = satisfactory; 1 = poor; 0 = fail

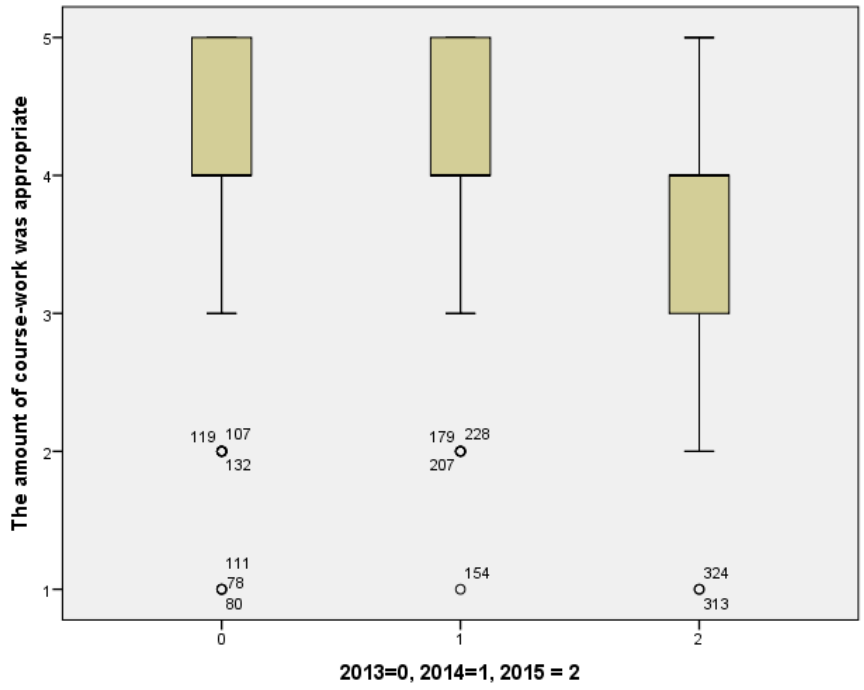


3.3.3 Course workload

Although we did not collect data on the time students put into course work, in all three years we checked whether students considered the amount of course work required in line with the number of credits the course gave. Students knew that one credit should corresponded to 27.5 hours of work but they were still reminded of this in the item: "*The amount of course-work was appropriate relatively to the number of the credits for this course (5 credits = 137.5 hours of course work*)". The three student populations had different perceptions of the course workload

(Kruskal-Wallis H test χ^2 (2) = 11.493, p = .003). Pairwise comparisons showed a significant difference both between the 2015 flip and the 2013 baseline and between the 2015 and the 2014 flips (see Table I and II in Appendix A and Figure 3). Even if the way the item was formulated does not allow to assess whether the students felt the workload was above or below 137.5 hours, comments by students suggest that the workload was perceived to be highest in the 2015 flip.

Figure 3. Perceived alignment of course workload with the number of credits from the course (Scale: 1 = 'totally disagree', 2 = 'disagree', 3 = 'neither agree nor disagree', 4 = 'agree', 5 = 'totally agree')



Note: Circular dots illustrate those data points that are more than 1.5 box-lengths but less than 3 box-lengths away from the edge of their box. These outliers are labelled with their case number.

3.3.4 Perceived difficulty of the course and the final exam

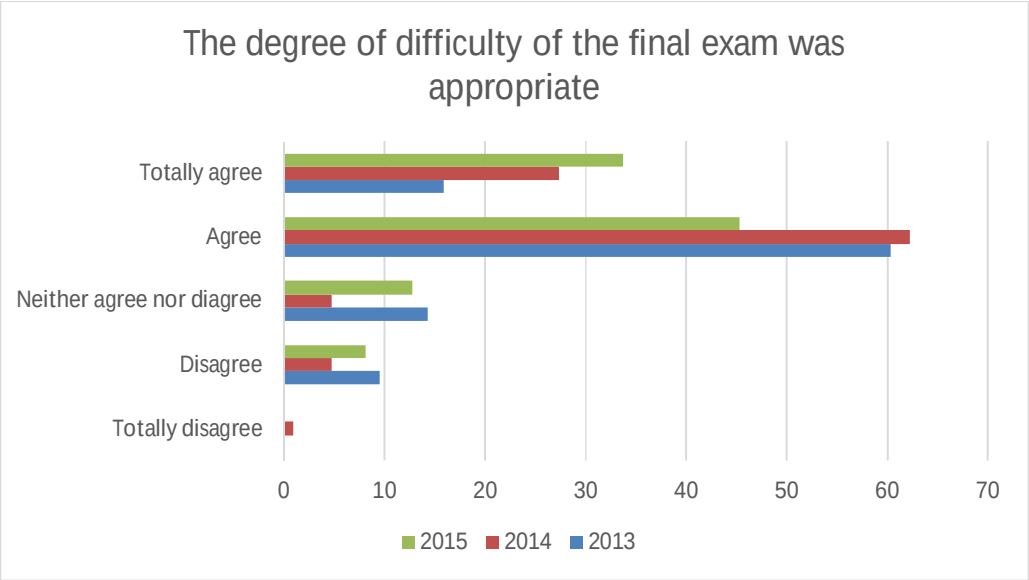
We tried to control for how demanding students felt the course was using the item *The course was demanding to the right degree*. As shown in Table I in appendix A, a Kruskal-Wallis H test showed that there was no statistically significant difference in how demanding students felt the courses were. However, there was a statistically significant difference in the perceived degree of

difficulty of the final exam relatively to the course as measured by the item *The degree of difficulty of the final exam was appropriate* (see Table I in Appendix A). We further explored this difference pairwise and found a statistically significant difference between the baseline course in 2013 and the 2014 flip (Adj.p =.015) while between 2013 and 2015 the adj. p value is just above the 5 % cutoff at 0.062 and the Mann-Whitney with Bonferroni correction is just above the corrected p value of .0167 with p =.029 (see Table 9).

Table 9. Pairwise comparisons of the distributions to the answers to the Likert item *The degree of difficulty of the final exam was appropriate*

Item	Pairwise comparison	Dunn	Adj-p	Mann Whitney u	p	Adj p 0.0167 Bonferroni
The degree of difficulty of the final exam was appropriate						
	2013-2014	-30.316	.015	5369.000	.003	*
	2014-2015	3.706	1.000	4487.500	.838	
	2013-2015	-26.610	.062	4547.000	.029	

Figure 4. Appropriateness of the degree of difficulty of the exam



Based on these statistics, the median perception of the appropriate degree of difficulty in the exam was not statistically significantly different between the baseline, lecture-based course 2013 and the full flip 2015, while it was different between the lecture-based course of 2013 and the partial flip 2014. Comments by the students suggest that the exam in 2013 might have been slightly easier. Figure 4 illustrate the distribution of the responses to this item.

3.3.5 Comparison between the flips according to students' evaluations

We investigated students' reactions to the different elements of the flip, that is, Moodle tests, videos, and group work to gain a deeper insight into what course practices the students felt supported well their learning. When comparing pairwise the 2014 and 2015 flip, we found no statistically significant difference between the two flips in the distribution of answers to the items *The instructor's lecture videos supported well my learning, the Moodle multiple choice-tests supported well my learning*, or *the course format supported well my learning* for which the median answer was agree/totally agree (see Table III in Appendix A). Comparing the Moodle tests and the video lectures, we found that the Moodle pre-class tests were perceived to be more supportive of learning compared to the video lectures (Wilcoxon signed-rank test = -4.524, p = .000).

There was no statistically significant difference in the percentage of video lectures students watched in the two flips (Mann Whitney test U = 5134.5, p = .136). Students reported watching on average 59.22 % (2014, N=103) and 64.54 % (2015, N=87) of all the video offered. Apparently the availability of the videos did not affect class attendance as the median to the item *The fact that lecture videos were available lead me to attend less classes* was "*disagree"* for both flips. Video lectures were a more popular study material than textbook: in 2015 11.2% and in 2014 13.1 % did not watch any of the video lectures prepared by the lecturer, a much smaller percentage compared to those students declaring that they did not use the textbook: 33 % in 2015 and 25 % in 2014.

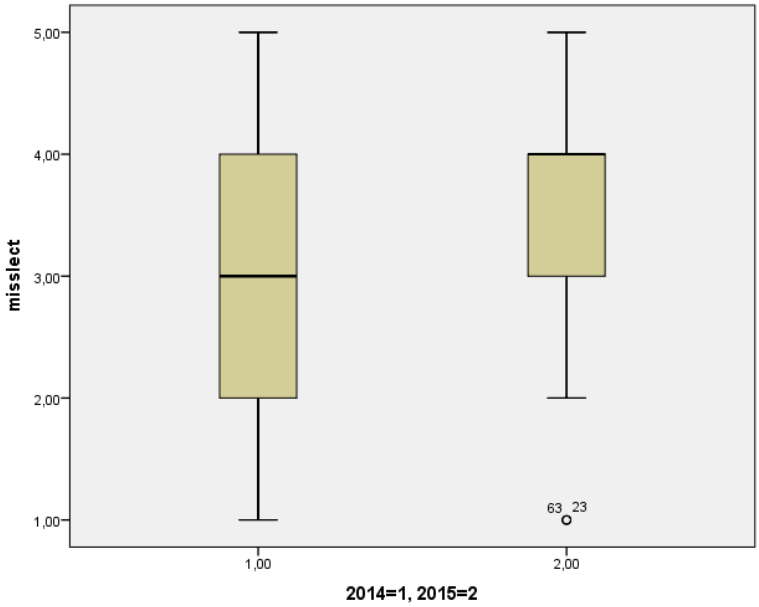
As making lecture videos requires a great investment of time and other resources, we wanted to see whether principles of microeconomics videos freely available on the internet would be a good substitute or complement for the videos made by the lecturer. Thus in 2015, the lecturer had indicated as additional materials the principle of microeconomics video lecture series in

English by the Marginal Revolution University (http://www.mruniversity.com/courses/principles-economics-microeconomics). Students in 2015 showed less agreement with the statement *The lecture videos in English supported well my learning* compared with the statement *The lecturer's lecture videos supported well my learning* (Wilcoxon signed rank test = -3.240, p = .001). In 2015, 16.9% did not watch any of the videos in English.

Students in the 2015 full flip showed more agreement with the item *I would have liked more lecturing in class* compared to the 2014 with a median in 2015 of "*4 = agree*" versus "*3 = neither agree nor disagree*" in the partial flip of 2014 (Mann-Whitney U = 6018.5, p = .001, see Table III in Appendix A and Figure 4).

To the open question "*How would you develop the course?*" seven students asked for more lecturing in class and one for less. Interestingly, students in the non-flipped active learning course agreed more with the item *In class there was a good interaction between the lecturer and students* than those in the flips (see Table III in Appendix A).

Figure 4. Item: *I would have liked more lecturing in class* (Scale: 1 = 'totally disagree', 2 = 'disagree', 3 = 'neither agree nor disagree', 4 = 'agree', 5 = 'totally agree')



The students' response to the items concerning group work did not statistically differ in the pairwise comparison between the two flips with the exception of the items *Group-work helped me to perceive the relationships between different concepts and models* and *The group-work assignment was meaningful and worthwhile* with median answer in 2014 "agree" and "neither agree nor disagree" in 2015 (see Table III in Appendix A). In 2015, in the student teaching evaluations 65 students answered the open question "*How would you develop the course?*" Of these, 33 students (51 %) mentioned group work, with suggestions ranging from making it elective or eliminating it altogether to reducing its size or moving out of class. The corresponding numbers for 2014 are 12 out of 57, that is 21 %. Thus group work emerges as the least favored element of the flipped course design by the students, especially in the case of the full flip.

4 DISCUSSION AND CONCLUSIONS

4.1 Learning outcomes

Students perceived a higher degree of achievement of the learning goals in both flipped courses compared to the active, non-flipped course as reported in the students' teaching evaluations. However, the linear regression analysis of model 1 suggests that only the partial flip was associated with better learning outcomes as measured by the final exam, while there was no statistically significant difference between the active, non-flipped course and the full flip. The lack of improvement in learning outcomes for the full flip is in line with Jensen et al. (2015) who did not find better outcomes from fully flipping the classroom when the control was a non-flipped, active learning classroom. Unfortunately, in our case it is not clear what drove the weaker learning outcomes of the full flip compared to the partial flip: was it the different degree to which the courses were flipped or an increase in workload? Although the instructor tried to design both flipped courses so that the workload would be the same as in the non-flipped, active learning course, according to student evaluations the workload was higher in the full flip compared to the other two courses. It has previously been suggested that flipping the classroom may unintentionally increase workload. Khanova et al. (2015) examined the student experiences when multiple flipped courses were offered within a single curriculum; they found that students

were concerned about the increased workload associated with the flips. Excessive workload can induce students to adopt surface-learning strategies (e.g. Lizzio et al., 2002; Baeten et al. 2010). Thus it is possible that the increased workload might be one factor explaining the weaker learning outcomes of the full flip.

As mentioned in the introduction, an important question is how flipping the classroom affects the outcomes of weaker students. The results of the binary logit regression suggest that the likelihood of getting a D or F or withdraw decreased with having took a flipped course. This result coupled with the lack of an overall improvement in learning outcomes for the full flip is in agreement with those of Ryan and Reid (2016), who found a reduction in Ds and Fs grades and in the withdrawal rate when flipping the classroom but no improvements in learning outcomes at the aggregate level. Interestingly, in our case the full flip performed slightly better than the partial flip with weaker students: enrollment in the partial flip decreased the odds of getting a D or F grade or of withdraw by 34.6 % while the full flip decreased them by 39 %. This might relate to the compulsory group work in the full flip, which might have promoted especially the weaker or less self-regulatory students to work harder than they would have done otherwise. Unfortunately our data does not allow us to identify disentangle the impact of compulsory group work on motivation and effort.

In both linear regressions, age was negatively associated with the percentage score in the final exam, although in model 2 the p-value for the age coefficient p = .106 was just above the 10 % cutoff rate for statistical significance. Moreover, age slightly increased the odds of getting a D or F grade or of withdraw in the binary logit regression. In future iterations of the flipped classroom, we will try to design data collection so as to gain better insight into what drives this link between age and learning outcomes and what can be done to better support older students' learning.

In model 1 being an economic major was negatively related to outcomes in the final exam but this result did not persist in model 2 which, limited to the flipped courses, controlled for entry level using the TUCE scores. Also economics major was not a significant predictor of D and F grades or withdraw in the binary logistic regression. Thus, it is not clear how robust this result is

nor how it should be interpreted. One possible interpretation is that students who are not economics major and thus take principles of microeconomics as an elective may be more motivated, an hypothesis we could test in the future by appropriately extending data collection to measure students' motivation and self-efficacy beliefs.

4.2 Students' satisfaction and the elements of the flipped course design

Students were least satisfied with the full flip and equally satisfied with the partial flip and the non-flipped, active learning course. Why was the full flip the least satisfactory to students? Our educated guess is the main driven of lower satisfaction in the full flip was making group work compulsory together with the way group work was organized. In fact, in the student teaching evaluations of the full flip in answering to the open question "*How would you develop the course?"* 51% of the respondents suggested major changes to group work while only 9 % asked for more lecturing. Nevertheless, decreased lecturing might also have had a role in reducing satisfaction. In the 2015 full flip the median answer to the item *I would have liked more lecturing in class* was "*4 = agree*" while in the partial flip of 2013 it was "*3 = neither agree nor disagree*" and the difference was statistically significant (Mann-Whitney U = 6018.5, p =.001, see Table III in Appendix A). Could it be then, that there is an optimal amount of lecturing in class, a middle way between a full flip and a traditional lecture course, which the partial flip came closer too? The fact that the partial flip yielded the best learning outcomes seems to support this hypothesis. Or is students request for more lecturing simply a sign of resistance to active-learning? The latter interpretation was suggested by Jensen et al. (2015), who also found that a significant percentage of their students expressed the desire that both the flipped and non-flipped, active learning course included more lecturing. An interesting issue is whether there are differences between freshmen and the other student in terms of how important they consider the role of in class lecturing and, more in general, in terms of their satisfaction with the flipped classroom, given that being a freshmen significantly increased the odds of a D and F grade or withdraw. Unfortunately, we could not explore this issue in our study since we did not ask about enrollment year in the teaching evaluations in order to guarantee full anonymity.

The pre-class multiple choice tests on Moodle were a clear students' favorite. In both flipped courses teaching evaluations, the vast majority of students strongly agreed that these tests supported well their learning. Moreover, in the open comments to the question "*What was good or even great in the course"*, the most frequent answer related to the Moodle tests as in "*The Moodle test, because I learned the materials during the course and not just before the exam*" and "*The Moodle-tests. When one did those during the course, there was almost no need to study for the exam*". It appears that Moodle tests provided more structure to the course by offering opportunities for frequent formative assessment. They helped students to better distribute their study time during the whole length of the course rather than massing it just before the exam. Increasing course structure has been shown to be highly supportive of learning: it reduces failure rates (Freeman et al. 2011) and increases course performance proportionately more for students from less privileged economic and social backgrounds (Eddy and Hogan 2014).

4.3 Cost of implementing the flip

The fixed cost of flipping the classroom were significant in our case. Most time consuming was creating the videos, which took the lecturer approximately 300 hours including the time needed to learn to use the recording and editing programs. Students expected the video lectures to be in their mother tongue, so we did not use existing principles of economics videos in English. Moreover, making own videos for the course provided a better fit with the textbook, lecture notes, and Moodle pre-class tests. However, for courses where the language of instruction is English, there is a wide choice of high quality videos illustrating principles of economics to choose from. If the classroom is flipped using these teaching materials, the cost of flipping the classroom can be reduced significantly and may even become negative if the use of ready teaching materials is coupled with a decrease in face-to-face class time as pointed out by Olitsky and Cosgrove (2016). In our case face-to-face in class time was not reduced as one objective of flipping was to free class time for more active learning.

4.4 Limitations of the study

Our data have some limitations. Firstly, data was not collected though a randomized controlled trial, as this was not possible. Secondly, there is no entry exam data for the non-flipped, active

learning course although we do have this data for the flipped courses. Thirdly, the final exams were not identical although they were meant to test the same abilities. In future studies, when faced with non-identical exams, one could try to assess exam equivalence using the Weighted Bloom's index developed by Freeman et al. (2011). Fourthly, the students' teaching evaluations are not perfectly comparable. In 2013, they were collected at the end of the semester jointly for both the principles of macro- and microeconomics as these were taught as a single course with separate exams. In 2014 and 2015, principles of microeconomics was taught as a separate course with its own students' teaching evaluation. However, since the lecturer and pedagogic approach were the same in 2013 for both principles of micro and macro, this should not be a major concern. Fiftly, model 2 is estimated using a non-random subset of course participants, as it includes only those who both took the entry exam and the final exam. Among the excluded from the sample, less motivated or weaker students may be over-represented. Also, in the specification of model 2, Moodle test average percentage score and participation to group work are included as explanatory variables of performance in the final exam. It is however possible that good students attempt Moodle tests more times and choose to attend group work when given the choice and not that attempting Moodle tests or participating in group work improves exam scores. In this study, we focused on content learning and did not attempt to measure the learning of generic competences and transferable skills. Had we measured them, the meaning and role of group work might look different.

4.5 Avenues for further research

This paper described only the first two iterations of a flipped classroom. When experimenting with a new teaching approach, instructors need several iterations to gain a good command of the new approach and to fine tune course design. Thus our results are in a sense preliminary, a tool to formulate new hypothesis and help develop successive implementations of the flipped classroom in the spirit of design-based research (Cobb et al. 2015). One interesting issue for further research concerns the relationship between the degree to which a course is flipped and its impact on learning outcomes and students' satisfaction. How robust is the result that a partial flip has better overall learning outcomes than a full flip? If it is robust, under which conditions and for which groups of students? When comparing the non-flipped active learning course and the

partial flip, we could ask: how much of the improvement in learning outcomes in the partial flip is due to the increase in structure that the bi-weekly multiple choice tests on Moodle provided to students? Would the non-flipped, active learning course reach the same learning outcomes of the partial flip, if integrated with the same multiple choice tests on Moodle, taken after the topic has been lectured in class? A third issue is the relationship between course workload, classroom flip and students' learning. Does the flip classroom increase students' workload? If it does, how does this affect learning outcomes in the flipped course as well as in other courses the students take in the same semester? A rigorous analysis of this issue would require the collection of reliable data on students' effort, a challenge for future iterations of our course. Finally, the cost of flipping the classroom continues to remain relatively unexplored: one could apply cost-effectiveness analysis to compare flipped and non-flipped classrooms considering the costs to faculty and administration as well as the costs to students.

Funding: This work was partially supported by the Teachers' Academy at the University of Helsinki and by the European Commission (KNORK 402765 543154-LLP-1-2013-1-FI-KA3-KA).

REFERENCES

Abeysekera, L. and Dawson, P. (2015) Motivation and cognitive load in the flipped classroom: definition, rationale and a call for research, *Higher Education Research & Development, 34*(1), 1-14, doi: 10.1080/07294360.2014.934336

Anderson, L., and Brennan, J. P. (2015). An experiment in "Flipped" teaching in freshman calculus. *PRIMUS, 25*(9), 861-875. doi:10.1080/10511970.2015.1059916

Baeten, M., Kyndt, E., Struyven, K., and Dochy, F. (2010). Using student-centred learning environments to stimulate deep approaches to learning: Factors encouraging or discouraging their effectiveness. *Educational Research Review, 5*(3), 243–260. doi:10.1016/j.edurev.2010.06.001

Baker, J. W. (2000). The ‘‘classroom flip’’: Using web course management tools to become a guide by the side. *Paper presented at the 11th International Conference on College Teaching and Learning*, Jacksonville, FL.

Bishop, J. L., and Verleger, M. A. (2013). The flipped classroom: A survey of the research. *Paper presented at the ASEE Annual Conference and Exposition*, Conference Proceedings, https://peer.asee.org/22585

Bloom, B.S. ed. (1956). Taxonomy of educational objectives: The classification of educational goals: *Handbook I, cognitive domain*. New York; Toronto: Longmans, Green.

Calimeris, L. and Sauer, K. M. (2015). Flipping out about the flip: All hype or is there hope? *International Review of Economics Education, 20*(1), 13-18. doi:10.1016/j.iree.2015.08.001

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., and Schauble, L. (2003). Design experiments in education research. *Educational Researcher, 32*(1), 9–13. doi:10.3102/0013189X032001009

Cobb, P., Jackson, K., and Dunlap, C. (2015). Design research: An analysis and critique. *Handbook of international research in mathematics education*: Third edition (pp. 481-503)

Design-Based Research Collective (2003). Design-based research: An Emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8. doi:10.3102/0013189X032001005

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics 6*, 241–252. doi:10.1080/004017 06.1964.10490181

Eddy, S. L., and Hogan, K. A. (2014). Getting under the hood: How and for whom does increasing course structure work*? CBE Life Sciences Education, 13*(3), 453-468. doi:10.1187/cbe.14-03-0050

Edelson, D. C. (2002). Design research: What we learn when we engage in design. *Journal of the Learning Sciences, 11*(1), 105–121.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., and Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America, 111*(23), 8410-8415. doi:10.1073/pnas.1319030111

Freeman, S., Haak, D., and Wenderoth, M. P. (2011). Increased course structure improves performance in introductory biology. *CBE Life Sciences Education, 10*(2), 175-186. doi:10.1187/cbe.10-08-0105

Giannakos, M. N., Krogstie, J., and Chrisochoides, N. (2014). Reviewing the flipped classroom research: Reflections for computer science education. *Paper presented at the Proceedings - CSERC 2014: Computer Science Education Research Conference*, 23-29. doi:10.1145/2691352.2691354

Goffe, W. L., and Kauper, D. (2014). A survey of principles instructors: Why lecture prevails. *Journal of Economic Education, 45*(4), 360-375. doi:10.1080/00220485.2014.946547

Khanova, J., Roth, M. T., Rodgers, J. E., and Mclaughlin, J. E. (2015). Student experiences across multiple flipped courses in a single curriculum. *Medical Education, 49*(10), 1038-1048. doi:10.1111/medu.12807

Jensen, J. L., Kummer, T. A., and Godoy, P. D. D. M. (2015). Improvements from a flipped classroom may simply be the fruits of active learning. *CBE Life Sciences Education, 14*(1), 1-12. doi: 10.1187/cbe.14-08-0129

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

Lizzio, A., Wilson, K., and Simons, R. (2002). University students' perceptions of the learning environment and academic outcomes: Implications for theory and practice. *Studies in Higher Education, 27*(1), 27-52. doi:10.1080/03075070120099359

McPherson, M. S. and Bacow, .L S.(2015) Online Higher Education: Beyond the Hype Cycle, *Journal of Economic Perspectives*, 29(4), 135–154, doi:10.1257/jep.29.4.135

O'Flaherty, J., and Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *Internet and Higher Education, 25*, 85-95, doi: 10.1016/j.iheduc.2015.02.002

O'Flaherty, J., Phillips, C., Karanicolas, S., Snelling, C., and Winning, T. (2015). Erratum: The use of flipped classrooms in higher education: A scoping review (the internet and higher education (2015) 25 (85-95)). *Internet and Higher Education, 27*, 90. doi:10.1016/j.iheduc.2015.05.001

Olitsky, Neal H. and Cosgrove, Sarah B. (2016). The better blend? Flipping the principles of microeconomics classroom, *International Review of Economics Education, 21*, 1-11, doi: 10.1016/j.iree.2015.10.004

Ryan, M. D., and Reid, S. A. (2016). Impact of the flipped classroom on student performance and retention: A parallel controlled study in general chemistry. *Journal of Chemical Education, 93*(1), 13-23, doi: 10.1021/acs.jchemed.5b00717

Touchton, M. (2015). Flipping the classroom and student performance in advanced statistics: Evidence from a quasi-experiment. *Journal of Political Science Education, 11*(1), 28-44. doi: 10.1080/15512169.2014.985105

Walstad, W. B., Watts, M. W., & Rebeck, K. (2007). Test of understanding in college economics: Examiner's manual, 4th edition. New York, NY: National Council on Economic Education.

Watts, M., & Schaur, G. (2011). Teaching and assessment methods in undergraduate economics: A fourth national quinquennial survey. *Journal of Economic Education*, *42*(3), 294-309. doi:10.1080/00220485.2011.581956

Appendix A Summary statistics of the students' teaching evaluations

Table I Comparison of the distributions to the answers to the Likert items in the students teaching evaluations of 2013-15 with scale 1 = 'totally disagree', 2 = 'disagree', 3 = 'neither agree nor disagree', 4 = 'agree', 5 = 'totally agree'

Item	Academic year	N	Mdn	Interquartile range	Kruskall-Wallis H $\chi^2(2)$	P
The learning objectives for the course were clearly stated					2.723	.256
	2015	90	5	1		
	2014	107	4	1		
	2013	137	5	1		
The topics dealt with in the course were interesting					1.158	.561
	2015	90	4	1		
	2014	107	4	1		
	2013	137	4	1		
I put sufficient effort in the course					16.778	.000
	2015	90	4	1		
	2014	107	4	1		
	2013	137	3	1		
The course was demanding to the right degree					4.426	.109
	2015	89	4	1		
	2014	106	4	1		
	2013	136	4	0		
I put enough effort in preparing for the final exam					2.611	.271
	2015	86	3	2		
	2014	107	3	2		
	2013	122	3	2		
The final exam was aligned with the course learning objectives, content and implementation					4.217	.121
	2015	87	4	1		
	2014	107	4	1		
	2013	126	4	1		

I achieved the learning objectives for the course					11.848	.003
	2015	88	4	0		
	2014	107	4	0		
	2013	137	4	1		
I learned a lot of new things in the course					7.503	.023
	2015	89	4		2015	89
	2014	107	4		2014	107
	2013	137	4		2013	137
The amount of course-work was appropriate relatively to the number of the credits for this course (5 credits = 137,5 hours of course work)					11.493	.003
	2015	89	4	1		
	2014	105	4	1		
	2013	137	4	1		
The degree of difficulty of the final exam was appropriate					9.343	.009
	2015	86	4	1		
	2014	106	4	1		
	2013	126	4	0		
Contact teaching helped me to understand key concepts					29.601	.000
	2015	86	4	1		
	2014	106	4	2		
	2013	128	5	1		
Contact teaching helped me to perceive the relationships between different concepts and models					18.814	.000
	2015	87	4	1		
	2014	106	4	2		
	2013	129	4	1		
Contact teaching improved my ability to apply economic concepts and models					22.977	.000
	2015	87	4	2		
	2014	106	4	2		
	2013	128	4	1		

In class there was a good interaction between the lecturer and students					28.876	.000
	2015	87	4	1		
	2014	106	4	2		
	2013	134	4	1		
Contact teaching supported well my learning					24.314	.000
	2015	87	4	2		
	2014	85	4	1		
	2013	125	4	1		
The use of clicker questions in class supported well my learning					47.843	.000
	2015	88	3	2		
	2014	106	3	2		
	2013	120	4	2		

Table II Pairwise comparisons of the distributions to the answers to the Likert items in 2013-15 with scale 1 = 'totally disagree', 2 = 'disagree', 3 = 'neither agree nor disagree', 4 = 'agree', 5 = 'totally agree'

Item	Pairwise comparison	Dunn	Adj-p	Mann Whitney u	p	Adj p 0,0167 Bonferroni
I achieved the learning objectives for the course						
	2013-2014	-28.892	0.017	6058.500	.007	*
	2014-2015	-4.203	1.000	4593.500	.724	
	2013-2015	-33.095	0.008	4821.500	.002	*
I learned in the course a lot of new things						
	2013-2014	21.568	0.194	6385.500	.068	
	2014-2015	-34.743	0.022	3762.500	.007	*
	2013-2015	-13.176	0.854	5619.500	.286	
The amount of course-work was appropriate relatively to the number of credits for this course (5 credits = 137,5 hours of course work)						
	2013-2014	6.130	1.000	6903.000	.557	
	2014-2015	33.114	0.027	3715.500	.007	*
	2013-2015	39.244	0.003	8679.000	.001	*
The degree of difficulty of the final exam was appropriate						
	2013-2014	-30.316	0.015	5369.000	.003	*
	2014-2015	3.706	1.000	4487.500	.838	
	2013-2015	-26.610	.062	4547.000	.029	
Contact teaching helped me to understand key concepts						
	2013-2014	62.312	.000	4184.000	.000	*
	2014-2015	-36.049	.020	3489.000	.003	*
	2013-2015	26.263	.672	4558.500	.022	

Contact teaching helped me to perceive the relationships between different concepts and models						
	2013-2014	49.442	.000	4779.500	.000	*
	2014-2015	-34.598	.016	3578.000	.005	*
	2013-2015	14.844	.672	5052.000	.183	
Contact teaching improved my ability to apply economic concepts and models						
	2013-2014	55.491	.000	4491.000	.000	*
	2014-2015	-35.818	.016	3529.500	.003	*
	2013-2015	19.673	.330	4833.000	.083	
In class there was a good interaction between the lecturer and students						
	2013-2014	61.090	.000	4496.000	.000	*
	2014-2015	-23.240	.213	3908.000	.054	
	2013-2015	37.850	.006	4432.000	.001	*
The use of clicker questions in class supported well my learning						
	2013-2014	69.569	.000	3819.000	.000	*
	2014-2015	3.252	1.000	4575.000	.813	
	2013-2015	72.823	.000	3047.000	.000	*
I give the course the grade						
	2013-2014	23.089	.104	6020.500	.028	
	2014-2015	26.500	.088	3758.000	.018	*
	2013-2015	49.589	.000	4247.000	.000	*

Table III Comparison of the two flipped courses with Likert scale 1 = 'totally disagree', 2 = 'disagree', 3 = 'neither agree nor disagree', 4 = 'agree', 5 = 'totally agree'.

Item	Year	N	Mdn	Inter-quartile range	Mann-Whitney U	p
The course format: lecture videos and Moodle tests at home, and activating methods and tailored lecturing in class supported well my learning					4374	.274
	2015	89	4	1		
	2014	107	5	1		
I would have liked more lecturing in class					6018.5	.001
	2015	90	4	1		
	2014	105	3	1		
The Moodle multiple choice-tests increased my understanding of the topics					4297.5	.258
	2015	87	5	1		
	2014	107	5	1		
The lecturer's lecture videos supported well my learning					4540.5	.546
	2015	89	4	1		
	2014	107	5	1		
In class there was a good interaction between the students					5592.0	.005
	2015	86	4	1		
	2014	86	3	1		
The fact that lecture videos were available lead me to attend less classes					4299	.809
	2015	86	2	2		
	2014	89	2	2		
Group-work helped me to perceive the relationships between different concepts and models						
	2015	87	3	2	2203.5	.007
	2014	67	4	2		
Group-work improved my ability to apply economic concepts and models					2482.5	.104
	2015	87	4	2		
	2014	67	4	2		

Group-work improved my understanding of key economic concepts				2354.5	.061
	2015	86	3	2	
	2014	66	4	2	
Group-work improved my ability to co-operate with other students				2990.0	.651
	2015	87	4	1	
	2014	66	3	3	
All members in the group participated actively and devoted effort to the group-work				3185.6	.250
	2015	86	4	3	
	2014	67	3	3	
The group-work assignment was meaningful and worthwhile				2207.5	.008
	2015	87	3	2	
	2014	67	4	2	