

A CASE STUDY EXAMINING THE IMPACT OF TIME OF DAY ON TEACHING EVALUATIONS *

*Ben Stephenson
Department of Computer Science
University of Calgary
2500 University Drive NW
Calgary, Alberta, Canada
ben.stephenson@ucalgary.ca*

*Flora Stephenson
Institute for Public Health
University of Calgary
2500 University Drive NW
Calgary, Alberta, Canada*

ABSTRACT

Many factors influence student evaluations of teaching. Some of these factors, such as enthusiasm and fairness of exams, can generally be directly influenced by the instructor. Other factors such as class size, time of day and classroom are often out of the instructor's control. In this study, we examine the impact that time of day has on student evaluations of teaching. Data is analyzed from four terms where the same instructor taught two sections of the same computer science course at different times. Our results show that nearly all of the statistically significant differences indicate that lower evaluations were earned in the later section. This difference was observed for both evaluation items that we expected to change with time of day, such as enthusiasm, and for evaluation items where consistent results were expected, such as fairness of evaluation methods.

1 INTRODUCTION

Most institutions of higher learning survey students about their experience in each course near the end of each semester. These survey results are used for a variety of purposes that include providing feedback so that courses can be improved, evaluating faculty members for tenure, promotion and merit pay, and reporting to prospective students, governments, donors and other stakeholders.

* Copyright © 2014 by the Consortium for Computing Sciences in Colleges. Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the CCSC copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Consortium for Computing Sciences in Colleges. To copy otherwise, or to republish, requires a fee and/or specific permission.

Despite their widespread use, the value of student evaluations of teaching is a contentious topic in some circles. Numerous studies, some of which are discussed in Section 2, have considered a wide variety of factors that may affect teaching evaluations. Some of these factors are within the control of faculty members including how enthusiastically the material is presented and whether or not exams are fair. Other factors like class size, class level, and classroom are often outside of the instructor's control. Yet, even though teaching evaluations are impacted by these uncontrollable factors, they may still be used to determine professors' merit pay and opportunities for advancement.

This case study examines the impact that time of day has on teaching evaluations by examining the differences in scores earned by the same instructor when teaching two sections of the same course in the same term at different times of day over a 4-year period. At most institutions, including our own, time of day is a factor that is outside of the instructor's control. Yet our study shows that time of day impacts student evaluations of the course for the instructor under consideration in this case study.

2 RELATED WORK

Previous studies have examined the impact of time of day on the teaching evaluations reported by economics and business students. For example, Hinkin examined how the evaluations for 8:00am classes compared to classes later in the day, finding no statistically significant differences [7]. Other studies that showed no statistically significant difference due to time of day include work by Koh and Tan [9], Liaw and Goh [10], and McPerson [12].

In contrast, a study by McCready involving business students concluded that there is a statistically significant difference in teaching evaluations based on time of day [11]. In particular, he notes that earlier courses generally received higher scores, and that no courses starting after 5:30pm received top evaluations. Badri, Abdulla, Kamali and Dodeen also report that daytime classes received better evaluations than evening classes [1], as did DeBerg and Wilson [5].

Statistically significant differences have also been reported showing that evening classes receive better evaluations than daytime classes, including work by Cranton and Smith [4]. Isely and Singh also found that evening classes received higher scores in two of their models [8].

3 STUDY DESCRIPTION

Our study controls for one of the largest variables in teaching evaluations, which is the instructor himself. We examined teaching evaluations from four consecutive fall terms where the same instructor taught two sections of the same introductory computer science course, one at 1:00pm and one at 3:00pm. The course sections were taught on the same days of the week, using the same PowerPoint slides and code examples. The assignments, exams, textbook and grading policies used in each course section were identical in the 1:00pm and 3:00pm sections each year.

Some changes were made to the course during the four-year study period to keep it up to date, respond to student feedback and incorporate additional participatory

elements. In addition, the university moved from identifiable to anonymous teaching evaluations during the time period under consideration. Total enrolment increased from approximately 150 students in 2009 to 300 students in 2012. Because of these differences we did not compare results across years.

3.1 Data Collection Instrument

Students complete the same multiple choice questionnaire about each course that they take. Both course sections were surveyed on the same day each year. However, the day of the week that the survey was administered may have varied from one year to the next. The following questions are used to assess the course and its instructor. Students responded to each question using a seven point Likert scale.

1. The overall quality of instruction was:
2. The course outline or other descriptive information provided enough detail about the course (e.g., goals, reading list, topics covered, assignments, exams, due dates, grade weightings).
3. The course as delivered followed the outline and other course descriptive information.
4. The course material was presented in a well-organized manner.
5. Student questions and comments were responded to appropriately.
6. The course content was communicated with enthusiasm.
7. Opportunities for course assistance were available (e.g., instructor office hours, out-of-class appointments, email, telephone, websites).
8. Students were treated respectfully.
9. The evaluation methods used for determining the course grade were fair.
10. Students' work was graded in a reasonable amount of time.
11. I learned a lot in the course.
12. The support materials (e.g., readings, audio-visual materials, speakers, field trips, equipment, software, etc.) used in this course help me to learn.

We hypothesized that questions 4 (organized presentation), 5 (response to student questions), 6 (enthusiasm), and 8 (students treated respectfully) might show differences due to increased instructor fatigue during the later section. We also hypothesized that question 11 (learned a lot in course) could show lower results in the later section due to lower attendance and increased student fatigue. Since we hypothesized that differences could occur in approximately half of the questions, we also hypothesized that a difference might occur in question 1 (overall instruction).

We hypothesized that there would not be any difference in question 2 (course outline provided sufficient detail), question 3 (course followed outline), question 7 (opportunities for assistance) or question 12 (support materials) because the same course outline, course website and PowerPoint slides and examples were used for each course section. The instructor held the same office hours for both sections of the course, and generally had no idea which class section a student was enrolled in when responding to student email.

Both course sections used the same assignments and exams. All of the students wrote the midterm exam simultaneously and the exam was returned on the same day in

both sections. While there may have been a small variation in when assignments were returned in the tutorial sections, the tutorials were open to students from both sections. As a result, even if one teaching assistant was particularly fast or particularly slow returning an assignment, it would have impacted students from both sections. Therefore we did not anticipate differences in questions 9 (evaluation methods fair) or 10 (worked graded in a reasonable amount of time).

3.2 Methods

A Chi-Square test was used to determine if the distribution of student responses differed between the sections. In order to perform the analysis, all of the negative and neutral responses were collapsed into a single category. This transformation was necessary due to the small number of responses at the lower end of the scale. The three positive response categories were not modified. Questions which showed a statistically significant difference for $p < 0.05$ are marked with a † in figures 1 through 4.

We also performed a t-test to determine whether or not the average score for each question differed across course sections. At a significance level of $p < 0.05$, this analysis identified an overlapping but different set of questions with statistically significant differences. In particular, there were several questions that showed statistically significantly different response distributions, but similar average scores that were not statistically significantly different. Similarly, two questions showed statistically significant average score differences without having statistically significantly different distributions, likely as a result of collapsing all of the negative and neutral responses into a single category when examining the distributions. Questions that showed statistically significant differences in the average score are marked with a * in figures 1 through 4.

4 RESULTS

Data for the 1:00pm and 3:00pm sections of the course from 2009 to 2012 are shown in figures 1 through 4. Error bars show the standard deviations of the responses.

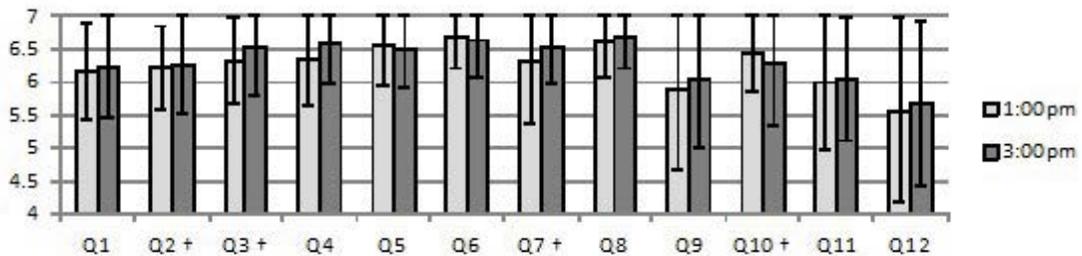


Figure 1: Average Student Response for each Section in 2009

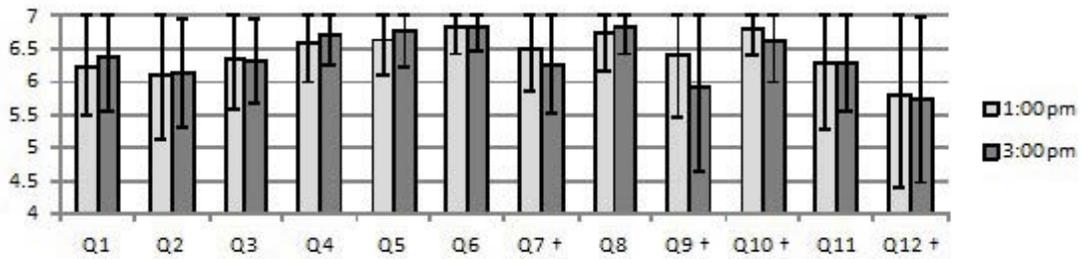


Figure 2: Average Student Response for each Section in 2010

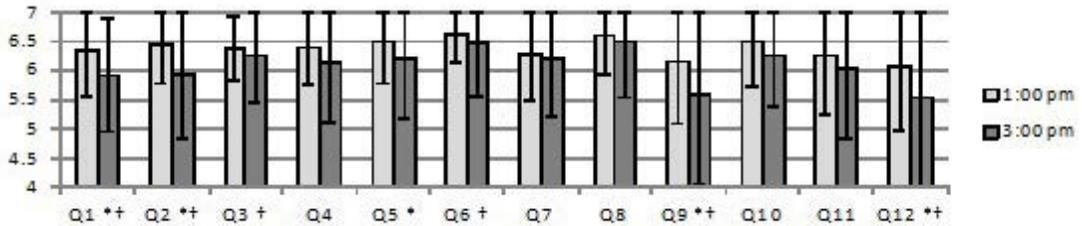


Figure 3: Average Student Response for each Section in 2011

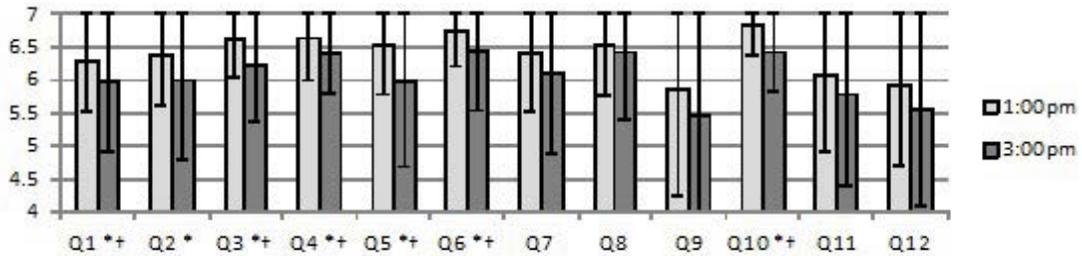


Figure 4: Average Student Response for each Section in 2012

Of the 48 comparisons performed, 10 showed statistically significant differences in both the distribution of the student responses and the average student response. In all cases, these differences indicate that lower scores were achieved in the 3:00pm class. An additional 10 data points showed statistically significant differences in the distribution of student responses without showing a statistically significant difference in the average result. Of these 10 differences, 7 indicate that lower scores were achieved in the 3:00pm class. All three of the questions that showed higher scores in the 3:00pm class occurred in 2009. Finally, there were two questions, one in 2011 and one in 2012, that showed statistically significant differences in the average score, without showing statistically significant differences in the distribution. In both cases, the difference in the average was relatively large at 0.30 to 0.38 respectively, and in each case the score achieved in the 3:00pm class was lower than in the 1:00pm class.

Overall, 22 of 48 questions showed a statistically significant difference. Of these 22 questions, 19 showed stronger results in the 1:00pm class. We discuss the causes and implications of these results in the following section.

5 DISCUSSION

We expected to observe some differences between the 1:00pm and 3:00pm sections because both the students and the instructor have changed during the two hour time difference. The instructor has taught an additional class, and may have attended a meeting, provided individual assistance to students, or answered a myriad of emails during the one hour gap between the classes. Similarly, the students may have attended additional lectures or tutorials, or they may have spent time on research, assignments, homework or employment. The instructor's informal self-assessment clearly indicated a greater level of fatigue for the 3:00pm class, and students may have experienced a similar phenomenon.

Our results, particularly from 2011 and 2012, show that time of day impacts both factors one would expect it to and factors that one wouldn't necessarily expect it to. In particular, we observed differences in excess of 0.48 in 2011 for several questions where the classes were treated identically including question 2 (enough detail in course outline), question 9 (evaluation methods fair) and question 12 (support materials helpful). Differences in excess of 0.35 were observed for the same questions in 2012, though the results for question 9 were not statistically significant that year. These results clearly indicate that lower scores are awarded in the later time slot for the same performance because both class sections used the same course outline, evaluation methods and support materials. This result is consistent with work conducted previously by Williams and Ceci which showed that simply increasing the amount of voice modulation and number of gestures improved student opinions for seemingly unrelated questions such as the quality of the textbook [13].

While we have noted that numerous statistically significant differences have occurred, it is also necessary to ask whether or not those differences are administratively significant. Are these statistically significant differences apparent to administrators? Will they impact administrative decisions including promotion, tenure, merit pay and workload assignments? These are difficult questions to answer, and likely vary from administrator to administrator. A survey of administrators would provide additional insight into how the evaluations are used, either confirming or dispelling fears that differences caused by factors beyond the instructors control could result in differential treatment.

6 CONCLUSION

From the data evaluated in this study, it is clear that time of day impacted the teaching evaluations for the instructor that was studied, especially in 2011 and 2012. Furthermore, it appears that this impact is not confined to factors one might expect such as enthusiasm. Other factors that were objectively identical in both lecture sections such as the level of detail in the course outline also received lower scores in the later class.

Despite these concerns, the authors believe that student evaluations should continue to be an important source of feedback for professors as they work to improve courses, for administrators as they make decisions regarding promotion, tenure and merit pay, and for other stakeholders. However, the student evaluations should not be the only source of information. Instead, as several other authors have noted previously [2, 6], student evaluations of teaching should be considered along with other data such as feedback from colleagues who have observed classes and reflective statements made by the instructor. Other work has also suggested that valuable feedback can be received from students through a facilitated discussion [3] lead by an arms-length faculty member. Our data supports the value of these initiatives, and provides further evidence that teaching should be evaluated by faculty members and administrators, in addition to students.

REFERENCES

- [1] M. A. Badri, M. Abdulla, M. A. Kamali, and H. Dodeen. Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAE. *International Journal of Educational Management*, 20(1):43-59, 2006.
- [2] T. Baldwin and N. Blattner. Guarding against potential bias in student evaluations: What every faculty member needs to know. *College Teaching*, 51(1):27-32, 2003.
- [3] M. Craig. Facilitated student discussions for evaluating teaching. In *Proceedings of the 38th SIGCSE technical symposium on computer science education, SIGCSE '07*, pages 190-194, New York, NY, USA, 2007. ACM.
- [4] P. A. Cranton and R. A. Smith. A new look at the effect of course characteristics on student ratings of instruction. *American Educational Research Journal*, 23(1):117-128, 1986.
- [5] C. L. DeBerg and J. R. Wilson. An empirical investigation of the potential confounding variables in student evaluation of teaching. *Journal of Accounting Education*, 8(1):37-62, 1990.
- [6] R. M. Felder and R. Brent. How to evaluate teaching. *Chemical Engineering Education*, 38(3):200-202, 2004.
- [7] T. R. Hinkin. The effects of time of day on student teaching evaluations: Perceptions versus reality. *Journal of Management Education*, 15(1):35-45, 1991.
- [8] P. Isely and H. Singh. Do higher grades lead to favorable student evaluations? *Journal of Economic Education*, 36(1):29-42, 2005.
- [9] H. C. Koh and T. M. Tan. Empirical investigation of the factors affecting set results. *International Journal of Educational Management*, 11(4):170-178, 1997.
- [10] S.-H. Liaw and K.-L. Goh. Evidence and control of biases in student evaluations of teaching. *International Journal of Educational Management*, 17(1):37-43, 2003.

- [11] D. J. McCready. Student evaluation of teaching. *The Canadian Journal of Higher Education*, 11(2):67-77, 1981.
- [12] M. A. McPherson. Determinants of how students evaluate teachers. *Journal of Economic Education*, 37(1):21-37, 2006.
- [13] W. M. Williams and S. J. Ceci. "How'm I doing?" Problems with student ratings of instructors and courses. *Change*, 29(5):12-23, 1997.