Flipped Statistics Class Results: Better Performance Than Lecture Over One Year Later

Jennifer R. Winquist
Kieth A. Carlson
Valparaiso University

Journal of Statistics Education Volume 22, Number 3 (2014),
www.amstat.org/publications/jse/v22n3/winquist.pdf

Key Words: Flipped class; Active learning; Flip teaching; Reverse teaching; Inverted class.

Abstract

In this paper, we compare an introductory statistics course taught using a flipped classroom approach to the same course taught using a traditional lecture based approach. In the lecture course, students listened to lecture, took notes, and completed homework assignments. In the flipped course, students read relatively simple chapters and answered reading quiz questions prior to class and completed workbook activities in class. The workbook activities consisted of questions (multiple choice, short answer, computation) designed to help students understand more complex material. Over one year after taking the course (median = 20 months), students took a standardized test of their knowledge of statistics as well as nine other content areas in psychology. Students in the flipped course outperformed the students in the lecture course on the statistics portion of the test (d =.43), but not on non-statistics portions of the test.

1. Introduction

After years of teaching statistics as a typical, lecture-based course, we decided to restructure the course in an attempt to improve student performance and attitudes toward statistics. A detailed description of our restructured course is below, but the revised course is essentially a version of a flipped course. Previously, we found that the flipped course improved students' attitudes toward statistics (Carlson and Winquist 2011), but we did not, at that time, have an objective measure of students' performance. In this paper, we discuss how we developed our flipped classroom and describe the long term effectiveness of this new course design on students' performance.

1.1 Old Course Structure and Problems

We each taught statistics for over ten years using a standard, lecture-based approach. Students had assigned readings from a textbook prior to class and then during class we lectured on the material in the book. We provided students with incomplete versions of the lecture slides and encouraged them to complete the slides as we lectured. After we finished lecturing over a chapter, students completed required homework assignments from the problems at the end of the book. To encourage self-correction, answers to the homework questions were posted online and students were able to view these answers while doing their work. Because the answers were posted, students had to show their work to receive credit for the assignment. Frequent exams were used to assess student learning, with multiple midterm exams and a cumulative final.

Overall, students were satisfied with the course structure. End of course student teaching evaluations were high and students frequently commented that they enjoyed the course far more than they had expected. Students were particularly complimentary regarding lectures, reporting that the lectures made the material clear. Although student satisfaction was high, there were two indications that the course was not meeting our expectations. One problem was students had poor long term retention of the material. Mean performance on the cumulative final exam was commonly 10% lower than mean performance on midterm exams. Generally, students' overall performance could be described as sporadic, sometimes quite good and sometimes poor. Additionally, instructors of subsequent courses in our curriculum indicated that students forgot a lot from their statistics course.

Upon reflection, there were also issues within the class that we had not recognized. Specifically, students often talked about how easy statistics seemed in class and how difficult it was at home. For a long time we interpreted these comments as compliments regarding the clarity of our lectures. Only after reflection were we grudgingly willing to admit that these were not compliments but rather an indication of serious problems with the course design. After all, our goal was to create students who could use statistics in their future classes and careers. If students were not retaining the lecture material long enough to complete homework assignments, they were almost certainly unable to use statistics in their future classes and careers.

A related problem was that some students rarely read the book either prior to or after class. Despite trying multiple incentive systems, students resisted reading the book prior to class, commenting that it was too difficult. Because of this, we lectured on the content they were supposed to read. Given this structure, it is not surprising that students rarely read the book after class, either. In fact, on end of course evaluations students frequently commented that the textbook was unnecessary. Consequently, we were inadvertently teaching students that they do not need to read.

1.2 New Course Structure Based on Principles from Cognitive Psychology

Because of these problems we decided to restructure the course with the primary goal of maximizing our students' long term retention of the material. Once we identified students' retention as a problem, we realized that, as psychologists, we should be able to apply memory research to our course design. Unfortunately, neither of us took this research into consideration when initially designing our course. We probably lectured because that seemed like the norm at our university, not because there was convincing evidence that lecture was the best pedagogical approach. To develop evidence-based pedagogies designed to improve students' retention, we turned to the cognitive psychology literature to find research we could apply to our course. A review of the memory literature revealed two promising phenomena that lead to significant improvement in long term memory experiments and that could be applied to a classroom in a straightforward way, specifically, the generation effect and the testing effect.

The generation effect is the experimental finding that generated material is recalled at higher rates than read material (Slamecka and Graf 1978). For example, students who have to generate an antonym to a word they are given (e.g., given *hot*, generate *cold*) remember the word pair better than students who simply read the word pair (e.g., hot-cold). The theoretical explanation for this effect is not yet clear but a recent meta-analysis found an impressive .40 effect size (*d*) across 86 different studies (Bertsch, Pesta, Wiscott, and McDaniel 2007).

Another way to improve long term memory is by using testing not just as a way to assess learning, but also as a way to improve students' learning. Studies have repeatedly demonstrated that participants who answer test questions prior to a final test outperform those who study for that same final test. The reported effect sizes for testing effects have typically been quite large, with researchers reporting effect sizes (d) greater than 1 (e.g., Roediger and Karkpicke 2006).

Given the robust findings for both the generation effect and the testing effect, our goal in redesigning the course was to have students generate as much of the material as possible while also answering test questions over the material. Clearly, students cannot generate all of the material. Thus, students were provided with chapters from a textbook (Carlson and Winquist 2014) to read prior to class that were intended to be relatively simple. These chapters contained the main points for a given statistic, but did not go into a great deal of depth. For example, the chapter on independent measures t-tests discussed the purpose of the test and showed students how to do the computations by hand and using SPSS. They were also shown how to set up the hypotheses, define the critical region, compute the effect size, and interpret the results. To increase cognitive engagement with the material while reading the chapters, students were required to answer multiple choice test questions over the main points in the chapter. These questions were embedded in the chapters so that there were at least one or two questions on every page of text. Students submitted answers to these questions prior to class via an on-line course management system.

In class, students completed workbook pages on which they performed computations and interpretations as described in the chapter, but they also answered questions designed to help them generate more complex material. For example, in the independent measures t-test chapter the questions were designed to help them generate the distributions of sample means for the null and research hypothesis. Once the distributions were generated, students used them to answer questions about Type I error, Type II error, effect sizes, and statistical power. By answering these questions, students generated new information about these important topics that should be easier for them to remember than if they had simply read that same information. For example, students were not shown what the distribution of sample means looked like if the null hypothesis is true, instead, they generated the distribution with the help of guided workbook questions.

Additionally, because research suggests that testing is more effective when the correct answer is provided to students after they generate answers (Kornell, Hays, and Bjork 2009), we provided answer keys for all in-class questions so students could check their own work immediately after completing each question. Students were told that their goal should be to understand why each answer is correct not to "get the correct answer."

Our revised course structure was similar to a flipped classroom with one noteworthy difference. Most flipped classes involve having students watch videos of lectures before class and then in class students work with the video-lecture material in hands-on activities. Although videos are certainly appropriate in some classes, we decided not to video tape lectures because we wanted our students to gain confidence in their ability to learn via reading. Thus, our course was flipped in the sense that students read material on their own before coming to class and then they did hands-on activities in class.

This particular structure not only capitalizes on the testing and generation effects, but also affords two other advantages. One advantage is that students generally came to class prepared. They received points for completing the reading quizzes prior to class and the vast majority (usually over 90%) of students did this. Another advantage is that because the chapters and the reading quizzes were relatively simple, most of the difficult material is encountered for the first time in class where students have the support of the instructor and/or fellow students. This means that students get individualized feedback exactly when they need it. This one-on-one interaction with students allowed us to know what material our students were struggling with so we could revise the activities for future semesters.

Although the generation and testing effects have been repeatedly demonstrated in the lab, their application in a classroom environment (essentially the flipped classroom) is less well documented. There are very few studies of the effect of flipping the classroom on student performance overall, and we were able to locate only one study of performance in a flipped statistics course. Wilson (2013) compared the performance of students who took statistics using a "flipped classroom" with videotaped lectures to students who took it with a traditional lecture course and found that students in the flipped classroom scored significantly higher on a standardized test at the end of the semester ($d = .57$). Although students were not randomly assigned to the two teaching methods, the students in the two groups did have similar scores on the pre-test given at the beginning of the course. These results suggest that the flipped classroom can be quite effective in improving performance in a statistics course. Because this is the only study we found on performance in a flipped statistics classroom, it is important that we replicate and extend this work.

There are a number of similarities between our class and Wilson's (2013). We both required students to read and answer reading questions online prior to class. We both minimized lecture time in class and instead had students spend the time doing hands-on activities. Although there are similarities, there are also some important differences. Wilson's students watched video-taped lectures prior to class while ours did not. Our course also differed in that our chapters were less comprehensive and so new material was introduced during class, requiring students to generate more information. Finally, our activities are different than those used by Wilson and we need to ensure that the success associated with flipping the classroom is not dependent upon one particular set of activities.

In addition to the differences in the class structures, we also used a different assessment measure than Wilson (2013). In her class, students completed an assessment at the end of the course using the Individual Development and Educational Assessment (IDEA) Center ratings (www.theideacenter.org). Our students completed a different standardized test, the Area Concentration Achievement Test (ACAT; www.collegeoutcomes.com). Furthermore, our students did not complete the assessment at the end of the course, but during their senior year as part of departmental assessment procedures. Therefore, the primary goal of this paper is to investigate the effects of a flipped statistics curriculum on the long term memory for the material, specifically, over a year after the completion of the course.

2. Methods

2.1 Participants

All participants were Psychology majors at Valparaiso University. All Psychology majors are required to complete an introductory statistics course and are also required to complete the ACAT during their senior year. Valparaiso University is a private, four-year, independent Lutheran University with about 3,000 undergraduate students. Most students are traditionally aged and about 66% live on campus. Overall, there were 30 males and 81 females included in our sample.

Students taking statistics from us between 2003 to 2008 took the lecture course and students taking the course between 2009 -2013 took the flipped course. For each student, we recorded the semesters between students' completing their statistics course and their taking the ACAT. Students at our University generally take statistics during their sophomore or junior year, but there is variability in this timeline. When comparing the delay interval experienced by the two groups, we discovered that several students in the lecture group took the course as freshmen, but no one in the flipped course group took the course as freshmen. In order to make the time delay between the groups more equivalent, we only used data from students who did not complete the course as a freshman (i.e., delay times between 2 and 7 semesters). With each semester being 4 months, the time delays range from 8 to 28 months. Consequently, we had a sample of 58 lecture students and 53 flipped course students.

2.2 Teaching Methods

2.2.1 Flipped Course

Prior to class students read a chapter from the textbook (Carlson and Winquist 2014), answered embedded reading questions online, and corrected their incorrect answers. Classes lasted 75 minutes and began with instructors answering questions about the reading as well as giving a brief (~10-20 minute) lecture to prepare students for that day's activity. Students spent the remainder of the class time working on the class activities. These activities presented new information that was not presented in the chapter and included questions about the reading and the new information. Activity questions were a mixture of multiple choice, fill in the blank, computation, short answer, and long answer questions. Students were provided with answers to all of the activity's questions, but they were encouraged to first answer each question and then immediately compare it to the key. If their answer was wrong they were told to find, and then correct, their error. A great deal of emphasis was placed on students understanding *why* the correct answer is correct. At the beginning of the semester we repeatedly emphasized the value in making and correcting mistakes as well as the importance of understanding the rationale for the answers to activity questions. Although students were repeatedly encouraged to work through the problems before looking at the answers, we had no way to verify that students attempted to answer the questions before looking at the answer key. It is possible that some students sometimes looked at the key and worked backwards.

Although we encouraged students to work with their peers, they were not *required* to work together, because we could not find a convenient way to enforce group work. Our students seemed to greatly prefer to work at their own pace and completed the activities primarily alone, but frequently consulted other students and the instructor when they had a question. Although we did not collect data on group work, our daily observations and interactions with students suggest that most students talked with other students and/or the instructor multiple times while completing an activity.

The course had four exams and a cumulative final exam. Before every exam students received practice exams with answers. Students received a small number of points for completing the reading questions, class activities, and practice exams. None of the reading, activity or practice test questions appeared on subsequent tests. The course covered frequency distributions, central tendency, variability, z scores, the distribution of sample means, hypothesis testing, single sample t, related sample t, independent t, confidence intervals, ANOVA, Factorial ANOVA, correlation, chi-square, and statistical assumptions.

2.2.2 Lecture Course

Prior to class, students were encouraged to read a chapter from the textbook (Gravetter and Wallnau 2003) but instructors did not verify that students completed the assigned reading. In class, the instructors lectured on the material in the book using PowerPoint slides. Students were given copies of the slides that were partially complete and were encouraged to complete the slides while listening to the lecture. Homework problems were assigned from the book and were turned in for a grade at the beginning of the next class. The computational problems were very similar to the problems used in the flipped class. However, there were fewer conceptual problems included in the homework for the lecture class than the flipped class. Students were provided with the answers for the homework assignments but needed to show their work to receive credit. Students were encouraged to work together on the homework assignments and to contact the instructor for help. There were a total of 6 midterm exams as well as a cumulative final. The midterm exams primarily focused on the most recent material, but included a cumulative component.

2.3 Measures

Students completed an online version of the Psychology Area Concentration Achievement Test (ACAT; collegeoutcomes.com) with 10 scales. One of the scales assessed knowledge of Statistics. The remaining scales assessed Abnormal, Animal Learning, Experimental Design, Social, Cognitive, Clinical/Counseling, Developmental, Physiological, and Personality. Scores on each scale can range between 200 and 800. For our purposes the Statistics scale was analyzed by itself. The remaining nine scales were averaged for an overall index of student performance in areas outside of Statistics.

2.4 Procedures

Students from a variety of disciplines take this general education Statistics course. Only psychology majors take the ACAT and so they are the only students included in this study. All psychology majors are required to complete the ACAT during their senior year. The test is not part of a class nor are students encouraged to study for the exam. Students are told that the department uses the exam to learn what students know about psychology when they leave the program. The tests are administered in small group testing sessions and are proctored. Up to two hours are permitted for completion of the test. We collected ACAT scales for all psychology majors.

3. Results

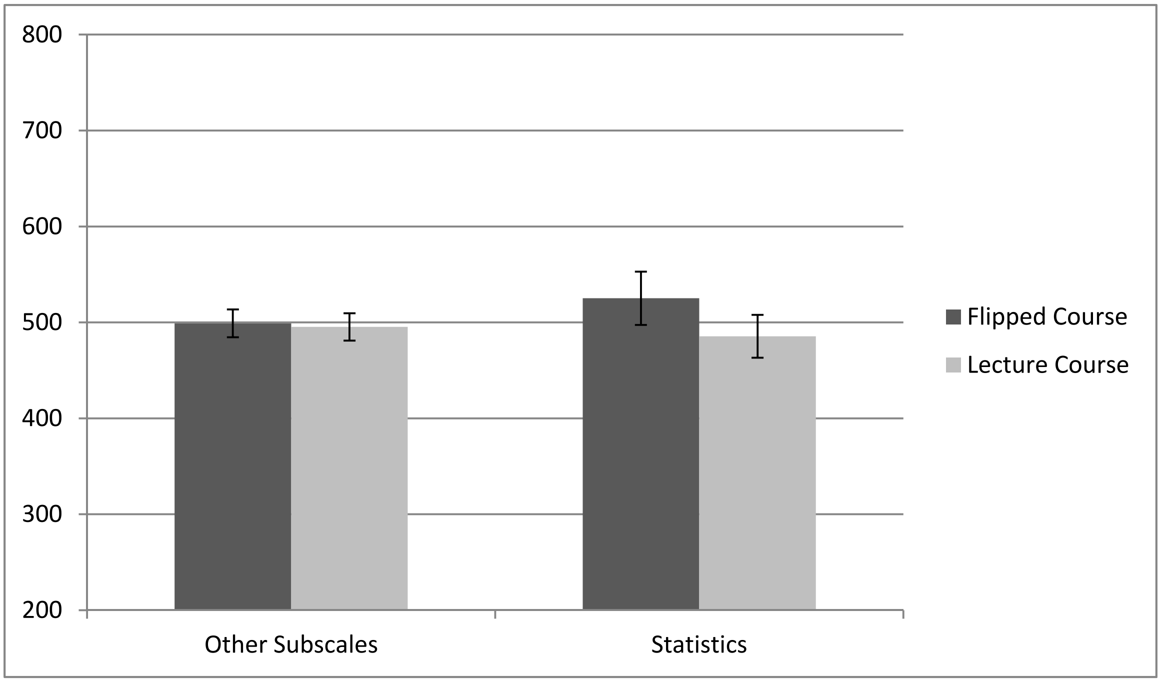
3.1 Sample characteristics

The median semester delay between completing the class and the ACAT was 6 for the lecture group and 5 for the flipped group. Since each semester was approximately 4 months, this translates into median delays of approximately 24 months and 20 months, respectively. A Mann-Whitney U test revealed that this difference was significant, (z = 3.44, p = .001). However, time delay was not correlated with scores on the Statistics subscale, r_s(109) = -.07, p = .50 or with the mean of the remaining nine scales, (r_s(109) = -.04, p = .71).This delay difference is not ideal. However, given that there was no correlation between the delay and performance and the common finding in the human memory literature that most forgetting occurs in the days and weeks immediately following learning (Wixted and Ebbesen 1991), there is good reason to conclude the delay periods experienced by the two groups were functionally equivalent.

3.2 ACAT Performance

We compared the ACAT performance of students who took the flipped statistics course to the performance of students who took the lecture course with a 2 (Teaching method: flipped vs. lecture) X 2 (ACAT Scale: statistics vs. average on other 9 scales) mixed factorial ANOVA. The Method X ACAT Scale interaction was significant, (F (1, 109) = 6.11, p = .02, η_p^2 = .05); the means are in Figure 1. The simple effects revealed that students in the flipped course scored significantly higher on the Statistics subscale (M = 525.17, SD = 100.54, *95% CI* [497.37, 552.97]) than students in the lecture course (M = 485.57, SD = 84.96, *95% CI* [463.24, 507.90], F (1, 109) = 5.05, p = .027, d = .43, 95% CI [.05, .80]). Importantly, the flipped group's better statistics performance is not due to their being better overall; their mean score on the other 9 scales (*M* = 499.00, *SD* = 52.52, *95% CI* [484.53, 513.47]) was not significantly different than the lecture group's performance (*M* = 495.29, *SD* = 54.50, *95% CI* [480.97, 509.61]) on the other nine scales, (*F* (1,109) = .13, *p* = .72, *d* = .07, 95% CI [-.30, .44]). Neither the main effect of subscale, (*F* (1, 109) = 1.28, *p* = .26, η_p^2 = .01) nor the main effect of teaching method (*F* (1, 109) = 3.04, *p* = .08, η_p^2 = .03), was significant.

Figure 1. Mean scores on ACAT subscales by teaching method with 95% confidence interval bars



4. Discussion

The students taking the flipped course did significantly better on the statistics portion of the ACAT but not on the other nine ACAT scales. The moderate effect size (*d* = .43) is particularly noteworthy when one recognizes that the ACAT exam was taken over a year (median = 20 months) after students had completed the flipped course, students were not encouraged to study for the exam, nor did we design the course specifically to prepare students for the ACAT exam. Clearly, the flipped course led to better long term performance than the lecture course. These results suggest that having students answer exam type questions and having them generate answers more frequently can lead to better performance even a year after completing a course.

Although students were not randomly assigned to the two teaching methods and the different teaching methods were used at different times, we were able to control for some individual difference variables by looking at the ACAT scores on the remaining nine subscales. The students in the lecture and flipped groups were not significantly different on the other subscales, suggesting that the improved performance on the statistics subscale was unlikely to be the result of student ability or motivation.

Of course, this study has important limitations as well. The two types of classes did not vary just in the class format, but in other important ways as well. For example, we used different books for the flipped and lecture classes. Both books were introductory statistics books intended for students in the behavioral sciences and the substantive content was very similar. There were differences between the books, but these differences are relatively minor compared to the changes in class structure between the lecture and flipped classes. It seems very unlikely that these minor differences in the content covered in the books would result in changes in performance on a standardized exam over one year later. It is much more likely that the improved performance was the result of changes in the course structure. Because students in the flipped classes were reading and understanding the material prior to class, we did not need to use our class time for lecturing on the assigned reading. Instead, we were able to use class time to have students work on workbook activities that covered more material than our lectures, but also incorporated generation effects and testing effects. The flipped format was far more efficient than our lecture course and because of this we were able to increase our coverage of the material without overburdening our students.

Another important concern has to do with the generalizability of these results to other instructors. It seems clear that *our* flipped approach is better than *our* lecture approach. Future research is needed to determine if this flipped approach interacts with instructor characteristics such that the approach is only effective for some instructors.

As stated above, we are assuming that our flipped approach was better than our lecture approach because of its greater reliance on the generation and testing effects. It is not that lectures cannot also enable these same effects, but, all too frequently, lectures require audiences to generate and answer their own questions while simultaneously listening. In some cases an audience can manage this heightened responsibility and difficulty and in some cases they can’t. Highly knowledgeable, motivated audiences that are inherently interested in the lecture topic likely do engage in this elaborative processing that leads to effective learning. With less knowledgeable and less motivated learners, teaching approaches that require every student to generate answers will probably be more beneficial than even brilliant lectures. Based on our data and experiences, it seems that a flipped, activity-based teaching approach is ideal for introductory statistics courses.

Acknowledgments

This research was supported by a Kapfer Award at Valparaiso University.

References

Bertsch, S., Pesta, B.J., Wiscott, R. & McDaniel, A. (2007), “The Generation Effect: A Meta-Analytic Review,” *Memory & Cognition*, 35, 201-210.

Carlson, K. A., & Winquist, J.R. (2011), "Evaluating an Active Learning Approach to Teaching Introductory Statistics: A Classroom Workbook Approach," *Journal of Statistics Education [Online]*, 19 (1). Available at http://www.amstat.org/publications/jse/v19n1/carlson.pdf.

Carlson, K. A. & Winquist, J.R. (2014), *Introduction to Statistics: An Active Learning Approach*, Los Angeles, CA: Sage.

Gravetter, F. J., & Wallnau, L. B. (2003), *Statistics for the Behavioral Sciences*, 6th edition, Belmont, CA: Thomson Wadsworth.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009), "Unsuccessful Retrieval Attempts Enhance Subsequent Learning," *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 35, 989-998.

Roediger, H.L., III, & Karpicke, J.D. (2006), "The Power of Testing Memory: Basic Research and Implications for Educational Practice," *Perspectives on Psychological Science*, 1, 181-120.

Slamecka, N.J., & Graf, P. (1978), "The Generation Effect: Delineation of a Phenomenon," *Journal of Experimental Psychology: Human Learning & Memory*, 4, 592-604.

Wilson, S. (2013), "The Flipped Class: A Method to Address the Challenges of an Undergraduate Statistics Course," *Teaching of Psychology*, 40(3), 193-199.

Wixted, J. T., & Ebbesen, E. B. (1991), "On the Form of Forgetting," *Psychological Science*, 2(6), 409-415.

Jennifer Winquist
Department of Psychology
1001 Campus Drive South
Valparaiso, IN 46383
Jennifer.Winquist@valpo.edu

Kieth Carlson
Department of Psychology
1001 Campus Drive South
Valparaiso, IN 46383
Kieth.Carlson@valpo.edu
