Teaching Mathematics and Its Applications (2016) **35**, 187–201
doi:10.1093/teamat/hrv019 Advance Access publication 29 December 2015

Flipping the calculus classroom: an evaluative study

WES MACIEJEWSKI*
Department of Mathematics, The University of Auckland, Auckland 1142, New Zealand
**Email: w.maciejewski@auckland.ac.nz*

[Submitted July 2015; accepted November 2015]

Classroom flipping is the practice of moving new content instruction out of class time, usually packaging it as online videos and reading assignments for students to cover on their own, and devoting in-class time to interactive engagement activities. Flipping has garnered a large amount of hype from the popular education media and has been adopted in a variety of contexts. Despite this high amount of interest, few studies have evaluated the effectiveness of classroom flipping on student academic outcomes. Specifically, no rigorous studies of the effects of flipping a mathematics course on students' mathematical understandings and achievement appear in the literature. This article reports results from a control group study of flipping a large (N = 690), first-year university calculus course for life sciences students. Students in the flipped course sections on average outperformed their counterparts in the traditional sections on the final exam, though only by approximately 8%. A more detailed analysis reveals the true beneficiaries in a flipped classroom—those with high basic mathematical ability and low initial calculus knowledge. Gains for this group are considerable: approximately 10% on the final, with an effect size of d = 0.56, and comparable gains on an independent measure of calculus concept mastery. This study positions classroom flipping as an effective practice in undergraduate mathematics and calls for further research into the mechanisms behind its effectiveness.

1. Introduction

Classroom flipping is a mode of course delivery where content instruction takes place outside of class time, while in-class time is devoted to conceptual practice and interaction. The classroom-flipping model acknowledges that most technical mastery can occur with little direct interaction with an instructor and should therefore be de-emphasized in student–instructor encounters. Concurrently, conceptual development is facilitated with social interaction, whether with peers or an instructor, and this ought to be the focus of class time.

Despite the surge of interest and research in classroom flipping, and interactive engagement methods more broadly, to date few studies analyse the effectiveness of classroom flipping in improving student achievement (Bishop & Vergeler, 2013; O'Flaherty & Phillips, 2015). In fact, no study on the effect of flipping an undergraduate mathematics course on students' achievement has appeared in the literature. The majority of studies have focused on student and instructor perceptions of flipping

(Bowers & Zazkis, 2012; Glover, 2015) and instructor accounts of implementing a flipped classroom model (McGivney-Burelle & Xue, 2013; Talbert, 2013; Eager *et al.*, 2014; Glover, 2015; Jungić et al., 2015). Though student attitudes towards flipping are mixed, they are generally positive; see, however, Sonnert *et al*. (2014) who report that 'ambitious teaching' may have a slight, negative impact on students' attitudes. However, asking about an entire flipping implementation may be too much to gain insight on students' views; as Toto and Nguyen (2009) report, students prefer in-person lectures to videos, but also prefer interactive class time to traditional lectures. Though students' perception of, and attitudes towards, the learning environment are important, and do influence their academic achievement (Prosser & Trigwell, 1999), there is a clear need to evaluate the effectiveness of classroom flipping in terms of various measures of learning.

A recent article (Code *et al.*, 2014) examines the effect of introducing interactive engagement methods for short periods of time in an existing course. Drawing from the study design of Deslauriers *et al.* (2011), a novice instructor was trained in interactive engagement methods and replaced the experienced course instructor for one week in each of two sections of a first-year calculus course. This switching of instructors happened in two different weeks and for two different topics, making for a more robust study design. In both instances, the students in the treatment section demonstrated greater gains in a topic-specific quiz than their traditional section counterparts.

Though the results of Code *et al*. (2014) are promising, I am mindful of the call—see, for example, Speer *et al.* (2010) and Stylianides & Stylianides (2013)—for more whole course studies and a greater understanding of university-level mathematics teaching practices. The current study is the first whole course, quasi-experimental, comparison group study of the effectiveness of classroom flipping in a first-year undergraduate calculus course. This study addresses the broad question, what impact does a flipped classroom model course have on students' academic outcomes in a first-year calculus course? This question could be addressed from many different perspectives. Indeed, classroom flipping is not a single instructional approach. Any flipping implementation consists of many components, the effectiveness of which could be analysed individually and a variety of theoretical perspectives be productively brought to bear. Based on variables important to instructors and identified in the literature above as indicators of student success, I have chosen to narrow the question above by considering three pertinent factors: students' (i) performance in the course, (ii) understanding of calculus concepts and (iii) expert-like orientations to mathematics. These factors lead to the following focused research questions:

(1) What effect does the flipped model of instruction have on students' performance in a first-year calculus class?
(2) Do students in a flipped calculus class exhibit a greater understanding of calculus concepts than their counterparts in a traditional lecture?
(3) What are the effects of the flipped classroom model on students' expert-like orientations to mathematics?

In answering these questions, a fourth question emerged from the data analysis:

(4) What groups of students, based on prior experience with and ability in mathematics, most benefit from a flipped classroom model of instruction?

2. A Review of the literature

Classroom flipping has experienced a surge of interest from teachers and instructors at all levels of formal education. At least two factors have led to the rise in popularity of classroom flipping. The first is a revolt, by some educators, against lecturing. Lecturing as the necessary, primary mode of

instruction has been called into question a number of times throughout its long history and the debate of lecturing's role in tertiary education, especially concerning mathematics, continues (Pritchard, 2010; Sonnert *et al.* 2014). Contemporary critiques of lecturing in tertiary education—see, for example, Gibbs (1981)—find their roots in the mid-20th century era of the opening up of higher education institutions. With a more liberal admissions policy came a greater diversity of students and a corresponding heterogeneity in background academic preparedness. This caused academics to re-evaluate their traditional teaching techniques, which were found to be increasingly inadequate (Prosser & Trigwell, 1999). The field of the Scholarship of Teaching and Learning (SoTL) emerged as a practitioner-led academic discipline focusing on understanding and improving tertiary education (Boyer, 1990). One main result in the SoTL literature is the acknowledgement that students perform better when 'actively engaged' in their studies. This engagement can result from internal motivation, but often needs support when developing. This support has come to take the form of non-traditional class time focusing on student interaction, a point that I return to later.

The second factor which has led to the increase in classroom flipping is a proliferation of cheap, flexible, advanced, and easy-to-use technology. Beginning in the early 2000s, universities began granting full Internet access to all students, at least in the form of campus computing facilities. Prices of personal computers began to drop and a greater diversity of devices became available. This led to almost all students owning at least one Internet-ready device; some universities even included a laptop computer in with the price of tuition. Software and peripheral devices also saw a dramatic increase in quality and decrease in price over this same time. This, coupled with an increasingly technology-savvy professoriate, resulted in viable modes of classroom flipping.

Though 'flipping' could mean any number of delivery modes—for examples in the context of undergraduate mathematics (see McGivney-Burelle & Xue, 2013; Talbert, 2013; Eager *et al.*, 2014; Jungić *et al*., 2015)—a common feature is that class time is devoted to interactive engagement activities. These can take the form of small group or whole class discussions (Springer *et al.*, 1999), or technology-intensive peer instruction (Mazur, 1997; Crouch & Mazur, 2001) involving personal response systems, or 'clickers' (Caldwell, 2007), among other methods (O'Flaherty & Phillips, 2015). Interactive engagement methods have been shown to be effective in a number of subjects. In the first large-scale study of interactive engagement in physics—indeed, in any subject—Hake (1998) reports pronounced learning gains for students in interactive engagement classrooms over their counterparts in traditional lecture-style classes. The study reports data from 62 introductory physics courses, with a total enrolment of 6542 students, from American high schools, colleges and universities. Criteria for inclusion in the study were pre- and post-test data for three measures of physics conceptual knowledge—the Halloun–Hestenes Mechanics Diagnostic (MD) Test (Halloun & Hestenes, 1985a,b), Force Concept Inventory (FCI; Hestenes *et al.*, 1992) and the Mechanics Baseline (MB) test (Hestenes & Wells, 1992)—and final exam scores. The author defines a *gain score* to account for varying levels of achievement among the participating courses; see Fig. 1. Traditional courses, those with little or no interactive engagement components, saw an average gain score of 0.23, while those with some appreciable degree of interactive engagement had a gain score of 0.48 on either MD or FCI tests. A further analysis of the correlation between MB and FCI and MD scores revealed a strong positive influence of interactive engagement methods on problem solving ability.

In a contemporary meta-analysis of 225 studies on interactive engagement, the main classroom component of flipping, across a variety of undergraduate Science, Technology, Engineering and Mathematics (STEM) disciplines, Freeman *et al*. (2014) find an overall large effect of flipping on student performance. Through a series of meta-analytical statistical techniques, the authors reach three main conclusions. First, the overall mean effect size was 0.47, reported as a weighted standardized mean difference of scores on comparable examinations/assessments and concept inventories. The

$$Gain\ Score = \frac{(Final\ Score) - (Initial\ Score)}{(Maximum\ Score\ Achieved) - (Initial\ Score)}$$

FIG. 1. The gain score is calculated by dividing the difference between an individual's initial and final scores by the difference between the maximum score achieved by all members of the comparison group and the individual's initial score. For example, when comparing course-wide CCI scores in this study, all students, regardless of treatment condition, are in the same comparison group and the maximum score achieved was 95.5%. Comparisons made in ability groups restrict the comparison groups to those defined by ability, each with their own maximum score achieved. Note that this definition of gain score differs from the original given by Hake (1998) in that the maximum score achieved is not assumed to be 100%. This modification accounts for the possibility that no student is able to solve all questions.

authors interpret this as students in interactive engagement classrooms achieving on average slightly under half a standard deviation higher than students in traditional classrooms. Secondly, failure rates in traditional classes were 1.5 times greater than those in interactive engagement classes. And third, assessment scores were on average 6% greater in interactive engagement classrooms than in traditional. The authors conclude by claiming that the empirical support for interactive engagement is so compelling that the use of traditional lecturing in undergraduate STEM teaching should be called into question. Moreover, the authors claim that support for lecturing is so thin that lecturing no longer serves as a useful datum to measure teaching innovations against, a sentiment echoed by at least one commentary to Freeman *et al*. (2014): 'If a new antibiotic is being tested for effectiveness, its effectiveness at curing patients is compared with the best current antibiotics and not with treatment by bloodletting. However, in undergraduate STEM education, we have the curious situation that, although more effective teaching methods have been overwhelmingly demonstrated, most STEM courses are still taught by lectures—the pedagogical equivalent of bloodletting' (Wieman, 2014).

Aside from interactive engagement, another key feature of the flipped classroom is that the students are assessed frequently, both formally, through quizzes, assignments and examinations, and informally, through voting on questions posed in class. This creates greater opportunities for the students to receive feedback on their understanding. Feedback, in many of its various forms, is known to have a profound effect on student learning and performance, ranking among the top most effective educational practices found in Hattie's (2008) extensive meta-analysis. Hattie & Timperley (2007) identify three effective types of feedback: feedback on the (i) task level (FT), (ii) process level (FP) and iii) self-regulation level (FR). Whereas feedback is almost entirely absent from standard first-year mathematics courses—indeed, it is a stretch to label assessment marks as feedback due to the relatively low amount of information they convey—classroom flipping accommodates all three types of effective feedback. Moreover, the feedback received by students in a flipped classroom is timely, adding to its effectiveness. For example, Hattie & Timperley (2007) conclude that immediate feedback (FT) during knowledge acquisition can increase uptake rates and accuracy. In the flipped classroom, this feedback takes the form of informal peer discussions of multiple-choice questions posed in class. Students first vote individually, the results are then displayed and a discussion ensues. The student response data help frame the student discussions; a student recognizes a portion of the class did not share their response and this disconnect must be navigated when arguing their solution. This provision of timely, focused feedback during knowledge acquisition is impractical for an instructor in a traditional large-lecture setting.

3. Methods

All seven sections of a first-year calculus course for life sciences majors at a research-intensive Canadian university participated in this study. A sole instructor taught each section (i.e. there was no 'co-teaching') and one instructor taught two of seven sections, for a total of six participating

instructors. The author was an instructor of a flipped section. Three of the sections acted as control (traditional) and four as treatment (flipped). Instructors were not randomly assigned to treatment and control sections. This was for two reasons: (i) the idea of flipping this course emerged over the summer months in conversations between the three instructors of the flipped sections, and (ii) the other three sections had yet to be assigned instructors. However, there was no a priori reason to expect great variation in student characteristics between sections. The instructors of the flipped sections, being two professors and a postdoctoral research fellow, had more experience teaching mathematics at the university level than the traditional section lecturers, comprised of two graduate students and a postdoctoral research fellow. However, the graduate students were required to complete training in education prior to being approved to teach a course. Also, there is no evidence from institutional records that suggests the academic achievement of students enrolled in a section taught by a graduate student is any different than those taught by professors. Anecdotally, graduate students at the site institution often perform well on student evaluations of instruction, often on par or better than faculty members.

During the first two weeks of the semester, all students wrote a Basic Skills Test (BST), developed at the site institution. This took the form of their first assignment but did not count towards their final grades. The questions on the BST examined the students' proficiency with elementary algebraic and arithmetic operations.

Concurrently, students were invited to complete the Mathematics Attitudes and Perceptions Survey (MAPS) online (Code *et al.*, manuscript under review). The MAPS instrument quantifies students' attitudes towards and perceptions of mathematics in seven categories: (1) the student's growth mind set, (2) ability to see connections between mathematics and the world around them, (3) confidence with engaging in mathematical tasks, (4) interest in working in mathematical situations, (5) persistence in mathematical problem solving, (6) their desire to make sense of solutions to mathematical problems and (7) their characterization of the nature of solutions to mathematical problems. The resulting student data is contrasted with data from practicing mathematicians to give a measure of how well aligned the students' conceptions of mathematics are with experts—students receive +1 for a question if it was answered in the same direction, positively or negatively, as the expert consensus and 0 otherwise. This yields an *overall expertise index*, that is, the average score for all questions, and corresponding subscale scores. Though the list above does not exhaust all categories of expert-like behaviour, these seven have been identified as being representative (Code *et al.*, manuscript under review). Also during the first two weeks, the Calculus Concept Inventory (CCI; Epstein, 2007, 2013) was administered in each class as a written, multiple choice test; this writing of the CCI is referred to as CCI (Begin) for the remainder of this article. The CCI is composed of questions on calculus concepts commonly found in first-year university curricula. These questions have been designed to not rely too heavily on the students' basic (arithmetic/algebraic) mathematical ability.

Classroom observations were made in each section twice during the semester, by the author primarily and, for the authors' course section, an additional research unaffiliated with the course. These were to verify the fidelity of the instructor to their assigned/chosen approach. Observations were performed with the Classroom Observation Protocol for Undergraduate STEM classroom practices (COPUS) instrument (Smith *et al.*, 2013). With this instrument, a record of the activity occurring in the classroom is made every two minutes. These observations are 'objective' in a sense—if the instructor is lecturing at one minute, then a check mark is put in the 'Lec.' box; a student asking a question at another minute gets a check in the 'SQ (Student Question)' box. The actual content of the lecturing or question is not recorded. Thus, the COPUS instrument gives a measure of the level of activity in the class—a variable that ought to be greater in a flipped classroom.

In the final two weeks of class, students wrote the CCI—referred to as CCI (End)—in class and the MAPS online, both with questions unaltered from the first offering. No other data collection was

performed at this stage. The students wrote their final exam and this was later collected and matched with their term survey data.

3.1 A description of the dataset

Initial total enrolment in the seven sections of the course was roughly 700. The course add/drop deadline was three weeks into the semester and some shuffling of students was observed during this time. At the conclusion of the course, a total of 690 students were enrolled. Results reported here correspond to different subsets of this total dataset. First, exam scores are reported for all students ($N = 690$) and compared across treatment conditions. Second, the CCI pre, post and gain scores are compared. There were 506 students that completed both CCI tests, 352 in flipped sections and 154 in traditional sections. This dataset is referred to as 'Linked'. Third, students are partitioned into 'prior ability groups' according to their BST and CCI scores. All students who completed pre-CCI tests also completed the BST, resulting in 650 students, 428 for flipped and 222 for traditional sections. The number of students in each ability group by treatment is given below.

Attendance was not mandatory in this course and students were free to attend classes in sections they were not registered in, though they were encouraged to register for the section they ultimately attended. No data on students' registered-for versus attended sections was collected. To account for potential student movement between sections, a student was considered attending a given section if their CCI response, gathered from in-class time, was from that section. That is, the data reported below was divided into treatment conditions according to what classes the students actually attended, not by their enrolment data. Although this is a proxy measure of section attendance, this is likely too close to the actual section attendance since there was no indication of substantial attendance differences between pre- and post-CCI data. In addition, there are two reasons to suspect that students did not attend classes from sections they were not registered in: (1) the room capacities were about the same as enrolment sizes, which often discourages inter-section movement; and, (2) clicker data from flipped sections indicated only rare participation from clickers registered in other sections.

MAPS data were not used in refining the overall dataset and were analysed separate from the CCI, BST and final exam data; the subset of the data comprised students who completed both pre and post instances of CCI and MAPS, in addition to the BST and final exam was relatively small, representing $<50\%$ of the overall population. The number of students to complete both the pre- and post-MAPS tests is 324, 209 in flipped sections and 115 in traditional sections. These student numbers are lower than those for the other instruments likely because MAPS was completed outside of class time and did not count directly towards the final grade. Though, students were incentivized to complete all instruments with a 1% completion mark towards their final grades.

3.2 Flipping implementation details

The general feature of a flipped class is that some or all of the new content instruction that would normally take place in class is relegated to out of class time. This frees time during class for use in interactive engagement activities. As some of the early proponents of flipping—'inversion' in their article—describe it, '[i]nverting the classroom means that events that have traditionally taken place inside the classroom now take place outside the classroom and vice versa' (Lage *et al.*, 2000). The details of flipping—the length and frequency of videos and resources, the number and timing of assessments, the nature of the in-class activities—vary by implementation. Those of the course in this study are as follows.

Prior to each class, the students watch a short video on that day's topics. The videos were all roughly 6 minutes in length, which is the optimal length for maintained student engagement (Guo *et al.*, 2013). Each video is followed by a 3–6 question quiz hosted on the department's WeBWorK server. These quizzes account for 10% of the overall course grade.

The main feature of the flipped section in-class time is a predominance of interactive engagement. Lecturing, though may still be present, is de-emphasized to the point of less than a quarter of in-class time. Though class time activities varied by section, the common theme was for class time to be spent posing multiple choice questions for students to vote on with personal response systems—in this case, clickers. Students were incentivized to vote with negligible participation marks, accounting for a fraction of a percent of the overall grade. Each instructor of a flipped section created their own questions and slides, though questions created in previous courses were available as templates. After voting on a question is completed, the instructor reveals the response distribution. If there is substantial disagreement, the instructor invites the students to 'convince their neighbours that their solution is correct'. After some student-level discussion, a subsequent vote is called. If the responses have converged on the correct response, the instructor moves on to the next question, often after soliciting reasons from the students for why the answer is correct. If substantial disagreement persists, either student will be called on to provide perspectives on the possible solutions or the instructor guides the students in constructing a correct solution.

Aside from the pre-lecture quizzes, students also wrote biweekly online (WeBWorK) assignments, five short written assignments, two paper-based, hour-long midterms and a two-hour, paper-based final exam. All assessments other than the pre-lecture quizzes were common to all sections.

3.3 A comparison of the flipped and traditional sections

The traditional course sections in the study were characterized by transmission-style lectures. The instructors wrote on the board or document camera and asked the occasional question to the entire class, but did not occasion class discussion. One of the traditional lecturers attempted to include some clicker questions in their class, but this activity neither constituted a sizable portion of the class time nor resulted in noticeably productive student discussion. This instructor diminished their use of clicker questions as the semester progressed. A further discussion of the differences between the flipped and traditional sections is included with the reporting of the COPUS data below.

The pre-lecture videos were available to all students, regardless of treatment condition. Two of the three instructors of the traditional sections required their students to complete the pre-lecture quizzes. The third included the pre-lecture questions on the weekly assignment and adjusted the assignment weights appropriately.

The course used a custom textbook written by one of the flipped section instructors. This was available online for free to all students. The course webpage included a detailed, day-level topic schedule with reference to learning goals, textbook pages and the corresponding online videos. All instructors stayed very close to the pace of the schedule throughout the semester.

4. Results

The data gathered during the semester turned out to be rich and nuanced. I therefore present results with increasing resolution, from broad to more focused conclusions. First, the COPUS data confirms that the level and type of student and instructor activity in the flipped and traditional sections were indeed different. These data are reported in Fig. 2. All data reported in Fig. 2 is averaged over two

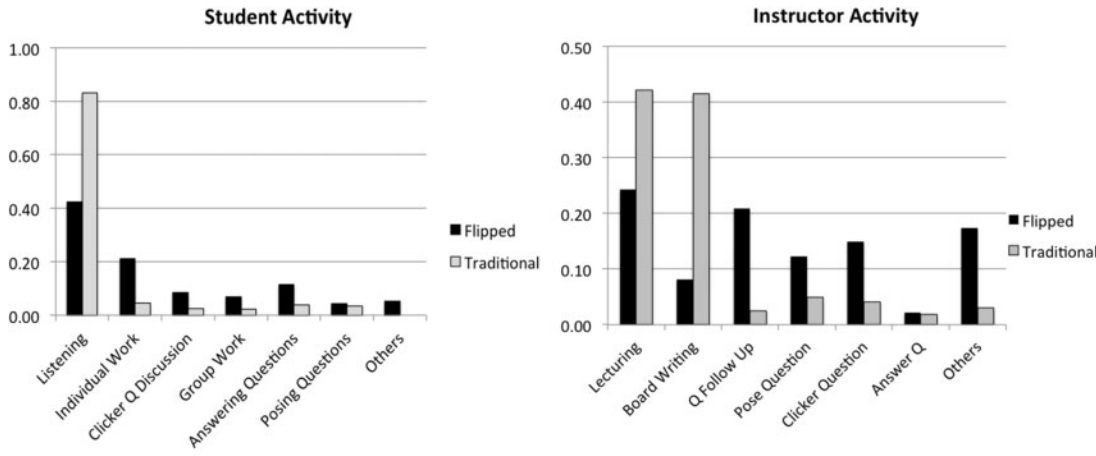

FIG. 2. Summarized COPUS classroom observation data.

classroom observations and treatment condition. The traditional and flipped classrooms differed in the activities engaged in by both instructors and students. Students in traditional sections spent a great percentage of their time (∼83%) in class listening to the instructor, while flipped students spent substantially less time (∼42%) on listening and more on different types of individual and group work. Instructors in the traditional sections spent most of their time lecturing and board writing (∼42% each, though these are not always exclusive activities—grouping these as a single activity yields roughly 87%). Flipped section instructors lectured for approximately one quarter of class time, with a clear focus on asking and following up questions. Interestingly, very little time (8% of class time) was spent on board, or document camera, writing by flipped section instructors.

I preface the following results with a caveat. Surprisingly, an analysis of BST, CCI and MAPS data revealed start-of-term differences in student characteristics by treatment condition: the traditional sections had lower CCI, BST and MAPS averages than the flipped sections. These differences are not statistically significant, but worth highlighting. Though there was no a priori reason to expect variation across sections, such variation was nonetheless observed. This could be due to at least two factors:

(1) The seventh section (a traditional section) was created just prior to the start of term to accommodate students on the lengthy course waitlist. Students with top high school marks are offered advanced admission, so there is reason to expect waitlist students to have achieved lower in high school.
(2) Students are free to withdraw from the course without penalty in the first week. This frees space in the, initially full, sections and often results in student migration between sections, effectively allowing some students to choose their treatment condition. Though the data received from withdrawn students was not included in the analysis, we do not have detailed data on student migration. Anecdotally, this is not expected to be significant, as was highlighted in Section 3.1 above.

I believe this difference in student characteristics between treatment conditions does not greatly impact the overall results. Though the initial result on student outcomes by treatment condition is not perfectly sensitive to the initial differences, the subsequent results do factor in these differences.

4.1 Overall results by treatment condition

The first result is that the treatment sections, on average, had higher final exam grades than the control section, ($M_{\text{Flipped}} = 64.81$, $M_{\text{Traditional}} = 56.61$, $t(399) = 4.50$, $p < 0.01$, $d = 0.39$). Furthermore, the flipped sections had an overall lower standard deviation in final exam scores than the traditional sections ($\text{SD}_{\text{Flipped}} = 18.85$, $\text{SD}_{\text{Traditional}} = 24.83$, $F(445, 243) = 0.76$, $p < 0.01$). That is, not only did the flipped sections have a higher average final exam score, the distribution of student scores in the flipped section are more tightly centred about the mean than the traditional section scores.

Treatment students also had higher, on average, CCI scores at the end of the course ($M_{\text{Flipped}} = 52.16$, $M_{\text{Traditional}} = 49.17$, $t(275) = 1.76$, $p = 0.04$, $d = 0.18$). However, there was no statistically significant difference between the flipped and traditional CCI gain scores. A summary of these results is reported in Table 1.

In terms of the MAPS data, all sections, regardless of treatment condition, saw declines in average MAPS scores. That is, all students moved further from expert-like conceptions of mathematics. This shift away from expert-like conceptions and attitudes corroborates previous results from MAPS and related CLASS surveys (Adams *et al.*, 2006; Barbera *et al.*, 2008; Gray *et al.*, 2008; Semsar *et al.*, 2011; Jolley *et al.*, 2012; Code *et al.*, manuscript under review) and other measures of student attitudes (Sonnert *et al.*, 2014). This was also observed for all MAPS subscales in both treatment conditions, though many of the scales did not exhibit a statistically significant decline.

One important observation about the aggregate MAPS data is that the flipped section ($N = 209$) MAPS scores *decreased less* than those from the treatment section ($N = 115$). There was no statistically significant difference between treatment and control pre-test MAPS overall expertise index ($M_{\text{Flipped}} = 55.13$, $M_{\text{Traditional}} = 54.63$). However, there was a difference between post-test scores, with the flipped sections having a higher average overall expertise index ($M_{\text{Flipped}} = 51.64$, $M_{\text{Traditional}} = 46.34$, $t(223) = 2.44$, $p = 0.01$, $d = 0.29$). This stands in contrast with previous reports of 'ambitious', non-transmission style teaching having a profoundly negative effect on students' attitudes (Sonnert *et al.*, 2014).

4.2 Results by 'initial ability' grouping

The initial results led to questioning the effects of flipping on different sets of students grouped by 'prior ability'. I segregated the student data in four groups, based on start-of-semester BST and CCI scores:

(1) low (below course BST average) BST, low (below course CCI average) CCI ($N_{\text{Flipped}} = 66$, $N_{\text{Traditional}} = 52$);
(2) high (above course BST average) BST, low CCI ($N_{\text{Flipped}} = 83$, $N_{\text{Traditional}} = 42$);

TABLE 1. *Summary of all data by treatment condition*

	Flipped	Traditional	Difference	*P*	Cohen's *d*
Final exam	**64.81 (18.85)**	**56.61 (24.83)**	**8.20**	**<0.01**	**0.39**
CCI (Begin)	13.51 (7.45)	12.57 (7.56)	0.93		
CCI (End)	**52.16 (16.71)**	**49.17 (17.93)**	**2.98**	**0.04**	**0.17**
CCI (Gain)	47.30 (20.03)	44.46 (20.88)	2.84		

Means are reported, followed by standard deviations in parentheses. Statistically significant results are bolded.

(3) low BST, high CCI ($N_{\text{Flipped}} = 115$, $N_{\text{Traditional}} = 55$); and,
(4) high BST, high CCI ($N_{\text{Flipped}} = 164$, $N_{\text{Traditional}} = 73$).

These results are presented in Table 2. The first observation to make is that all four of these ability groups achieved higher on the final exam in the flipped classroom than their counterparts in the traditional classrooms. However, this result is statistically significant for only the two groups with low initial CCI scores (low BST/low CCI: $M_{\text{Flipped}} = 54.89$, $M_{\text{Traditional}} = 47.77$, $t(102) = 2.06$, $p = 0.02$, $d = 0.39$; high BST/low CCI: $M_{\text{Flipped}} = 67.73$, $M_{\text{Traditional}} = 57.66$, $t(92) = 3.22$, $p < 0.01$, $d = 0.56$). Note that average final exam scores for the low BST/low CCI are substantially lower than those for the high BST/low CCI group. A potentially confounding factor for this group is their low basic mathematical ability—it is possible that they did acquire calculus knowledge, a possibility confirmed by the CCI data, but their low basic mathematics skills did not improve over the semester and prevented them from achieving higher on the final exam.

A closer look reveals a likely cause of the non-statistical significance of exam score differences between conditions in the high initial CCI groups: the traditional sections had higher standard deviations in final exam scores than the flipped sections. That is, even though there are no statistically significant differences between the means for both high CCI groups, the flipped sections had lower variation in exam scores, though this is only significant for the high BST/high CCI group ($F(163, 72) = 0.63$, $p < 0.01$).

In terms of CCI scores, those with low initial CCI scores had greater final CCI scores in the flipped classes than in the traditional (low BST/low CCI: $N_{\text{Flipped}} = 50$, $M_{\text{Flipped}} = 44.10$, $N_{\text{Traditional}} = 34$, $M_{\text{Traditional}} = 36.90$, $t(67) = 2.39$, $p < 0.01$, $d = 0.54$; high BST/low CCI: $N_{\text{Flipped}} = 100$, $M_{\text{Flipped}} = 48.91$, $N_{\text{Traditional}} = 36$, $M_{\text{Traditional}} = 42.93$, $t(69) = 2.24$, $p = 0.01$, $d = 0.41$). The two groups with initially higher CCI scores did not have a statistically significant difference between final CCI scores in the flipped and traditional classes. Since the flipped and traditional classes had initially different CCI averages, we also highlight differences in CCI gain scores. Those with initially low CCI scores had a higher gain in the flipped sections than in the traditional sections (low BST/low CCI: $N_{\text{Flipped}} = 50$, $M_{\text{Flipped}} = 57.09$, $N_{\text{Traditional}} = 34$, $M_{\text{Traditional}} = 47.00$, $t(70) = 2.34$, $p = 0.01$, $d = 0.52$; high BST/low CCI: $N_{\text{Flipped}} = 100$, $M_{\text{Flipped}} = 56.29$, $N_{\text{Traditional}} = 36$, $M_{\text{Traditional}} = 48.67$, $t(69) = 2.13$, $p = 0.02$, $d = 0.39$). There was no statistically significant difference between CCI gain scores for the high CCI groups. A summary of CCI gain score results is in Table 2.

An analysis of the MAPS data divided into the four ability groups described above was performed. As was the case for the aggregate data, all average overall expertise index scores decreased for both treatment conditions and all ability groups. There were no statistically significant differences between

TABLE 2. *Summary of results by prior ability group*

Ability Group	Measure	Flipped	Traditional	Difference	p	Cohen's d
Low BST/low CCI	**Final exam**	**54.89 (17.18)**	**47.77 (19.77)**	**7.12**	**0.02**	**0.39**
	CCI (Gain)	**57.09 (19.28)**	**46.97 (19.52)**	**10.12**	**0.01**	**0.52**
Low BST/High CCI	Final Exam	58.11 (18.99)	54.08 (25.05)	4.03		
	CCI (Gain)	46.59 (24.73)	41.77 (23.48)	4.82		
High BST/Low CCI	**Final Exam**	**67.73 (16.80)**	**57.66 (20.02)**	**10.06**	**<0.01**	**0.56**
	CCI (Gain)	**56.29 (19.94)**	**48.67 (17.83)**	**7.61**	**0.02**	**0.39**
High BST/High CCI	Final Exam	71.69 (17.35)	68.24 (27.37)	3.45		
	CCI (Gain)	51.46 (22.06)	56.04 (22.51)	−4.58		

Means are reported, followed by standard deviations in parentheses. Statistically significant results are bolded; p-values are reported only for statistically significant results.

flipped and traditional average pre-test MAPS overall expertise index scores in any ability group. Although the flipped sections had consistently higher post-test expertise index scores, this difference was statistically significant for only one ability group, the high BST/low CCI group ($N_{\text{Flipped}} = 60$, $M_{\text{Flipped}} = 52.48$, $N_{\text{Traditional}} = 32$, $M_{\text{Traditional}} = 41.65$, $t(66) = 2.71$, $p < 0.01$, $d = 0.59$). Note that the number of students from each ability group that wrote both instances of the MAPS survey was relatively small, which may account for the non-significance of the majority of observed differences.

5. Discussion

Students enrolled in sections of a first-year university calculus for life sciences course using a flipped model of learning outperformed their counterparts in the traditional, transmission-style sections on the final exam and on the Calculus Concept Inventory. Flipped section students also exhibited less of a shift away from expert-like conceptions of mathematics than those in traditional sections. Segregating the students into prior ability groups reveals that those students with good basic mathematics skills and low calculus knowledge benefit most from the flipped classroom model, outperforming the traditional section students on both the final exam and the Calculus Concept Inventory. Though the students' expert-like conceptions of mathematics declined over the semester in both treatment conditions, there was less of a decline in the flipped sections.

The results of this study establish classroom flipping as an effective practice in first-year undergraduate calculus. This adds to the burgeoning evidence on the effectiveness of certain non-traditional, interactive engagement-based modes of instruction in university (Hake, 1998; Freeman *et al.*, 2014). This study also identifies who most benefits from classroom flipping—regardless of level of basic mathematical ability, those with little prior calculus experience benefit more from a flipped classroom than from a traditional lecture. Those with solid basic mathematical ability and little or no prior calculus experience receive the greatest benefit. It must be emphasized that, due to ethical concerns, the pre-class videos were available to all students and the pre-lecture quizzes were used in all but one traditional class. Therefore, keeping in mind that some extraneous variables—time of class, or other unforeseen variables—were not controlled for, the results observed between treatment conditions is likely due to only what happened in class time. A future study may be designed to further evaluate the relative benefit of in and out of class activities.

Numerous studies have established aspects of classroom flipping as a viable and effective mode of education at the tertiary level (Freeman *et al.*, 2014). However, not all implementations of classroom flipping have improved student academic gains; see, for example, Overmyer (2014) and Willis (2014) who report null effects of flipping on student achievement. Future studies should move on from evaluating the effectiveness of classroom flipping in different contexts and uncover the features and mechanisms of flipping that make it effective. Taking a more nuanced approach to studying classroom flipping will aid in explaining why most implementations have been effective and ought to identify the cause of some negative or null results. I suggest a number of features of classroom flipping that are worthy of further study.

The first is the occasioning of different types learning other than the passive reception of information. Whereas traditional lectures limit student classroom activity to listening and note taking, with the very rare opportunity to respond to a question, interactive engagement-type classes offer opportunities to engage in additional modes of learning—discussions, group and individual work, and movement and gestures, among others. All of these are established in the mathematics education literature as promoting deeper mathematics learning. What is more, all of these activities violate the social and socio-mathematical norms of traditional lectures (Yackel *et al.*, 2000). The culture of a flipped mathematics classroom is fundamentally different than that of a lecture-based classroom, not just in how the

material is transferred from expert to student, but in how mathematical understanding develops in students, both individually and collectively. A beneficial study would identify what modes of learning students are employing and how these affect their understandings.

Second, and more pragmatically, is the timing and frequency of assessed and un-assessed practice. In our classroom flipping implementation, students completed a quiz before attending each class. This quiz was graded instantly, generating timely feedback, and counted towards the students' overall grade. The students then completed a bi-weekly assignment, again online and graded instantly. These two assessments likely prepared the students for the larger, more heavily weighted and less frequent exams—two midterms and a final. This assessment scheme was available to all sections, flipped and traditional, except on traditional section in which the instructor opted out of daily quizzes with weight transferred to the bi-weekly assignments. The key difference between treatment conditions in terms of assessment was the daily informal assessment experienced by flipped section students. This frequent informal assessment during knowledge acquisition should act to correct misunderstandings before they become entrenched (Hattie & Timperly, 2007). The feedback received by students in their small discussion groups is timely and should aid in the identification and rectification of misunderstandings. Identifying what statements made by the student, the feedback they are given and how they receive this information should be a priority in flipped classroom research.

The third is the social aspect of a flipped classroom. The development of mathematical understanding through social interaction has a long history in mathematics education research, dating back to at least the work of Vygotsky (1978, 1986), though independent developments appeared before the availability of Vygotsky's work outside of the Soviet Union; see Lerman (2000) for a brief review. In the flipped classrooms in this study, after students respond to a multiple-choice question in class they break into informal groups and discuss their answers. What happens during these interactions is almost entirely unexplored. The interactions appear, anecdotally in this study and quantifiably elsewhere (Mazur, 1997; Crouch & Mazur, 2001; Knight & Wood, 2005), to have a profound effect on the responses of the students, but do they affect longer term understanding? Is it just that the most perceived-to-be-intelligent, or suave, or loudest group member convinces the others to vote as they did? Smith *et al.* (2009) argue that students do learn from brief discussions with their peers. The authors demonstrate this by having biology students, after a discussion and re-voting episode, respond to a novel-yet-isomorphic follow-up question. A high proportion of students were able to solve the new question, even those who did not solve the initial question correctly; an indication of learning from small-group discussions. However, Smith *et al*. (2009) do not examine the exchanges taking place between and among students, just the product of these exchanges. Essentially, is there *actual* mathematical learning taking place during these bursts of social activity? I suspect there is, but noticing it and leveraging it for educational gains ought to pose a substantial challenge to mathematics education researchers.

As a final point, classroom flipping has emerged as a practitioner-led innovation and, as such, is not necessarily grounded in the extant mathematics education literature. The connections are present, but ought to be made more explicit. The current study was intended to be a coarse-grained evaluation of classroom flipping at the course level. Future studies should take a nuanced approached and be couched in the language of existing mathematics education theories. Such an approach would allow for a deeper understanding of the working mechanisms of flipping and strengthen research-practice ties.

Acknowledgements

A sincere thanks is due to the members of the Carl Wieman Science Education Initiative and the Department of Mathematics at the University of British Columbia, especially Eric Cytrynbaum and Leah Keshet, for their support and guidance with this work.

References

Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D. & Wieman, C. E. (2006) New instrument for measuring student beliefs about physics and learning physics: the Colorado learning attitudes about science survey. *Phys. Rev. ST Phys. Educ. Res.*, 2, 010101.

Barbera, J., Adams, W.K., Wieman, C.E. & Perkins, K.K. (2008) Modifying and validating the Colorado learning attitudes about science survey for use in chemistry. *Chem. Educ. Res.*, 85, 1435–1439.

Bishop, J. L. & Vergeler, M. A. (2013) The flipped classroom: a survey of the research. Paper presented at the *120th ASEE Annual Conference & Exposition* (Paper ID #6219), Atlanta, Georgia, USA.

Bowers, J. & Zazkis, D. (2012) Do students flip over the ''flipped classroom'' model for learning college calculus? *Proceedings of the 34th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (L. R. Van Zoest, J.-J. Lo & J. L. Kratky eds). Kalamazoo, MI: Western Michigan University, pp. 849–852.

Boyer, E. L. (1990) *Scholarship Reconsidered: Priorities of the Professoriate*. Princeton, NJ: Carnegie Endowment for the Advancement of Teaching.

Caldwell, J. (2007) Clickers in the large classroom: current research and best-practice tips. *CBE – Life Sci. Educ.*, 6, 9–20.

Code, W., Maciejewski, W., Merchant, S. & Thomas, M. (manuscript under review) The Mathematics Attitudes and Perceptions Survey: a new tool to assess expert-like behaviour among undergraduate mathematics students.

Code, W., Piccolo, C., Kohler, D. & MacLean, M. (2014) Teaching methods comparison in a large calculus class. *ZDM—Math. Educ.*, 46, 589–601.

Crouch, C. H. & Mazur, E. (2001) Peer instruction: ten years of experience and results. *Am. J. Phys.*, 69, 970–977.

Deslauriers, L., Schelew, E. & Wieman, C. (2011) Improved learning in a large-enrolment physics class. *Science*, 332, 862–864.

Eager, E., Peirce, J. & Barlow, P. (2014) Math bio or biomath? Flipping the mathematical biology classroom. *Lett. Biomath.*, 1, 139–155.

Epstein, J. (2007) Development and validation of the Calculus concept inventory. *Proceedings of the Ninth International Conference on Mathematics Education in a Global Community* ((D. K. Pugalee, & A. Schinck eds). Charlotte, NC, pp. 165–170.

Epstein, J. (2013) The calculus concept inventory—measurement of the effect of teaching methodology in mathematics. *Notices AMS*, 60, 1018–1026.

Freeman, S., Eddy, S., McDonnough, M., Smith, M., Okoroafor, N., Jordt, H. & Wenderoth, M. (2014) Active learning increases student performance in science, engineering, and mathematics. *PNAS*, 111, 8410–8415.

Gibbs, G. (1981) Twenty terrible reasons for lecturing. *SCED Occasional Paper No. 8, Birmingham*. (http://www.brookes.ac.uk/services/ocsd/2_learntch/20reasons.html). [accessed Jun 2015].

Glover, E. (2015) A mathematician's experience flipping a large lecture calculus course. *Proceedings of the 18th Research in Undergraduate Mathematics Education conference (forthcoming)*, Pittsburgh, Pennsylvania, USA.

Guo, P. J., Kim, J. & Rubin, R. (2014) How video production affects student engagement: An empirical study of MOOC videos. *Proceedings of the First ACM Conference on Learning @ Scale conference*. New York: ACM Press, pp. 41–50.

Gray, K. E., Adams, W. K., Wieman, C. E. & Perkins, K. K. (2008) Students know what physicists believe, but they don't agree: a study using the class survey. *Phys. Rev. ST Phys. Educ. Res.*, 4, 020106.

Hake, R. (1998) Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys.*, 66, 64–74.

Halloun, I. & Hestenes, D. (1985a) The initial knowledge state of college physics students. *Am. J. Phys.*, 53, 1043–1048.

Halloun, I. & Hestenes, D. (1985b) Common sense concepts about motion. *Am. J. Phys.*, 53, 1056–1065.

Hattie, J. (2008) *Visible Learning: A Synthesis of over 800 Meta-analyses Relating to Achievement*. London: Routledge.

Hattie, J. & Timperley, H. (2007) The power of feedback. *Rev. Educ. Res.*, 77, 81–112.

Hestenes, D. & Wells, M. (1992) A mechanics baseline test. *Phys. Teach.*, 30, 159–166.

Hestenes, D., Wells, M. & Swackhamer, G. (1992) Force concept inventory. *Phys. Teach.*, 30, 141–158.

Jolley, A., Lane, E., Kennedy, B. & Frappé-Sénéclauze, T.-P. (2012) SPESS: a new instrument for measuring student perceptions in earth and ocean science. *J. Geosci. Educ.*, 60, 83–91.

Jungić, V., Kaur, H., Mulholland, J. & Xin, C. (2015) On flipping the classroom in large first year calculus courses. *Int. J. Math. Educ. Sci. Technol.*, 46, 508–520.

Lage, M., Platt, G. & Treglia, M. (2000) Inverting the classroom: a gateway to creating an inclusive learning environment. *J. Econ. Educ.*, 31, 30–43.

Lerman, S. (2000) The social turn in mathematics education research. *Multiple Perspectives on Mathematics Teaching and Learning* (J. Boaler ed.). Westport, CT: Ablex Publishing, pp.19–44.

Knight, J. & Wood, W. (2005) Teaching more by lecturing less. *Cell Biol. Educ.*, 4, 298–310.

Mazur, E. (1997) *Peer Instruction: A User's Manual*. Upper Saddle River, NJ: Prentice Hall.

McGivney-Burelle, J. & Xue, F. (2013) Flipping calculus. *PRIMUS*, 23, 477–486.

O'Flaherty, J. & Phillips, C. (2015) The use of flipped classrooms in higher education: a scoping review. *Internet Higher Educ.*, 25, 85–95.

Overmyer, J. (2014) The flipped classroom model for college algebra: effects on student achievement. *Unpublished Doctoral Thesis*, Colorado State University, Fort Collins, Colorado.

Pritchard, D. (2010) Where learning starts? A framework for thinking about lectures in university mathematics. *Int. J. Math. Educ. Sci. Technol.*, 41, 609–623.

Prosser, M. & Trigwell, K. (1999) *Understanding Learning and Teaching: The Experience in Higher Education.* Berkshire: Open University Press.

Semsar, K., Knight, J. K., Birol, G. & Smith, M. K. (2011) The Colorado learning attitudes about science survey (CLASS) for use in biology. *CBE – Life Sci. Educ.*, 10, 268–278.

Smith, M., Jones, F., Gilbert, S. & Wieman, C. (2013) The classroom observation protocol for undergraduate STEM (COPUS): a new instrument to characterize university STEM classroom practices. *CBE – Life Sci. Educ.*, 12, 618–627.

Smith, M., Wood, W., Adams, W., Wieman, C., Knight, J., Guild, N. & Su, T. (2009) Why peer discussion improves student performance on in-class concept questions. *Science*, 323, 122–124.

Sonnert, G., Sadler, P., Sadler, S. & Bressoud, D. (2014) The impact of instructor pedagogy on college calculus students' attitude toward mathematics. *Int. J. Math. Educ. Sci. Technol.*, 46, 370–387.

Speer, N. M., Smith, J. P & Horvath, A. (2010) Collegiate mathematics teaching: An unexamined practice. *J. Math. Behav.*, 29, 99–114.

Springer, L., Stanne, M. & Donovan, S. (1999) Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: a meta-analysis. *Rev. Educ. Res.,* 69, 21–51.

Stylianides, A. J. & Stylianides, G. J. (2013) Seeking research- grounded solutions to problems of practice: classroom-based interventions in mathematics education. *ZDM—Int. J. Math. Educ.*, 45, 333–341.

Talbert, R. (2014) Inverting the linear algebra classroom. *PRIMUS*, 24, 361–374.

Toto, R. & Nguyen, H. (2009) Flipping the work design in an industrial engineering course. In Frontiers in Education Conference, FIE 2009, 39th IEEE. San Antonio, Texas: IEEE, pp. 1–4.

Vygotsky, L. (1978) *Mind in Society*. Cambridge, USA: Harvard University Press.

Vygotsky, L. (1986) *Thought and Language*. Cambridge, USA: MIT Press.

Wieman, C. (2014) Large-scale comparison of science teaching methods sends clear message. *Proc. Natl Acad. Sci.*, 111, 8319–8320.

Willis, J. (2014) The effects of flipping an undergraduate precalculus class. *Unpublished Doctoral Thesis*, Appalachian State University, Boone, North Carolina.

Yackel, E., Rasmussen, C. & King, K. (2000) Social and sociomathematical norms in an advanced undergraduate mathematics course. *J. Math. Behav.*, 19, 275–287.

Wes Maciejewski obtained his Ph.D. in mathematical biology in 2012 from Queen's University, Canada. Since then, his research has focused on mathematics education at the tertiary level. He always welcomes unsolicited emails from potential collaborators.